A NOVEL DYNAMIC FEATURE SELECTION AND PREDICTION ALGORITHM FOR CLINICAL DECISIONS INVOLVING HIGH-DIMENSIONAL AND VARIED PATIENT DATA

SHERINE NAGY MOUSTAFA SALEH

Doctor of Philosophy

ASTON UNIVERSITY

March 2016

ASTON UNIVERSITY

A novel dynamic feature selection and prediction algorithm for clinical decisions involving high-dimensional and varied patient data

Sherine Nagy Moustafa Saleh

Doctor of Philosophy
March 2016

Abstract

Predicting suicide risk for mental health patients is a challenging task performed by practitioners on a daily basis. Failure to perform proper evaluation of this risk could have a direct effect on the patient's quality of life and possibly even lead to fatal outcomes.

Risk predictions are based on data that are difficult to analyse because they involve a heterogeneous set of patients' records from a high-dimensional set of potential variables. Patient heterogeneity forces the need for various types and numbers of questions to be asked regarding the individual profile and perceived level of risk. It also results in records having different combinations of present variables and a large percentage of missing ones. Another problem is that the data collected consist of risk judgements given by several thousand assessors for a large number of patients. The problem is how to use the associations between patient profiles and clinical judgements to generate a model that reflects the agreement across all practitioners.

In this thesis, a novel dynamic feature selection algorithm is proposed which can predict the risk level based only on the most influential answers provided by the patient. The feature selection optimises the vector for predictions by selecting variables that maximise correlation with the assessors' risk judgement and minimise mutual information within the ones already selected. The final vector is then classified using a linear regression equation learned for all patients with a matching set of variables. The overall approach has been named the Dynamic Feature Selection and Prediction algorithm, DFSP.

The results show that the DFSP is at least as accurate or more accurate than alternative gold-standard approaches such as random forest classification trees. The comparison was based on accuracy and error measures applied to each risk level separately ensuring no preference to one risk over the other.

Keywords: Data mining, Missing data, Healthcare, Suicide risk, Assessment, Prediction

Acknowledgements

First of all, I would like to show my deepest gratitude to Dr. Christopher Buckingham. If it were not for your impeccable supervision and strong enthusiasm, I definitely would not have delivered such a thesis. I would like to thank the GRiST team members for delivering such an inspiring project and providing me with the data to work on.

I would also like to dedicate this work to all my family members whom have supported me through out the years. Endless love, support and guidance have always been provided by my parents through all the stages of my life. An amazing husband who has always put my study a priority, he was and will always be my rock. My son who is the light of my path and can put a smile on my face no matter what. My sister, brother-in-law and their angels for just being by my side and encouraging me through out the days.

Finally, I would like to thank the Arab Academy for Science and Technology (AAST) in Egypt for funding my PhD and giving me such an opportunity to learn at Aston University.

Table of contents

List of figures 8

List of tables 11

List of algorithms 13

List of Abbreviations 14

1 Introduction 15

1.1 Understanding suicide risk prediction 15

1.1.1 Automated data analysis 16

1.2 Approaches to suicide risk prediction 17

1.3 Predictive data mining tools 19

1.3.1 Feature selection 19

1.3.2 Missing data 20

1.3.3 Risk prediction 21

1.4 GRiST support system 21

1.5 Thesis contributions 22

1.6 Thesis outline 23

2 Background 25

2.1 Introduction 25

2.2 Mental health risk and safety 26

2.2.1 Clinical problems 27

2.2.2 Approaches to risk assessment 28

2.2.3 Current tools 30

2.3 Clinical decision support systems 31

2.4 The GRiST CDSS 32

2.4.1 Initiation of GRiST 32

2.4.2 The Galatean model 34

2.4.3 User interface 36

2.4.4 Problems to be handled by GRiST 39

2.5 Summary and conclusions 41

3 Prediction Methods 43

3.1 Introduction 43

3.2 Missing data 45

3.2.1 Why is data missing? 45

3.2.2 Types of missing data 46

3.2.3 Handling missing data 48

3.3 Recent work on missing data 51

3.4 Feature selection 52

3.4.1 General feature selection approaches 53

3.4.2 Methods of feature selection 56

3.4.3 Recent work on feature selection 60

3.5 Predictive data mining algorithms 61

3.5.1 Decision trees 62

3.5.2 Random forest 63

3.5.3 Regression 63

3.6 Summary and conclusions . . . 65

4 Dynamic Feature Selection and Prediction 67

4.1 Introduction . . . 67

4.2 Rationale . . . 68

4.3 Basic DFSP algorithm . . . 68

4.3.1 Evaluating features . . . 69

4.3.2 Filter evaluation criteria . . . 70

4.3.3 Assessing patient records . . . 71

4.4 DFSP algorithm . . . 72

4.4.1 Defining parameters . . . 72

4.4.2 Learning models . . . 73

4.4.3 Testing models . . . 73

4.4.4 Parametrised DFSP . . . 74

4.5 Summary and conclusion . . . 77

5 Experiments 78

5.1 Introduction . . . 78

5.2 Experiment 1: Methods and parameters . . . 79

5.2.1 Choosing the proper method . . . 80

5.2.2 Setting *DV-Thresh* parameter . . . 87

5.2.3 Setting *IV-Thresh* parameter . . . 91

5.3 Experiment 2: Applying multinomial logistic regression . . . 94

5.4 Experiment 3: Testing GLM . . . 95

5.5 Experiment 4: Adapting output . . . 97

5.6 Experiment 5: Comparison to other standard approaches . . . 101

5.6.1 Applying standard approaches . . . 102

5.7 Summary and conclusions . . . 107

6 Implementing the DFSP within GRiST 109

6.1 Introduction 109

6.2 Assessing speed 110

6.3 Enhancing DFSP algorithm 110

6.3.1 Enhancing feature selection 112

6.3.2 Enhancing classification 113

6.3.3 Performance evaluation 114

6.4 Speed comparison with decision trees 115

6.5 Summary and Conclusions 115

7 Generalising the DFSP for all mental health risks 117

7.1 Introduction 117

7.2 Determining parameters 118

7.3 Comparison to others 128

7.4 Summary and conclusions 132

8 Summary, conclusions and future work 133

8.1 DFSP algorithm summary 135

8.2 Current limitations and future work 136

8.2.1 Parameter estimation 136

8.2.2 Generic application on different datasets 137

8.2.3 Learning from missing information 138

8.3 Final conclusion 138

References 139

List of figures

2.1 Usage of membership grades to calculate suicide risk intent, reproduced from (Buckingham, 2002) 35

2.2 Snapshot of the user interface showing filter questions 37

2.3 Snapshot of the user interface showing filter and non filter questions 38

2.4 Distribution of the number of records per risk category 39

3.1 Missing map of the GRiST dataset with the colour blue indicating missing values and red for existing ones 47

3.2 General steps of feature selection 54

5.1 Accuracy and shifted accuracy for suicide risk levels 0 patients (Figure a) and 1 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for mutual information *MI*, correlation *Corr* and minimum redundancy maximum relevance *mRMR* criteria. 81

5.2 Accuracy and shifted accuracy for suicide risk levels 2 patients (Figure a) and 3 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for mutual information *MI*, correlation *Corr* and minimum redundancy maximum relevance *mRMR* criteria. 82

5.3 Accuracy and shifted accuracy for suicide risk levels 4 patients (Figure a) and 5 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for mutual information *MI*, correlation *Corr* and minimum redundancy maximum relevance *mRMR* criteria. 83

5.4 Accuracy and shifted accuracy for suicide risk levels 6 patients (Figure a) and 7 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for mutual information *MI*, correlation *Corr* and minimum redundancy maximum relevance *mRMR* criteria. 84

5.5 Accuracy and shifted accuracy for suicide risk levels 8 patients (Figure a) and 9 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for mutual information *MI*, correlation *Corr* and minimum redundancy maximum relevance *mRMR* criteria. 85

5.6 Accuracy and shifted accuracy for risk 10 using feature counts ranging from 1 to 20. The results are shown for mutual information *MI*, correlation *Corr* and minimum redundancy maximum relevance *mRMR* criteria. 86

5.7 The clinical judgement versus mean risk prediction 97

5.8 Mean absolute error for each risk level 98

5.9 Mean absolute error for each risk level with and without the equation 99

5.10 Mean absolute error for each risk level when using DFSP as opposed to other off the shelf methods. 106

7.1 Accuracy and shifted accuracy for HTO risk levels 0 patients (Figure a) and 1 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for correlation *Corr* criterion. 119

7.2 Accuracy and shifted accuracy for HTO risk levels 2 patients (Figure a) and 3 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for correlation *Corr* criterion. 120

7.3 Accuracy and shifted accuracy for HTO risk levels 4 patients (Figure a) and 5 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for correlation *Corr* criterion. 121

7.4 Accuracy and shifted accuracy for HTO risk levels 6 patients (Figure a) and 7 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for correlation *Corr* criterion. 122

7.5 Accuracy and shifted accuracy for HTO risk levels 8 patients (Figure a) and 9 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for correlation *Corr* criterion. 123

7.6 Accuracy and shifted accuracy for risk 10 using feature counts ranging from 1 to 20. The results are shown for correlation *Corr* criterion. 124

7.7 Mean absolute error for each risk level 131

List of tables

5.1 Results of testing the algorithm with *DV-Thresh* ranging from 0.5 to 0.1: the *ACC* column is the percentage of predictions exactly matching the clinical judgement and the *S.ACC* column is the percentage that either match the judgement or are one away from it. 88

5.2 Results of testing the algorithm with *DV-Thresh* 0.2, 0.1, 0.05, 0.04 and 0.03 while setting the feature limit to 12: the *ACC* column is the percentage of predictions exactly matching the clinical judgement and the *S.ACC* column is the percentage that either match the judgement or are one away from it. . 89

5.3 Results of comparing the effect of the addition of a 12 feature limit to records previously chose more than 12 using a *DV-Thresh* of 0.1. 90

5.4 Results of accuracy and shifted accuracy for *IV-Thresh* ranging between 0.4 and 0.9 as compared to the results achieved by having no *IV-Thresh*. All results are based on setting the *DV-Thresh* to 0.04 and feature limit to 12 features. 92

5.5 Confusion matrix for predicting clinical risk judgements using 0.8 *IV-Thresh*, 0.04 *DV-Thresh* and 12 feature limit 93

5.6 Results for using logistic regression versus linear regression inside the DFSP algorithm 95

5.7 Results of applying GLM versus linear regression 96

5.8 Confusion matrix for predicting clinical risk judgements after the application of the equation. 100
5.9 Result of DFSP with and without the equation 101
5.10 Results of comparing accuracy when classifying suicide risk in GRiST dataset to different off the shelf methods 104
5.11 Results of comparing shifted accuracy when classifying suicide risk in GRiST dataset to different off the shelf methods 106

6.1 Time taken for assessing a single patient on average using the DFSP and total time for all 71,024 records. 110
6.2 Time taken for assessing a single patient on average using the DFSP and total time for all 71,024 records before and after the enhancement. 115
6.3 Time taken for assessing a single patient on average and total time for all 71,024 records for the DFSP and decision trees *DT*. 115

7.1 Results of testing the algorithm with DV-Thresh ranging from 0.04 to 0.09 while setting the feature limit to 8: the *ACC* column is the percentage of predictions exactly matching the clinical judgement and the *S.ACC* column is the percentage that either match the judgement or are one away from it. . 126
7.2 Results of accuracy and shifted accuracy for *IV-Thresh* ranging between 0.4 and 0.9. All results are based on setting the *DV-Thresh* to 0.01 and *Count-Limit* to 8 features. 127
7.3 Results of comparing accuracy when classifying violence risk in GRiST dataset to different off the shelf methods 129
7.4 Results of comparing shifted accuracy when classifying violence risk in GRiST dataset to different off the shelf methods 130
7.5 Time taken for assessing a single patient on average and total time for all records for the *DFSP* and *mean+DT* 131

List of Algorithms

1 Basic version of the DFSP algorithm 68

2 Generating *MI-IV-DV*, *Corr-IV-DV* and *IV-IV* matrices 69

3 Parametrised version of the DFSP algorithm 75

4 Final version of the DFSP feature selection stage 111

5 Final version of the DFSP classification stage 113

6 Proposed exhaustive parameter evaluation methodology 137

List of Abbreviations

Acronyms / Abbreviations

CDSS	Clinical Decision Support Systems
CFS	Correlation-based Feature Selection
DFSP	Dynamic Feature Selection and Prediction
DV	Dependent Variable
EM	Expectation Maximization
GLM	Generalized Linear Models
GRiST	Galatean Risk Screening Tool
HTO	Harm To Others
IV	Independent Variable

Chapter 1

Introduction

The World Health Organization presented a fact sheet concerning suicide in August 2015 which stated that on average there are 800,000 suicide cases in the world and there are many more suicide attempts. Furthermore, they categorised suicide as the second main cause of death in the 15 to 29 age group based on an assessment performed in year 2012 (WHO, 2015). Suicide is thus a significant risk that needs to be addressed.

In this chapter, suicide risk prediction will be introduced along with common approaches used for predicting it. An introduction to predictive data mining and the methods to handle common mental health data will then be described. A brief description of clinical support systems will be given before finishing with an outline to the rest of the thesis chapters.

1.1 Understanding suicide risk prediction

Suicide risk prediction is a very important area of research for many reasons, starting from improving the quality of life to trying to reduce the average number of deaths caused by it. The prediction and management of suicide risk is sometimes conducted by assessors who have not received specialist training. In such cases, human assessments will be unreliable and the same risk profile will result in varied evaluations of risk level. Also the experience upon which the evaluation is based differs from one assessor to the other. In many situations, suicide risk assessments are performed by emergency personnel who are not mental health professionals. In such cases, incorrect prediction and management could be lethal (Dawes et al., 1989).

1.1.1 Automated data analysis

Clinical data mining attempts to find patterns and analyse relationships within a dataset. The outcome is used for either description or prediction, where description is involved with showing the end-user the patterns in a comprehensible way and prediction is involved with determining future values or behaviour of patients. Data mining provides useful tools for making decisions in the healthcare domain (Berner, 2007; Weitschek et al., 2013).

Clinical Decision Support Systems (CDSS) are specially designed software for helping make decisions. They are computer programs that manage the data, models, knowledge engine and user-interface. CDSSs often use rules to make decisions that are stored in the knowledge system. The problem with this type of CDSS is that it needs to have all possible combinations of information stored in the knowledge system, which might not be feasible given large sized datasets. However, a combination of data mining and CDSSs add power of finding and analysing patterns to storage and retrieval of users records (Berner, 2007).

Sometimes, the data collected have missing information. For instance, patients might choose to withhold information they consider private. In other cases, the specified questions in the interface do not relate to the particular case of the patient to be assessed. For common CDSSs that use knowledge engines, missing data can be a big obstacle especially, if they represent a large percentage of the dataset (Berner, 2007).

Previous surveys (Chesin and Stanley, 2013; Lotito and Cook, 2015; Silverman and Berman, 2014) have mentioned that currently there are no structured tools to assess and explain suicide risk. Such tools could help in prediction and management of the risk of suicide and reduce death rates. For these reasons creating a data mining algorithm that would fit inside a CDSS to provide support for assessing risks is the challenge addressed by this thesis.

1.2 Approaches to suicide risk prediction

There are different approaches to assessing risk such as clinical, actuarial or structured. Clinical assessments are based on decisions made by the clinicians without the use of any decision aids. In order to assess suicide risk, in clinical approaches, a practitioner asks the patient many questions. The questions are usually not planned before the assessment and are oriented according to the answers provided by the patient. The practitioner then provides an assessment to the answers of the questions based on experience and intuition (Bouch and Marshall, 2005; Dawes et al., 1989).

Actuarial approaches are based on statistical algorithms that provide a numeric judgement to risk. The algorithms deduce relations between the data and the outcome of interest. Meehl (1954) was one of the people who initially surveyed multiple research studies and concluded that actuarial tools are an improvement on or are at least equivalent to clinical assessment. However, his research stated that clinical assessments were indispensable and were the base from which the actuarial tools were developed. The disadvantage of using actuarial methods in isolation is that it forces the assessor to provide specific information in order to make the assessment and cannot handle any criteria outside the specified set (Douglas and Kropp, 2002).

Structured professional approaches are also numerical methods that add explanatory output to explain the calculated risk. This method was created to combine the merits of both clinical and actuarial approaches. The reason behind it was to allow non-experts to perform assessments by guiding them through the steps to be performed and the minimum set of information required. This method has been shown to be more flexible than actuarial methods since it can handle extra information not specified in the minimum set (Bouch and Marshall, 2005; Dawes et al., 1989; Douglas and Kropp, 2002).

Accordingly, converting answers provided by the patients into a risk is not just a simple calculation and building a tool to perform such an evaluation is complicated. There are many obstacles that need to be addressed, such as:

1. Mental health records contain a diversity of heterogeneous sets of patients profiles. For example, there are large variations between patients who have a history of previous attempts compared to those who do not. Similarly, patients with no current intention for making a suicide attempt have fewer questions needed and are perceived as lower risk, which impacts on the motivation for asking additional questions in more general areas such as their social context. Given these differences in the subgroups of patients, each needs to be treated according to different criteria (Gopalkrishnan and Babacan, 2015).

2. There are missing data and the percentage differs from one dataset to another. The reasons for them vary, from patients choosing not to share what they consider private issues to assessors not asking questions because they think they do not relate (Ding and Simonoff, 2010; Enders, 2010; García-Laencina et al., 2010).

3. The electronic health records may suffer from having redundant information among each record which causes processing to be more complicated (Weiskopf et al., 2013).

4. Quantifying risk, for example suicide risk, can involve different number of levels. Jones et al. (2003) divided suicide risk into low, moderate and high while Bryan and Rudd (2006) recognised risk as one of five levels: non existent, mild, moderate, severe and extreme.

Such obstacles require proper methodologies to solve them. The next section will consider some of the mathematical tools for addressing the data challenges.

1.3 Predictive data mining tools

Data mining provides the basic tools to process and extract information from data. There are different tools available in this domain depending on the problem to be addressed. In the following subsections, the probable solutions to the problems mentioned in the previous section will be briefly and individually discussed.

1.3.1 Feature selection

Based on the previously described nature of electronic health records, the heterogeneous nature of the data forces the need for many features to cover the diversity of questions addressed. Also, redundancy among the features is a challenge forced by the nature of the data. Therefore, feature selection is required to reduce the number of features while maintaining the amount of needed information and reducing redundancy.

Feature selection methods choose a subset of the available questions that would best resemble the dataset. Feature selection methods include filter, wrapper or embedded. Filter evaluation is merely dependent on the relationship between the feature and the output categories with no relation to the classifier used. Wrapper and embedded methods involve the use of a classification algorithm to evaluate subsets of features. The subsets are regenerated until one that fulfils a predefined constraint is achieved (Bolón-Canedo et al., 2014; Chandrashekar and Sahin, 2014).

The choice of the method of feature selection is restricted by the data at hand. Filters are usually simpler methods than wrapper and embedded because there are multiple measures that could be used to evaluate the features. The most common are correlation and mutual information measures. Correlation measures the linearity between two features while mutual information measure the amount of general dependence between them (Bolón-Canedo et al., 2014; Chandrashekar and Sahin, 2014).

The terminology dynamic feature selection was previously exploited in other research but representing different methodologies than that used in this thesis. For instance, one method involved the application of feature selection using Markov decision process which defined states representing features selected, remaining ones and the goal set. The features were added until the goal was achieved (He et al., 2012). Another research that used the same terminology was Fern et al. (2006) where a hardware predictor was proposed to choose the most influential features from the dataset. Those techniques are not relevant to the problem addressed in this research since they require the existence of all the values in the dataset which is not the case addressed here.

1.3.2 Missing data

Missing data represent a challenging problem in data mining. The lack of a piece of information can stop a classifier from working. Thus there are common methods to handle missing data. The most simple is to delete records that are lacking any pieces of information. If the number of records with missing data is high, there is a potential for preference towards one class over the other assuming they are not a random subset of the population. This method is only applicable if the data for population members is generally expected to be complete and missing information is rare. For the mental health domain, this is not the case. (Cheema, 2014).

As the percentage of missing data increases, imputation methods were proposed. Imputation replaces the missing information by using different methods. These range from simple random or mean imputation to more complex multiple imputation. It is an effective method provided for handling missing data. The problem is assessing a risk based on imputed values could lead to giving an explanation based on answers that were not even provided in the first place. This causes problems of confidence in the outcome for assessors (Cheema, 2014; Young et al., 2011).

1.3.3 Risk prediction

Assuming the problems of selecting appropriate features and handling missing data have been resolved, there are many algorithms for classifying a record as a particular level of risk. Common ones are linear regression and decision trees. Linear regression (Neter et al., 1996) calculates the weights of each feature thus resulting in the equation that calculates the risk. Decision trees, build classification trees that are then used to assess risk (Yoo et al., 2012).

Choosing a classification method depends on a number of factors. For example, the type of data, whether it is categorical or numerical, is important. Another factor is how the output needs to be presented to the end user and, in particular, whether there must be a clearly understood explanation for it. Speed and accuracy are also usually main influences on the choice (Witten and Frank, 2005).

Selecting the best approach depends on a good understanding of both the data itself and how it is to be used for decision support. For this research, the data were collected by the Galatean Risk and Safety Tool, GRiST, a CDSS for mental health risk and safety management, which will be introduced in the next section.

1.4 GRiST support system

GRiST is a web-based software that collects information about risks of suicide, self-harm, harm-to-others, self-neglect and vulnerability. Ethics approval was obtained for analysing the results from the Department of Health as well as Aston University. The GRiST project has provided an interface to assessors to input answers to questions along with their assessment of risk levels 0 (no risk) to 10 (maximum risk) (Buckingham, 2002; Buckingham et al., 2013).

The CDSS was designed to handle different populations, currently including:

- children and adolescents;

- working-age adults;
- older adults; and
- people with learning disabilities.

It is also able to accommodate different assessment contexts, such as forensic services.

The GRiST project represented expert practitioners' psychological representation of mental health risks by a mind map, which defined the hierarchical knowledge structures they use (Buckingham et al., 2007, 2008). This hierarchy or "tree" for each risk consists of higher-level concepts such as previous history of suicide attempts, current intention, feelings and emotions. It is then divided into lower-level ones such as the pattern of suicide attempts. Finally, the lowest level leaf nodes or data are reached, such as the date of the most recent suicide attempt. For each individual patient, there will be different subsets of the tree that are relevant to them. Hence each one has a varied set of answers, which is the origin of the data analysis challenges discussed earlier. There will be a section explaining the GRiST project in more details in the following chapter.

1.5 Thesis contributions

This thesis describes the development and implementation of a new algorithm for predicting suicide risk using a database of patients' information profiles and associated clinician risk judgements. This Dynamic Feature Selection and Prediction (DFSP) algorithm accounts for high dimensional data that varies by which elements are applied to particular patients. Although it was motivated by the specific GRiST data set, the same data properties can be found in many other decision domains. Examples of such domains are survey data (Henry et al., 2013), medical data (Yoo et al., 2012) and observational studies (Xing et al., 2003). Hence there is generic applicability of the algorithm beyond mental health risk.

The DFSP algorithm needed to handle a dataset with some very challenging specifications. The dataset had a high percentage of missing data, a large number of features, unbalanced categories and was collected from a variety of assessors. The algorithm was not only expected to provide better accuracy than other off the shelf methods but also to select a feature set that could be used to explain to the assessor how the output risk was derived for each individual patient. It can then be built into the GRiST DSS to support the judgements of mental health practitioners, acting as a safety net in case they make mistakes, and to provide risk evaluations with advice for non-mental health assessors who do not have the expertise to make accurate judgements themselves.

The novelty presented in this algorithm lies in how features are selected to satisfy each patient's provided information instead of ones fitting all patients. This can be marked as a new technique to handle exceedingly large percentage of missing data and at the same time choose a feature set to represent each patient. The algorithm also presents an optimisation method that aims to speed up the assessment process. Finally it has the capability to provide the practitioner with an explanation of the reasoning behind the predicted risk.

1.6 Thesis outline

The thesis will consist of the following chapters:

- Chapter 2 will provide an introduction to the risk domain, specifically mental health risk, along with common approaches to assessing risk. It will also introduce the GRiST mental health risk, safety assessment decision support system and the problems inherent within its data set.

- Chapter 3 will give a general background on the missing data problem, feature selection methods and the predictive data mining algorithms. This will provide the essential background and context to the proposed DFSP.

- Chapter 4 will present a comprehensive explanation and rationale of the proposed DFSP algorithm.
- Chapter 5 describes the different experiments that were used to test and improve the DFSP during its evolution to the final state. It will include comparisons with alternative common approaches.
- Chapter 6 will discuss the performance of the DFSP and the methods used to improve it. It will define the final version of the algorithm.
- Chapter 7 will test the DFSP on risk of harm to others to assess its ability to generalise across risks.
- Chapter 8 will summarise the research and consider next directions.

Chapter 2

Background

2.1 Introduction

Everyday people take risks as part of their normal life such as climbing stairs, driving a car or playing sports. These risks are weighted up against their benefits. For example, driving to work can be risky but there is no other way to earn a salary. Therefore, risks are intuitively balanced against the benefits that justify them (Flewett, 2010).

Mental health risks are different. They are not so much choices as products of mental health problems. Assessors have to judge whether the problems are creating risks and, if so, what level of risks these are and how should they be managed. There are different forms of mental health risks such as harm to others, neglect or self-harm, which includes trying to kill oneself (complete suicide). Sometimes these assessments are performed by clinicians with limited experience, resulting in a high percentage of uncertainty when predicting outcomes. Assigning the wrong treatment to a high risk suicide patient could end up with the patient succeeding in completing suicide. Similarly, setting a high risk of violence patient free could cause harm to others (Dixon and Oyebode, 2007).

Minimising mental health risk is a necessity for both patients and healthcare providers. For patients it is needed to maintain their well being and in some cases the safety of people surrounding them. For the health provider, it reduces costs and improve one's reputation in managing risk. Any effective mental healthcare system has to have good assessors of risk in order to be able to manage mental health properly (Flewett, 2010). Research in this domain is not expected to replace the work of human assessors but to assist and provide good explanations in order to aid the risk management stage and eventually reduce the risk involved (Department of Health, June 2007).

In the next sections, mental health and safety will be discussed, including different approaches to assessing risk, clinical difficulties, and current available tools for helping with it. The conception and development of the Galatean Risk and Safety Tool (GRiST) will be discussed in detail because it is the one that generates data used in this research. The idea is to implement an algorithm for predicting risk within GRiST, which means considering the knowledge structure, user interface, and challenges generated by the nature of the data.

2.2 Mental health risk and safety

One of the main objectives stated by the UK in the Department of Health (2011) is to improve the mental health service provided by the government. Mental health problems have a negative effect not only on the people facing them, but on the families, carers and their communities. Inaccurate assessment and management of mental health can lead to making life even harder. For example, people might be considered unqualified to find and preserve jobs which would result in earning less money. This scenario becomes more complicated if patients then cannot support themselves which puts them at risk of becoming homeless.

The Department of Health (2011) stated the objectives for mental health management as:

1. Improving mental health for a wide range of people.

2. Aiding people with mental health problems to overcome them.

3. Decreasing the rate of death caused by mental illness.

4. Better care and support coverage.

5. Better behaviour towards people with mental health problems.

The Department of Health (June 2007) defined a guide for practitioners to make proper risk analysis and how different organizations should choose mental health tools. The document builds a framework for practitioners, healthcare providers and service users to make sure that communication among different health care providers is well established.

Risk assessment systems are intended to help practitioners make decisions, not to replace them. If care providers combine their knowledge about what a risk assessment system has to offer then the provided service would have a much improved standard (Department of Health, June 2007). However, managing the risk of mental health is not an easy task (Flewett, 2010). The objectives set by the Department of Health (2011) cannot be met while depending only on the power of clinicians and mental health providers. Having a proper system to guide an assessment will eventually lead to improvement of the mental health service. Such a system would provide guidance for less qualified personnel to make proper management decisions and will result in a wider range of coverage to people in different locations.

2.2.1 Clinical problems

Mental health care faces many problems when it is being practised. Although patients are always at risk of a wrong assessment and management which could be fatal, assessors and the mental health service providers also face different kinds of risk (Flewett, 2010). Clinicians may need to handle many problems, such as (Vail et al., 2012):

- difficulty in assessing some patients for having a different language or culture;

- risk of being attacked from a violent patient; and

- risk of earning a bad reputation for making a wrong judgement.

Problems faced by clinicians not only affect themselves but also have an effect on patients and mental health providers. Mental health providers on the other hand have to handle difficulties such as (Flewett, 2010):

- cost of providing the service;
- legal costs in case of mistakes;
- need for more practitioners to handle more cases; and
- risk of earning bad reputations for wrong assessments, or for neglect of critical cases.

The key to addressing these problems is the early detection of a mental health issue and its association with a good treatment plan: proper assessment leads to proper management. Non-professionals providing emergency services sometimes have to perform assessments. They are more likely to give a wrong assessment, which can lead to inappropriate risk management and potentially fatal consequences (Department of Health, June 2007; McCrone et al., 2008; Yang and Lester, 2007).

2.2.2 Approaches to risk assessment

Finding the optimal method to assess a mental health patient is usually dependent on the experience and accuracy of the practitioner (Dawes et al., 1989). The mental health practitioner collects information by asking the patient different questions and then evaluates the answers provided. There are three methods of assessing risk: clinical, actuarial and structured professional judgement (Bouch and Marshall, 2005; Dawes et al., 1989; Dixon and Oyebode, 2007).

Clinical approach

Mental health judgement is based on the clinician's experience and intuition. This approach is usually criticised for being based on feelings rather than evidence. The disadvantage of this approach is that evaluation of risk is merely dependent on the assessors. Given differences in their expertise, it may vary from one to the other. Therefore the same patient may be given different risk evaluation assessments when reviewed by more than one clinician (Bouch and Marshall, 2005; Godin, 2004; Littlechild and Hawley, 2010).

Actuarial approach

Assessment is based on algorithmic methods and known statistical procedures for risk classification. This could be reliable if the clinician is lacking experience but it does not give the assessor proper feedback on the reasoning behind such an evaluation. The approach analyses data previously provided on a group level and cannot handle exceptional individual cases. Another problem of this approach is that it cannot note the fluctuations in the factors affecting risk, as they change, since the evaluation is based on static factors. Finally, limited amount of epidemiological data would hinder the creation of a comprehensive actuarial tool (Bouch and Marshall, 2005).

Structured professional approach

The structured professional approach is a combination of clinical and actuarial approaches but aims to overcome their shortcomings. It operates dynamically to identify the factors affecting the risk, starting with common factors but changing as information is added. The merit of this methodology is that it allows clinical intervention to re-evaluate risk accordingly. The structured approach can also provide guidance to the clinician to produce more accurate judgement with proper reasoning underpinning it (Bouch and Marshall, 2005; Brown and Singh, 2014).

2.2.3 Current tools

Currently there is a limited number of tools for handling mental health risk. The Department of Health (June 2007) discussed some of the tools ranging from specialist ones addressing risk of violence only to others addressing multiple risks including violence, sexual violence, antisocial behaviour, suicide and self-neglect. Most of the available tools help in structuring the information collected from clinical approaches.The actuarial tools were mostly related to violence and sexual violence risks. The document shows that there are no tools that can evaluate the level of suicide risk of mental health patients, which is the main concern of this thesis.

Silverman and Berman (2014) discussed how risk is assessed and accordingly management is planned. The document gave a historical view of the work done on suicide risk assessment through the years. The historical view was followed by an explanation of different suicide risk levels ranging from low to high. This research reviewed work that has been done to build methods to assess suicide risk. They concluded the need for more research that would lead to a clinical tool to assist clinicians in assessing suicide risk.

