An Efficient Remand Risk Assessment Tool based on Machine Learning Techniques

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Mahsa Azizi

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Abstract

The criminal justice system in Saskatchewan is challenged by the large population of people who are charged with committing crimes and are waiting to be summoned, the so-called pretrial population. Although some of these people are released until their trial, others are remanded in custody. The two most common reasons people are remanded are: (i) probable failure to appear for their trial and (ii) risk to public safety. A large pretrial population leads to increased expenses for both the government and the defendants. The pretrial population may be reduced using a remand risk assessment tool (RRAT). The goal of the RRAT is to lower the number of unnecessary remands by determining which defendants are likely to not appear or pose a risk to public safety while they are on release. This study uses the Saskatchewan Primary Risk Assessment (SPRA) as an assessment to measure general recidivism in both male and female adult offenders under the jurisdiction of the Ministry of Corrections and Policing. The SPRA, comprised of 15,117 offenders information in the form of 15 questions, is considered as the input to the RRAT.

In this thesis, the use of machine learning models is proposed for the RRAT to predict which defendants should be remanded, potentially achieving a reduction in pretrial population size. In the first step, to choose the best machine learning model, several classification models, including the support vector classifier, decision tree classifier, random forest classifier (RFC), naive Bayesian classifier, and extreme learning classifier (ELC), are compared in terms of classification performance. According to the simulation results, the ELC outperformed all other models in the comparison considering all existing features followed by the RFC. The two models of the ELC and the RFC achieved the lowest false positive rate and the highest accuracy, precision, specificity, and area under the curve compared to the other explored models. In the second step, to identify the best features from the SPRA, the ELC is used in conjunction with binary particle swarm optimization (BPSO) and the result is compared to the RFC. The ELC–BPSO has shown high superiority to increase the accuracy of the ELC model by using only seven features of the SPRA data. The ELC–BPSO is able to achieve an accuracy of around 74% using the SPRA data.

Acknowledgements

First and foremost, I deeply thank God Almighty for the blessings He has bestowed upon me and for giving me the strength and wisdom to achieve this dream. This master thesis is due to the support and encouragement of many people. It is a pleasure to express my sincere thanks to all those who helped me for the success of this study.

I would like to express my sincere gratitude to my supervisor Prof. Raymond J. Spiteri for the continuous support of my M.Sc. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my M.Sc. study.

Also, I would like to thank my lovely mother, who supported and encouraged me a lot throughout my entire life to achieve this goal. In addition, I would like to thank my dear husband, Benyamin, who supported me spiritually during my married life especially my M.Sc. study.

To my family and my beloved father who is sadly no more.

Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
BBN	Boosted Bayesian Networks
BPSO	Binary Particle Swarm Optimization
CSV	Comma-Separated Value
CV	Cross-Validation
DTC	Decision Tree Classifier
ELC	Extreme Learning Classifier
FN	False Negative
FNR	False Negative Rate
FP	False Positive
H	High
IG	Information Gain
L	Low
LD	Levenshtein Distance
LR	Logistic Regression
M	Medium
ML	Machine Learning
NBC	Naive Bayesian Classifer
NNs	Neural Networks
PRA	Primary Risk Assessment
PSO	Particle Swarm Optimization
RFC	Random Forest Classifier
RNR	Risk-Need-Resposivity
ROC	Receiver Operating Characteristic
RRAT	Remand Risk Assessment Tool
SLFNs	Single-Hidden-Layer Feed-Forward Neural Networks
SPRA	Saskatchewan Primary Risk Assessment
SVC	Support Vector Classifier
TN	True Negative
TP	True Positive
TPR	True Positive Rate
VH	Very High
WCCS	Wisconsin Case Classification System

Chapter 1

Introduction

One of the most important challenges related to the growth of the prison population is the pretrial population, those who are charged for committing a certain crime and are waiting to be summoned. Although some of these people are released during an initial hearing, others are remanded for the duration of their trial. The two most common reasons why someone may be remanded are: (i) likely failure to appear for the hearing of their case and (ii) risk to public safety by recidivism prior to their trial date.

Figure 1.1 depicts a recent statistics in Canada that compares the average daily number of offenders who were held in remand and sentenced custody from 2004/2005 to 2014/2015 (*Correctional Services Program*, 2017). Sentenced custody is known as serving in a provincial or federal correctional facility depending on the duration of the sentence after considering to be guilty (*The Government of British Columbia*, 2019). Offenders with sentences of less than two years, serve in a provincial correctional facility, whereas offenders with sentences of two years or more, serve in a federal correctional facility. The aforementioned statistics includes only the data of offenders who served in a provincial correctional facility, which their sentence duration was less than two years (*Correctional Services Program*, 2017). Also, the study excludes the Prince Edward Island and Alberta data due to the unavailability of data for the full period (*Correctional Services Program*, 2017).

According to the statistics shown in Figure 1.1, over the years 2014/2015, there were more than 11,000 offenders on average per day held in remand while waiting for their trial. This number is compared to more than 9,000 offenders on average per day who were held in sentenced custody. Compared with the year before, 2012/2013, the average daily number of offenders who were in sentenced custody decreased around 6%, while the average daily population of offenders in remand was almost steady.

However, over a 10-year time period, the remand population has been increasing. From Figure 1.1, it can be observed that between the years 2004/2005 and 2014/2015, there was approximately a 39% increase in the average daily number of offenders who were remanded, whereas the average daily number of offenders in sentenced custody increased by approximately 7%. Moreover, since 2004/2005, the number of offenders who were in remand exceeded the number of offenders in sentenced custody. The gap between these two groups widened steadily from that point until the years 2009/2010 when 57% of the sentenced custody population was made up of remanded offenders. After 2009/2010, the gap has moderately narrowed.

There are many challenges for the provincial and territorial correctional systems when the pretrial pop-