Recently Lotito and Cook (2015) searched the latest articles that were released in MEDLINE journals using suicide and risk keywords. They reached the same conclusion that there is no tool for assessing suicide risk. One is needed that provides evidence for the reasoning behind an assessment is arising but this remains a challenge. Chesin and Stanley (2013) confirmed the same problem of a current lack of actuarial or structured tools to assess suicide risk and have discussed the presence of such tools in assessing violence.

In brief, suicide risk assessment is currently dependent on the clinician’s expertise. Some emergency situations require people without enough expertise to perform such an assessment. In these cases, lacking tools that would aid in making risk decisions could be fatal. Building such a tool would require examining data provided by mental health patients then trying to construct a methodology generating an accurate evaluation of risk. GRiST provides a sophisticated tool that attends to mental health risk understanding and provides the data used in this research. The upcoming section will give more details of GRiST initiation, structure, user interface and data problems to be handled by it.

2.3 Clinical decision support systems

Clinical Decision Support Systems (CDSSs) initiated as computer systems trying to imitate the way humans think. Eventually the simulation of human thinking evolved into assistance in making decisions. Currently, CDSSs can guide the user with information on how to reach a specific assessment. There are two types of CDSSs, those with or without knowledge. Knowledge based CDSSs often use a knowledge base of stored if-then rules, an inference engine to combine these rules and a communication technique to make the user interact with it. In other cases, the knowledge base might store the different symptoms associated with their probable diagnosis (Berner, 2007).

On the other hand, non-knowledge based CDSS does not require a knowledge base. Instead it learns from previous data and creates models to handle new input. There are different and sophisticated non-knowledge based CDSSs that are constructed using neural networks and genetic algorithms (Berner, 2007). An example of such CDSSs is the Clinical Assessment of Risk Decision Support (CARDS) (Watts et al., 2004) which can assess and help manage risk of violence.

Despite the strength of CDSSs, they are not yet widely employed in healthcare practice. The reason is partly the difficulty of integrating them with the customary way employees and organisations work. Another reason is that it is not possible to add information that would exceed the predefined and usually limited scope of the CDSS. GRiST addresses these problems by representing information in an easy format that is comprehensible to users and by building an interface that can be flexible enough to tolerate addition of textual notes for adding extra information when required (Buckingham et al., 2013; Shibl et al., 2013). The next section introduces GRiST in detail.

2.4 The GRiST CDSS

The GRiST CDSS is based on a psychological model that was developed by Buckingham (2002). It is a sophisticated web-based CDSS used to evaluate risk (GRiST, 2016). It was founded on learning how experienced practitioners use information provided by mental health patients and convert it into a risk assessment (Buckingham, 2002).

The future aim of GRiST is to provide the benefits of (Buckingham, 2007):

- advance discovery of mental health issues, which would lead to better intervention;
- enhanced evaluation of risk;
- ability to make appropriate referrals;
- patients assessing themselves; and
- an improvement in mental health quality within society.

2.4.1 Initiation of GRiST

The research began with 46 fully qualified practitioners having a minimum of two years experience. Multiple disciplines were included to gain the experience of the consensus, those disciplines involved psychiatric nursing, psychiatry, social workers, general practitioners and psychologists. Through the evolution of GRiST, more experts were enrolled which resulted in a total exceeding 100 clinicians and service users (Buckingham et al., 2007).

The creation of GRIST was partly based on interviewing the practitioners and using the transcripts of interviews to build the knowledge structure. Another part was based on discussions with focus groups and revising the system with feedback from practitioners using it. Eventually the data needed were recognised along with their relationship to each other levels of risk within the practitioners' conceptual model of suicide risk. GRiST introduced a structured model representing clinical judgement by extracting and combining the expertise of multiple experts followed by presenting it in a language that they can understand (Buckingham, 2007; Buckingham et al., 2008).

Each interview with an expert was converted into a mind map, then all interviews were combined into one marked by expert's identification numbers so that the original interview can be referenced when needed. The mind maps were stored as XML, thus the information was defined in a structured format that can be easily handled by different machines and programming languages. The XML enabled the easy accessibility to information by XSLT queries. A query could generate the sub tree for a specific risk or a specific population. (Buckingham et al., 2007).

The combined mind map represented a large tree of the combined knowledge of the consensus, but size was a drawback. Initially the number of nodes in the complete structure was 7210 from which 1439 were unique. The unique nodes were divided into 477 concepts (i.e having children nodes) and 962 leaf nodes. Clinicians prefer using assessment tools that would not involve the use of too many questions; therefore tree pruning was the next stage. The pruning of the tree was done on multiple stages by annotations to the XML by focus groups with the Flash interface visually recording their opinions. After the tree was pruned it resulted in 394 nodes in total with 124 concept nodes and 228 unique leaves. The final version of the XML hierarchy was the building block on which the GRiST CDSS was built. The GRiST CDSS provided an interactive interface for collecting the data that is needed for the assessment of mental health risk (Buckingham et al., 2007).

2.4.2 The Galatean model

In GRiST, trees (named galateas) were initially constructed to represent the flow of knowledge for different risk levels where each root of a tree is a specific risk that is partitioned through different levels down to the leaf nodes, which represent the input data. The input data can be represented in any format but is converted to a membership grade (MG) with a value ranging between 0 and 1. This membership is assigned by the assessor as a quantification of the answer provided by the patient to a certain question. Thus the MGs are a representation of the data provided by the patients in the real world but in a numerical format that can be used in modelling input (Buckingham, 2002; Buckingham et al., 2013). MGs have a monotonically increasing relationship to risk: the higher the MG, the greater it contributes to the level of risk.

This model shows how a mental health practitioner uses the information provided accompanied by the output risk's conditional probabilities to assess risk. An advantage of using such a model is that it aims at boosting the assessment accuracy by finding the connection between practitioners expertise and data models. The model initially represented a large number of questions as nodes in the tree. Each node having a value to represent its relative influence (RI) on the output assessment. Expecting the practitioners to supply all the information required was not possible. At this stage, the information represented at the leaf nodes was represented by the MGs which were deduced from the expert's entry. The aim was to propagate these membership grades through the tree according to different node influences (relative influence) so that the risk is finally deduced. Therefore, further research was performed to provide the calculation of the relative influence of the different nodes (Hegazy and Buckingham, 2009).

Figure 2.1 shows how MGs are propagated to calculate the suicide intention of a patient from a substructure of GRiST, given three datum components: seriousness, realism and steps taken. The patient's value for each component is multiplied by the RI to calculate the

Fig. 2.1 Usage of membership grades to calculate suicide risk intent, reproduced from (Buckingham, 2002)

resulting MG. The calculated MG is passed up the tree to calculate the intention MG. For example, the patient's steps taken value in Figure 2.1 was evaluated by an expert as 8 (given that 0 is lowest and 10 highest) which provides MG value 0.8. The MG value is multiplied by its RI value of 0.4 to produce a MG 0.32. At the same time, the realism was assessed to be a value of 7 which results in a MG of 0.7. The produced value from its multiplication with its RI is 0.42 which is added to the 0.32 produced from the steps taken to produce a result of 0.74. This represents the MG for the plan/method concept which has a RI of 0.3 resulting in a value of 0.222. For the seriousness datum, a MG of 0.6 is multiplied by RI of 0.7 to pass a value of 0.42. This MG is finally is added to that produced by the plan/method concept to produce the MG of the intention which is 0.642. This makes the patient having an approximate intent of 6 to complete suicide given a range of 0 to 10 (Buckingham, 2002).

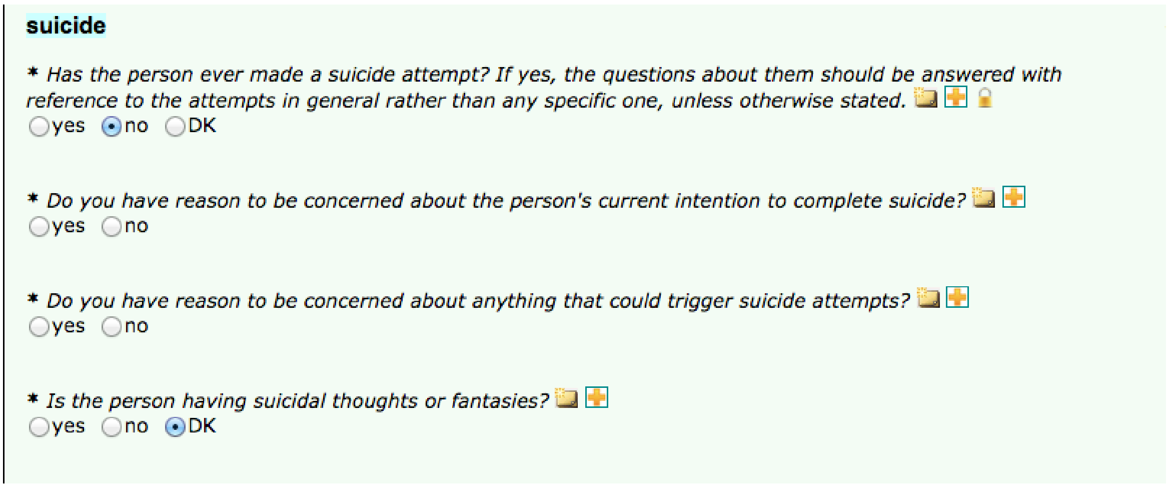
2.4.3 User interface

The main concept behind the galatean CDSS was to ease the interaction between the computer and assessors. The building blocks of the system were three resources:

1. a database of patients' features with experts' risk analyses;
2. analysis of the dataset using statistical pattern recognition tools; and
3. an expert risk model that would describe how features affect risk.

The outcome of combining the building blocks was an interactive tool providing a graphical understanding of the clinical expertise. The interface shows how the input features are converted into risk level evaluation (Buckingham, 2002).

The interface of the GRiST tool (GRiST, 2016) is shown in Figures 2.2 and 2.3. The concept nodes are represented as filter questions and leaf nodes are non-filter questions. The filter (see Figure 2.2) questions, written in italic, are questions that produce subsequent ones depending on the provided answer. For example, a question like "Has the person ever made a



suicide

* *Has the person ever made a suicide attempt? If yes, the questions about them should be answered with reference to the attempts in general rather than any specific one, unless otherwise stated.*
○ yes ● no ○ DK

* *Do you have reason to be concerned about the person's current intention to complete suicide?*
○ yes ○ no

* *Do you have reason to be concerned about anything that could trigger suicide attempts?*
○ yes ○ no

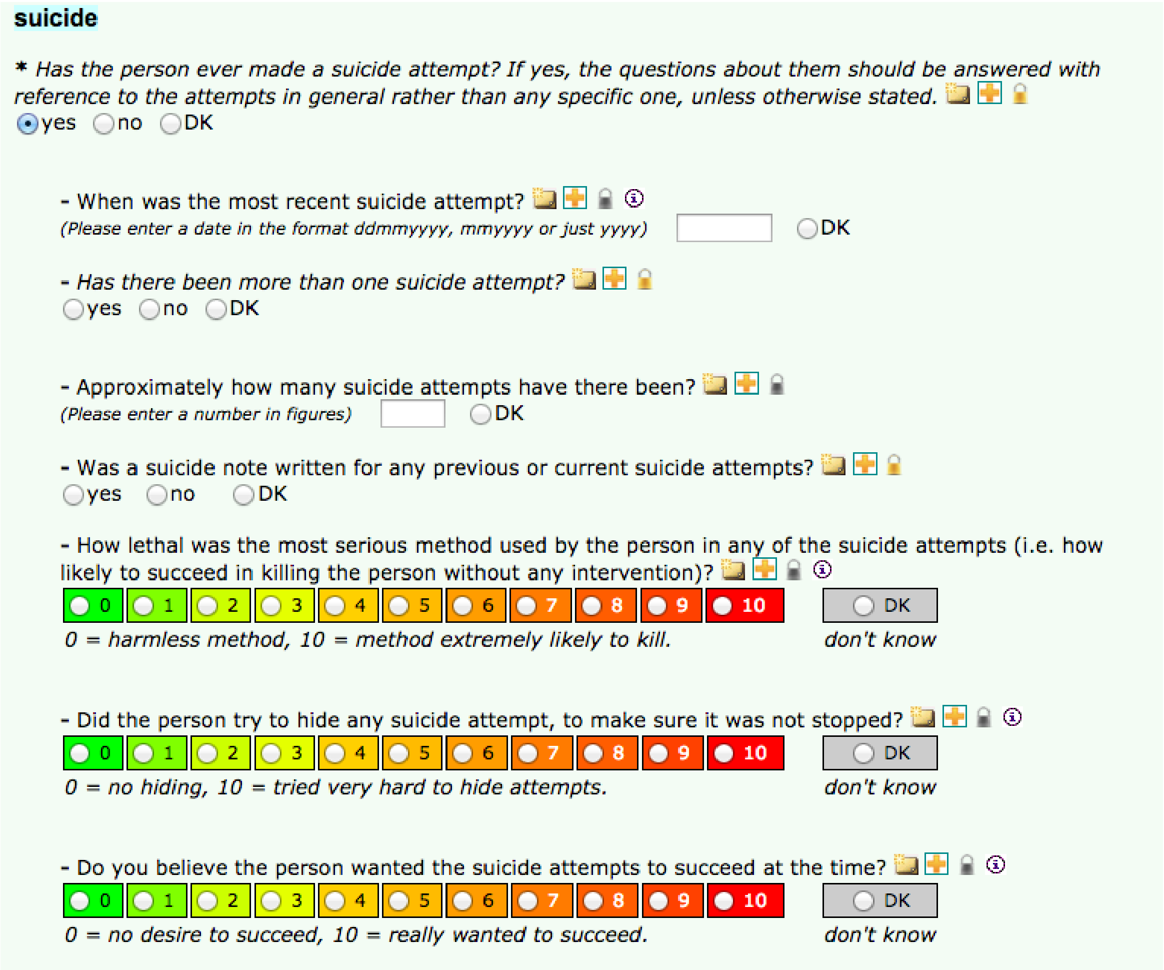
* *Is the person having suicidal thoughts or fantasies?*
○ yes ○ no ● DK

Fig. 2.2 Snapshot of the user interface showing filter questions

suicide attempt?" would not have any subsequent question if it was answered with a "no" or "don't know", but if it was answered by a "yes" then more details about the attempts would be required as shown in Figure 2.3.

Non-filter questions have a variety of answers depending on the type of question asked. Some require a "yes", "no" or "DK" answer while others require a value ranging from 0 to 10 where 0 is the least influence and 10 is the highest along with a "DK" option. All this can be seen in Figure 2.3

This interface was used by clinicians to collect information regarding each patient individually and add to it the assessed level of risk. The data were collected by clinicians using the GRiST clinical decision support system as part of their normal practice. All risks are automatically anonymous as part of the collection and data storage implementations. The data collected by the GRiST interface, specifically for suicide risk, are the data analysed and used to create the proposed algorithm presented in the upcoming chapters.



suicide

* *Has the person ever made a suicide attempt? If yes, the questions about them should be answered with reference to the attempts in general rather than any specific one, unless otherwise stated.*
● yes ○ no ○ DK

- When was the most recent suicide attempt?
(Please enter a date in the format ddmmyyyy, mmyyyy or just yyyy) ☐ ○ DK

- *Has there been more than one suicide attempt?*
○ yes ○ no ○ DK

- Approximately how many suicide attempts have there been?
(Please enter a number in figures) ☐ ○ DK

- Was a suicide note written for any previous or current suicide attempts?
○ yes ○ no ○ DK

- How lethal was the most serious method used by the person in any of the suicide attempts (i.e. how likely to succeed in killing the person without any intervention)?
○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 ○ DK
0 = harmless method, 10 = method extremely likely to kill. *don't know*

- Did the person try to hide any suicide attempt, to make sure it was not stopped?
○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 ○ DK
0 = no hiding, 10 = tried very hard to hide attempts. *don't know*

- Do you believe the person wanted the suicide attempts to succeed at the time?
○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 ○ DK
0 = no desire to succeed, 10 = really wanted to succeed. *don't know*

Fig. 2.3 Snapshot of the user interface showing filter and non filter questions

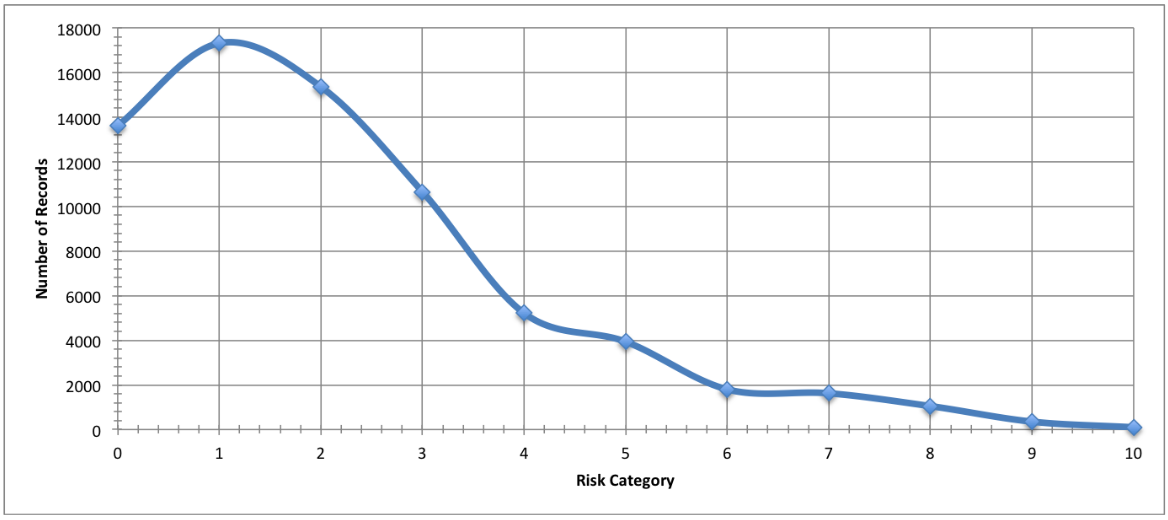


Fig. 2.4 Distribution of the number of records per risk category

2.4.4 Problems to be handled by GRiST

Although the creation of GRiST solved the difficulty in communication across service boundaries and provided a language to address and understand risk, the dataset generated had some complications that needed to be addressed. The design of the interface using filter questions where the subsequent questions are shown only if the answer is yes provided an advantage for the users but caused problems for analysis.

The first problem is having a large number of questions that result in many features needing to be analysed. The increase in the number of questions might be considered a method to obtain more information but the problem is that this increase will result in more processing and computational requirements.

At the same time, the distribution of patients among the different risk levels from 0 to 10 is not equal. Figure 2.4 shows the number of patients in each risk level with a big difference between the count of records in high and low-risk patients. The problem here is that classifiers can be baised towards low risk simply because they are more probable where the average value of the risk is 2.186.

The clinicians collected patient data and provided their own judgement of the suicide risk, giving a value between 0 and 10, (one of 11 levels). The data were collected from almost 3000 different physicians working in various organizations, with different training regimes and patient population characteristics. This heterogeneity is testimony to the difficulty in building a single model that can accurately represent classification behaviour across all patients and clinicians.

Another problem is that there are two types of absent data. A certain piece of information could be missing based on the choice of a patient not to answer a question or could be blank because it is not relevant to the case. For example, the history of suicide episodes for someone who never attempted it is not relevant. Alternatively, information maybe relevant but not have been supplied for the assessment. Analysing the GRiST dataset has shown that there is not a single feature that does not have missing cells. The missing percentage among all features range from 0.23% to 98.55% of all patient records. Only 20 features from 176 have missing data percentage less then 10% of all the records. Using only these 20 features, if records with missing cells were deleted, 37% of them would be omitted. This shows that there is no easy way to match patients to a set of features that are guaranteed to be present, then an alternative solution had to be found.

To summarise, the GRiST suicide data set is challenging because the percentage of missing data is 67.04% where none of the patients records are complete. The risk population is not equally distributed and is divided into 11 levels. The number of features in the data set has very high dimensionality of 177, including the risk level variable. In mitigation, the total number of patients records is 71,024 and is increasing by 500 a week. Although GRiST is a very powerful CDSS, it does not generate its own risk assessment for a patient. An algorithm is needed to provide a complete tool to assessors of different domains and at the same time cope with all the problems expected in a mental health dataset like the ones in GRiST. The goal of this PhD is to create an algorithm that can exploit the GRiST database of experts judgements to generate a reliable prediction of its own for non experts.

2.5 Summary and conclusions

There are multiple risks that exist in the mental heath domain. Such risks include, harm-to-others, harm-to-self or neglect. Clinicians are expected to evaluate patients and give an assessment of the risk. This evaluation is done based on human expertise but in some cases its deficiency could have fatal outcomes. In other scenarios, assessors having the required expertise could use a second opinion or guidance on certain patients.

Many clinical problems are involved with mental health. If mental health problems are not detected early, then there are threats that would not only involve the patient but also practitioners and mental health providers. The early detection and management of mental health leads to a better quality of life to society.

Different approaches are used to assess risk. A common one is the clinical approach where evaluation is done only based on assessors' experience and intuition with no aid from any external tools. The problem with this approach is that a different assessment might be given to the same patient if different clinicians were employed. Another approach is actuarial methods where assessment is based on mathematical and statistical formulas. This method has a shortage in explanatory power, thus an assessor would be faced with a risk but no reasoning on why this risk is the most suitable. Also assessment in actuarial approach is based on static factors and would not accept the dynamic introduction of new information that could be considered vital. Overcoming the downsides of both methods is the structured professional approach where clinical assessment would be combined with computational power. This approach provides reasoning behind assessments, allows for clinical intervention and is more flexible for using dynamic data than actuarial approaches. Currently, in order to assess suicide risk, there is no definitive actuarial or structured professional tool available.

This thesis is based on the GRiST project, which was initiated with the aim of improving mental health services, giving a better coverage for early detection of mental health problems by assessors and eventually providing a tool that enables patients to conduct self assessments. The research achieved in GRiST initially created a tree that would mimic the understanding of clinicians on how to evaluate risk. The trees represent knowledge flow from risk assessment levels down to leaf nodes.

The data collected by GRiST suffers from multiple challenges.The large number of features require heavy processing and computational powers, including addressing missing or irrelevant data. The distribution among different risk levels is skewed to lower risk which makes lower risk classes more probable and higher risk harder to classify. Finally classifying risk into 11 levels is difficult because of the high granularity (subtle distinctions between risk levels) .

Assessment of suicide risk is a challenging and important area of research that needs to be further explored. Silverman and Berman (2014), Lotito and Cook (2015) and Chesin and Stanley (2013) have discussed the current lack of actuarial or structured professional tools that would provide an untrained assessor well explained guidance. GRiST is a strong support system that could meet this need but only if it addresses the challenges of its dataset. In the next chapter these challenges will be explored, including background on missing data, feature selection and prediction algorithms then a new algorithm will be presented and explained.

Chapter 3

Prediction Methods

3.1 Introduction

Data mining provides different techniques for use in analysis and prediction. It discovers relationships among data and specifies tools that could be used to create models of large numbers of patients records but it is not yet fully explored (Bellazzi and Zupan, 2008; Shibl et al., 2013; Weitschek et al., 2013). Yoo et al. (2012) stated that the application of data mining on healthcare emerged in the 1990's.

Data mining is driven by the data being addressed. The specifications of data determine the required steps. For instance, datasets having a large number of features might require a reduction in this number if, for example, some of them represent redundant information. This problem is addressed by feature selection where a subset of features is chosen without changing their format. Another method is feature extraction where the existing set is transformed into smaller sized vectors representing them but in a different, often incomprehensible, format (Chandrashekar and Sahin, 2014; Krzysztof et al., 2007).

The choice between feature selection and feature extraction depends on the application itself. Some application domains prefer to keep features in their original format and not apply alterations to it (Guyon and Elisseeff, 2003; Jain et al., 2000). In the work presented in this thesis, the methodology was based on feature selection since the format of the features was preserved for reasons that will be discussed in following chapter.

Another specification of the dataset addressed is absence of information. Medical datasets have large proportions of missing values in their records. Analysing reasons behind missing data is important because they show the impact of bias that may be caused by ignoring their existence. The methodology of handling missing data is vital (Ding and Simonoff, 2010; Enders, 2010; García-Laencina et al., 2010; Myrtveit et al., 2001) and will be discussed in more detail in this chapter.

The last stage, after missing data are handled and features selected, is the classification stage where patient records are classified into a specific category according to a model built on a training set. This model is expected to represent data reliably without causing over-fitting or performance degradation. (Yoo et al., 2012).

Predictive data mining is a subdomain of data mining, which uses a large amount of data to predict a certain piece of information that would provide support to the CDSS. Predictive data mining can be used in different fields of the medical domain like prognosis or diagnosis or even plan a course of treatment for patients. (Bellazzi and Zupan, 2008; Weitschek et al., 2013). Bellazzi and Zupan (2008) presented how predictive data mining methods should be compared according to multiple factors such as:

- how missing data is treated;
- types of variables handled (categorical or numerical);
- level of understanding of generated models;
- feature selection methodology;
- computational cost;
- possibility of explaining output decisions; and

- comparisons of results when applied on different datasets.

In this chapter, the basic steps of predictive data mining are explained. First there will be a subsection about the problem of missing data along with reasons behind this problem and common ways to solve it. The second part will discuss feature selection as well as common methods of selection. Third part will discuss predictive data mining, which will include a brief explanation of common predictive tools such as regression and decision trees.

3.2 Missing data

Choosing a treatment methodology for missing data is dependent on the problem and how it is affected by the advantages and disadvantages of the selected methodology. Some datasets may work well with the deletion of records at the price of losing information or by using imputation at the price of increasing the computational cost (García-Laencina et al., 2010; Myrtveit et al., 2001). In the upcoming subsections, reasons behind missing data will be explained followed by their types and how they are usually handled.

3.2.1 Why is data missing?

Osborne and Overbay (2012) divided the reasons behind missing data into legitimate and illegitimate. A legitimate reason is, for example, sub questions that would concern an event that did not happen. For example a question like "did you commit suicide before?" If the answer is "yes", then probably it will be followed by questions involving the number of attempts or the last date of attempted suicide but if it was a "no" then those sub questions would be legitimately discarded. Another example would be if a certain patient has moved to another country and thus stopped attending follow up sessions, which means part of their record would be legitimately missing. In these cases, replacing missing data would not be appropriate.

On the other hand illegitimately missing data may be caused by a number of reasons. For example, it might be caused by an electronic failure from sensors or any data entry device. Another reason would be patients choosing not to answer a certain question that they consider private or have forgotten a certain piece of information. Also, it could be caused by a question that might have been forgotten by an interrogator or a respondent (Cheema, 2014; Osborne and Overbay, 2012).

3.2.2 Types of missing data

Knowing the mechanism of missing data is very important since it gives a clue on how to handle them. There are three types of missing data: Missing Completely at Random (MCAR); Missing At Random (MAR); and Missing Not At Random (MNAR). Data MAR are not linked by relationships or patterns between items in the dataset. Instead, they could be MAR due to external influences, which might only affect certain variables but not in any systematic way. MNAR data have a relationship between missing items and the sample set, such as assessors skipping a certain question for a particular group of patients because of consequences of the answers. Finally, data MCAR are those where there is no relationship between missing data items or the output classes (Allison, 2001; Enders, 2010; Little, 1988; Little and Rubin, 2002; Myrtveit et al., 2001).

Both MCAR and MAR represent missing data that could be ignored. On the other hand, MNAR cannot be ignored and needs to be modelled to represent knowledge of missing patterns (Little, 1988; Little and Rubin, 2002; Osborne and Overbay, 2012). Since the GRiST dataset is collected from multiple experts on a large variety of patients, missing data can hardly be considered MNAR. Collecting the data by multiple experts reduced the effect of having certain assessors biasing towards factors such as race or gender even if it existed. This was also noted by the scatter in the missing data through out the whole dataset in all features and all records. Figure 3.1 represents the missing data plot for the complete dataset with the

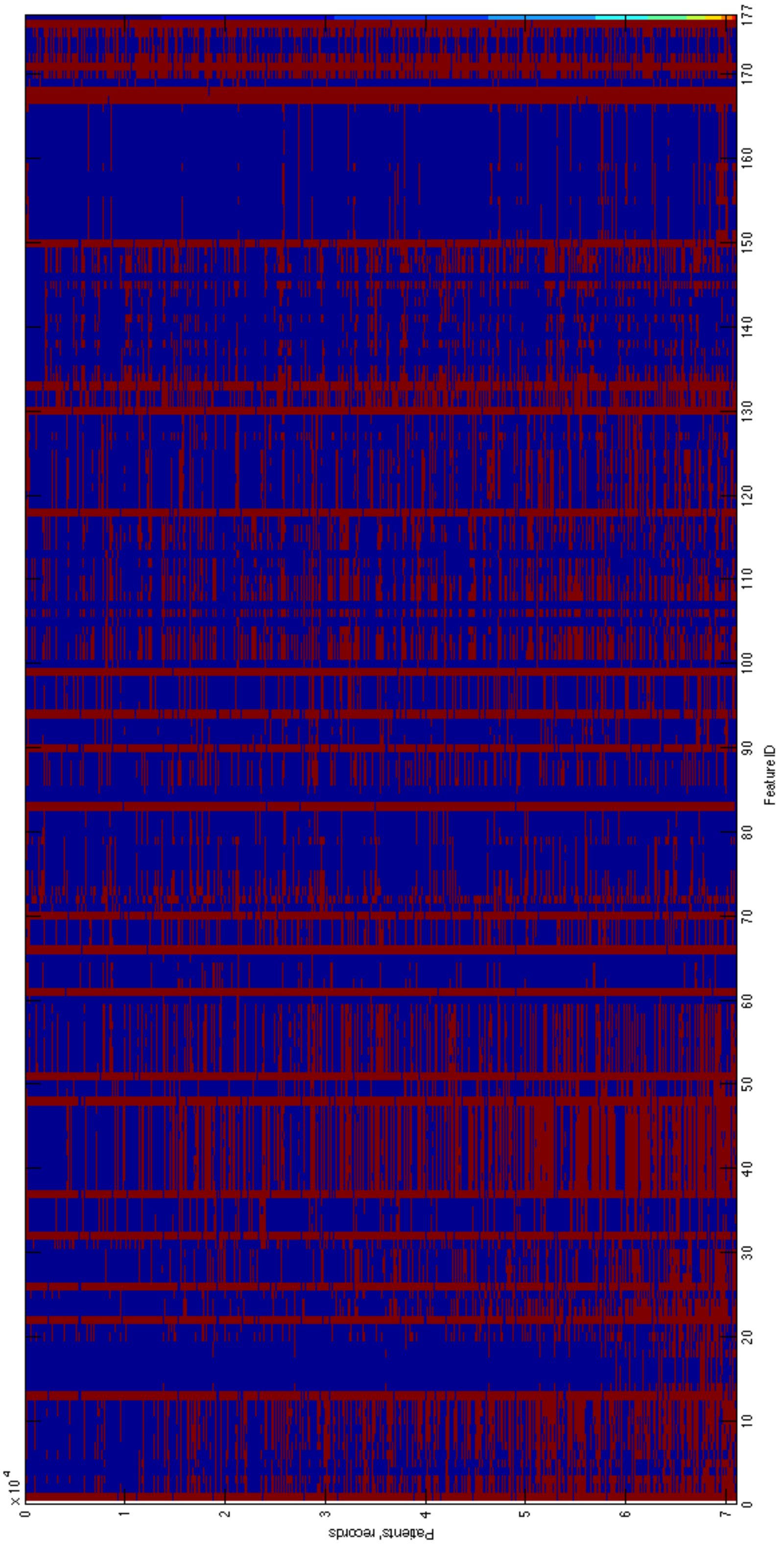


Fig. 3.1 Missing map of the GRiST dataset with the colour blue indicating missing values and red for existing ones

colour indications of blue for missing and red for existing values. The rows represent all the 71024 patient records and the columns represent the IDs of the 177 features. In this figure, the rows are ordered according to the risk level column (having feature ID of 177) to be able to view patients having similar risk together. Blue colour indicated risk 0 and the colour changes as risk elevates. Analysing each level of risk separately presented the same scatter of missing features among different records. There is not a single feature that is missing for a specific risk and existing for all others. Thus in this research, the choice of missing data handling methodology will be based on the assumption that data is either MCAR or MAR where in both cases it is ignorable.

3.2.3 Handling missing data

There are two main categories into which missing data methods are divided, either case-wise deletion or imputation (Cheema, 2014). Both methods will be discussed in this subsection.

Case-wise deletion

There are two ways to delete records from a dataset, list-wise deletion and pair-wise deletion. List-wise deletion is a more traditional way of handling missing data (also known as complete case method). It involves the removal of incomplete cases (Baraldi and Enders, 2010; Cheema, 2014). Young et al. (2011) stated that methods that involved removing missing data are regarded as the simplest and computationally least expensive. On the other hand, the disadvantage of list-wise deletion is that it could cause a bias if the dataset size is small since deletion of records could disturb the population distribution. Thus they concluded that this method works best when the size of the dataset is very large and have a small percentage of missing data that is MCAR or MAR.

Pair-wise deletion, on the other hand, does not delete the whole record whenever a cell is missing. It can be applied when the calculation required is between two variables at a time.

All the complete cases of the two variables addressed are used regardless of the missing values in other variables (Baraldi and Enders, 2010; Cheema, 2014; Graham, 2009). When compared to list-wise deletion, for some of the attributes, it computes a less biased and more precise statistic since it can use more information that would have been omitted by list-wise deletion if other variables of the same records were missing. Therefore, it uses more of the information available in the dataset. On the other hand, pair-wise deletion has limited usage when defining methods since it removes the missing instances for two variable, at a time, to calculate relationship between them. Therefore it cannot be applied on methods that work with multiple features at a time. Example of using it is in finding correlations between variables (Graham, 2009; Young et al., 2011).

Imputation methods

Young et al. (2011) divided imputation methods into three groups: non-statistical, statistical and machine learning methods. Non-statistical methods do not require computations to impute the missing values.

Examples of non-statistical methods are Last Observation Carried Forward (LOCF), hot deck imputation and random imputation. LOCF completes missing values by duplicating values of the same variable from the previous sample. It best works when the records are ordered in a logical format such that there is little difference between a sample and its previous. The LOCF is also best used when the variation within a variable's values is small (Cheema, 2014; Young et al., 2011).

Hot deck imputation involves finding the closest set of donors that match a record with a missing value based on common predefined characteristics which could be found by random selection or using deterministic metrics. A donor from this set is randomly chosen and the value for this feature replaces the missing one in the record (Krzysztof et al., 2007). There are a couple of variations of hot-deck imputation explained in Cheema (2014) but it has not yet fully evolved compared to other methods and requires the existence of close matches of samples within the dataset.

Random imputation is, as the name implies, a random pick for the missing variable from all the other samples. This method cannot work on data having a high percentage of missing because it will result in a distortion among the variables since it does not put into consideration relationships with other existing features (Young et al., 2011).

Imputation methods could also be statistical such as mean imputation. Mean imputation substitutes the missing feature by the average of other existing values of it. This method is computationally simple when it comes to imputation but lowers the standard deviation of the dataset, which affects the importance of other attributes (Baraldi and Enders, 2010; Krzysztof et al., 2007; Young et al., 2011).

More complicated methods include multiple imputations, resulting in several datasets with alternative imputed values for the missing variables. Each set is passed to the classifier and results are merged to produce a consensual classification. Another example is the K nearest neighbour algorithm (KNN) where a subset of the dataset is chosen according to a distance measure to the sample with missing data. The replacement value is provided according to a predefined criterion, sometimes the mode for discrete values or mean for continuous, and other times by applying weights according to their distances. These methods are computationally more expensive and their effect is dependent on the problem structure (Acuna and Rodriguez, 2004; Baraldi and Enders, 2010; García-Laencina et al., 2010).

Machine learning methods are considered the most powerful and computationally expensive. One of the commonly used methods is Expectation Maximization (EM). It repetitively tries to find the maximum likelihood of a missing value which is then used to infer latent variables in following iterations. The aim behind the approach is not to add any knowledge to the data but to try and retain relationships between samples (Graham, 2009; Young et al., 2011).

The GRiST dataset has a high percentage of missing data which represents 67.04% of the whole dataset. Applying list-wise deletion will result in deleting the whole dataset, knowing that not a single record in GRiST is complete. Imputing missing values would mean that an answer to a question that the patient might have chosen not to answer is manufactured. This cannot be accepted since one of the objectives of the system is to explain the output evaluation of risk and this cannot be based on imaginary data. The problem here is that applying standard methods discussed in this section is not possible. Recent research need to be reviewed.

3.3 Recent work on missing data

Choosing the methodology of handling missing data depends on many factors. Osborne and Overbay (2012) performed a survey with the American Psychological Association on how missing data were handled. The result of this survey was that 61% did not discuss the presence of missing data which is supposedly not rare in this domain. The authors expected that the data were cleaned by removing any records with missing cells or were handled by an imputation method that was not reported. On the other hand, the ones that did report the presence of missing data mostly handled it by list-wise deletion or simple imputation because their data usually had a very low percentage of missing items. Problems with handling missing data occur when common statistical techniques expect a complete dataset or can only handle minimal amount of incomplete records (Osborne and Overbay, 2012).

Schlomer et al. (2010) surveyed multiple research on missing data and concluded that experts have not yet defined the percentage of missing data that is considered problematic if exceeded. Researchers have varied this percentage to range from 5% (Schafer, 1999) to 20% (Peng et al., 2006). Schlomer et al. (2010) also added that there are a lot of researchers who fail to report the existence of missing data in the first place and how it was handled.

Another survey about handling missing data in survey-based research was presented by Karanja et al. (2013). They confirmed what was stated by Schlomer et al. (2010) by reviewing 749 surveys in the period between 1990 and 2010 and concluding that most researchers fail to discuss the presence of missing data and methods of handling them. When missing data happened to be mentioned, researchers tended to use simple methods like list-wise and pair-wise deletion. Karanja et al. (2013) presented a table comparison between different methods of handling missing data along with their advantages, disadvantages and recommendations. They recommended using expectation maximization and avoiding the usage of list-wise deletion, pair-wise deletion, mean imputation and hot-deck imputation.

Nanni et al. (2012) and Ghannad-Rezaie et al. (2010) handled a percentage of missing data of 30% using multiple imputation and expectation maximization but tested the methods on datasets not exceeding 1000 records and with a small feature count. If this work was implemented on datasets of larger sizes, the computational power required would be very high.

Cheema (2014) presented the most recent survey about methods of handling missing data in education research. The survey reviewed multiple studies in the topic of missing data in different fields but concluded that there is a need for a comprehensive study in the domain of handling missing data.

Accordingly, the research performed on missing data could be expanded to acquire higher percentages and newer methods of handling them. The research addressed here has a high focus on the problem of missing data and how a novel methodology to handle them could be proposed.

3.4 Feature selection

A feature or attribute is a characteristic of a certain entity that is being examined. Any process will have multiple features to describe it that can then be used by classification algorithms in order to distribute different records into classes. Given there are different applications and fields where the concept of features exist, the number of features and their meanings vary. The problem is that as the number of features increase the existence of non informative and redundant features increase as well. These features present a load on the classification algorithms and usually lead to a degradation in classification accuracy (Bolón-Canedo et al., 2014; Chandrashekar and Sahin, 2014; Jain et al., 2000).

By needing to select the best features to represent a dataset, feature selection became essential to better understanding of data, reducing computational complexity, building better classification models and improving their accuracy. Feature selection aims at finding the smallest subset of features with the highest amount of information (Bolón-Canedo et al., 2014; Jain et al., 2000).

3.4.1 General feature selection approaches

Feature selection is an important preprocessing task since it produces a subset that best represents the whole dataset. It reduces the number of features by removing redundant and irrelevant ones to enhance the performance of classifiers. Figure 3.2 shows the general steps of feature selection.

Subset generation

To find the best subset, one approach is to generate and evaluate everyone. If the dataset has features of count N then the total number of possible subsets that can be created would be 2^N (Chandrashekar and Sahin, 2014) which means the brute force approach is not tractable. The different methods to handle this problem are to find ways of generating the subset of features without going through all the possible combinations. This includes complete, heuristic and random (Dash and Liu, 1997) methods.

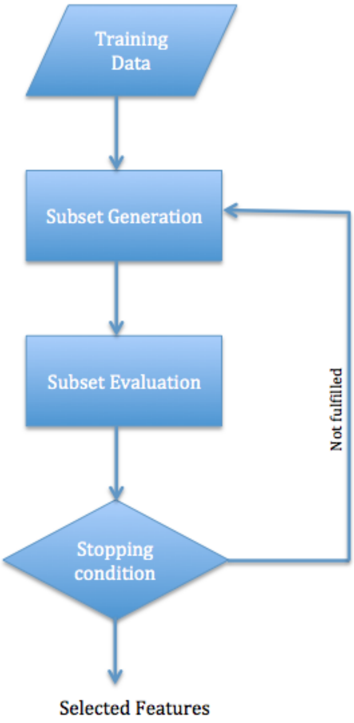


Fig. 3.2 General steps of feature selection

A subset of features is chosen according to two criteria: the starting point and the search strategy. Feature selection should start with an empty set and have features gradually added to it (forward selection) or start with all the features and have ones gradually removed (backward selection). A combined approach could be used where it starts with either method and features are added or removed according to a bidirectional measure (Krzysztof et al., 2007; Liu and Yu, 2005). The search strategy, which aims for choosing the best sample could be complete, sequential or random (Krzysztof et al., 2007; Liu and Yu, 2005).

- A complete search assures the production of an optimal subset given the right evaluation criterion. Although aiming for an optimal subset could mean passing through all the available ones (exhaustive search), this does not have to be the case since there are heuristic functions that can reduce the number of subsets to be examined while assuring the optimal set would be found. Examples of these search functions are the branch and bound, best first search, and beam search (Krzysztof et al., 2007; Liu and Yu, 2005).

- A sequential search uses a greedy hill climbing methodology where features are added or removed according to an evaluation criterion until a stopping condition is fulfilled. Examples are sequential forward selection and bidirectional selection. The advantages of this strategy are speed and simplicity (Krzysztof et al., 2007; Liu and Yu, 2005).
- A random search starts with a random subset of features and either continues in the same manner as sequential search or keeps producing other random subsets until the best subset that fulfilled the stopping condition is met (Krzysztof et al., 2007; Liu and Yu, 2005).

The choice of which strategy to use is restricted with data domain specifications and the extent of complexity that could be tolerated (Guyon and Elisseeff, 2003). The complete search is computationally expensive so it would work best with a restricted number of features. Pudil et al. (1994) specified that if the feature count exceeded 30, complete search using heuristic functions would still be computationally expensive. Since data addressed in this research have a high feature count of 177 and the speed of selection is important, the sequential forward selection strategy was chosen.

Subset evaluation

The evaluation of subsets is done depending on whether a wrapper or filter model is used. The difference is that the wrapper uses dependent criteria, where it evaluates the model based on the performance of a classifier to see whether the selected features suit the chosen classification algorithm or not. The filter model uses independent criteria that evaluate the subset without knowing which classifier will be used for classification. Independent criteria include correlation measures between features and classes or information measures to calculate the information gained from the addition (or removal) of a feature. A hybrid model that combines both filter and wrapper methods can be used with a mining algorithm and independent measures to evaluate subsets (Doquire and Verleysen, 2012; Liu and Yu, 2005).

Stopping criteria and result validation

Specifying the stopping condition that defines acceptable performance follows subset evaluation. It could be that all the subsets are evaluated, or the evaluation could stop when addition or deletion of new subsets fails to make any improvements. Or there could be an acceptable error threshold and the first subset that meets it is chosen. Finally, the selected features could be compared to ones already determined by experts or by comparing results produced by the mining algorithm on the selected features and on the whole set of features (Liu and Yu, 2005).

3.4.2 Methods of feature selection

Feature selection is divided into filter and wrapper methods. Filter methods are considered a preprocessing step that ranks features and selects from them according to the rank. In wrapper methods, the performance of data mining algorithms is used to determine how good a feature subset is in an iterative manner. Wrapper methods are highly computational given a large number of features because they require repeat evaluations with prediction algorithms in order to reach the feature set. Too many iterations can cause this method to tend to over fit (Bolón-Canedo et al. (2014); Chandrashekar and Sahin (2014); Tuv et al. (2009). The GRiST dataset has too many features for a wrapper approach and so the work presented in this research is based on filter methods.

The main idea behind filter methods is to rank features according to a predefined criterion, which is simple yet often successful in various practical applications (Bolón-Canedo et al., 2014; Chandrashekar and Sahin, 2014). The filter method is applied before classification to enhance the quality of the subset or data provided (Chandrashekar and Sahin, 2014; Tang et al., 2014). The idea behind improving the quality of the provided subset is to choose one that is most relevant to the output class. Relevance is measured as its dependence on the class labels. Hence features with low relevance to the class are discarded (Chandrashekar and Sahin, 2014; Tang et al., 2014).

Redundancy and irrelevant features are additional problems for filter methods. An irrelevant feature is one that is not related to the output variable and a redundant one does not introduce any additional information compared to those already selected. The presence of irrelevant or redundant features degrades the created classification models and may need to be removed to improve the speed of learning algorithm as well as produce a better classification model (Bolón-Canedo et al., 2014; Chandrashekar and Sahin, 2014; Tang et al., 2014). The most common filter criteria used are correlation and information measures which are explained next.

Correlation criteria

A criterion that is commonly used to measure the relationship between two variables is correlation. Equation 3.1 shows the Pearson correlation coefficient r between feature x_i and class Y (Chandrashekar and Sahin, 2014; Krzysztof et al., 2007)

$$r = \frac{cov(x_i, Y)}{\sqrt{var(x_i) * var(Y)}}, \tag{3.1}$$

where x_i is the i_{th} feature; Y is the class, $cov(x_i, Y)$ is the covariance between the i_{th} feature and the class; and $var(x_i) * var(Y)$ is the multiplication of variance of the i_{th} feature by variance of the class.

Correlation-based Feature Selection (CFS) is a filter-based algorithm that uses correlation to select a subset of features. The CFS applies a forward selection method and the subset evaluation is the correlation criteria.

First, two matrices are created, feature-feature correlation and feature-class correlation, from the training dataset. CFS uses these matrices to choose features that maximise correlation with the class vector and minimise correlation between each other (Hall, 2000; Senliol et al., 2008). This would reduce the redundancy between selected features and thus increase the quality of the chosen subset as a representation of the dataset (Hall, 2000; Senliol et al., 2008). The proposed methodology will be based on correlation and mutual information methods, it will be compared to CFS approach in chapter 5.

Pearson correlation is a measure of the strength of linearity between one variable and another. It is a simple and fast calculation but is not useful if linearity does not exist (Chandrashekar and Sahin, 2014; Krzysztof et al., 2007). Mutual information is a measure that is not restricted to linear relation and is described next.

Mutual Information

Chosen features should be highly informative of various classes of the dataset. Mutual information measures the amount of information between two random variables X and Y by calculating entropy. Shannon's entropy $H(Y)$ is defined as a measure of the uncertainty of a certain random variable. It is a representation of how much information is absent when the outcome if not known. This measure shows how many bits of information can be used by all combinations of a random variable Y by using Equation 3.2 where $p(y)$is the probability of occurrence of y (Chandrashekar and Sahin, 2014; Liu et al., 2014).

$$H(Y) = -\sum_{y} p(y) log_2 p(y) \tag{3.2}$$

After calculating the uncertainty about Y, the conditional entropy $H(Y|X)$, which is the amount of information gained from Y given the value of X, is represented by Equation 3.3, where $p(x,y)$is the probability of co-occurrence of x and y; and $p(y|x)$ is the probability of occurrence of y given the value of x.

$$H(Y|X) = -\sum_{x}\sum_{y} p(x,y) log_2 \, p(y|x) \tag{3.3}$$

$$I(Y;X) = H(Y) - H(Y|X) \tag{3.4}$$

Thus the mutual information between variable X and the class Y is defined by Equation 3.4. If X is completely independent of Y then the mutual information between them will be zero because the uncertainty of Y is not changed by knowing X, which is completely uninformative. The advantage of mutual information is that it is independent of the linear relation between the two variables. It is commonly used to evaluate the redundancy of features (Doquire and Verleysen, 2012; Peng et al., 2005; Yu and Liu, 2004).

Minimum redundancy maximum relevance

The Minimum Redundancy Maximum Relevance (mRMR) method combines two constraints in order to select a candidate feature. Both constraints are based on mutual information. Maximum relevance $D(S,c)$ tries to find a feature that increases the average mutual information of the selected subset, S, with the output class, c, when added as shown in Equation 3.5. $|S|$ is the number of features chosen and $I(x_i;c)$ represent the mutual information between feature x_i from the chosen subset and class c (Peng et al., 2005).

$$max\, D(S,c),\; D = \frac{1}{|S|}\sum_{x_i \varepsilon S} I(x_i;c) \tag{3.5}$$

The other constraint is minimum relevance, $R(S)$, which attempts to add a feature that has the least average mutual information among the chosen set of features thus reducing redundancy as shown in Equation 3.6.

$$min\, R(S),\; R = \frac{1}{|S|^2}\sum_{x_i, x_j \varepsilon S} I(x_i, x_j) \tag{3.6}$$

Maximizing the difference between the maximum relevance and the minimum redundancy constraints produced the mRMR criteria shown in Equation 3.7. This equation is intended to choose the mth feature from the set of available features denoted $X - S_{m-1}$.

$$\underset{x_j \varepsilon X - S_{m-1}}{max} \left[I(x_j;c) - \frac{1}{m-1}\sum_{x_i \varepsilon S_{m-1}} I(x_j;x_i) \right] \tag{3.7}$$

In this research, correlation, mutual information and mRMR will be applied and tested for the selection of features. More details of their experiments and method of use will be presented in chapters 4 and 5.

3.4.3 Recent work on feature selection

Chandrashekar and Sahin (2014) surveyed the common feature selection techniques and advised that comparison of methods needed to be applied for a single dataset to understand the behaviour of each algorithm. They also concluded that feature selection techniques prove that more information is not always an advantage. The application of feature selection methods provide an inevitable advantage for the speed of classification, understanding data better and making classifiers build simpler models. A survey by Bolón-Canedo et al. (2013) on comparing different feature selection methods using synthetic data recommended the use of filter methods over wrapper methods for their simplicity, independence, speed and ability to generalise.

Tomar and Agarwal (2013) surveyed different methods used specifically in healthcare and emphasised the importance of the use of data mining to find patterns and manage healthcare. They confirmed that the type of algorithm used depends on the data available. The next section will discuss different predictive data mining algorithms along with their common advantages and disadvantages.

3.5 Predictive data mining algorithms

Data mining involves two broad categories of algorithms, descriptive or predictive. Descriptive, which is also known as unsupervised, divides the data into clusters according to the similarity between them. It is used to explain relationships among the data. Predictive, also known as supervised, are algorithms that build models to classify or regress the data in order to determine the classes to which unclassified samples belong to. The output classes are then compared to the predefined values for evaluation (Yoo et al., 2012).

Whether a dataset is of known class labels or not is defined as supervised versus unsupervised classification. Supervised is learning from data with predefined labels and evaluating the results on those labels while unsupervised is involved with clustering samples together according to commonalities among their features (Jain et al., 2000). Semi-supervised is a hybrid where only a part of the data is labelled and the learning process is a mix between labelled and unlabelled records. Choosing between supervised, unsupervised and semi-supervised depends on the problem domain and the dataset at hand (Witten and Frank, 2005).

In this work, the main interest is supervised classification since the application domain of the proposed research is on labelled data.

Data are partitioned into training set and testing set. From the training set, models are built to describe the relationship between attributes and predefined categories. These models are then applied to the testing data to estimate their categories and compare results with the predefined values. There are different algorithms used for data mining such as linear regression, decision trees and random forest (Yoo et al., 2012).

Kononenko (2001) reviewed classification algorithms for medical diagnosis. They noted that physicians require output diagnoses to have explanatory reasons. Both Kononenko (2001) and Yoo et al. (2012) emphasised that decision trees were the only classifier to have built in feature selection capabilities.

The algorithm proposed in this thesis is built using linear regression. It will be compared to results produced by classification algorithms that have been shown to excel in healthcare, which are decision trees and random forest (Yoo et al., 2012). All these classifiers will be explained in the following subsections.

3.5.1 Decision trees

Initially, decision trees were introduced by Quinlan et al. (1979) and were known as the Iterative Dichotomiser 3 (ID3). Fourteen years later it was succeeded by the C4.5 (Quinlan, 1993) which is commonly known as decision trees. The idea behind decision trees is to build a tree by an iterative process where the data is partitioned multiple times according to values of variables until the remaining subset of the data is of the same class. The top of the tree represents the root node and leaf nodes represent the classes (Bellazzi and Zupan, 2008; Larose, 2014; Tomar and Agarwal, 2013).

Decision trees are internally feature selective and built trees can provide an explanation of why a specific decision was taken. They provide visualisations of reasoning behind the choice of a specific class by going down the trees (Duda et al., 2012; Tomar and Agarwal, 2013).

Advantages of using decision trees as explained by Han et al. (2011) and Tomar and Agarwal (2013) are that they:

1. do not require much knowledge about the domain;

2. can work with large numbers of features;

3. produce representations that are comprehensible by humans; and

4. produce acceptable accuracy.

On the other hand the disadvantage is the complexity of resulting trees if the number of features is too high (Han et al., 2011; Tomar and Agarwal, 2013).

3.5.2 Random forest

The random forest method is an application of ensemble classification. Ensemble classification is based on the concept that combining the use of multiple classifiers would be stronger than using a single one. Therefore the idea is to work with multiple classifiers in parallel, then combine the generated results by different methodologies like voting. In the case of random forests, the random classifiers are decision trees. Nodes are generated randomly and used to build a forest of trees. This forest will be used for classifying a single record and combining the answer to reach a specific classification (Witten and Frank, 2005). Yoo et al. (2012) showed that the use of ensemble classifiers enhances the performance results and they will be used for comparison with the proposed algorithm.

3.5.3 Regression

Regression involves building a mathematical model that defines the correlation between different variables using the available training data. The output class is known as the

Dependent Variable (DV) and the predictor variables are the Independent Variables (IVs). A DV can be either numerical or categorical depending on the dataset addressed. Regression finds the mathematical equation that models the relationship between IVs and the DV (Tomar and Agarwal, 2013).

Linear regression

With linear regression, the relationship between a single IV and the DV is represented by a straight line. This kind of regression accepts the IV and DV to be only numerical values. Linear regression was extended to accommodate multiple IVs, which is called multiple linear regression (Cohen, 2008; Han et al., 2011).

Multinomial logistic regression

Logistic regression is a non-linear regression that can handle a categorical DV. It uses the logit function to predict the probability of belonging to a certain class (Hosmer Jr et al., 2013). The logit distribution restricts the output to be in a range of 0 to 1 which is then interpreted as a probability (Witten and Frank, 2005).

Generalised linear models

Generalised Linear Models (GLM) are a general case of linear regression. When the relationship between the DV and the predicted values is not linear, a link function is used to transform them. There are different examples of link function such as identity and log. The identity is used with a normal distribution and the log (which represents the natural logarithm) is used with a poisson distribution (Witten and Frank, 2005).

3.6 Summary and conclusions

Predictive data mining is a subdomain of data mining. It has been recently applied to the healthcare domain but is constrained by the kind of data addressed. For instance, missing information needs addressing and a large number of features needs reducing. There are different ways to deal with missing data such as deletion of records or different methods of imputations. Choosing the methodology depends on the percentage of missing data along with the application domain. Several publications have claimed that missing data is usually not reported by researchers and most of the time ignored. On the other hand, publications that reported it, applied their methodologies on datasets of very small size with a small number of features.

Recalling from the previous chapter, the mental health research addressed here created a suicide dataset (GRiST) that has a large percentage of missing values (67.04%) in varying amounts per sample. Applying list-wise deletion on the whole dataset would have been impossible since there are no complete records available. Imputation would be impractical because of the computational cost, the degradation of accuracy in imputing such a large percentage of missing records and above all that introducing values to questions that patients did not answer would be unacceptable. Along with the missing data challenge, there is a very large number of features and 11 risk levels each having a completely different number of records.

After missing data is handled, features need to be selected to best represent the dataset. This requires selecting a subset of features that is most relevant and least redundant. There are different methods such as filter and wrapper. Filter methods are most commonly used because they are simple, do not rely on a prediction algorithm, and allow the classification to be performed on a real-time basis. They are most suited to mental health data collected by the GRiST decision support system.

Finally different algorithms can be used to predict the category to which a patient's record belongs to. The ones applied in this research create models that are easiest to explain: regression, decision trees and random forest.

The aim of this PhD research is to provide a risk prediction element to GRiST. It will need to overcome all the challenges of the GRiST dataset, be able to predict the risk of suicide, and provide an explanation of how this risk was predicted without fabricating information that a patient did not provide. In the upcoming chapter, the proposed algorithm will be thoroughly explained.

Chapter 4

Dynamic Feature Selection and Prediction

4.1 Introduction

Assessing the risk of suicide is a very difficult task achieved by assessors from different domains, some within mental health services and some outside. Assessors need to make a strong justification for categorizing a patient as suicidal that is not based on intuition. The justification has to be based on information that was actually provided by the patient rather than imputed information. In this context, and given the lack of tools for guiding assessors (Chesin and Stanley, 2013; Lotito and Cook, 2015; Silverman and Berman, 2014), an algorithm is needed that both helps make an assessment and explains its derivation.

In this chapter, the derivation of a novel algorithm that addresses the problems of missing data and high-dimensionality inherent in mental health risk data is proposed. It will be thoroughly explained in the upcoming sections. First the rationale behind it will be presented. The basic idea will then be explained and finally updated to a parametrised one that represents the initial version of the proposed Dynamic Feature Selection and Prediction (DFSP) algorithm. It will be tested in chapter 5 and the final version presented in chapter 6.

4.2 Rationale

The proposed algorithm is expected to provide practitioners with the strongest and most relevant parameters having least redundancy and maximising their explanatory power. The selection of the dynamic set of features will be based on having high correlation value with the risk while preserving the least mutual information among the selected features. In short, predicting the risk is not the only aim of this research; it is also to explain the predictions to assessors and to do so in real time during an assessment. In the following section, the basic idea of the DFSP will be presented.

4.3 Basic DFSP algorithm

Algorithm 1 Basic version of the DFSP algorithm

```
for every patient record do
    Ignore all absent features
    Choose the feature with highest relationship with the risk
    while number of features chosen <= defined count do
        Find next feature with highest relationship with the risk
        if Candidate feature is a filter question and a descendant was pre chosen then
            discard feature
        else if Candidate feature is a descendant and filter exists then
            remove filter and add feature
        else add feature
        end if
    end while
    Find a complete subset of the training records for the chosen features
    while Size of subset <= 100 * number of chosen features do
        remove last added feature.
        Regenerate a complete subset of the training records
    end while
    Build regression model for patient and assess record
end for
```

The DFSP algorithm (Saleh and Buckingham, 2014) is intended to select features from only the answers provided in a patient's record without the use of common methods of imputing missing features. The selection of a unique set of features per patient record is a novel approach that was not accounted for in the literature. The selected features are used to build a regression model that will assess this record. The basic steps of the proposed DFSP are shown by Algorithm 1. It is intended to be tested multiple times using different number of selected features.

From only the existing information supplied by a patient, features having the highest relationship to the risk level are chosen while fulfilling the filter criteria and not exceeding the defined count. The chosen set is then used to build a classification model for the assessment. The explanation of each part of the algorithm will be discussed in the following subsections.

4.3.1 Evaluating features

Algorithm 1 has introduced the proposed feature selection algorithm. Measuring the relationship between a feature and the risk level will be tested using mutual information and correlation which were previously explained in section 3.4.2. Using each method, a matrix will be generated to represent the relationship between a feature (independent variable, IV) and the output risk level (dependent variable, DV). Another matrix will be generated to represent the mutual information between the IVs.

Algorithm 2 Generating *MI-IV-DV*, *Corr-IV-DV* and *IV-IV* matrices

```
Require: TrainRecords
  for each feature $feat1$ in TrainRecords do
      Calculate pairwise mutual information between $feat1$ and DV
      Place value in MI-IV-DV
      Calculate pairwise correlation between $feat1$ and DV
      Place value in Corr-IV-DV
      for each feature $feat2$ in TrainRecords do
          Calculate pairwise mutual information between $feat1$ and $feat2$
          Place value in IV-IV
      end for
  end for
  Scale MI-IV-DV, Corr-IV-DV and IV-IV
  Store MI-IV-DV, Corr-IV-DV and IV-IV
```

Algorithm 2 generates the three matrices: *Corr-IV-DV*; *MI-IV-DV*; and *IV-IV*. *Corr-IV-DV* matrix represents the correlation between each IV and the DV. It resulted from applying equation 3.1 to calculate each IV's pairwise correlation with the DV. *MI-IV-DV* represents the mutual information between each IV and the DV. Each cell is calculated by applying equation 3.4 on the rows with non missing information for the addressed IV and the DV. Finally, *IV-IV* matrix represents the mutual information among the IVs. After the matrices are created, they are scaled so that all their values range between 0 and 1. Correlation results range between -1 and 1 while mutual information values have a minimum of zero and a maximum value that differs based on the data addressed. To standardise the values addressed such that the thresholds applied would be tested in a specific range, the scaling was required.

In chapter 5, there will be an experiment that uses the three generated matrices to compare between the use of different methods of feature selection and choose the appropriate one for the DFSP. The methods compared will be: minimum redundancy maximum relevance (*mRMR*) which was previously explained in section 3.4.2; Correlation (*Corr*) ;and mutual information (*MI*). The application of the three methods will be on Algorithm 1.

4.3.2 Filter evaluation criteria

In chapter 2, it was explained that a filter question opens up its descendants if answered in the affirmative. This means that if a filter question was chosen along with one of its descendants then the generated complete subset of the training set will have the filter question as the same value of 1 in all records (the value for 'yes'). So a filter question would be more valuable if its descendants were not found in the chosen features. This forced a restriction to be added that would check every candidate feature against the pre-selected set. If the candidate feature is a filter and none of its descendants is in the set then add it, otherwise do not add it. On the other hand, it is a descendant, then check the set for the parent. If the parent was chosen, remove it and either way add the descendant. The following subsection describes how the relationship between features is calculated.

4.3.3 Assessing patient records

After the features are chosen, a complete subset of records from the training set is chosen. This would represent similar patients whom answered the same set of questions. At this stage, the number of records chosen have to exceed a certain limit so that the coefficient parameters are accurately calculated. Harrell (2006) set this limit to a minimum of 10 times the number of features addressed. Another range was specified by Schmidt (1971) to be 15 to 20 times the number of features. In order to avoid the risk of in accurate calculation of coefficients, this research fixed the limit to be a minimum of 100 times the number of features. If the generated number of records is less than 100 times the number of features, it would be considered small. This would cause the algorithm to drop the last selected features and regenerate the subset until a proper sized one is produced. Forcing subset size restriction would assure that the classifier would have enough data to deduce its parameters. This subset will then be used to build a model for classification.

Algorithm 1 showed the basic methodology of the proposed algorithm but did not handle certain measures. For instance, when practitioners choose the best informative features to build the patient analysis on, they are not supposed to be constrained by a defined count. Defining a specific count of features would force the system to use non informative ones, for some patients, just to fulfil the required number. If this did not affect the accuracy, it would affect the algorithm's capability to provide a good explanatory set of features. From this concept, the algorithm was updated to include a threshold that would define the limit onto which features are to be added. This threshold would be based on the feature's relationship with the risk level.

The redundancy among the selected features needed to be considered. A situation that could occur is when two of the selected features represent a large amount of shared information thus the existence of both would result in redundant information. This would have an effect on the algorithm's accuracy and on the explanation provided to the assessor. For these reasons, two thresholds were introduced to the basic algorithm. One threshold to specify the relationship with the risk level and the other to control the accepted level of shared information among the features. These thresholds will be introduced in the following subsection where the parametrised version of the DFSP algorithm will be presented and explained.

4.4 DFSP algorithm

The proposed DFSP algorithm is a feature selection algorithm that can dynamically choose different features for each patient in order to build its classification model. The DFSP chooses the best set of features that can represent a patient from the answers provided and does not impute any missing features. The selected features have the highest relationship with the risk level and the least redundancy among each others. In this section, the proposed feature selection algorithm will be thoroughly explained. First the thresholds required to make the feature selection of the DFSP dynamic will be introduced. Secondly the methodology for learning models will be discussed along with how they are tested. Finally the parametrised and detailed version of the algorithm will be presented and explained.

4.4.1 Defining parameters

Algorithm 2 has described the generation of the matrices needed for determining the relationship between an IV and the DV or between two different IVs. After the three different measures (*Corr*, *MI*, *mRMR*) are experimented, one will be chosen. Given the chosen measure, two threshold parameters needed to be defined to dynamically select the features: *DV-Thresh* and *IV-Thresh*. *DV-Thresh* is a threshold that causes all chosen features to have a relevance value with the DV higher than or equal to it. *IV-Thresh* restricts that no feature is to be added if it exceeds this threshold as mutual information with features already selected in order to limit redundancy.

4.4.2 Learning models

After the set of features for the patient to be assessed are selected, a model has to be learned. This model is built based on other records. As previously explained, linear regression attempts to find the best weights to describe the linear relationship between a set of IVs and the DV. In the DFSP algorithm, the IVs represent the selected set of features and the DV represents the assessed risk. One of the reasons for using linear regression is how it relates to the GRiST knowledge tree, which is effectively a hierarchical regression model if the weights were to be learned. This means there is a natural affinity between the psychological model and linear regression. Another advantage of using linear regression over other methods is that it will satisfy one of the main objectives of this work which is to provide proper comprehensible explanation.

4.4.3 Testing models

After the model is learned, it is applied to its test record and compared against expert opinion for verification. The testing phase would show how accurately the proposed algorithm is performing. The results from different experiments will be discussed in the following chapter. Evaluating the results of the different experiments will be based mainly on two criteria: accuracy and shifted accuracy.

A patient record in the GRiST dataset can be classified into one of 11 risk levels which is considered to be a large number of classes. Addressing such a specification usually involves grouping the risk levels into low, medium and high groups. For example, low risk would be defined by risks 0, 1, 2 and 3 patients while medium risk would be 4, 5 and 6. The problem with this approach is the classification of patients on the risk group boundaries. For example, a risk 3 patient would be acceptably classified as risk 2 but when classified as risk 4 that would be considered an error. Therefore dividing the risks into groups was not acceptable.

Another specification addressed was the numbers of records among all risk levels, where they are not equally distributed. There is a big difference from one risk level to another, which was shown previously in Figure 2.4.

As a proposed solution to those difficulties, the evaluation mechanism of the results is based on a modified method. Instead of addressing the total accuracy of the whole classification, accuracy is measured per risk level. This is to assure that the classification among all risk levels is evaluated and higher count risk levels did not dominate the results. Not only the accuracy is calculated but also the shifted accuracy per risk level. This means that there is a tolerance in the calculation of accuracy. For example predicting a patient that was assessed by experts as a risk 4 as a risk 3 or risk 5 patient would be accepted. This would handle the granularity specification by giving all risk levels the same 1 level shift tolerance in both directions. This measure is used in the following chapter and is referred to as the *Shifted Accuracy*. In the following subsection the parametrised DFSP will be introduced.

4.4.4 Parametrised DFSP

The parametrised version of the DFSP is shown in Algorithm 3. It requires setting two main parameters in order to work. Those parameters were described earlier and introduced the dynamic functionality of the algorithm where the selected features are chosen until certain criteria is met. The addition of a new feature is restricted by having relevance with the DV higher or equal to a certain threshold.

Algorithm 3 Parametrised version of the DFSP algorithm

```
Require: TrainRecords, TestRecords
Require: DV-Thresh and IV-Thresh
  for Every SingleTestRecord in TestRecords do
      Initialise temp with non missing features from SingleTestRecord and S to empty
      Find feature with highest relevance in temp and place it in candidate
      while candidate.val >= DV-Thresh do
          if candidate.ID is a filter question and S includes a descendant or more then
              discard candidate
          else if candidate.ID is a descendant and filter exists then
              Remove filter and add candidate to S
          else
              if Verify candidate.ID low relevance against all features in S < IV-Thresh then
                  Add Candidate to S
              end if
          end if
          Remove candidate from temp
          Find feature with highest value in temp and place it in candidate
      end while
      Find a complete subset in TrainRecords for features in S
      while Number of records <= 100 * number of chosen features do
          remove last added feature
          Regenerate a complete subset from the training records
      end while
      Build the regression model accordingly and classify records
  end for
```

The data provided to the algorithm needs to be divided into a training set (*TrainRecords*) and a test set (*TestRecords*). *TrainRecords* has a large number of samples and features, each row represents a sample, each column a feature and the last column the risk level. The training set will be used to estimate the parameters and create the models used for prediction. *TestRecords* is a different subset of the data set that represents the test records and is used to test the algorithm.

For each *SingleTestRecord* in *TestRecords*, the features need to be selected. A variable $temp$ is created to contain all the existent features in *SingleTestRecord*. Starting with an empty set of features *S*, where *S* is the feature set to be chosen per record. The algorithm then finds $candidate$ in $temp$ which currently has the highest relevance with the DV.

For each $candidate$, its value is checked against *DV-Thresh*. If the value is higher or equal, then it is checked for the filter criteria. Now the filter criteria checks whether $candidate$ is a filter question and *S* has a descendant then it is discarded. Otherwise, if it is a descendant and a filter parent exists then remove it and add $candidate$. In this scenario, there is no need to check $candidate$ for redundancy against features in *S*. The reason behind this is that the high level of mutual information lies between descendants of the same parent. Since the filter parent was in *S* then it is definite that there was no descendant sibling. Finally if neither conditions is met then the algorithm checks for redundancy by testing the relevance against the features in *S* to be lower than *IV-Thresh*.

After the loop is terminated, *S* has a complete set of chosen features that represent *SingleTestRecord*. A subset of the *TrainRecords* with the same features in *S* but having complete rows will be extracted. Before the subset is used, the number of records in it is tested to make sure that it is at least 100 times the number of features in *S*. If it is not, then a feature is dropped and subset is re-evaluated. This subset will then be used to build the model needed to classify the test record.

The generated set of features is the first piece of information that will be output to the assessor to show what are the important questions that the patient answered. The second piece will be the assessment of the suicide risk level. In the following chapter, different experiments will be presented and discussed.

4.5 Summary and conclusion

In this chapter, the DFSP algorithm was introduced and explained. The idea behind the DFSP is that for each patient test record, a dynamic number of features that were answered are selected. This subset will have features with higher relevance values with the DV than a specific threshold limit. A condition was added to assure that the subset of features would not include a filter along with its descendant. Finally the selected features will have mutual information among them that is lower than a certain threshold to ensure the reduction of redundant features.

This set of features is then used to build a new regression model based on a subset of the training set of patients who answered the very same questions like the ones chosen for this record. The new regression model is then used for predicting the risk for this patient.

The different experiments concerning the application of the DFSP algorithm will be presented in the following chapter. These experiments will include how the predictor evaluators were chosen, and methods of determining the thresholds. This algorithm will be further enhanced and compared to other off the shelf methods.

Chapter 5

Experiments

5.1 Introduction

The DFSP required a lot of experimentation starting from the basic idea, enhancing it to finalise the proposed algorithm. Throughout this chapter; multiple experiments will be presented along with their results. First an experiment was performed to identify the best method to use for selecting features. There were three candidate methods: Correlation (*Corr*), Mutual Information (*MI*) and Minimum Redundancy Maximum Relevance (*mRMR*). The aim behind this experiment was to explore the application of the three methods and analyse the accuracy results of each risk level patients separately using different feature counts. After the method was chosen, the optimal parameter setting needed to be learned. Learning them required experimenting with many values and thorough interpretation of results.

The replacement of linear regression by multinomial logistic regression was tested in Experiment 2. Multinomial logistic regression would perform the classification into discrete classes, which would then test whether the continuous results produced by the linear regression could be enhanced by discrete classification. Experiment 3 was then performed to test the application of GLM regression instead of linear regression and study whether the results can be affected by mapping the results of the linear regression onto a different distribution. Finally an equation was proposed to enhance the prediction of the DFSP. The final results were then presented and analysed.

All experiments were applied using the GRiST data set. The data was initially ordered by the client identification number. Different assessments of the same patient would be in a sequential order. To make sure that training and testing phases are not affected by such order, the records were first randomised.

Given the skewed population of the GRiST data and large differences between sizes of the risk levels, the experiments were applied in a 10 fold classification. It divides the data into 10 portions then uses one at a time as a test subset versus the remaining 9 as a train subset. Each fold would have a proportional number of records from all the risk levels. For example, there are 99 records in risk 10, then the first 9 folds would have 10 records each and the last one would have 9. This ensures that the training and testing sets are of the same distribution as the whole dataset. The 10 different results obtained are finally averaged. Using 10 fold cross validation adds the advantage of avoiding over fitting and bias to certain risk levels (Breiman and Spector, 1992; Friedman et al., 2001). All the work presented was implemented on a core i-7 Mac-Book Pro with a 2.2 GHz processor and 8 GB memory. The algorithm was implemented using MATLAB.

In the upcoming sections, different experiments will be presented that informed the structure of the proposed DFSP algorithm. Initially exploring the methods used and parameter settings. This experiment will be followed by testing different prediction models. A proposed equation is then introduced to enhance the prediction method. Finally a comparison to alternative approaches is presented and discussed.

5.2 Experiment 1: Methods and parameters

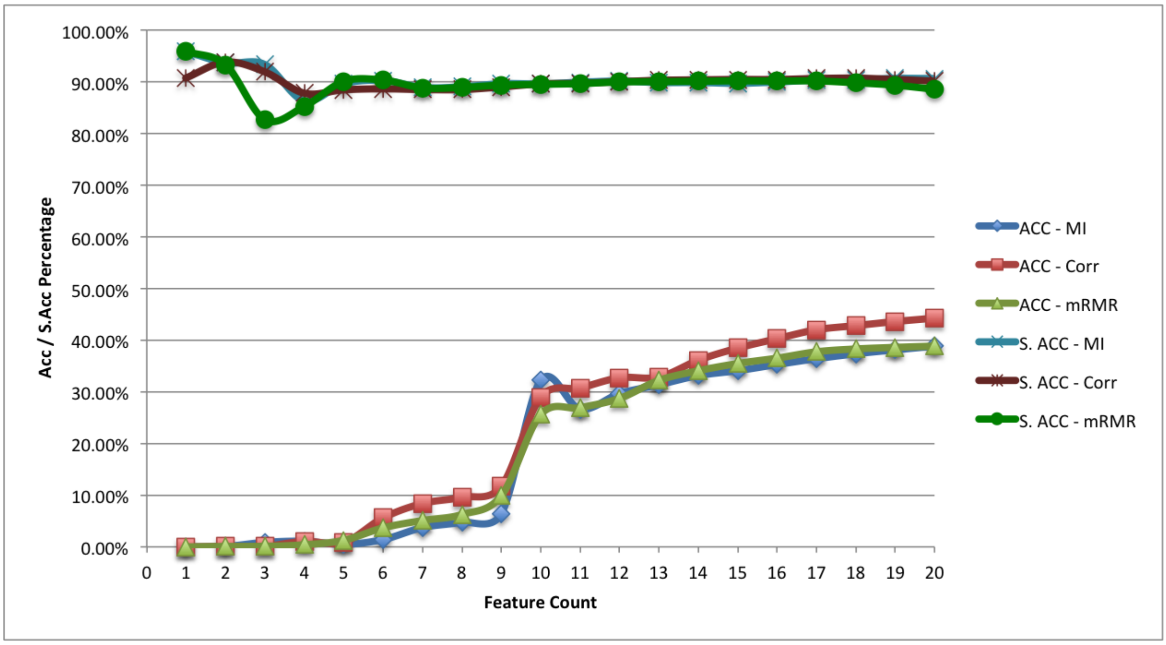
This experiment aims to test different methods and choose the best one. Also, it shows how to select the parameters that were introduced in Algorithm 3 in chapter 4. First the choice of the method will be discussed. The second part will discuss the setting of the *DV-Thresh*. Finally the *IV-Thresh* will be selected.

5.2.1 Choosing the proper method

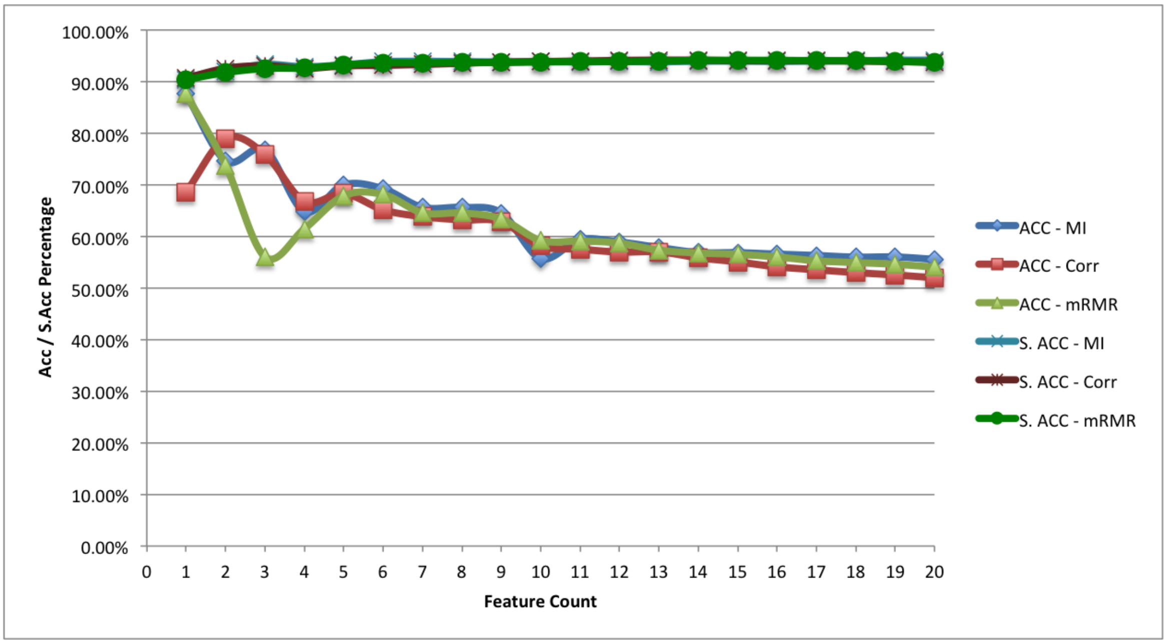
The first part of the experiment tests the influence of different methods using multiple feature count on each risk level separately. This experiment was applied using Algorithm 1 in chapter 4, where the different thresholds were not included. The tested methods are *Corr*, *MI* and *mRMR*. All three methods were previously discussed in section 4.3.1.

The results are shown in Figures 5.1, 5.2, 5.3, 5.4, 5.5 and 5.6. The figures included 11 charts, one for each risk level. Each figure presents two curves per method, one showing its accuracy (Acc) and the other its shifted accuracy (S. Acc) giving a total of six curves per figure. For each curve, 20 different runs were performed, one for each feature count shown by the x-axis, resulting in either the Acc or S. Acc results for a specific risk. This means that each point on the curve shows the result of a risk level when running Algorithm 1 using the defined method (identified by its colour) and the number of features defined by the x-axis.

Analysing Figure 5.1a shows that for risk 0 applying feature counts 1 to 5 have not produced a significant number of patients' records. This improved when feature count exceeded 6. Correlation showed the best accuracy results for this risk level using different feature counts. Shifted accuracy results were almost constant with 7 features or more but a slight degradation can be noticed as feature count reached 18 in the *mRMR* curve. This chart shows that as more features are used, better classification for the zero risk patients occurs for both *Corr* and *MI*.

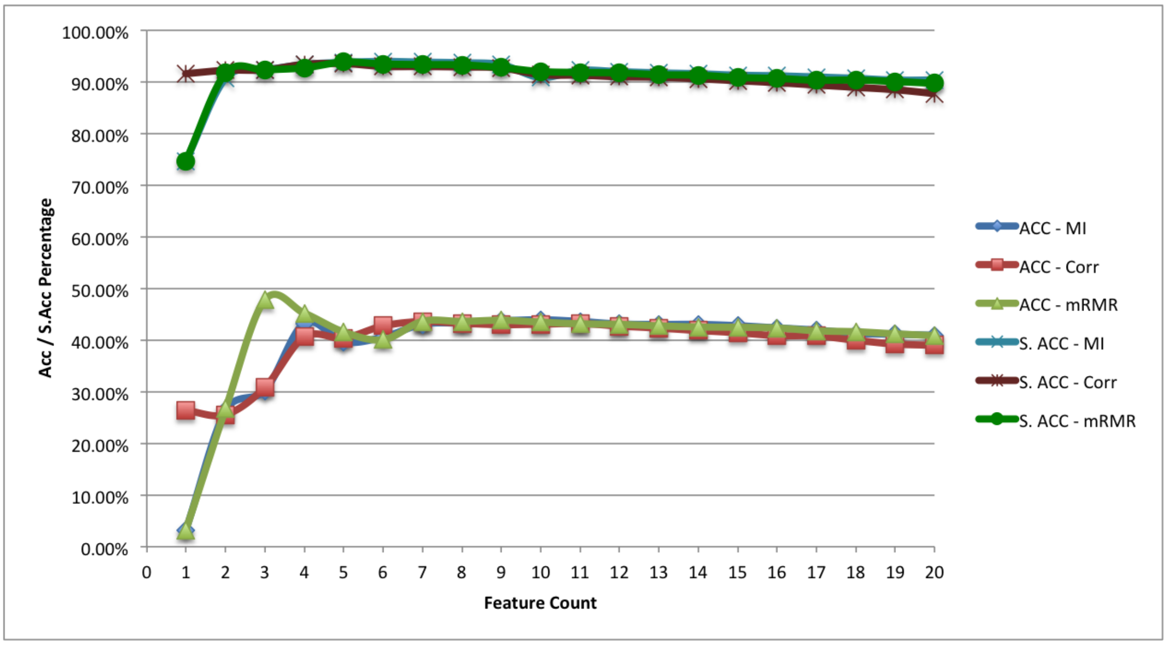


(a) Risk 0 results

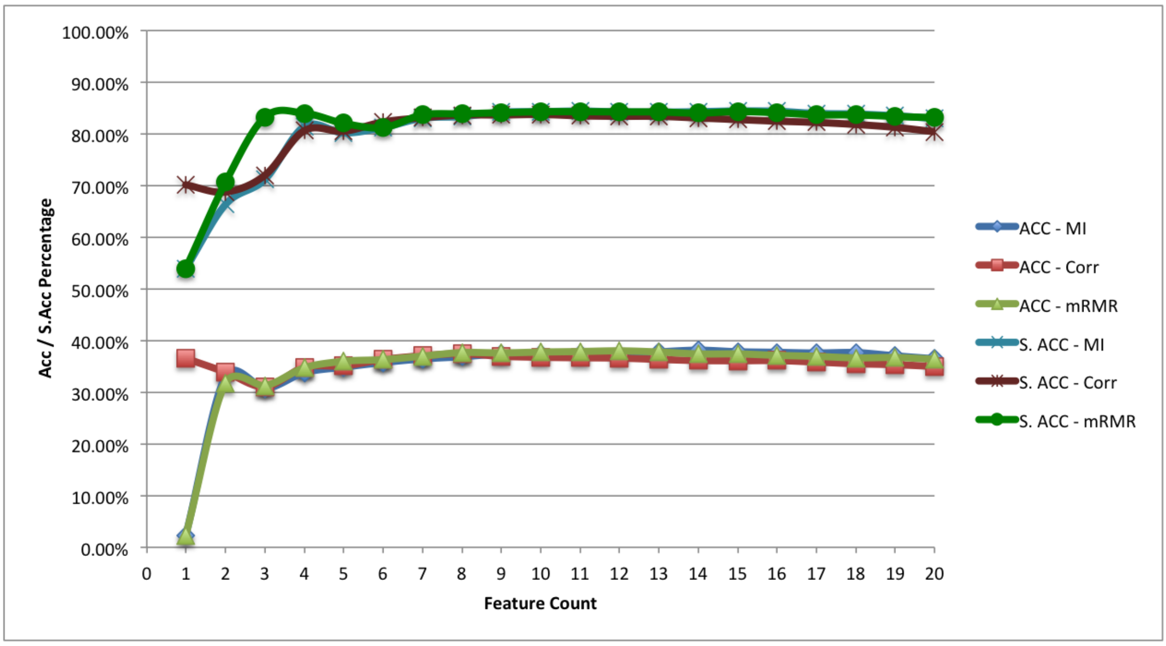


(b) Risk 1 results

Fig. 5.1 Accuracy and shifted accuracy for suicide risk levels 0 patients (Figure a) and 1 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for mutual information *MI*, correlation *Corr* and minimum redundancy maximum relevance *mRMR* criteria.

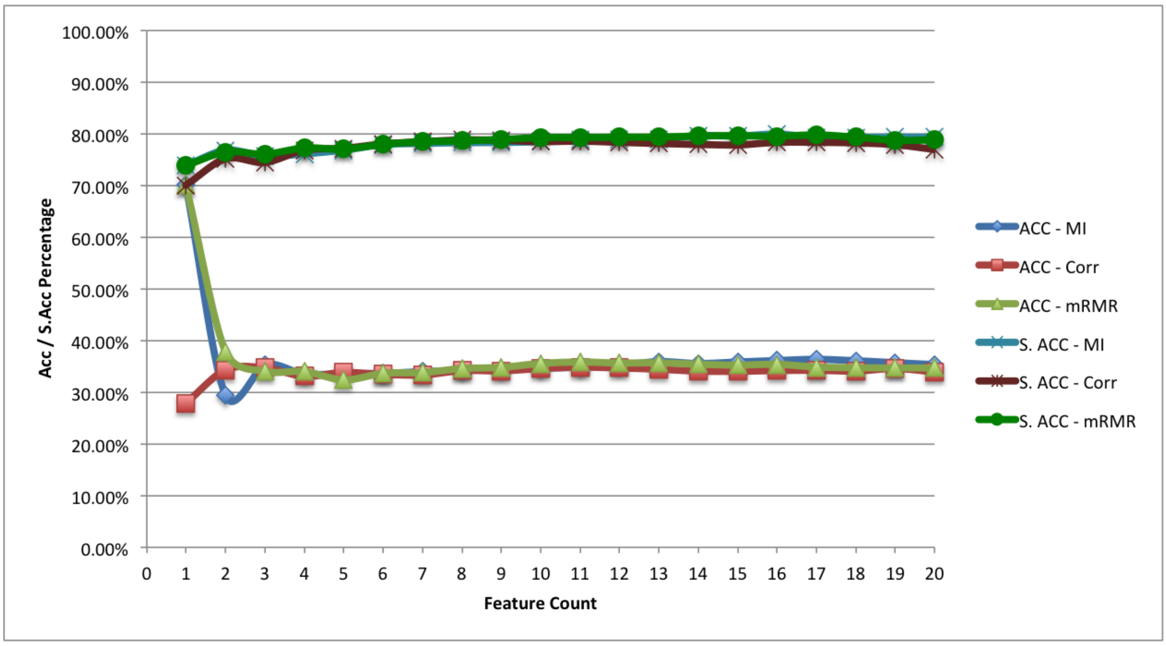


(a) Risk 2 results

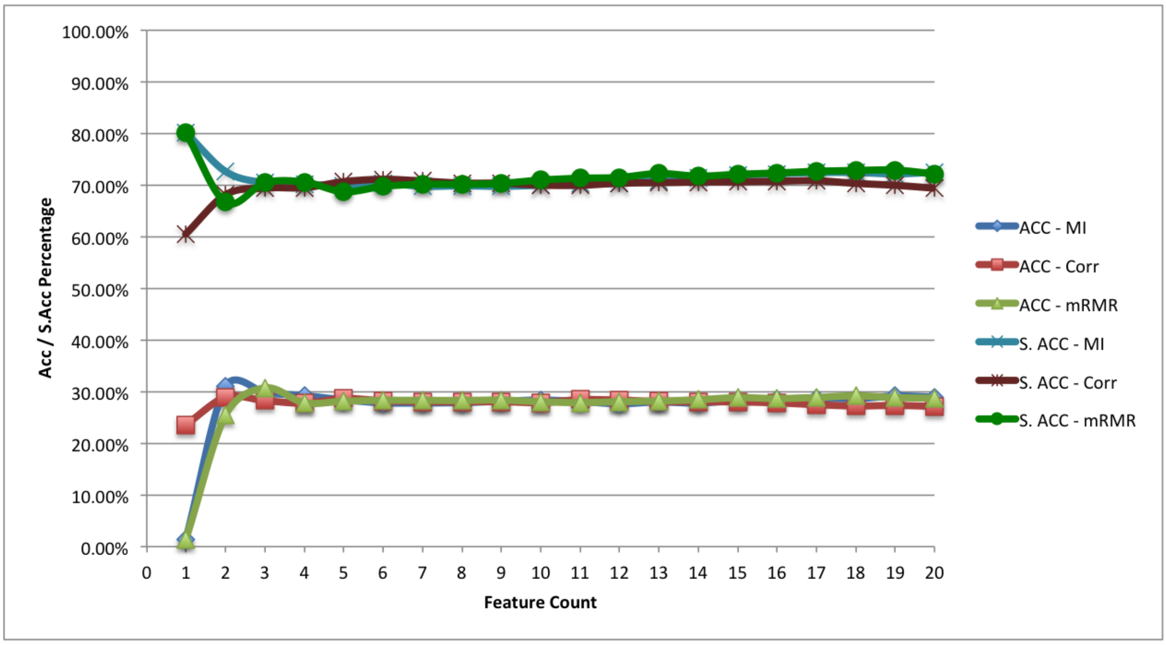


(b) Risk 3 results

Fig. 5.2 Accuracy and shifted accuracy for suicide risk levels 2 patients (Figure a) and 3 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for mutual information *MI*, correlation *Corr* and minimum redundancy maximum relevance *mRMR* criteria.

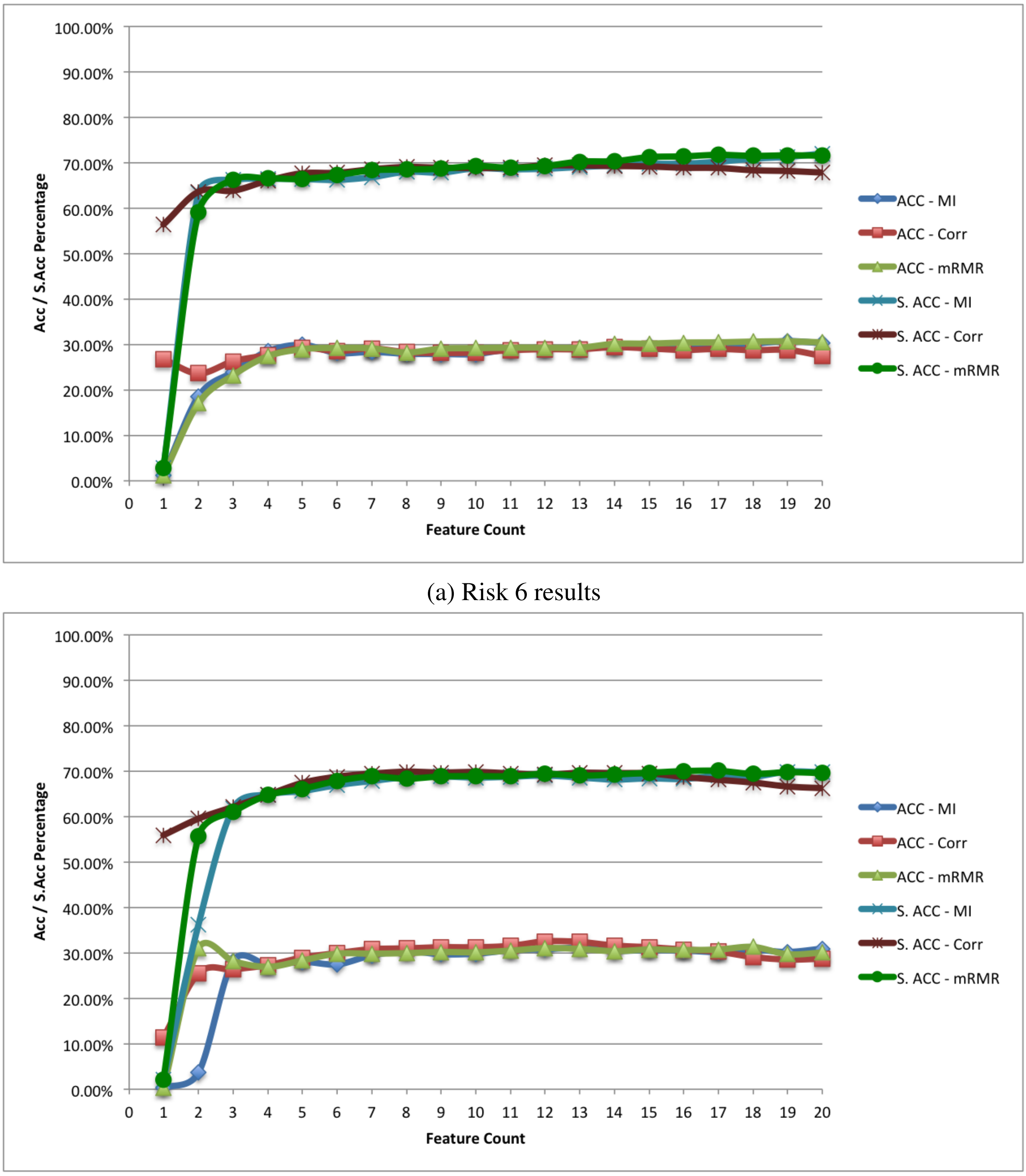


(a) Risk 4 results



(b) Risk 5 results

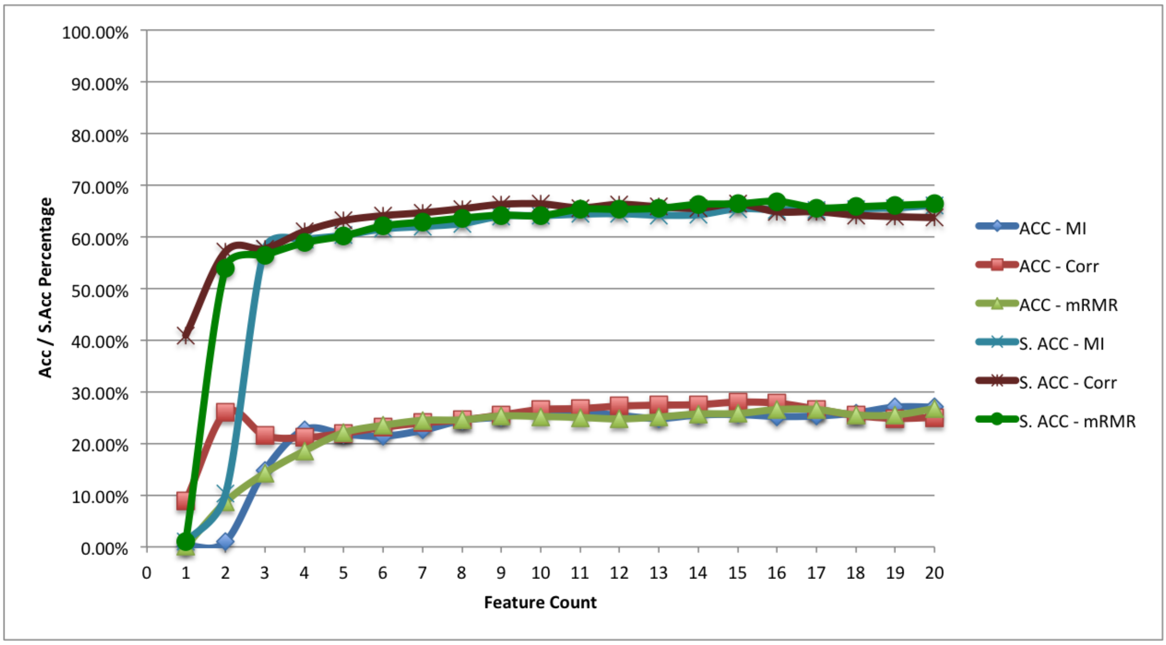
Fig. 5.3 Accuracy and shifted accuracy for suicide risk levels 4 patients (Figure a) and 5 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for mutual information *MI*, correlation *Corr* and minimum redundancy maximum relevance *mRMR* criteria.



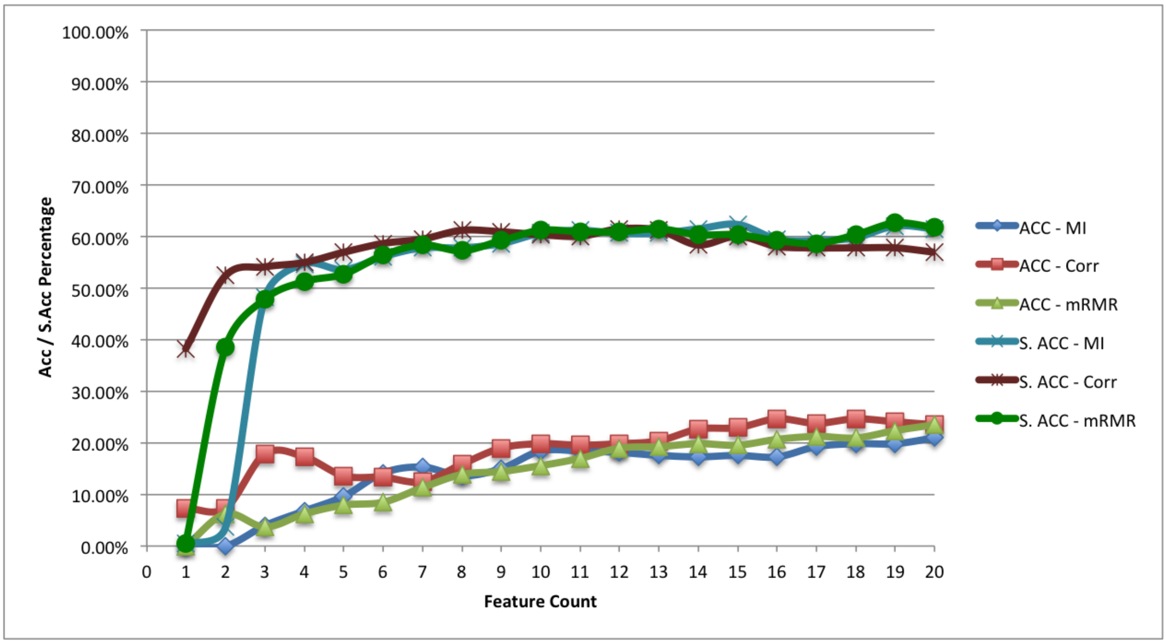
(a) Risk 6 results

(b) Risk 7 results

Fig. 5.4 Accuracy and shifted accuracy for suicide risk levels 6 patients (Figure a) and 7 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for mutual information *MI*, correlation *Corr* and minimum redundancy maximum relevance *mRMR* criteria.



(a) Risk 8 results



(b) Risk 9 results

Fig. 5.5 Accuracy and shifted accuracy for suicide risk levels 8 patients (Figure a) and 9 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for mutual information *MI*, correlation *Corr* and minimum redundancy maximum relevance *mRMR* criteria.

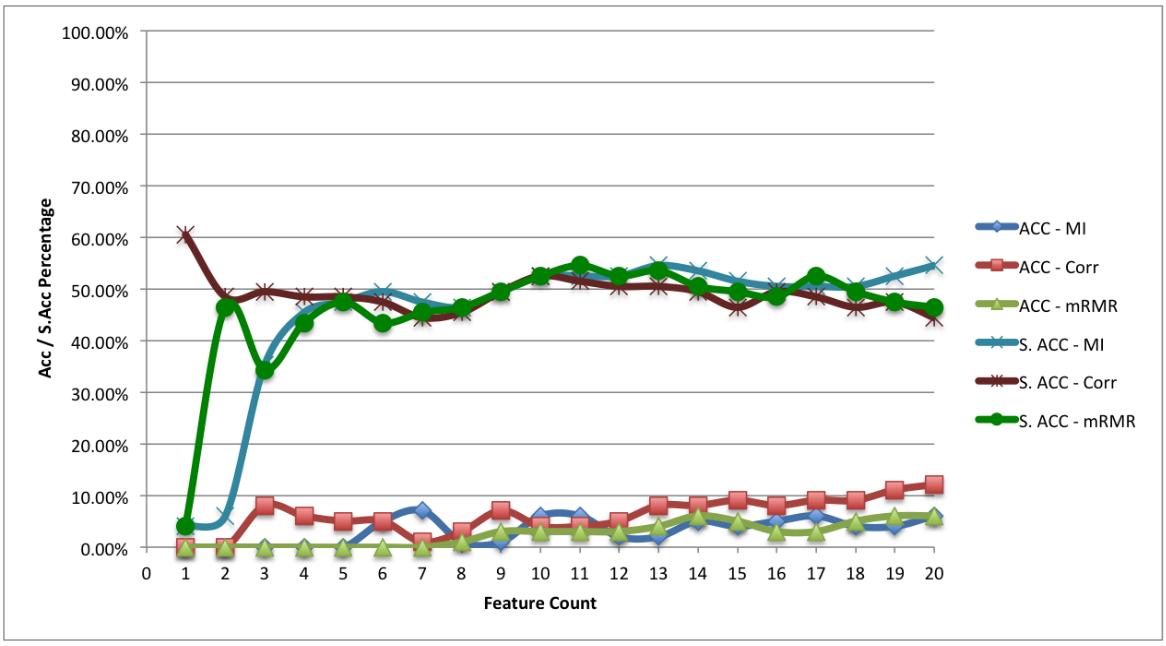


Fig. 5.6 Accuracy and shifted accuracy for risk 10 using feature counts ranging from 1 to 20. The results are shown for mutual information *MI*, correlation *Corr* and minimum redundancy maximum relevance *mRMR* criteria.

Figure 5.1b shows that using a single feature, the algorithm was able to classify risk 1 patients correctly with a high accuracy. There was a decrease in the accuracy as more features were added. On the other hand, the S.Acc curves were almost constant along most feature counts. This shows that some of the records initially classified as risk 1 (whether they were actually risk 1 or not) are now classified as risk 0. Thus showing the improvement in Figure 5.1a versus the degradation in Figure 5.1b. The shifted accuracy curves were almost constant with more than 6 features indicating that any alterations are within the 1 class tolerance.

Figures 5.2a to 5.5a show that risks ranging between 2 and 8 were almost constant using the three methods when 7 features or more were chosen. The difference between the three methods were minimal when it comes to accuracy and shifted accuracy. For the 7 figures, using a single feature, correlation has produced the best accuracy results. Thus *Corr* was stronger than *MI* and *mRMR* in identifying a stronger initial feature in the feature set.

Risk 9 and 10 results are shown in Figures 5.5b and 5.6. The two curves show an obvious variation for each addition of a feature. The reason behind this is that risk 9 population is represented by 353 records and risk 10 by 99 records resulting in a significant change in the chart with minor changes in the results. This was not obvious in lower risk levels because they have much higher records count, such as risk 0 having 13627 records and risk 1 having 15358 records. Although the higher risk patients are a much smaller count that the lower risk, there is an increased importance of getting their assessment correct.

To sum up, 0 risk might need more variables to be accurately classified but fewer variables are better for minimising the missing information, increasing effectiveness of regression and maximising speed. Correlation showed a small improvement in the accuracy in risk levels 0 and 10. It chooses the features with highest linearity with the risk level, which would explain why using a single feature in *Corr* curves was better than *MI* and *mRMR*. Another advantage of correlation is that it produced the best results given feature count lower than 10, which could be the case for some patients whom did not provide enough information. Finally correlation managed to have the most stable results through the different feature counts therefore it is the method to be used in the rest of the experiments.

5.2.2 Setting *DV-Thresh* parameter

The second part of the experiment aims to test the use of the *DV-Thresh*, which was introduced in Algorithm 3 in chapter 4. This threshold parametrised the feature selection algorithm. The features are selected when their correlation with the risk level exceeds the threshold defined. This makes the selected variables having a significant influence on predicting the DV, where the level of influence is *DV-Thresh* controlling when a variable is not selected. At this stage, the *IV-Thresh* is not used so that the influence of *DV-Thresh* alone could be monitored and understood.

Since very few features had a correlation with the DV higher than 0.5, thresholds were tested for values ranging from 0.5 to 0.1 as shown in Table 5.1. The table shows the accuracy (*ACC*) and shifted accuracy (*S.ACC*) for each threshold value. The *Mean num of feats* represents the average number of features selected for patient records given the threshold defined. This number increases as the threshold decreases since more features are found complying to the limit. *Mean sample size* defines the average size of the generated training subset for patient records. The sample size of the selected records from the training set decreases as more features are added because it becomes harder to find records not missing the selected features. *drop-outs* are the number of records that did not have any features with the given threshold and so were dropped out of the analysis.

	DV-Thresh	0.5		0.4		0.3		0.2		0.1	
	Mean num of feats	2.162		3.327		4.806		7.969		12.508	
	Mean sample size	12135		10874		5746		4844		3051	
	drop-outs	3365		2555		2210		74		40	
		ACC	S.ACC	ACC	S.ACC	ACC	S.ACC	ACC	S.ACC	ACC	S.ACC
Risk Categories	0	0.91%	94.90%	0.01%	93.55%	0.42%	93.26%	4.36%	90.83%	12.42%	89.47%
	1	80.47%	92.73%	79.07%	93.59%	78.19%	93.76%	69.67%	93.79%	62.99%	93.60%
	2	25.78%	91.11%	30.17%	93.66%	31.60%	93.98%	40.08%	93.74%	41.91%	91.79%
	3	32.88%	68.81%	33.57%	70.90%	33.01%	71.55%	35.94%	81.62%	35.96%	82.21%
	4	37.09%	76.74%	32.87%	76.06%	32.28%	75.21%	32.92%	77.38%	32.39%	75.64%
	5	29.71%	73.03%	29.97%	70.70%	28.78%	68.77%	27.76%	68.66%	26.20%	67.29%
	6	27.05%	66.84%	29.45%	69.41%	27.68%	68.22%	28.25%	66.52%	24.75%	60.79%
	7	26.38%	63.60%	29.77%	68.53%	31.02%	69.24%	29.80%	67.30%	22.60%	58.25%
	8	20.77%	59.45%	23.22%	63.41%	25.66%	65.89%	25.05%	63.33%	18.10%	52.29%
	9	13.78%	54.55%	14.45%	56.94%	18.73%	59.94%	24.08%	58.36%	24.36%	55.81%
	10	5.32%	52.13%	2.11%	47.37%	5.26%	52.63%	11.11%	47.47%	18.18%	44.44%

Table 5.1 Results of testing the algorithm with *DV-Thresh* ranging from 0.5 to 0.1: the *ACC* column is the percentage of predictions exactly matching the clinical judgement and the *S.ACC* column is the percentage that either match the judgement or are one away from it.

Table 5.1 shows that lowering the threshold results in an improvement as more features are selected. The accuracy for the higher values of the threshold were calculated on smaller number of records since the dropped cases reduced the test records' sample size. Lowering the threshold caused the *Mean sample size* to decrease and the average number of features to increase. Given the specifications of the dataset and the high percentage of missing data, finding a proper subset of training data to create a model for a high feature subset count will be impossible. This way, features will need to be removed to increase the sample size of the training subset. To overcome this dilemma, the need to add a feature count limit emerged.

Looking back at Figures 5.1 to 5.6, it was deduced that the accuracy and shifted accuracy after 9 features were selected was almost constant for all risk levels. In order to test the current theory that a limit to the number of features was needed, feature count value was chosen to be 12. The value 12 would assure that, for all risk levels, the number of chosen features would be low enough to reduce complexity and at the same time high enough to avoid the disruptions in the lower count features therefore providing good accuracy.

	DV-Thresh	0.2		0.1		0.05		0.04		0.03	
	Mean Num of feats	7.71		10.23		11.67		11.83		11.87	
	Mean Sample Size	5099		4142		3556		3555		3371	
	Dropouts	74		40		7		7		7	
		ACC	S.ACC	ACC	S.ACC	ACC	S.ACC	ACC	S.ACC	ACC	S.ACC
Risk Categories	0	4.33%	90.82%	11.73%	89.34%	32.76%	90.09%	33.15%	90.12%	33.00%	90.15%
	1	69.53%	93.79%	63.19%	93.67%	57.37%	94.04%	57.11%	94.04%	57.33%	94.04%
	2	40.11%	93.69%	42.98%	92.56%	41.94%	90.90%	41.92%	90.88%	41.90%	90.92%
	3	36.05%	81.63%	36.64%	83.56%	36.10%	83.11%	36.05%	83.12%	36.08%	83.19%
	4	33.07%	77.21%	34.37%	78.29%	34.14%	77.89%	34.20%	78.04%	34.29%	78.08%
	5	27.78%	68.96%	27.62%	70.75%	27.70%	70.65%	27.67%	70.70%	27.67%	70.70%
	6	29.20%	68.08%	28.92%	67.91%	29.09%	67.91%	28.98%	67.91%	29.03%	67.85%
	7	31.47%	68.23%	30.85%	67.98%	30.83%	68.06%	30.83%	68.12%	30.83%	68.18%
	8	26.86%	64.38%	28.00%	64.00%	27.71%	63.90%	27.71%	64.00%	27.71%	63.90%
	9	22.10%	58.07%	23.51%	59.21%	23.51%	59.49%	23.51%	59.77%	23.51%	59.49%
	10	10.10%	48.48%	11.11%	48.48%	11.11%	48.48%	11.11%	48.48%	11.11%	48.48%

Table 5.2 Results of testing the algorithm with *DV-Thresh* 0.2, 0.1, 0.05, 0.04 and 0.03 while setting the feature limit to 12: the *ACC* column is the percentage of predictions exactly matching the clinical judgement and the *S.ACC* column is the percentage that either match the judgement or are one away from it.

Setting the feature limit to 12, the *DV-Thresh* test was repeated with thresholds 0.2, 0.1, 0.05, 0.04 and 0.03 as shown in Table 5.2. Analysing the difference between the 0.1 columns in both Tables 5.1 and 5.2 shows that limiting the feature count introduced a few improvements. For instance, most risks have improved accuracy and shifted accuracy. To ensure the need for the feature limit, the records for the patients for which the algorithm previously chose more than 12 features using the 0.1 *DV-Thresh* were extracted and their results were compared with and without the limit. The number of records extracted were 32,394 out of the complete dataset of 71,024 in all 10 folds. The results are shown in Table 5.3 where it can be seen that the addition of the feature limit improved most of the accuracy and shifted accuracy results. This is because, the higher feature count causes a smaller training set size therefore resulting in less accurate results.

Risk	Without feature limit		With feature limit	
	Accuracy	Shifted Accuracy	Accuracy	Shifted Accuracy
0	17.96%	77.77%	13.51%	76.92%
1	45.67%	88.67%	46.39%	88.89%
2	45.88%	88.88%	48.15%	90.47%
3	40.64%	86.58%	41.68%	88.70%
4	35.67%	79.82%	38.35%	83.41%
5	28.37%	71.34%	30.17%	75.70%
6	25.79%	62.89%	30.72%	71.32%
7	23.79%	60.87%	33.22%	71.99%
8	18.53%	55.37%	30.04%	68.86%
9	26.47%	59.15%	25.49%	63.07%
10	20.93%	48.84%	12.79%	53.49%

Table 5.3 Results of comparing the effect of the addition of a 12 feature limit to records previously chose more than 12 using a *DV-Thresh* of 0.1.

Extra analysis was performed to understand the reason behind the 7 drop outs in the lower thresholds. The 7 records represented patients that answered at most 9 features out of the 176. The questions answered were general questions like age, gender, martial status, sharing accommodation and ethnicity. The correlation of such features with the risk level ranged between 0 and 0.0107. All those 7 patients were clinically assessed for risk ranging between 1 and 3. From this it was deduced that there is no need to decrease the threshold for such cases since it will decrease the *Mean sample size* which will reduce the quality of the regression models.

To conclude, as the threshold decreased improved accuracy is introduced especially for the risk 0 patients but a feature limit was required to enhance the *Mean sample size*. From Table 5.2, the *DV-Thresh* was chosen as 0.04 since lowering it further is not introducing much improvement.

5.2.3 Setting *IV-Thresh* parameter

The feature selection performed until this stage considered only the relationship between IVs and the DV. In this subsection, the introduction of the *IV-Thresh* is tested. The existence of redundant features among the selected subset could cause degradation in the classification accuracy. Since this research is not only interested in the accuracy, but in delivering a set of features to the assessor that best described the patient. This set of features would not be considered a good output to the assessor if it included redundant features.

Based on the previous experiment, *DV-thresh* is set to 0.04 and the feature limit to 12 as they produced the best results. The method used here is mutual information which will measure the amount of shared information among the features. As stated earlier, the mutual information is a common method for limiting redundancy since it measure the dependency. In Table 5.4, values ranging from 0.4 to 0.9 were tested. The final column shows the results achieved from the previous test where the feature limit is 12 and the *DV-thresh* is 0.04. From the table, it can be seen that as the threshold increased, the *mean number of features* increased and the *mean sample size* increased.

	IV-Thresh	0.4		0.5		0.6		0.7		0.8		0.9		No IV-Thresh effect	
	Mean Num of feats	11.69		11.747		11.805		11.82		11.829		11.829		11.829	
	Mean Sample Size	2430.7		2430.7		2698.9		2978.6		3415.3		3516.9		3562.8	
		ACC	S.ACC	ACC	S.ACC	ACC	S.ACC	ACC	S.ACC	ACC	S.ACC	ACC	S.ACC		
Risk Levels	0	34.20%	91.29%	34.68%	90.84%	35.94%	90.73%	34.70%	90.49%	33.48%	90.06%	33.28%	90.07%	33.15%	90.12%
	1	59.27%	94.36%	57.82%	94.20%	57.07%	94.17%	56.99%	93.98%	56.96%	94.01%	57.02%	94.07%	57.10%	94.04%
	2	39.27%	90.15%	40.24%	90.08%	40.96%	90.20%	41.14%	90.51%	41.77%	90.85%	41.92%	90.92%	41.93%	90.88%
	3	33.50%	80.43%	34.77%	81.95%	34.74%	82.23%	35.02%	82.59%	36.30%	83.22%	36.22%	83.12%	36.04%	83.12%
	4	30.63%	74.40%	32.47%	76.31%	32.97%	77.14%	33.35%	77.62%	34.35%	78.48%	34.31%	78.21%	34.20%	78.04%
	5	26.79%	66.48%	28.26%	68.44%	28.16%	70.37%	27.85%	70.93%	28.16%	71.16%	27.80%	70.52%	27.67%	70.70%
	6	26.86%	66.30%	26.75%	68.13%	26.86%	67.91%	26.97%	67.69%	28.36%	67.96%	28.81%	68.13%	28.98%	67.91%
	7	25.23%	63.14%	26.58%	65.17%	28.31%	66.46%	28.06%	67.51%	30.22%	68.68%	30.58%	68.43%	30.83%	68.12%
	8	23.14%	58.86%	23.62%	60.48%	23.90%	61.33%	26.00%	62.86%	27.81%	64.38%	28.00%	64.38%	27.71%	64.00%
	9	20.40%	59.77%	20.68%	59.49%	21.53%	59.77%	22.10%	58.07%	25.50%	59.21%	23.51%	60.06%	23.51%	59.77%
	10	11.11%	44.44%	10.10%	49.49%	11.11%	47.47%	11.11%	48.48%	10.10%	46.46%	11.11%	47.47%	11.11%	48.48%

Table 5.4 Results of accuracy and shifted accuracy for *IV-Thresh* ranging between 0.4 and 0.9 as compared to the results achieved by having no *IV-Thresh*. All results are based on setting the *DV-Thresh* to 0.04 and feature limit to 12 features.

The results show that for risks 0 and 1, better results were produced with lower values of *IV-Thresh*. Risk 2 performed better with either *IV-Thresh* having a value of 0.9 or without using it. Risks 3, 4 and 5 performed best at *IV-Thresh* value of 0.8. The shifted accuracy was best for risks 6 and 7 with threshold 0.8. Both accuracy and shifted accuracy improved for risk 8 when using *IV-Thresh* of 0.8. Risk 9 had an improved accuracy with *IV-Thresh* of 0.8 and improved shifted accuracy with *IV-Thresh* of 0.9. The risk 10 patients had best shifted accuracy results when *IV-Thresh* was equal to 0.5.

There is an improvement in different risk classes using *IV-thresh* as 0.8 and 0.9 over the *No IV-Thresh effect* column. Since redundancy is a factor that degrades the performance of classifiers and can cause part of the resulting explanation to be not very informative, it was then decided to use 0.8 as a threshold.

Actual	Predicted Risk										
Risk	0	1	2	3	4	5	6	7	8	9	10
0	4728	7603	1106	147	32	8	3	0	0	0	0
1	2300	9866	4104	823	163	42	11	4	0	0	0
2	551	4860	6319	2721	681	179	35	9	2	0	1
3	89	1188	3314	3730	1754	449	96	25	5	1	1
4	16	182	682	1524	1739	784	231	42	12	1	1
5	3	65	251	658	1194	1096	501	133	33	1	0
6	3	8	35	134	360	520	485	212	37	3	1
7	4	11	12	65	135	291	489	456	152	9	1
8	1	3	5	23	52	96	203	334	273	53	7
9	0	0	1	5	6	17	41	78	122	78	5
10	0	1	1	1	3	0	9	13	23	37	11

Table 5.5 Confusion matrix for predicting clinical risk judgements using 0.8 *IV-Thresh*, 0.04 *DV-Thresh* and 12 feature limit .

The confusion matrix of the results obtained with *IV-Thresh* of 0.8, *DV-Thresh* of 0.04 and feature limit of 12 is shown in Table 5.5. The matrix shows that the results are closer to the diagonal and minimal records were classified in the extreme corners. The values along the diagonal represent the records that were correctly classified for each risk level. Having the values closer to the diagonal means that the misclassified records were quite close to their actual value. Further analysis of the mean absolute error will be shown in a later experiment.

To sum up these experiments, multiple tests were performed to: determine the proper method to use; determine the value of *DV-thresh* and the feature limit; and determine the value for *IV-thresh*. However, the tests have used linear regression where the output value is on a continuous distribution from 0 to 10. When comparing accuracy of prediction, this continuous value has to match a discrete number. To avoid this, logistic regression could be used where it classifies into one of the 11 values. This raised the question of whether linear regression should be replaced by multinomial logistic regression and was investigated next.

5.3 Experiment 2: Applying multinomial logistic regression

Experiment 2 explored the performance of multinomial logistic regression. The aim was to test the use of a regression methodology that could handle discrete classes and see how it compares with the current version of the DFSP. The results are shown in Table 5.6. They do show an improvement in the accuracy of risks 0 and 10 but there was a substantial decrease in the accuracy of risks 1, 2, 4, 6 and 9. There was also a large degrade in the results of the shifted accuracy of risks 2, 3, 4, 5, 6 and 7. The results show that logistic regression has a tendency to favour certain risks given this dataset. For example, for the medium risk ranging between 4 and 6, logistic regression tends to choose risk 5 more often thus improving its accuracy while degrading the other two. When it comes to risks 0 and 10 it performed well but with the cost of degrading the accuracy and shifting accuracy of multiple risks in between. On the other hand, linear regression is better than logistic as the results are more consistent across the different levels.

Risk	DFSP using logistic regression		DFSP using linear regression	
	Accuracy	Shifted Accuracy	Accuracy	Shifted Accuracy
0	81.81%	94.17%	33.48%	90.06%
1	25.93%	96.70%	56.96%	94.01%
2	29.39%	62.38%	41.77%	90.85%
3	37.62%	64.16%	36.30%	83.22%
4	11.57%	67.51%	34.35%	78.48%
5	31.74%	43.71%	28.16%	71.16%
6	0.83%	59.68%	28.36%	67.96%
7	38.52%	53.42%	30.22%	68.68%
8	38.10%	72.00%	27.81%	64.38%
9	7.37%	70.25%	25.50%	59.21%
10	20.20%	48.48%	10.10%	46.46%

Table 5.6 Results for using logistic regression versus linear regression inside the DFSP algorithm

The GRiST dataset is defined by 11 risk levels, ranging from 0 to 10. The problem here is that linear regression expects the distribution of the DV to be normal meaning that it should be varying infinitely in both directions (Strickland, 2015). The GRiST DV is cut-off at both extremes by defining a specific range. This presented the need to examine the use of GLM instead of linear regression. GLM does not obligate the normal distribution of the DV therefore could produce better results on the upper and lower extremes as they will map the results to the distribution. The GLM experiment is shown in the following section.

5.4 Experiment 3: Testing GLM

The GLM is a generalised form of linear regression, which handles data having a distribution that is not normal. It uses a link function to map the output from the ordinary linear regression models to the data distribution (Witten and Frank, 2005). In this test, the linear regression was replaced by the GLM with a poisson distribution and a log link to test if the results would improve around the lower and upper extremes. The choice of poisson was based on the distribution of the risk levels shown previously in Figure 2.4 which resembled the curve of poisson and has similar specification as in being non negative and discrete. The default link is the log since poisson regression is based on the assumption that the logarithm of the mean of the DV can be modelled by linear regression. The log link also fits the choice since it links non negative integer response.

The results are shown in Table 5.7 and they show that the accuracy and shifted accuracy were improved for risk levels 2, 3, 4, 9 and 10 but degraded in the rest. The results on some risk levels did show a slight improvement but it was a degradation for others. The GLM did improve the results of risks 9 and 10 patients but degraded the results in risks 0 and 1. This introduced the need for a more robust methodology that would handle the distribution and improve the current results of the DFSP. In the next subsection, analysing the clinical versus the mean predicted judgement will be addressed. Also the mean absolute error for each predicted risk will be analysed. This will lead to the proposal of an equation that will adapt the output produced by the DFSP to the limitations of the linear regression. Further analysis will be presented along with a proposed solution.

Risk	GLM		DFSP	
	Accuracy	Shifted Accuracy	Accuracy	Shifted Accuracy
0	0.04%	79.99%	33.48%	90.06%
1	48.11%	93.32%	56.96%	94.01%
2	52.79%	95.25%	41.77%	90.85%
3	40.91%	90.60%	36.30%	83.22%
4	36.38%	81.22%	34.35%	78.48%
5	26.07%	68.28%	28.16%	71.16%
6	26.03%	64.57%	28.36%	67.96%
7	26.65%	64.06%	30.22%	68.68%
8	24.29%	61.81%	27.81%	64.38%
9	27.20%	62.61%	25.50%	59.21%
10	32.32%	59.60%	10.10%	46.46%

Table 5.7 Results of applying GLM versus linear regression

5.5 Experiment 4: Adapting output

Since the previous experiments did not provide the required improvement to the results achieved by DFSP in Experiment 1, further analysis was performed to understand the nature of the resulting judgement. Figure 5.7 shows the relationship between the clinical judgement supplied by the practitioners and the mean predicted risk produced by the algorithm. Analysing the figure shows that the 0 and 1 risk patients tend to be classified higher than their clinical judgement and starting risk 3 the mean predicted risk has a shift downwards that is propagated to the extent that risk 10 records are classified on average as risk 8. This is interpreted as the result of the regression being shifted towards the higher accumulation of the population (risk 1) where risk lower than it is classified upwards and risk higher than it is classified downwards. This population distribution of GRiST was shown previously in Figure 2.4 where the larger counts of the population were close to risk level 1 and decreased as risk increased to 10.

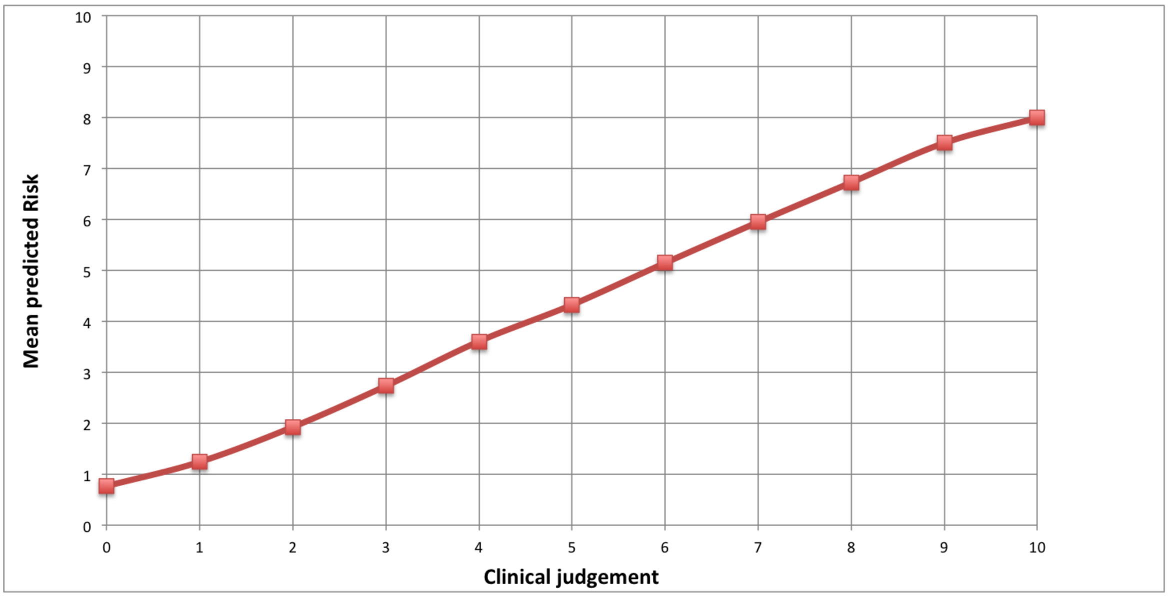


Fig. 5.7 The clinical judgement versus mean risk prediction

Another analysis was shown in Figure 5.8, this chart represents the mean absolute error at every risk level. The mean absolute error shows the average of the absolute difference between the clinical and predicted risks. The figure shows that risk 1, which has the highest population, has the least mean absolute error. The error increases as the risk increases where the size of the risk population decreases. The problem is, based on both Figures 5.7 and 5.8, there is a need to push the prediction values away from the centre of the population and extend them to reach the further ends of the upper and lower risks, namely risks 10 and 0. This introduced the need to find a solution that would enhance the resulting prediction by pushing it towards the extremes.

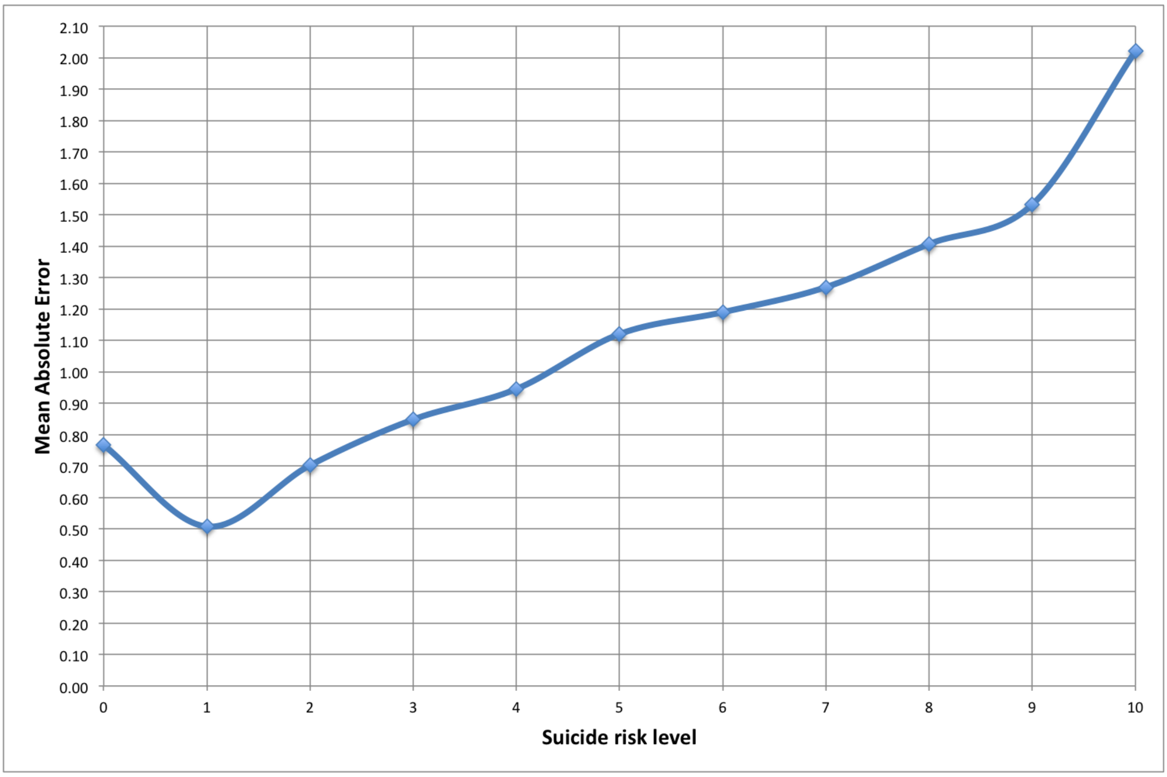


Fig. 5.8 Mean absolute error for each risk level

In GRiST, the data are cut off at the extremes, which means the regression algorithm cannot provide weights that work in the middle as well as at the extremes. This is caused by the properties of the features that are bounded in a specific range. This can be shown by the distribution of prediction errors, which increase at both poles. Equation 5.1 was designed to shift the prediction towards the extremes, the further it was away from the mean. This equation is considered another contribution in the thesis that enhanced the results produced by the DFSP.

$$NewPred = OldPred + \frac{prediction - mean}{par * (upper - lower)} \tag{5.1}$$

In equation 5.1, $OldPred$ stands for the result of the regression prediction, $mean$ is the average value of the prediction risk calculated from the training set. $upper$ and $lower$ represent 10 and 0 respectively where the difference between them specifies the range of the risk predictions. par is a user fed value that is used to adapt the result. The equation adds to the predicted risk a value that is either positive or negative. This value represents the amount of pulling away from the mean value that it is required depending on both the $Oldpred$ and the par. The resulting prediction is named $NewPred$.

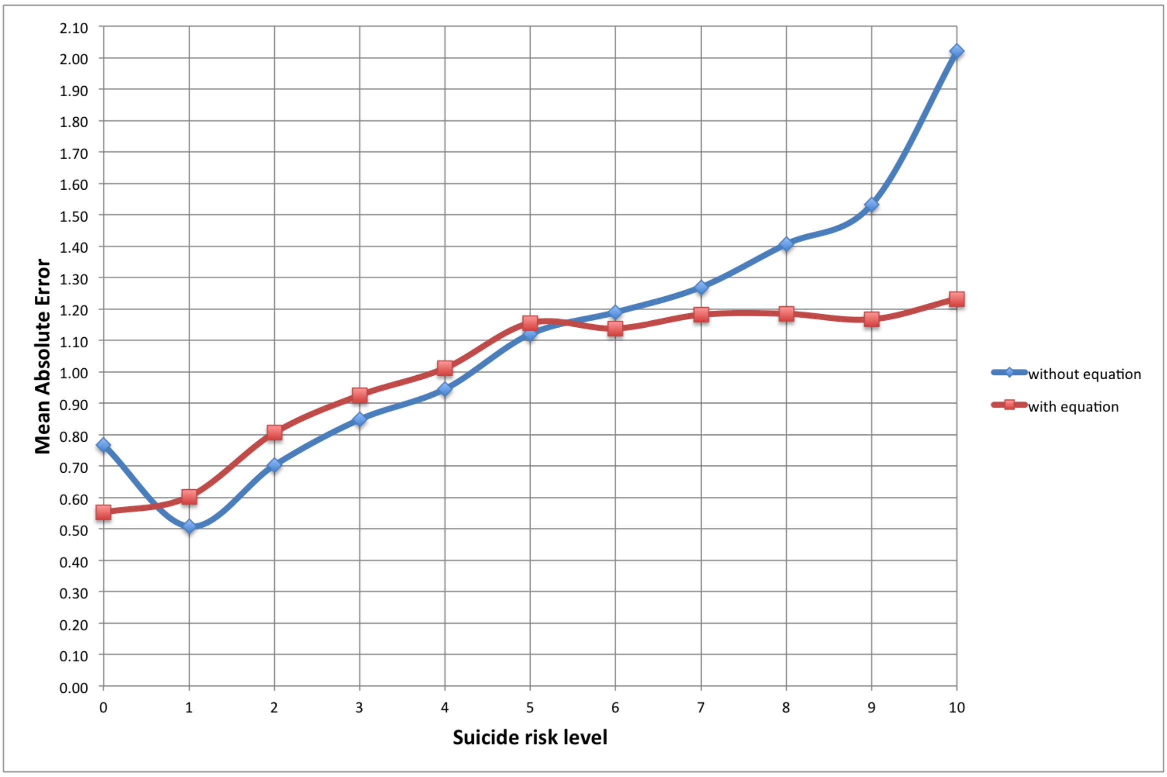


Fig. 5.9 Mean absolute error for each risk level with and without the equation

Different values were tested for *par* and the best results were produced at value 0.75. Figure 5.9 shows the improvement to the mean absolute error when the equation was applied with *par* value 0.75. The figure shows that there is an improvement in the 0 risk level error and in all levels above 5 with a small degradation in the range 1 to 4 that was caused by the spread to both poles.

Table 5.8 show the confusion matrix after applying the equation. This table is intended to confirm that the classification of all risks is still close to the diagonal and the amount of misclassification of lower risk towards upper and the opposite did not increase extremely. Comparing the accuracy of risks 0 and 1 in this matrix to that of Table 5.5, the value of risk 0 has improved from 4728 to 7566 and the value of risk 1 has improved from 11 to 43. This shows that the equation did improve the results on both extremes. Examining the remaining values of the diagonal, risks 1 to 5 showed a drop in the accuracy caused by the shifting towards the extremes.

Actual Risk	Predicted Risk										
	0	1	2	3	4	5	6	7	8	9	10
0	7566	4891	957	161	37	10	5	0	0	0	0
1	4328	8362	3538	807	196	55	18	9	0	0	0
2	1186	4669	5695	2713	780	229	66	14	4	2	0
3	198	1248	2904	3543	1929	630	150	34	11	4	1
4	30	198	603	1266	1670	961	362	97	23	3	1
5	8	73	223	544	993	1067	694	246	72	12	3
6	3	7	39	90	265	414	516	334	102	27	1
7	3	9	14	59	93	201	338	483	340	75	10
8	2	2	4	18	28	71	139	206	325	219	36
9	0	0	1	4	5	11	21	48	81	110	72
10	1	0	1	1	2	0	5	4	11	31	43

Table 5.8 Confusion matrix for predicting clinical risk judgements after the application of the equation.

Finally the results of the DFSP with and without the equation are shown in Table 5.9. The improvement for risks above level 5 is shown in both accuracy and shifted accuracy.

Also there was a large improvement in the risk level 0 accuracy. This represented the final stage of experimenting with the DFSP. The results after the application of the equation are compared to gold standard approaches in the following experiment.

Risk	Without Equation		With Equation	
	Accuracy	Shifted Accuracy	Accuracy	Shifted Accuracy
0	33.48%	90.06%	54.66%	91.30%
1	56.96%	94.01%	48.72%	93.75%
2	41.77%	90.85%	37.37%	85.53%
3	36.30%	83.22%	33.48%	78.93%
4	34.35%	78.48%	32.26%	74.93%
5	28.16%	71.16%	27.22%	70.22%
6	28.36%	67.96%	29.20%	70.41%
7	30.22%	68.68%	30.34%	71.02%
8	27.81%	64.38%	31.71%	71.81%
9	25.50%	59.21%	33.14%	73.94%
10	10.10%	46.46%	42.42%	74.75%

Table 5.9 Result of DFSP with and without the equation

5.6 Experiment 5: Comparison to other standard approaches

The purpose of this experiment is to compare DFSP with standard approaches of feature selection, missing data handling and classification. The approach addressed for missing data handling is mean imputation and expectation maximisation. For feature selection, the algorithm used, which has highest resemblance to the DFSP, is the CFS which was previously discussed in chapter 3. Finally the prediction algorithms used are linear regression, decision trees and random forest.

5.6.1 Applying standard approaches

Given the specifications of the GRiST dataset, if methods with extensive calculations were to be applied, they would be computationally very expensive. Thus there is quite a limitation on the methods that could be applied on datasets with such a large percentage of missing data. The computational expensive methods cannot be applied directly onto the datasets themselves so only the simple methods like mean imputation could be used. To apply more powerful methods, the size of the dataset needs to be reduced first by feature selection. This results in minimising the amount of missing data that needs replacement and the information available for the imputation.

The gold standard approaches for prediction are decision trees and random forests. These two algorithms internally treat missing data by splitting the value of the missing feature in the test record according to a probability weighting scheme that is determined from the training set. This means that for a given node, the probability of the different values that it can be split into is the assigned value for each branch. The split parts are input to branches of the tree and then the results are combined using weights at the leaf level (Witten and Frank, 2005).

Also linear regression is to be used in the comparisons since it is a building block of the DFSP. The difference between this linear regression and the one within the DFSP is the features input to it. The input features to the regression will be applied in two ways: once using the whole feature set while imputing the missing values with the mean; and another by applying the CFS algorithm and following it with EM to impute missing values. This will show whether the DFSP was an improvement to the performance of the linear regression or not.

Different combinations of missing data handling, feature selection and prediction algorithms were used for the comparisons. The abbreviations used for the combinations in the tables are:

- DT: Decision Trees.

- RF: Random Forest.
- Mean + LR: Mean imputation followed by Linear Regression.
- CFS+EM+LR: Correlation based Feature Selection followed by Expectation Maximisation imputation and Linear Regression.
- Mean+DT: Mean imputation followed by Decision Trees.
- CFS+EM+DT: Correlation based Feature Selection followed by Expectation Maximisation and Decision Trees.
- CFS+DT: Correlation based Feature Selection followed by Decision Trees.

The first reason behind the choice of these methods was to show a comparison to strong and effective classifiers like decision trees and random forest. Second reason is to show the effect of imputation methods on the classification. They could be simple mean imputation that would be applied on the whole dataset or more sophisticated imputation like expectation maximization which has to be applied after a feature selection algorithm since the percentage of missing values is very high. Finally the effect on the use of a feature selection algorithm on the decision trees, which handles missing data internally. The results of the different methods were performed using the WEKA 3.7 software (Hall et al., 2009). The analysis was performed on the same dataset, using 10 fold cross validation.

Table 5.10 shows the comparison between DFSP and the other methods when it comes to accuracy of correctly classifying a risk. For each column of the other methods the results are shifted towards lower risk more than upper except for *DT* and *mean+DT*. Yet, both were much weaker than the DFSP when it comes to the risks 1, 2 ,3 ,4 , 6, 9 and 10.

Risk	Different Methods							
	DFSP	DT	RF	Mean + LR	CFS + EM +LR	Mean + DT	CFS + EM + DT	CFS+DT
0	55.52%	69.35%	63.37%	64.78%	66.82%	57.93%	59.64%	67.98%
1	48.30%	46.95%	60.06%	46.06%	42.88%	45.23%	46.55%	44.11%
2	37.08%	35.69%	44.91%	32.62%	20.80%	37.75%	32.27%	28.14%
3	33.26%	41.12%	34.31%	33.69%	38.91%	32.73%	33.40%	36.77%
4	32.03%	16.76%	14.33%	3.26%	0.02%	23.99%	13.81%	9.11%
5	27.12%	29.05%	22.13%	16.98%	29.82%	23.38%	16.73%	23.79%
6	28.70%	8.84%	8.23%	0.22%	0.00%	16.02%	8.18%	1.17%
7	29.72%	34.22%	31.14%	20.43%	1.66%	24.68%	21.86%	27.08%
8	30.95%	29.71%	27.43%	22.10%	0.00%	29.62%	23.90%	16.57%
9	31.16%	24.36%	8.78%	0.00%	0.00%	25.21%	9.35%	3.40%
10	43.43%	32.32%	6.06%	0.00%	0.00%	25.25%	0.00%	0.00%

Table 5.10 Results of comparing accuracy when classifying suicide risk in GRiST dataset to different off the shelf methods

RF has provided good results but not for risks 4, 6, 9 and 10. General thing to note is that most classifiers did not function their best at the risks 4, 6 and 9. Analysing the results of the different methods shows that there is a tendency to classify towards lower risk. The misbehaviour in risks 4 and 6 for most classifiers can be explained as them representing boundary points. Anything under 6 is not considered high risk and above 6 is going to be treated as high risk. In effect, the 4, 5 and 6 risks are being treated as medium risk where there is a high similarity between the patients' records. This makes it even harder for a classifier to pick out the exact right risk level and preferring one over another. This was not the case for DFSP.

Table 5.11 showed that DFSP excelled over all the other methods for risk levels ranging 3 to 10 using the shifted accuracy. The second best was the *DT* classifier followed by the *mean+DT*; this shows that the decision tree's internal treatment of missing data is better than that of the mean imputation. The addition of the CFS to the decision trees caused a huge degradation compared to the results of the *DT*. This is because the data is missing a large percentage, so cutting down on the information available by feature selection caused a deficiency. This was not the case for the DFSP because the features selected for each record were the most suitable from the non-missing information supplied and the training subset resembled it. This means that the patient records were not constrained with a specific set of selected features like any ordinary feature selection algorithm.

When the EM imputation was added to the CFS subset then followed by decision trees, better results were produced. Those results were better than following the CFS by decision trees. This shows that EM is a strong missing data imputation algorithm. Although EM could have shown good results if applied without feature selection, trials to apply it had failed because it is computationally very expensive. Graham (2009) has explained that EM works best when the addressed number of features is less than 100 since it requires to make thousands of parameter estimates. Another factor mentioned is the percentage of missing data, high percentage would make the EM much slower. Given that GRiST has a missing data percentage of 67.04% and the number of addressed features are 177. That makes EM not suitable to be applied directly on the complete GRiST dataset.

As a final analysis, the mean absolute errors of all the methods were plotted as shown in Figure 5.10. The DFSP has shown to produce the least amount of mean absolute error across all methods for risk levels higher than 3. For risk levels lower than 4 the result across all methods were very close. The *DT* was the second best classifier after the DFSP.

Risk	Different Methods							
	DFSP	DT	RF	Mean + LR	CFS + EM +LR	Mean + DT	CFS + EM + DT	CFS+DT
0	91.41%	95.05%	95.93%	95.05%	95.59%	86.20%	91.98%	95.16%
1	93.73%	95.82%	98.05%	96.97%	93.69%	91.05%	92.84%	93.54%
2	85.15%	85.05%	91.19%	82.25%	78.59%	80.82%	85.57%	83.88%
3	78.63%	76.15%	78.04%	68.61%	60.16%	72.05%	71.33%	68.43%
4	74.74%	71.83%	62.04%	57.52%	66.13%	64.88%	61.53%	66.17%
5	69.99%	45.16%	35.58%	22.59%	29.90%	52.78%	37.17%	36.19%
6	70.30%	58.40%	44.77%	40.88%	52.95%	51.67%	41.71%	51.28%
7	71.45%	54.28%	43.45%	31.08%	1.66%	54.46%	45.94%	36.43%
8	71.43%	67.24%	56.95%	46.67%	1.52%	57.90%	50.76%	46.48%
9	74.50%	62.89%	42.21%	34.56%	0.00%	56.37%	44.48%	34.28%
10	74.75%	52.53%	9.09%	0.00%	0.00%	46.46%	23.23%	9.09%

Table 5.11 Results of comparing shifted accuracy when classifying suicide risk in GRiST dataset to different off the shelf methods

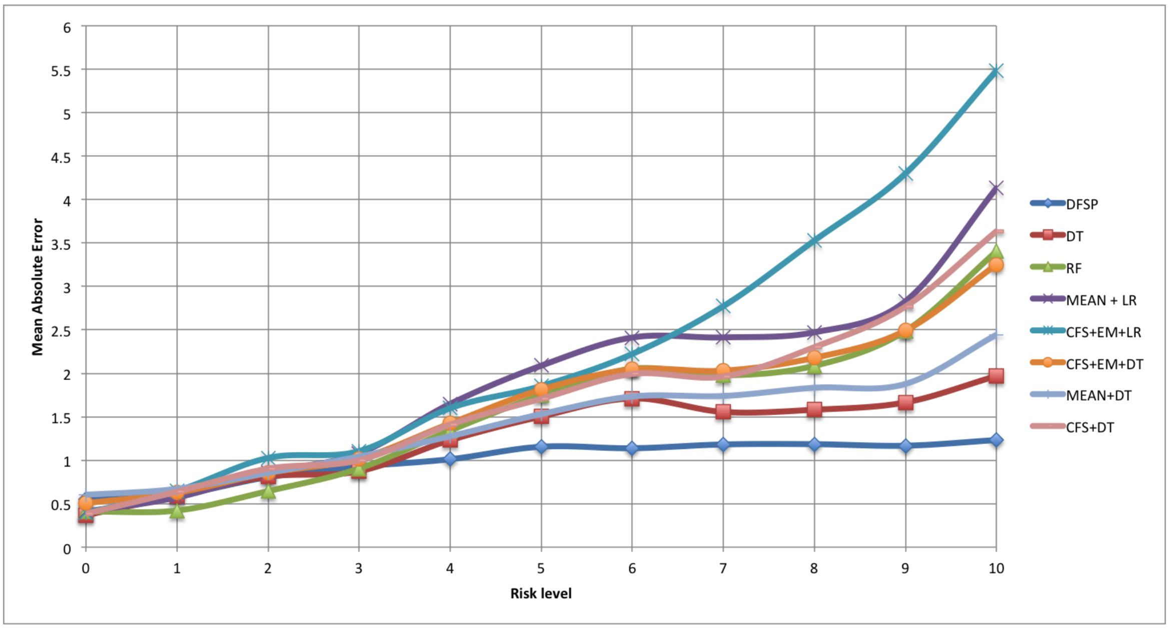


Fig. 5.10 Mean absolute error for each risk level when using DFSP as opposed to other off the shelf methods.

5.7 Summary and conclusions

Multiple experiments were presented in this chapter. The first experiment was to choose an appropriate method for feature selection. The results have shown that *MI*, *Corr* and *mRMR* have high resemblance in accuracy and shifted accuracy given a feature count above 9. Correlation was chosen since it produced the most stable results among all the different risk levels using any feature count.

After the feature selection method was decided, another test was performed to show the effect of applying a threshold to the correlation of selected features and the output risk. The test showed that as the threshold decreased, the results improved since more features were selected. There was a need to decrease the threshold beneath the value 0.1 but this was not possible since it would lead to choosing vectors of very large size and would result in smaller sized training subsets. This introduced the need for a feature count limit.

Feature count limit was tested and showed an improvement in the results. The comparison was using same patients records once with large vector sizes and another with the limit specified. After the feature limit was introduced, more tests to obtain the threshold for the correlation with the output risk were performed and a value of 0.04 was reached. Now features are added according the DFSP algorithm until either the *DV-Thresh* or the feature limit is sustained.

At this stage redundancy among the selected features needed to be tested in order to enhance the results. Further tests were performed to measure the influence of not adding features having a high mutual information with pre selected ones. The results showed an enhancement using the 0.8 threshold.

The results produced formed the baseline to which different experiments were then introduced to improve. The following experiment tested whether the replacement of the linear regression in the DFSP with logistic regression would enhance the results. The merit behind the logistic regression is that it would apply a discrete classification which would fit the GRiST dataset. The output results degraded.

Another experiment was introduced based on the need to push the results towards upper and lower extremes since they restricted between the 0 and 10 range. This forced the experimentation of the GLM classifier using a poisson distribution and a log link. The results did not show any improvement on the low risk but a slight improvement on the upper.

Since the logistic and the GLM failed to improve the results, an equation was introduced to provide a stretch to the predicted risk. The equation managed to push the prediction beyond the cut off points of the linear regression thus enhancing its accuracy. This showed the final results of the DFSP algorithm. The results were then compared to different methods of missing data handling, feature selection and gold standard classification methods. It has shown an improvement over all the methods. In the following chapter, issues regarding the implementation within GRiST CDSS will be discussed.

Chapter 6

Implementing the DFSP within GRiST

6.1 Introduction

The DFSP has managed to handle the difficulties of the GRiST dataset which included: large percentage of missing data; unequal class distributions; a large number of output risk levels; and high dimensionality. At the same time, it compared well with alternative methods. Implementing the DFSP within GRiST is a crucial step to the success of the algorithm. This involves assessing speed of generating prediction and addressing the possibility of enhancing it. GRiST is a real-time support system, which means the speed of assessing patients has to be fast. This has been one of the driving factors in developing the algorithm.

In this chapter, a speed analysis of the current DFSP is described. This informs decisions about how to implement the DFSP within GRiST to optimise its speed. Proposed performance enhancements to the algorithm will be presented. A comparison with decision trees will be shown and evaluated in the light of pragmatic considerations for real-time dynamic predictions .

6.2 Assessing speed

For each patient, in the GRiST dataset, features were selected and a linear regression model was built. Initially the DFSP was intended to run completely online. The feature selection and assessment model creation were performed per patient record. Table 6.1 shows that the total time taken for all records of the dataset was 134,945.6 seconds and the average time per patient was 1.9 seconds. Since the risk assessment is intended for a real-time system, the possibility of improving the average time taken for assessing a single record has to be addressed. In the following section, the algorithm will be analysed with a view to enhancing its speed.

	single record	all records
Assessment time	1.9s	134,945.6s

Table 6.1 Time taken for assessing a single patient on average using the DFSP and total time for all 71,024 records.

6.3 Enhancing DFSP algorithm

Although the DFSP is comparable with other standard approaches in accuracy measures, the possibility of enhancing its speed needed to be addressed. The algorithm will then be divided into two main sections to ease the illustration. It will be divided into a feature selection and a classification stage. Each subsection will show a part of the final version of the DFSP algorithm. The changes performed will not have any effect on the accuracy results but will speed up the feature selection and classification.

Algorithm 4 Final version of the DFSP feature selection stage

```
Require: TestRecords, Corr-IV-DV, IV-IV.
Require: DV-Thresh, IV-Thresh and Count-Limit.
  Sort Corr-IV-DV in descending order.
  Rearrange TestRecords columns with same order as Corr-IV-DV
  for Every SingleTestRecord in TestRecords do
      Initialise $temp$ to Corr-IV-DV, S to empty and ind=1
      Remove all empty features from $temp$ and SingleTestRecord
      while $temp(ind)$<= DV-Thresh & number of features in S<= Count-Limit do
          if $temp(ind)$ is a filter question and S includes a descendant or more then
              discard $temp(ind)$
          else if $temp(ind)$is a descendant and filter question $\in$ S then
              remove filter and add $temp(ind)$ to S.
          else
              if All values of $temp(ind)$ and S in IV-IV <= IV-Thresh then
                  Add $temp$ to S
              end if
          end if
          ind=ind+1
      end while
      Find a complete subset in TrainRecords for features in S
      while Number of records <= 100 * length(S) do
          remove last added feature in S.
          Regenerate a complete subset from the training records
      end while
  end for
```

6.3.1 Enhancing feature selection

Algorithm 4 presents the enhanced feature selection stage of Algorithm 3. It also includes the modifications resulting from the analysis performed in Chapter 5. Algorithm 4 now requires *DV-Thresh*, *IV-Thresh* and *Count-Limit*, which Chapter 5 showed how to determine. It also requires *TestRecords*, *Corr-IV-DV* and *IV-IV*. *TestRecords* is the records in the testing dataset that are going to be assessed, *Corr-IV-DV* is the correlation between each IV and the DV, which in Chapter 5 produced better accuracy results than MI and mRMR. *IV-IV* represents the mutual information among the IVs.

In Algorithm 3 $temp$, which was initialised by all the non-missing values in the patient's record, was searched for the highest value ($candidate$) in every iteration intending to add a new feature to the subset. This is replaced in algorithm 4 by a single descending sort of the *Corr-IV-DV* matrix that is performed once initially before any test patient record is addressed. The sorted *Corr-IV-DV* is now the initialisation of $temp$. Also $TestRecords$ were ordered in the same manner, this makes the first column in $TestRecords$ having the feature with the highest correlation with the DV and the last column having the one with the least. This makes it easier to handle *SingleTestRecord* such that the values missing in it would be in the same position as those in $temp$ thus removed simultaneously. Variable *ind* is introduced to point to the highest value in $temp$ in each feature iteration in a patient record. It is initialised to 1 as the highest value is the first element in $temp$. In every iteration it is incremented to point to the following one. This means that a search operation inside the loop was replaced by an increment instruction which would enhance speed.

For each patient record, a feature set *S* is initialised to be empty and each selected feature is added to it. The ordered set of existing features in *SingleTestRecord* is iterated until either the candidate feature's correlation with the DV is lower than *DV-Thresh* or the number of features in feature set, *S*, is greater than *Count-Limit*. The candidate feature is checked for the filter question and the IV redundancy criteria before being added to *S*. After all the features are added to *S*, a final check is performed to ensure that the selected subset of features would generate a training subset of at least 100 times the number of its features. If it does not, then remove last added feature to *S* and retest. This algorithm shows the final version of the feature selection stage of the DFSP but without the model creation and classification steps which will be discussed next.

6.3.2 Enhancing classification

The drawback of algorithm 3 was that a model needed to be created per patient given the chosen subset of features. After analysing the features chosen for the patients, some feature sets were repeatedly generated among patients. There is no need to regenerate the model because clearly it will select the same feature set and training subset. Hence it makes sense to store the model for each feature set and reuse it.

The algorithm was updated to introduce a new approach using a database of selected features linked with their models. Whenever a patient is assessed, first the selected features would be searched for in the database. If the set exists then the associated model can be directly applied for classification. Otherwise a new model has to be built as described but then stored with its feature set. The cost of sharing regression models far outweighs the cost of learning them and they are not a drain on memory.

Algorithm 5 Final version of the DFSP classification stage

```
Require: $TrainRecords$, $TestRecords$, $FSDB$ and $S$
  if the current set of features is in $FSDB$ then
      Retrieve regression model indexed by $FSDB$
  else
      Find records in $TrainRecords$ that have value for the set of features
      Build the regression model accordingly
      Add the new feature set and the new regression model to $FSDB$
  end if
  Classify the feature set according to the new model
  Apply equation 5.1 to the result
  Modify regression weights using equation 6.1 and show explanation to assessor
```

The idea, as shown in algorithm 5, is to have a database $FSDB$ that would store all the unique feature sets along with their classification models regression equations. Feature set S is generated by the feature selection stage. If the selected features are already in the $FSDB$ then the regression model is directly retrieved.

The stored regression weights are based on the subset of the training set produced from training data using the selected features. Equation 5.1 enhanced the assessment but did not change the regression weights. This makes the regression equation not compatible with the output risk. Since these weights are part of the explanation provided to the assessor, it must match the output risk. A simple modification to the weights of the linear model is introduced that is shown in Equation 6.1 where *OldLReq* is the linear regression model output from the DFSP, *NewPred* is the assessment value after applying the equation 5.1, *OldPred* is the assessment value before applying the equation and *LReq* is the modified linear regression model.

$$LReq = \frac{NewPred}{OldPred} * OldLReq \tag{6.1}$$

The selected features are then used as a reasoning to the assessor for the predicted risk assessment and the weights in *LReq* show the importance of each one. The merit behind the introduction of this concept in the classification is that as the DFSP continues to assess new patients, the database will start to grow. Eventually it will become almost instantaneous to assess and explain a new patient record.

6.3.3 Performance evaluation

Table 6.2 shows the difference between the results of the DFSP algorithm before and after the enhancement. The calculated time here represents how long it took to initiate all the algorithm stages on a whole new set of patients. The improvement resulting from the alterations introduced in this chapter is more than 10 multiples the time taken before. Given the nature of the proposed algorithm, it is expected that the time taken will improve as the *FSDB* grows.

	Average time for single record	all records
DFSP (before)	1.9s	134,945.6s
DFSP (after)	0.158s	11,221.8s

Table 6.2 Time taken for assessing a single patient on average using the DFSP and total time for all 71,024 records before and after the enhancement.

6.4 Speed comparison with decision trees

The final version of the algorithm was compared against decision trees, which showed to be the second best classifier in the previous chapter after the DFSP. The results can be seen in table 6.3. The DFSP did provide a speed enhancement over the decision trees and is expected to be even faster as the database stores more models. Therefore, the DFSP proved to be higher in accuracy, shifted accuracy, speed and lower in mean absolute error.

	single record	all records
DFSP	0.158s	11,221.8s
DT	0.183s	12,971.4s

Table 6.3 Time taken for assessing a single patient on average and total time for all 71,024 records for the DFSP and decision trees *DT*.

6.5 Summary and Conclusions

Implementing the DFSP within GRiST required optimising the process so that it delivers real time advice. This chapter presented methodologies to enhance the speed accordingly. Assessment of the speed showed that the current version of the algorithm was too slow. Two enhancements were proposed. The first was to replace a search instruction which was inside an iteration with a single sort outside the loop. The second was to introduce a database where the different classification models are stored along with the selected feature sets. New feature sets are only generated by the feature selection stage after searching the database to check for whether it already exists. If it does, then the model linked to it will be used for assessment and explanation. Otherwise, a complete subset of the training data with the same features as in the chosen set will be used to create a model. This model is then used to assess and explain the current record. It will also be stored for future use with resembling patients.

The results of the final version of the algorithm have shown a significant speed improvement over decision trees. The DFSP is expected to become faster when it's database grows bigger as more patients are assessed and more unique feature sets are stored. In the following chapter the application of the final version of the DFSP on the harm-to-others dataset will be presented to determine how well the approach can be generalised to new datasets.

Chapter 7

Generalising the DFSP for all mental health risks

7.1 Introduction

In order to test the algorithm's ability to generalise, it was also applied on the Harm To Others (HTO) dataset collected by the GRiST project. The HTO dataset is more challenging than that of the suicide risk since it has a higher percentage of missing values. The suicide dataset had 67.04% missing cells while the HTO has 70.76%. The HTO dataset has 11 risk levels like the suicide ranging from 0 to 10. Also the HTO dataset has 182 DVs as opposed to the 176 in the suicide. Therefore, applying the DFSP onto this dataset is considered a bigger challenge. A bigger problem is the confusion about what risks are being measured. HTO involves many different types of harm where suicide is just one outcome if it succeeds. Hence the connection between input data and output risks could be obscured if the different types of risk are not separated.

This chapter will demonstrate the applicability of the approach to the HTO dataset by comparison with the same alternatives as were used for suicide. The purpose is to evaluate relative performance rather than whether better results could be obtained by discriminating different types of HTO risks. This will be a future project.

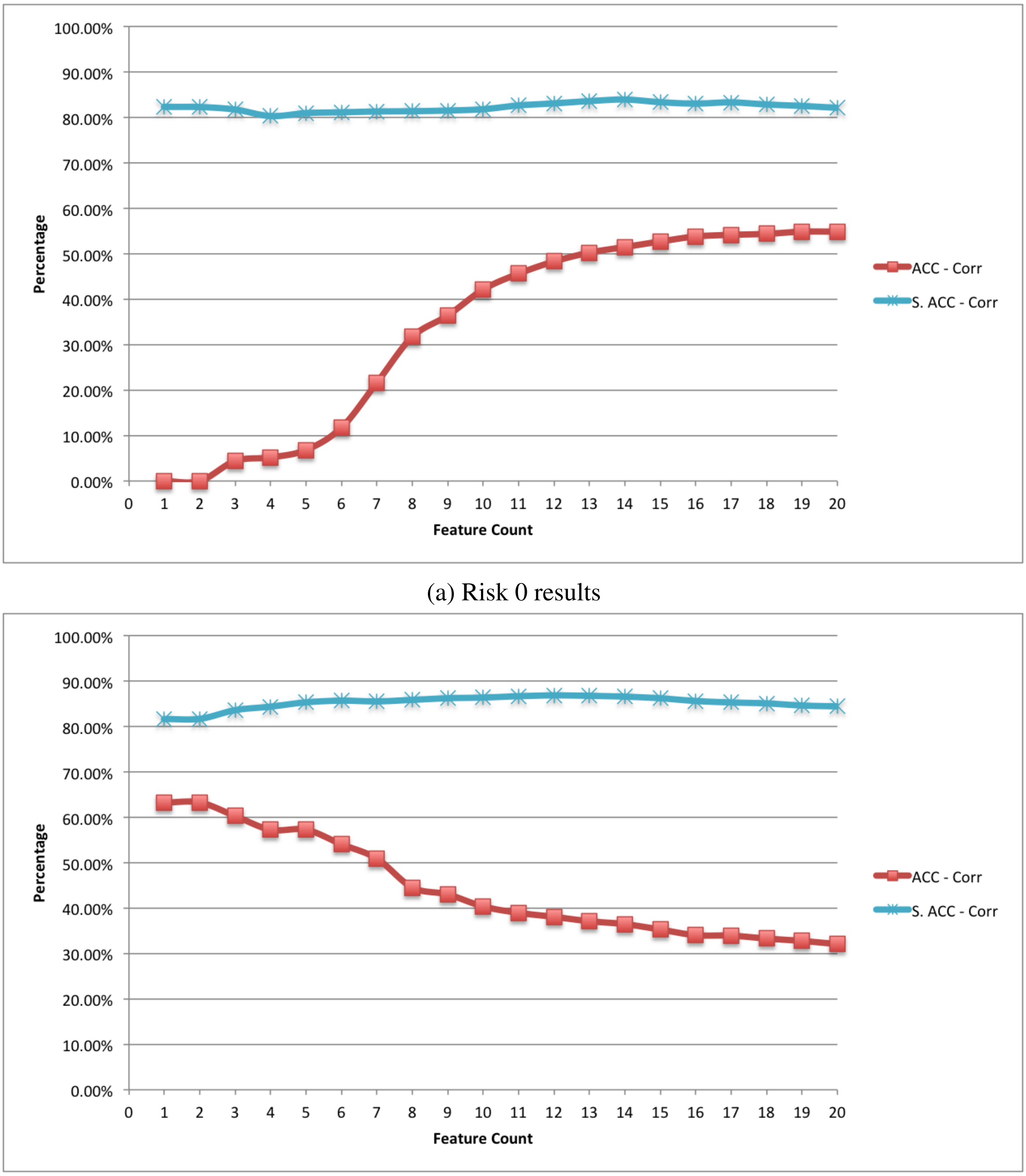
7.2 Determining parameters

The HTO dataset is expected to require different values than the ones used for suicide dataset for the parameters specified by the DFSP. In this section, the values for *Count-Limit*, *IV-Thresh* and *DV-Thresh* will be determined.

The first parameter to be determined is the *Count-Limit*. Analysing the data to estimate this parameter provides an understanding of the minimum and maximum acceptable number of features that the DFSP can use to produce good results. The aim is to find a low count such that speed is not jeopardised while producing acceptable accuracy. Determining *Count-Limit* is a step towards estimating the values of *IV-Thresh* and *DV-Thresh*.

Comparisons of accuracy and shifted accuracy for the 11 risk levels using feature counts 1 to 20 are shown in Figures 7.1 to 7.6. Different values for variable *par*, which was needed to tune the amount of stretching the prediction away from the mean in equation 5.1, were tested. Given the tough nature of the data, a low *par* was required to give more stretch to the results. In all the results of this chapter, the value for *par* is 0.25.

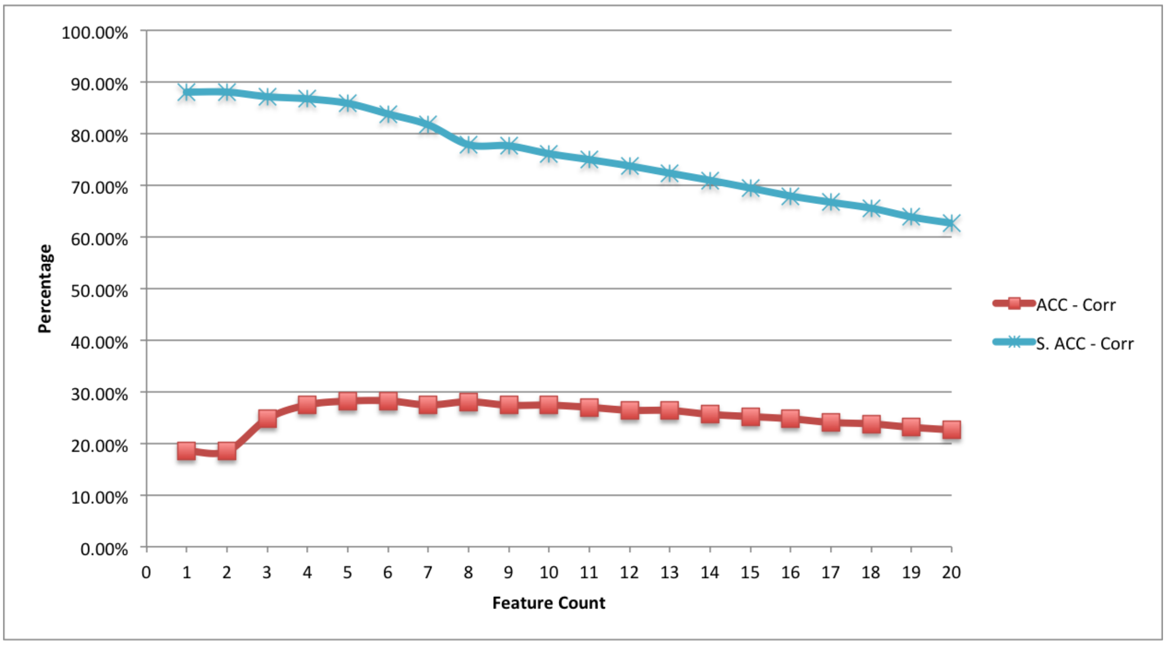
Figure 7.1 shows that the shifted accuracy in both risks 0 and 1 is almost constant among all feature counts. For risk level 0, the accuracy started at 0% with 1 feature then increased gradually as more features were added. On the other hand it started at 63% for risk level 1 and decreased as more features were added. Since the shifted accuracy was not affected by the accuracy's increase in risk 0 and decrease in risk 1, this means that as more features are added, some of the records that were classified as risk 1 are now classified as risk 0. This means that the excess addition of features introduced information that shifted the results towards risk 0 which proves the need for a feature limit.



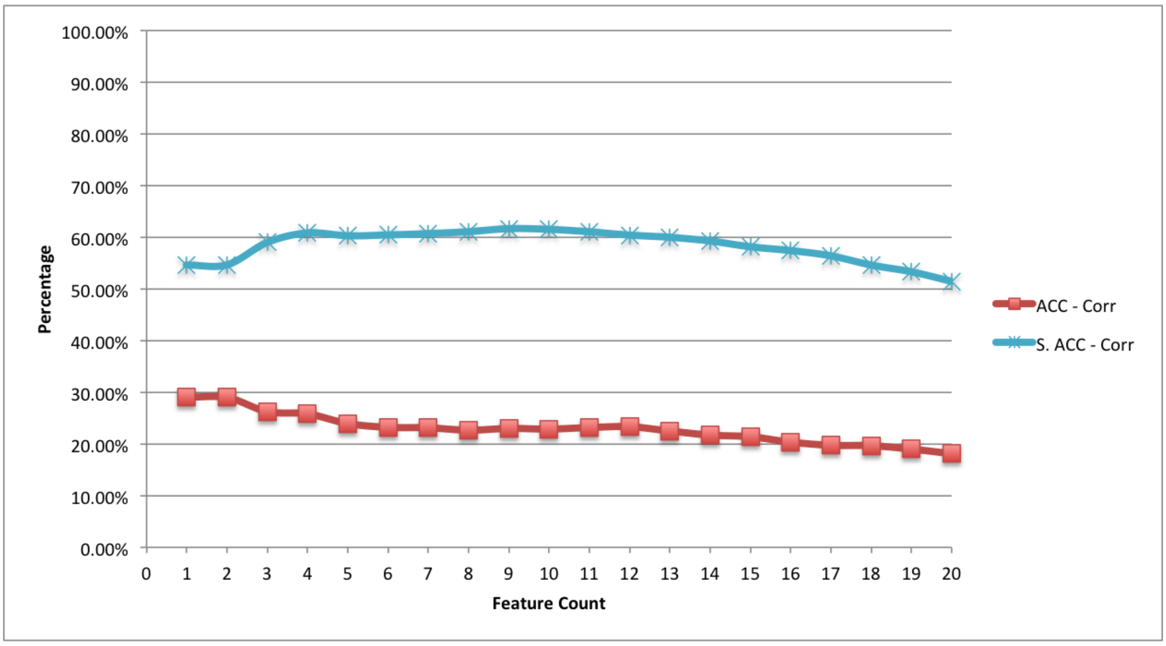
(a) Risk 0 results

(b) Risk 1 results

Fig. 7.1 Accuracy and shifted accuracy for HTO risk levels 0 patients (Figure a) and 1 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for correlation *Corr* criterion.

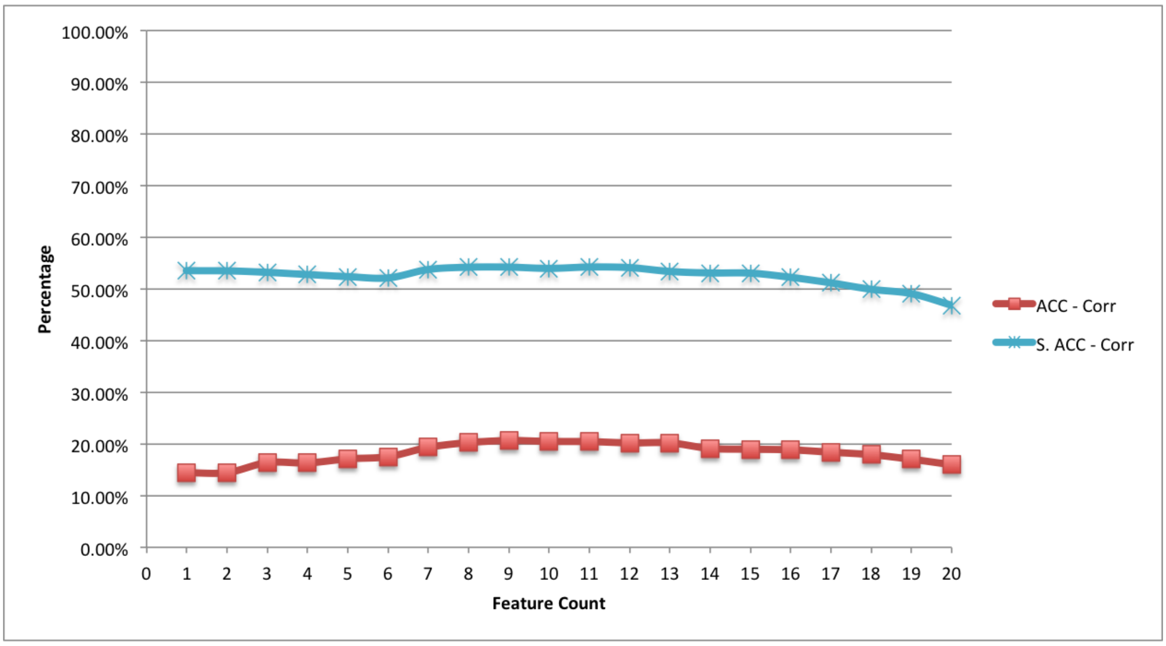


(a) Risk 2 results

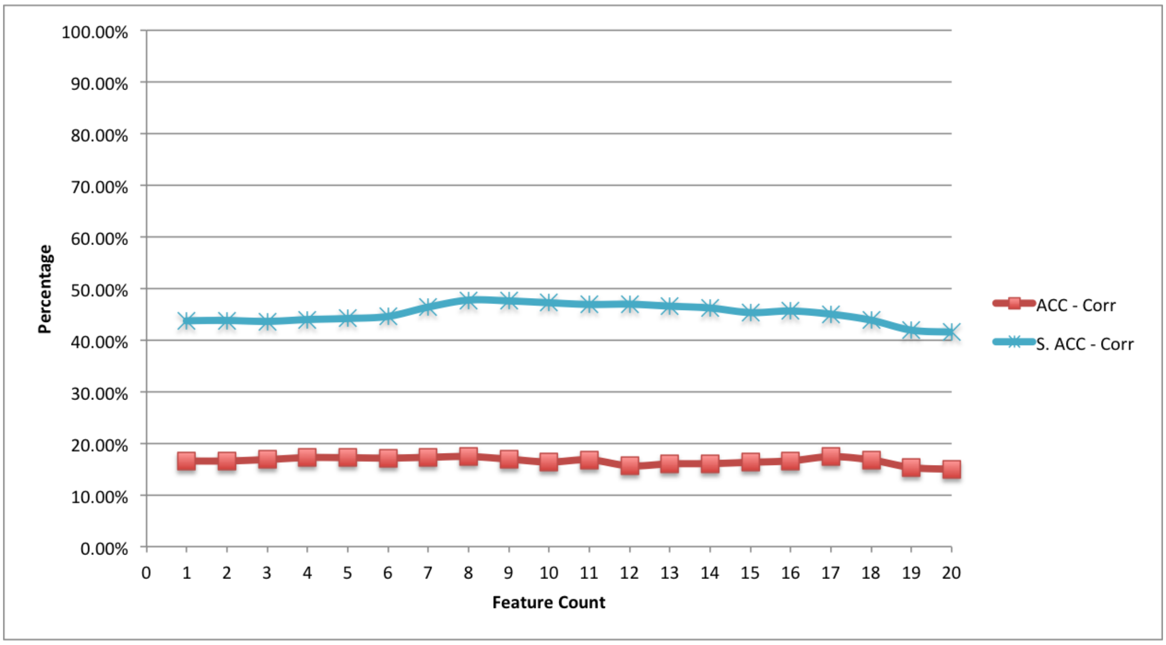


(b) Risk 3 results

Fig. 7.2 Accuracy and shifted accuracy for HTO risk levels 2 patients (Figure a) and 3 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for correlation *Corr* criterion.

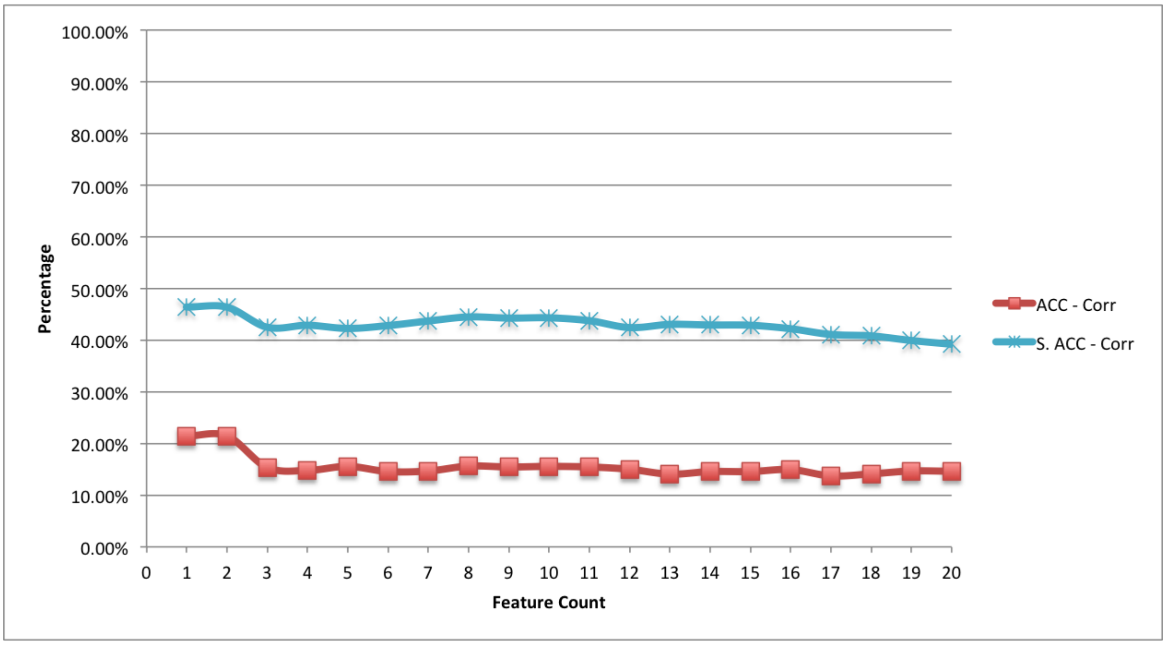


(a) Risk 4 results

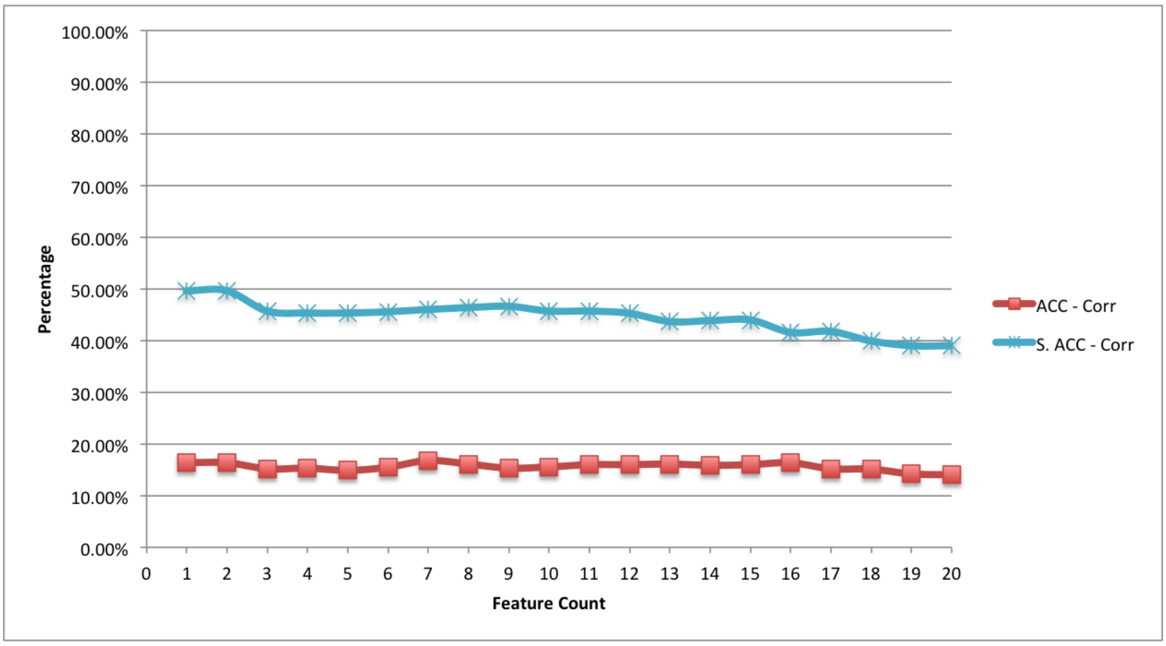


(b) Risk 5 results

Fig. 7.3 Accuracy and shifted accuracy for HTO risk levels 4 patients (Figure a) and 5 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for correlation *Corr* criterion.

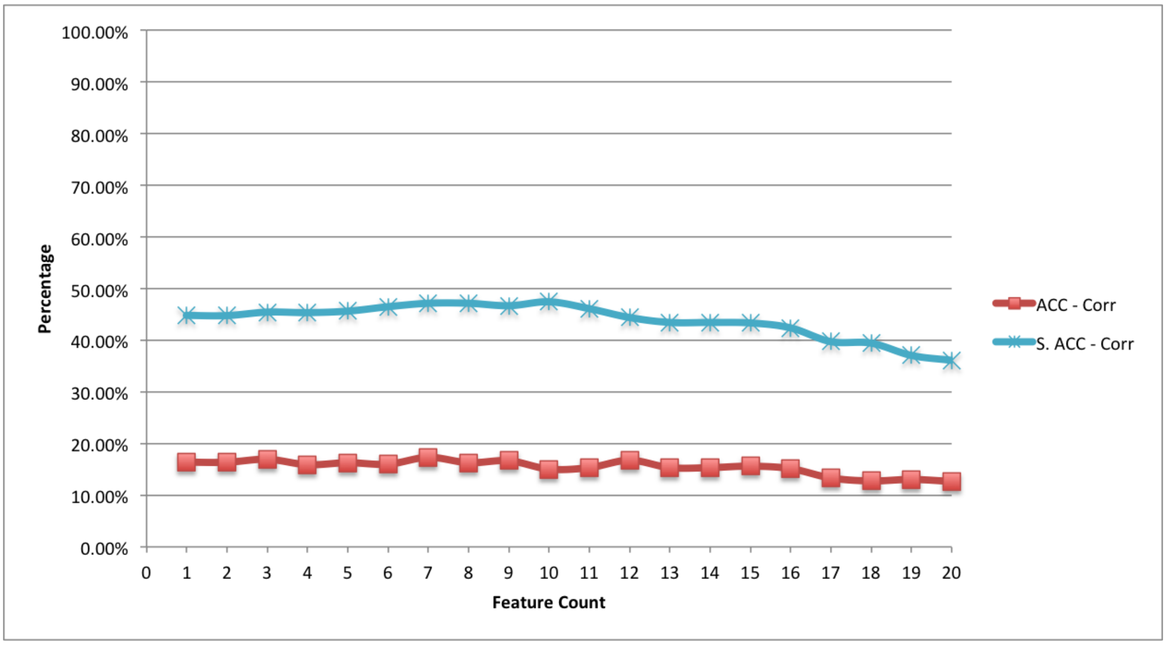


(a) Risk 6 results

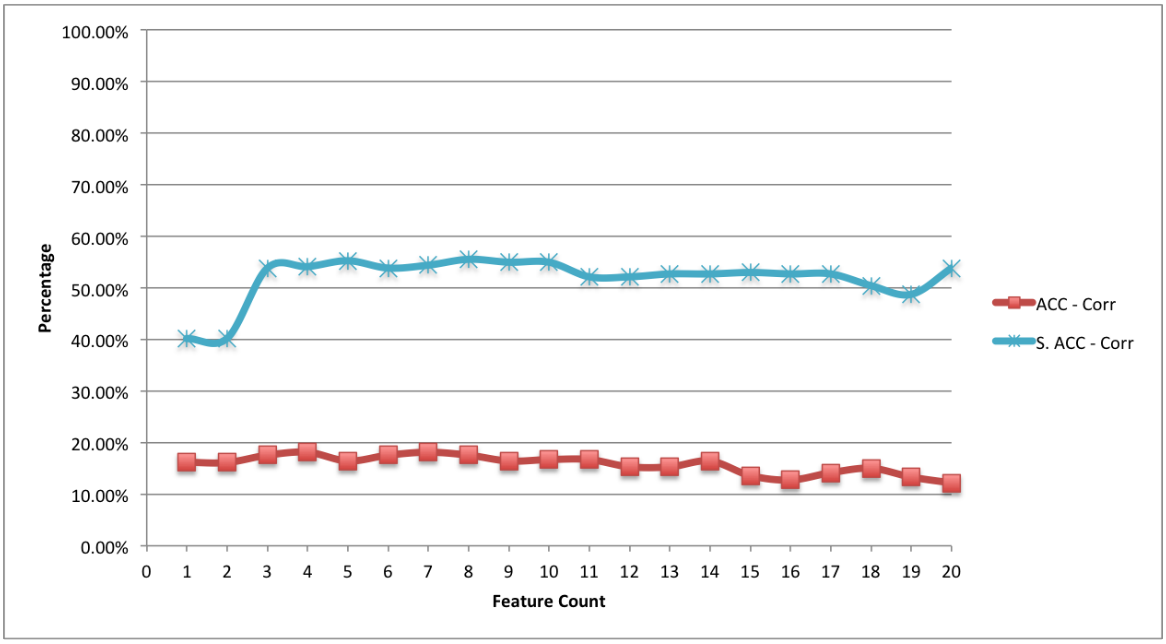


(b) Risk 7 results

Fig. 7.4 Accuracy and shifted accuracy for HTO risk levels 6 patients (Figure a) and 7 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for correlation *Corr* criterion.



(a) Risk 8 results



(b) Risk 9 results

Fig. 7.5 Accuracy and shifted accuracy for HTO risk levels 8 patients (Figure a) and 9 patients (Figure b) using feature counts ranging from 1 to 20. The results are shown for correlation *Corr* criterion.

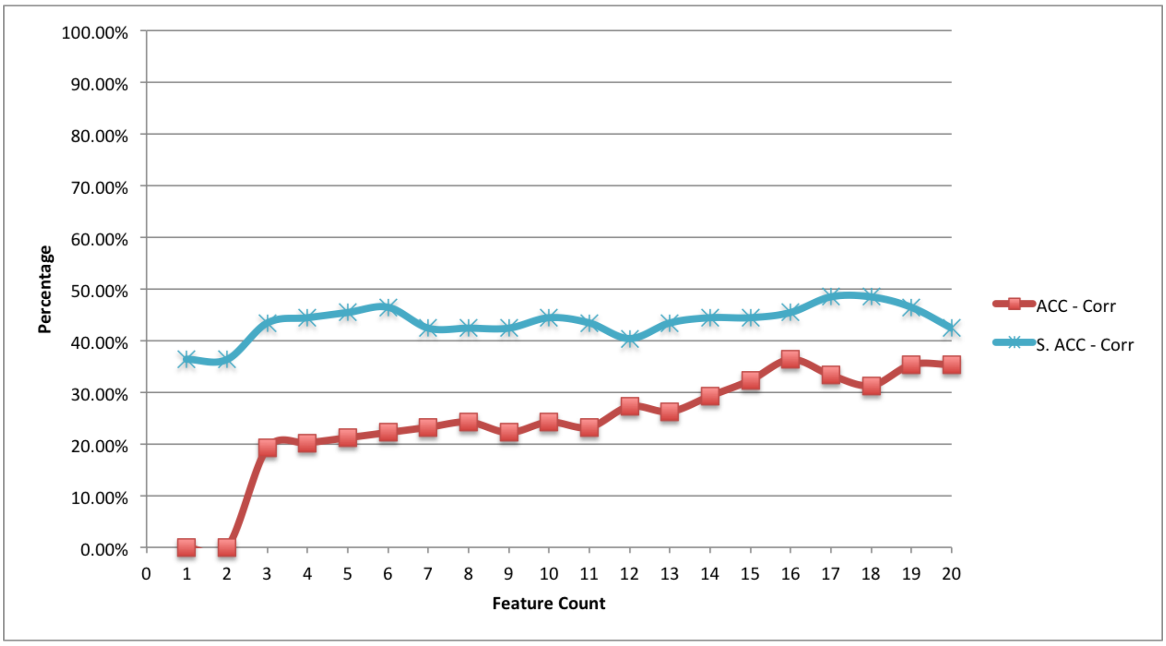


Fig. 7.6 Accuracy and shifted accuracy for risk 10 using feature counts ranging from 1 to 20. The results are shown for correlation *Corr* criterion.

Figure 7.2a shows that the shifted accuracy for risk 2 degrades quickly as more features are added. This is caused by the increase of misclassified risk 1 records as risk 0 when more features are added. For the accuracy, the results degrade slowly as more features are added for both risks 2 and 3.

Figures 7.3 to 7.5 show that starting with 3 features the accuracy and shifted accuracy is almost constant with a slight degradation when the feature count exceeds 17. Finally Figure 7.6 shows that starting feature count 3, the accuracy and shifted accuracy starts to increase gradually.

In order to choose the number of features to limit the algorithm, a few points need to be noted. When the count of features started to exceed 12, the accuracy starts to degrade for risks 1, 2, 3, 4, 6, 8 and 9. At the same feature count, shifted accuracy also starts to degrade for risks 1, 2, 3, 4, 5, 6, 7, 8 and 9. The only two risks that show big improvement with more than 12 features are risks 0 and 10 but with the downside of degrading most of the other risks.

Since the DFSP aims to give good accuracy for all risk levels, 12 features is the upper bound for the choice.

Risks ranging between 1 and 10 produced good accuracy and shifted accuracy results with only 3 features but this was the case for risk 0. Risk 0 started to give better results using 7 features. This makes the more promising range for the feature count between 7 and 12. Therefore, limit 8 was chosen such that it causes a lower percentage of misclassification by producing good accuracy results. Also the lower the feature count, the faster new models can be created.

After the feature count was decided, the next step was to estimate the value for *DV-Thresh*. Table 7.1 shows the comparison between the accuracy and shifted accuracy among all risk levels for different values of *DV-Thresh*. The difference among the results of the accuracy and shifted accuracy using different thresholds accuracy is minimal. The 14 drop outs in thresholds 0.01 and 0.009 are records with the most correlated answered feature having relevance with the DV of less than 0.004. The information provided for these records was minimal and the output risks assigned were less than level 3. Therefore, it was not necessary to decrease the threshold any further. *DV-Thresh* was then given a value of 0.01 such that the lowest number of patient records are dropped, the mean sample size is large and the accuracy was better.

	DV-Thresh	0.04		0.03		0.02		0.01		0.009	
	Mean Num of feats	7.935		7.951		7.973		7.986		7.989	
	Mean Sample Size	6027.7		5869.9		5743.7		5566		5561.5	
	Dropouts	37		37		26		14		14	
		ACC	S.ACC	ACC	S.ACC	ACC	S.ACC	ACC	S.ACC	ACC	S.ACC
Risk Categories	0	31.61%	81.39%	31.62%	81.35%	31.65%	81.36%	31.72%	81.37%	31.72%	81.37%
	1	44.63%	85.80%	44.58%	85.83%	44.58%	85.83%	44.53%	85.84%	44.53%	85.85%
	2	28.02%	77.86%	28.00%	77.85%	28.05%	77.88%	28.04%	77.86%	28.03%	77.86%
	3	22.73%	61.10%	22.66%	61.05%	22.63%	61.06%	22.69%	61.07%	22.69%	61.07%
	4	20.16%	54.20%	20.21%	54.14%	20.23%	54.22%	20.25%	54.22%	20.25%	54.22%
	5	17.51%	47.73%	17.48%	47.73%	17.43%	47.73%	17.43%	47.73%	17.43%	47.73%
	6	15.63%	44.49%	15.63%	44.49%	15.63%	44.49%	15.63%	44.49%	15.63%	44.49%
	7	16.06%	46.34%	16.06%	46.34%	16.06%	46.34%	16.06%	46.34%	16.06%	46.34%
	8	16.29%	47.05%	16.29%	47.05%	16.29%	47.05%	16.29%	47.14%	16.29%	47.14%
	9	17.56%	55.52%	17.56%	55.52%	17.56%	55.52%	17.56%	55.52%	17.56%	55.52%
	10	24.24%	42.42%	24.24%	42.42%	24.24%	42.42%	24.24%	42.42%	24.24%	42.42%

Table 7.1 Results of testing the algorithm with DV-Thresh ranging from 0.04 to 0.09 while setting the feature limit to 8: the *ACC* column is the percentage of predictions exactly matching the clinical judgement and the *S.ACC* column is the percentage that either match the judgement or are one away from it.

After choosing the values for *Count-Limit* and *DV-Thresh*, the final step was to determine *IV-Thresh*. Table 7.2 shows the results of applying different values of *IV-Thresh* using the defined values for *DV-Thresh* and *Count-Limit*. The *IV-Thresh* controls the accepted level of redundancy in the selected subset of features. The table shows the accuracy and shifted accuracy for risks 0 and 1 improve with lower values of *IV-Thresh* which makes these risks the most sensitive to redundancy. It also shows that risks 3, 4, 5 and 9 produced better accuracy results and risks 2, 3, 5, 7 and 8 showing better shifted accuracy using *IV-Thresh* as 0.8. Risks 2, 6 and 9 produced better shifted accuracy and risk 8 produced better accuracy using *IV-Thresh* as 0.9. Finally the accuracy of risks 2, 6, 7 and 10 along with the shifted accuracy of 10 were best without *IV-Thresh*. Putting into consideration that most levels were affected by the use of *IV-Thresh* proving the existence of redundant features. The chosen

	IV-Thresh	0.4		0.5		0.6		0.7		0.8		0.9	
		ACC	S.ACC	ACC	S.ACC	ACC	S.ACC	ACC	S.ACC	ACC	S.ACC	ACC	S.ACC
Risk Categories	0	35.37%	83.22%	34.43%	83.16%	35.32%	82.97%	33.80%	82.31%	32.42%	81.55%	32.37%	81.53%
	1	45.20%	87.17%	45.77%	87.21%	45.58%	87.43%	45.46%	86.63%	44.96%	86.10%	44.86%	86.03%
	2	27.12%	77.66%	27.05%	78.10%	27.20%	78.04%	27.16%	78.42%	27.23%	78.08%	27.33%	78.02%
	3	22.17%	60.29%	22.31%	60.27%	22.62%	60.46%	22.57%	60.77%	22.93%	61.12%	22.93%	61.08%
	4	18.76%	53.49%	18.51%	53.51%	19.18%	53.78%	20.18%	54.03%	20.50%	54.10%	20.46%	54.09%
	5	15.83%	45.92%	15.86%	45.90%	15.40%	45.82%	16.70%	46.61%	17.26%	47.98%	17.15%	47.93%
	6	14.07%	42.21%	14.18%	42.16%	13.68%	42.27%	14.74%	43.44%	15.63%	44.44%	15.57%	44.49%
	7	16.86%	44.31%	16.31%	44.06%	16.68%	43.57%	15.82%	44.37%	15.57%	46.40%	15.75%	46.28%
	8	14.86%	43.62%	15.05%	43.62%	14.57%	42.29%	14.76%	45.52%	16.57%	46.86%	16.57%	46.48%
	9	15.01%	54.11%	15.01%	53.82%	14.73%	52.69%	17.85%	52.41%	17.28%	54.96%	17.56%	55.24%
	10	26.26%	43.43%	27.27%	43.43%	26.26%	43.43%	27.27%	43.43%	26.26%	44.44%	26.26%	44.44%

Table 7.2 Results of accuracy and shifted accuracy for *IV-Thresh* ranging between 0.4 and 0.9. All results are based on setting the *DV-Thresh* to 0.01 and *Count-Limit* to 8 features.

value for *IV-Thresh* was 0.8 to even out the need for low value of *IV-Thresh* from risks 0 and 1 with the need for high value from the remaining risks.

The results of applying the DFSP algorithm on the HTO dataset using the estimated parameters is achieved. The next step was to compare them to other methods. This is presented in the following section.

7.3 Comparison to others

The same set of off-the-shelf methods that were applied to the suicide database are performed on the HTO dataset. Table 7.3 shows the results of comparing the accuracy and shifted accuracy of the HTO dataset with each one. In this table, the best performance for risks 4, 6, 9 and 10 is the proposed DFSP algorithm. Risks 4 and 6 are considered harder to classify as they are both representing the borders of the medium risk range thus most algorithms fail to discriminate them from level 5. This can be noted from the table since all the methods other than DFSP seemed to classify risk 5 better than 4 and 6. Risks 9 and 10 are represented by the least number of records in the population which makes them also very challenging to classify. This was not the case for the DFSP.

For each of the other classifiers, the accuracy is very good for some risk levels but bad for others, for example DT produced very high risk 0 and 1 accuracy but very low accuracy for levels 4, 6, 9 and 10. Analysing the remaining results in Table 7.3, *RF* performed similar to DT with sudden drops within the risks. $mean + DT$ seemed to have solved the degradation within the different risks but has not achieved the DFSP results. Also, the accuracy of classifying risks above 3 was weak. Both alterations to the linear regression whether it was $mean + LR$ or $CFS + EM + LR$ have resulted in the tendency to classify towards lower risks as both cases failed to classify any patients in risks 9 and 10. $CFS + EM + DT$ and DT also resulted in drops in different risks. Both methods tended to give low results in risks ranging from 4 to 6 along with risks 9 and 10.

Risk	Different Methods							
	DFSP	DT	RF	Mean + LR	CFS + EM +LR	Mean + DT	CFS + EM + DT	CFS+DT
0	32.82%	49.73%	49.43%	49.02%	47.00%	47.28%	54.25%	59.75%
1	46.54%	49.55%	59.07%	47.58%	48.04%	37.49%	44.97%	40.92%
2	28.66%	27.93%	37.71%	32.36%	31.18%	31.73%	24.22%	24.39%
3	23.55%	20.53%	19.83%	16.79%	11.97%	24.49%	25.04%	21.49%
4	20.66%	3.61%	8.67%	1.69%	0.00%	16.24%	7.78%	5.70%
5	16.09%	14.82%	9.78%	6.48%	8.10%	16.26%	8.71%	9.20%
6	14.40%	0.06%	5.90%	0.17%	0.00%	12.07%	6.86%	1.67%
7	16.62%	23.14%	12.37%	9.48%	0.00%	14.09%	17.78%	20.80%
8	16.48%	22.48%	9.05%	6.57%	0.00%	14.86%	13.49%	18.95%
9	17.28%	1.13%	6.52%	0.00%	0.00%	8.50%	10.76%	9.07%
10	20.20%	5.05%	8.08%	0.00%	0.00%	5.05%	0.00%	0.00%

Table 7.3 Results of comparing accuracy when classifying violence risk in GRiST dataset to different off the shelf methods

The second analysis is the comparison between results of the shifted accuracy of the DFSP and off-the-shelf methods. Table 7.4 shows the results of the shifted accuracy comparison between the DFSP and the other methods. The shifted accuracy results show that some methods might outperform the DFSP in classifying risk levels 0, 1 and 2 but none of them achieved the results of DFSP in all the remaining risks.

Comparing the two Tables 5.10 and 7.3 shows how the HTO data is much harder to handle data than the suicide data but the DFSP managed to produce the best results when compared to other methods for both sets. This is demonstrated by the mean absolute error for the different methods in Figure 7.7. The chart shows that the error is almost similar in risks ranging 0 to 3 and starting risk level 4, the DFSP has the least error when compared to all other methods. The average error difference between the DFSP and the second best method for risk 10 is close to 1.5. From this chart and the previous tables, the second best classification after the DFSP is the $mean+DT$. It outperformed $CFS+EM+DT$ and

$CFS+DT$ in classifying the risks ranging from 4 to 6. The $mean+DT$ managed to get a uniform lower average error among all the different risk levels when compared to all the other off-the-shelf methods.

Risk	Different Methods							
	DFSP	DT	RF	Mean + LR	CFS + EM +LR	Mean + DT	CFS + EM + DT	CFS+DT
0	81.78%	90.47%	90.71%	89.15%	89.37%	74.53%	89.58%	92.04%
1	87.39%	94.80%	96.94%	96.95%	97.89%	83.90%	92.33%	94.82%
2	80.69%	80.23%	87.81%	79.36%	81.09%	72.27%	76.57%	71.72%
3	62.80%	53.50%	59.21%	54.48%	50.82%	59.09%	52.81%	49.70%
4	55.06%	39.72%	33.37%	32.64%	28.34%	47.78%	42.83%	38.51%
5	46.48%	19.97%	16.24%	9.25%	8.10%	36.11%	22.02%	17.71%
6	43.44%	31.31%	16.74%	15.85%	17.40%	33.93%	27.66%	25.14%
7	44.49%	35.57%	19.20%	13.60%	0.00%	32.18%	32.53%	33.66%
8	44.29%	47.33%	19.24%	21.24%	0.00%	33.71%	38.85%	43.48%
9	48.44%	38.53%	17.85%	10.48%	0.00%	28.33%	35.41%	35.98%
10	36.36%	9.09%	9.09%	0.00%	0.00%	13.13%	19.19%	21.21%

Table 7.4 Results of comparing shifted accuracy when classifying violence risk in GRiST dataset to different off the shelf methods

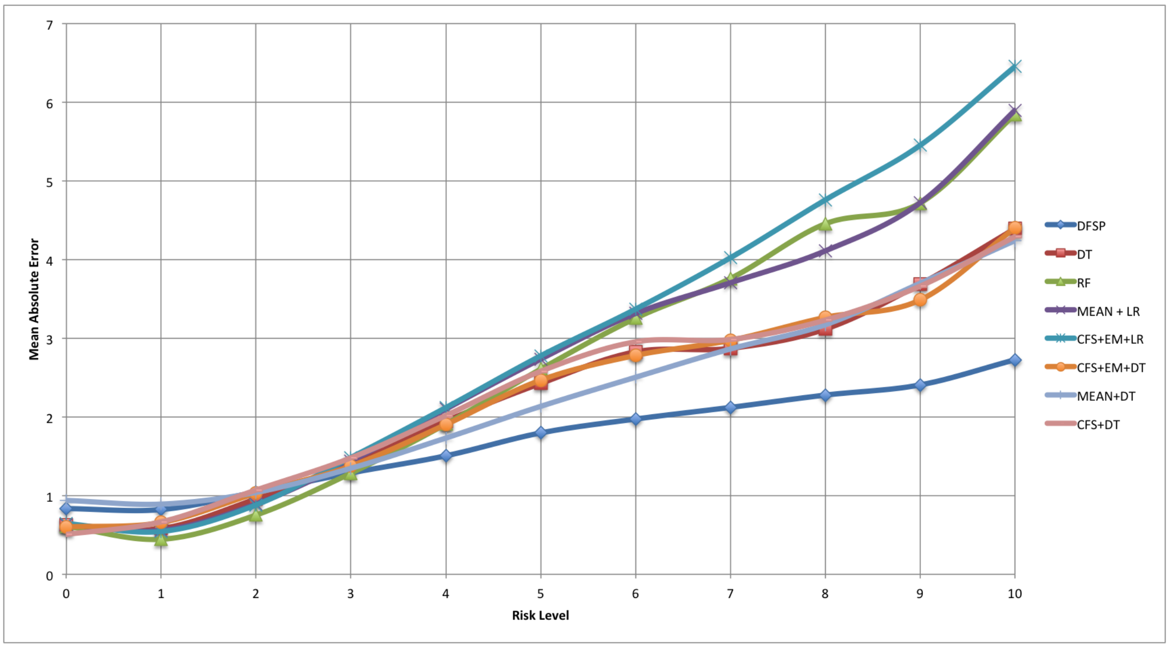


Fig. 7.7 Mean absolute error for each risk level

Finally, comparing the difference in speed with the *mean+DT*. The *mean+DT* was faster by an average of 0.014 seconds per patient than the DFSP. This would have minimal effect on the assessment speed for the end user. The DFSP performed better and is expected to have an instantaneous assessment when the database of patients grows bigger. As more patients are assessed, more feature sets are learned such that the need for building new models for unique patients will be minimal.

	single record	all records
DFSP	0.08	5684
Mean+DT	0.066	4734

Table 7.5 Time taken for assessing a single patient on average and total time for all records for the *DFSP* and *mean+DT*

7.4 Summary and conclusions

This chapter aimed to show the effect of generalizing the DFSP on a different dataset than suicide. The algorithm was applied to the HTO dataset which has a higher percentage of missing and more DVs than the suicide. First the different parameters were determined, after which the results were compared to the same off-the-shelf methods that were applied to the suicide dataset.

The results have shown an improvement in accuracy, shifted accuracy and mean absolute error for almost all risk levels over off-the shelf methods used commonly for handling challenges in such datasets. The DFSP algorithm has produced good results in general and is a stronger classifier when it comes to classifying high risk patients.

When comparing the classification time, the DFSP was not as fast as the mean imputation followed by decision tree classification. The speed comparison was including the time taken for the instantiation of the database which is an initial stage therefore it is expected to enhance over time. The future classification of a record will depend on its existence in the database of models. If it exists then the assessment will be a direct calculation from the stored model otherwise it will be equal to the time required to build a single model. In the following chapter an overall summary of the thesis will be given, including the limitations of the current system and the future work required.

Chapter 8

Summary, conclusions and future work

In this thesis the Dynamic Feature Selection and Prediction (DFSP) algorithm was introduced. This algorithm is used to assess the mental health risk of a patient and provide an explanation to practitioners about how the risk was derived. It handled different problems that could occur in mental health datasets which have a large percentage of missing data, varied frequencies of records in different risk levels, a large number of features and a diversity of assessors participating in the data collection, each with different clinical experiences.

A brief introduction to the topic of suicide risk assessment along with common approaches of prediction were mentioned in Chapter 1. The general problems of risk prediction were considered and how data mining tools can be used to overcome them. Finally the GRiST data set was introduced along with the idea behind the DFSP algorithm.

The second chapter explored mental health risks in detail along with current approaches and tools to assess them. The chapter detailed the initiation, structure and user interface of GRiST along with the problems handled by it. These problems include the high dimensionality, the heterogeneity of the risk levels and the large percentage of missing data.

The prediction methods were discussed in Chapter 3 where there was a section for the different domains addressed in this thesis. A section that discussed missing data problem and how it is commonly handled. This was followed by a discussion of feature selection methods and its general approaches. Finally, common predictive data mining algorithms were introduced.

Chapter 4 introduced the DFSP algorithm starting with the rationale behind it and how it relates to common problems in medical health data. The basic version was defined and then extended into a dynamic version with two thresholds one to control the correlation with the DV and the other to control redundancy between predictors (independent variables).

The dynamic version of the algorithm was applied to the GRiST suicide dataset in Chapter 5. Optimal values for different parameters were deduced first. The matching of continuous risk values produced by the linear regression to discrete risks was compared with the production of discrete risks directly using logistic regression. The results were degraded. A second analysis of the results produced by linear regression showed that its behaviour had performance problems at the end of the fixed scale where linearity was violated. This motivated an investigation of GLM to see if it could be cured. Yet no improvement was introduced. Thus an equation was introduced that solved the problem of having data cut off at the extremes by stretching the prediction away from the mean. The results produced were compared with other approaches that are commonly used to handle missing data, feature selection and prediction. The results showed the DFSP to outperform all the other methods in accuracy, shifted accuracy and mean absolute error.

Chapter 6 discussed the practical issues of implementing the DFSP within GRiST. The basic version of DFSP was very slow because a regression model was needed to be built per patient. In this chapter, enhancements were introduced to improve the performance. The main one was the introduction of a database that would store unique feature sets along with their linear regression models. When a feature set is selected for a patient's record, it is checked first against the database. If it exists then the model is directly used otherwise a new model is built based on a complete subset of the training set. The final version of DFSP was presented in this chapter.

Chapter 7 introduced the ability of the DFSP to generalise to different datasets by applying it to HTO. This dataset has a larger percentage of missing data than that of suicide, more features and different types of output risk levels. The parameters were first determined, then a comparison with others was applied. The comparison showed that the DFSP excelled over other methods in accuracy, shifted accuracy and mean absolute error. The speed of the DFSP was 0.014 second slower than the mean imputation followed by decision trees which can hardly be noticed by the end-user.

8.1 DFSP algorithm summary

The DFSP algorithm represented a novel method to handle a very common problem that can occur in any data collected for assessments, which is missing data. For instance patients sometimes choose to withhold information they consider private. Other times the questions do not relate to the patient examined. This usually results in records with many incomplete cells. Commonly large percentages of missing data are handled by imputation. Providing an explanation to the assessor using imputed data would undermine its credibility because the assessors would know they had not provided that data.

In the DFSP, the prediction model was only from features that patients answered. A subset of independent variables was selected according to correlation with the DV along with mutual information between IVs to reduce redundancy. The number of features to be selected were limited because this showed improvement to the results. Also, a threshold limit was applied to correlation between a candidate IV and the DV so that no weak predictors would be added. These unnecessarily expand the size of the regression model, which degrades the performance.

The nature of the GRiST data collection interface ensures child questions will mandate a single value of the parent question, which means the parent is useless. This was handled by the filter/non-filter criteria. After passing this criteria, the candidate variable has its mutual information checked against all the pre selected to determine the redundancy criterion on a pairwise basis. If this value exceeds a certain threshold then the feature is not added. The feature already chosen has a higher correlation with the DV than the candidate, which is why it is kept and the candidate is discarded.

The risk levels in the GRiST data have values ranging from 0 to 10. The nature of regression means weights are pulled away from the extremes because it is not possible to distinguish patients who are just at the highest value from those who might have been put even higher with an asymptotic distribution, which is what regression needs. Equation 5.1 was introduced to solve this problem by stretching the prediction away from the mean and beyond the extremes. The selected subset of features along with its modified weights is then output to the assessor as an explanation of the assessed risk.

Finally, the DFSP introduced an optimisation methodology that stores the regression weights along with the selected feature set. This aims at enhancing the speed of assessment of patients as the time progresses. The selected features for a patient would be mapped against the stored sets, if it exists then the stored weights are applied, otherwise a new model is created and stored.

8.2 Current limitations and future work

The proposed algorithm has some advantages over other methods both for prediction accuracy and by only using the data collected for the patient rather than fabricating missing data. It also has some limitations, which will be discussed along with proposed future enhancements.

8.2.1 Parameter estimation

In the DFSP, thresholds were estimated manually through multiple tests. There are 4 parameters that required estimation: *DV-Thresh*, *IV-Thresh*, *par* and *Count-Limit*. There is a need for an algorithm that automatically tests different values of the parameters and optimises them.

Algorithm 6 shows a proposed exhaustive methodology to do this by generating and testing all the possible combinations of the four. This method would be the most accurate, yet slowest. However, it could be implemented offline. For this algorithm to work, the upper and lower bounds of each parameter has to be specified along with the granularity required. Further research can be addressed to this problem to find faster means of estimating the parameters.

Algorithm 6 Proposed exhaustive parameter evaluation methodology

```
for every value of $par$ do
    for every value of $featCount$ do
        for every value of $IVThresh$ do
            for every value of $DVThresh$ do
                Apply DFSP algorithm
                Evaluate Results
                if Current results > best achieved then
                    store current results and parameters
                end if
            end for
        end for
    end for
end for
```

8.2.2 Generic application on different datasets

The DFSP has produced promising results when applied to the HTO dataset which suggests it will be applicable more generically. Future work is needed to explore this for dataset that have similar problems of high dimensionality, high missing data and heterogeneous members of the population that have varying relevance of data from within the complete set.

8.2.3 Learning from missing information

Learning from the questions that patients and assessors chose not to answer is another area of research to be addressed. This will require marking those questions by the assessor. Since the data currently has a lot of missing cells, it is impossible to know which questions were actually discarded by the assessor and which ones patients chose not to answer. This might help learn what kind of information some patients of different levels of suicide risk choose to hide. It might also help improve the predictive capabilities of the assessing algorithm.

8.3 Final conclusion

This thesis has presented and analysed the novel DFSP algorithm. This algorithm is going to be an essential part of the GRiST CDSS where it will introduce the functionality of assessing risk. For every patient, the assessment will be based only on the information provided. An explanation of the assessment will be also provided based on the regression model applied.

References

Edgar Acuna and Caroline Rodriguez. The treatment of missing values and its effect on classifier accuracy. In *Classification, Clustering, and Data Mining Applications*, pages 639–647. Springer, 2004.

Paul D Allison. *Missing data*, volume 136. Sage publications, 2001.

Amanda N Baraldi and Craig K Enders. An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1):5–37, 2010.

Riccardo Bellazzi and Blaz Zupan. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, 77(2):81–97, 2008.

Eta S Berner. *Clinical decision support systems: theory and practice*. Springer Science & Business Media, 2nd edition, 2007.

Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. A review of feature selection methods on synthetic data. *Knowledge and information systems*, 34(3): 483–519, 2013.

Verónica Bolón-Canedo, Iago Porto-Díaz, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. A framework for cost-based feature selection. *Pattern Recognition*, 47(7): 2481–2489, 2014.

Joe Bouch and John James Marshall. Suicide risk: structured professional judgement. *Advances in Psychiatric treatment*, 11(2):84–91, 2005.

Leo Breiman and Philip Spector. Submodel selection and evaluation in regression. the x-random case. *International statistical review/revue internationale de Statistique*, pages 291–319, 1992.

Jerrod Brown and Jay P Singh. Forensic risk assessment: A beginner's guide. *Archives of Forensic Psychology*, 1(1):49–59, 2014.

Craig J Bryan and M David Rudd. Advances in the assessment of suicide risk. *Journal of clinical psychology*, 62(2):185–200, 2006.

Christopher D. Buckingham. Psychological cue use and implications for a clinical decision support system. *Medical Informatics and the Internet in Medicine*, 27(4):237–251, 2002.

Christopher D. Buckingham. Improving mental health risk assessment using web-based decision support. *Health Care Risk Report*, 13(3):17, 2007.

Christopher D. Buckingham, Abu Ahmed, and Ann Adams. Using XML and XSLT for flexible elicitation of mental-health risk knowledge. *Informatics for Health and Social Care*, 32(1):65–81, 2007.

Christopher D. Buckingham, Ann Adams, and Chris Mace. Cues and knowledge structures used by mental-health professionals when making risk assessments. *Journal of mental health*, 17(3):299–314, 2008.

Christopher D. Buckingham, Arif Ahmed, and Andrew Adams. Designing multiple user perspectives and functionality for clinical decision support systems. In *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on*, pages 211–218. IEEE, 2013.

Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

Jehanzeb R Cheema. A review of missing data handling methods in education research. *Review of Educational Research*, 84(4):487–508, 2014.

Megan Chesin and Barbara Stanley. Risk assessment and psychosocial interventions for suicidal patients. *Bipolar disorders*, 15(5):584–593, 2013.

Barry H Cohen. *Explaining psychological statistics*. John Wiley & Sons, 2008.

Manoranjan Dash and Huan Liu. Feature selection for classification. *Intelligent data analysis*, 1(1):131–156, 1997.

Robyn M Dawes, David Faust, and Paul E Meehl. Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674, 1989.

DOH Department of Health. No health without mental health: a cross-government mental health outcomes strategy for people of all ages, 2011.

DOH Department of Health. Best practice in managing risk, June 2007.

Yufeng Ding and Jeffrey S Simonoff. An investigation of missing data methods for classification trees applied to binary response data. *The Journal of Machine Learning Research*, 11:131–170, 2010.

Mandy Dixon and Femi Oyebode. Uncertainty and risk assessment. *Advances in Psychiatric Treatment*, 13(1):70–78, 2007.

Gauthier Doquire and Michel Verleysen. Feature selection with missing data using mutual information estimators. *Neurocomputing*, 90:3–11, 2012.

Kevin S Douglas and P Randall Kropp. A prevention-based paradigm for violence risk assessment clinical and research applications. *Criminal Justice and Behavior*, 29(5): 617–658, 2002.

Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

Craig K Enders. *Applied missing data analysis*. Guilford Publications, 2010.

Alan Fern, Robert Givan, Babak Falsafi, and TN Vijaykumar. Dynamic feature selection for hardware prediction. *Journal of Systems Architecture*, 52(4):213–234, 2006.

Tom Flewett. *Clinical risk management: An introductory text for mental health clinicians*. Elsevier Australia, 2010.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

Pedro J García-Laencina, José-Luis Sancho-Gómez, and Aníbal R Figueiras-Vidal. Pattern classification with missing data: a review. *Neural Computing and Applications*, 19(2): 263–282, 2010.

Mostafa Ghannad-Rezaie, Hamid Soltanian-Zadeh, Hao Ying, and Ming Dong. Selection–fusion approach for classification of datasets with missing values. *Pattern recognition*, 43 (6):2340–2350, 2010.

Paul Godin. 'you don't tick boxes on a form': A study of how community mental health nurses assess and manage risk. *Health, Risk & Society*, 6(4):347–360, 2004.

Narayan Gopalkrishnan and Hurriyet Babacan. Cultural diversity and mental health. *Australasian Psychiatry*, 23(6 suppl):6–8, 2015.

John W Graham. Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60:549–576, 2009.

GRiST. Galatean risk and safety tool. www.egrist.org, 2016.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

Mark A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 359–366, San Francisco, CA, USA, 2000. Morgan Kaufmann Inc.

Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Elsevier, 2011.

Frank E. Harrell, Jr. *Regression Modeling Strategies*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387952322.

He He, Hal Daumé III, and Jason Eisner. Cost-sensitive dynamic feature selection. In *ICML Inferning Workshop*, 2012.

S. E. Hegazy and C. D. Buckingham. A method for automatically eliciting node weights in a hierarchical knowledge based structure for reasoning with uncertainty. *International Journal On Advances in Software*, 2:76–85, 2009.

Antonia J Henry, Nathanael D Hevelone, Stuart Lipsitz, and Louis L Nguyen. Comparative methods for handling missing data in large databases. *Journal of vascular surgery*, 58(5): 1353–1359, 2013.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

Anil K Jain, Robert PW Duin, and Jianchang Mao. Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):4–37, 2000.

Jana E Jones, Bruce P Hermann, John J Barry, Frank G Gilliam, Andres M Kanner, and Kimford J Meador. Rates and risk factors for suicide, suicidal ideation, and suicide attempts in chronic epilepsy. *Epilepsy & Behavior*, 4:31–38, 2003.

Erastus Karanja, Jigish Zaveri, and Ashraf Ahmed. How do mis researchers handle missing data in survey-based research: A content analysis approach. *International Journal of Information Management*, 33(5):734–751, 2013.

Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

J Cios Krzysztof, W Swiniaraski Roman, and Lukasz Kurgan. *Data Mining: A knowledge discovery approach*. Springer, 2007.

Daniel T Larose. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.

Roderick Little. A Test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, Vol. 83(No. 404):1198–1202, 1988.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2002.

Brian Littlechild and Christopher Hawley. Risk assessments for mental health service users ethical, valid and reliable? *Journal of Social Work*, 10(2):211–229, 2010.

Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):491–502, 2005.

Huawen Liu, Xindong Wu, and Shichao Zhang. A new supervised feature selection method for pattern classification. *Computational Intelligence*, 30(2):342–361, 2014.

Megan Lotito and Emmeline Cook. A review of suicide risk assessment instruments and approaches. *Mental Health Clinician*, 5(5):216–223, 2015.

Paul R McCrone, Sujith Dhanasiri, Anita Patel, Martin Knapp, and Simon Lawton-Smith. *Paying the price: the cost of mental health care in England to 2026*. King's Fund, 2008.

Paul E Meehl. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press, 1954.

Ingunn Myrtveit, Erik Stensrud, and Ulf H Olsson. Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods. *Software Engineering, IEEE Transactions on*, 27(11):999–1013, 2001.

Loris Nanni, Alessandra Lumini, and Sheryl Brahnam. A classifier ensemble approach for the missing feature problem. *Artificial intelligence in medicine*, 55(1):37–50, 2012.

John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.

Jason W Osborne and A Overbay. *Best practices in data cleaning*. Sage, 2012.

Chao-Ying Joanne Peng, Michael Harwell, Show-Mann Liou, Lee H Ehman, et al. Advances in missing data methods and implications for educational research. *Real data analysis*, pages 31–78, 2006.

Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

Pavel Pudil, Jana Novovičová, and Josef Kittler. Floating search methods in feature selection. *Pattern recognition letters*, 15(11):1119–1125, 1994.

J. Ross Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993. ISBN 1-55860-238-0.

J Ross Quinlan et al. *Discovering rules by induction from large collections of examples*. Expert systems in the micro electronic age. Edinburgh University Press, 1979.

Sherine Nagy Saleh and Christopher D Buckingham. Handling varying amounts of missing data when classifying mental-health risk levels. *Studies in health technology and informatics*, 207:92–101, 2014.

Joseph L Schafer. Multiple imputation: a primer. *Statistical methods in medical research*, 8 (1):3–15, 1999.

Gabriel L Schlomer, Sheri Bauman, and Noel A Card. Best practices for missing data management in counseling psychology. *Journal of Counseling psychology*, 57(1):1, 2010.

Frank L Schmidt. The relative efficiency of regression and simple unit predictor weights. *Educational and Psychological Measurement*, 31:699–714, 1971.

Baris Senliol, Gokhan Gulgezen, Lei Yu, and Zehra Cataltepe. Fast correlation based filter (FCBF) with a different search strategy. In *Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on*, pages 1–4, 2008.

Rania Shibl, Meredith Lawley, and Justin Debuse. Factors influencing decision support system acceptance. *Decision Support Systems*, 54(2):953–961, 2013.

Morton M Silverman and Alan L Berman. Suicide risk assessment and risk formulation part i: A focus on suicide ideation in assessing suicide risk. *Suicide and Life-Threatening Behavior*, 44(4):420–431, 2014.

Jeffrey Strickland. *Predictive analytics using R*. Lulu. com, 2015.

Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, page 37, 2014.

Divya Tomar and Sonali Agarwal. A survey on data mining approaches for healthcare. *International Journal of Bio-Science and Bio-Technology*, 5(5):241–266, 2013.

Eugene Tuv, Alexander Borisov, George Runger, and Kari Torkkola. Feature selection with ensembles, artificial variables, and redundancy elimination. *The Journal of Machine Learning Research*, 10:1341–1366, 2009.

Laura Vail, Ann Adams, Eleanor Gilbert, Alice Nettleingham, and Christopher D. Buckingham. Investigating mental health risk assessment in primary care and the potential role of a structured decision support tool, grist. *Mental health in family medicine*, 9(1):57, 2012.

David Watts, Jonathan Bindman, Mike Slade, Frank Holloway, Adrienne Rosen, and Graham Thornicroft. Clinical assessment of risk decision support (cards): The development and evaluation of a feasible violence risk assessment for routine psychiatric practice. *Journal of Mental Health*, 13(6):569–581, 2004.

Nicole G Weiskopf, George Hripcsak, Sushmita Swaminathan, and Chunhua Weng. Defining and measuring completeness of electronic health records for secondary use. *Journal of biomedical informatics*, 46(5):830–836, 2013.

Emanuel Weitschek, Giovanni Felici, and Paola Bertolazzi. Clinical data mining: problems, pitfalls and solutions. In *Database and Expert Systems Applications (DEXA), 2013 24th International Workshop on*, pages 90–94. IEEE, 2013.

WHO. World health organization. http://www.who.int/mediacentre/factsheets/fs398/en/, 2015. accessed October 4th.

Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

Chao Xing, Fredrick R Schumacher, David V Conti, and John S Witte. Comparison of missing data approaches in linkage analysis. *BMC genetics*, 4(Suppl 1):S44, 2003.

Bijou Yang and David Lester. Recalculating the economic cost of suicide. *Death Studies*, 31 (4):351–361, 2007.

Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, and Lei Hua. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems*, 36(4):2431–2448, 2012.

W. Young, G. Weckman, and W. Holland. A survey of methodologies for the treatment of missing values within datasets: limitations and benefits. *Theoretical Issues in Ergonomics Science*, 12(1):15–43, 2011.

Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004.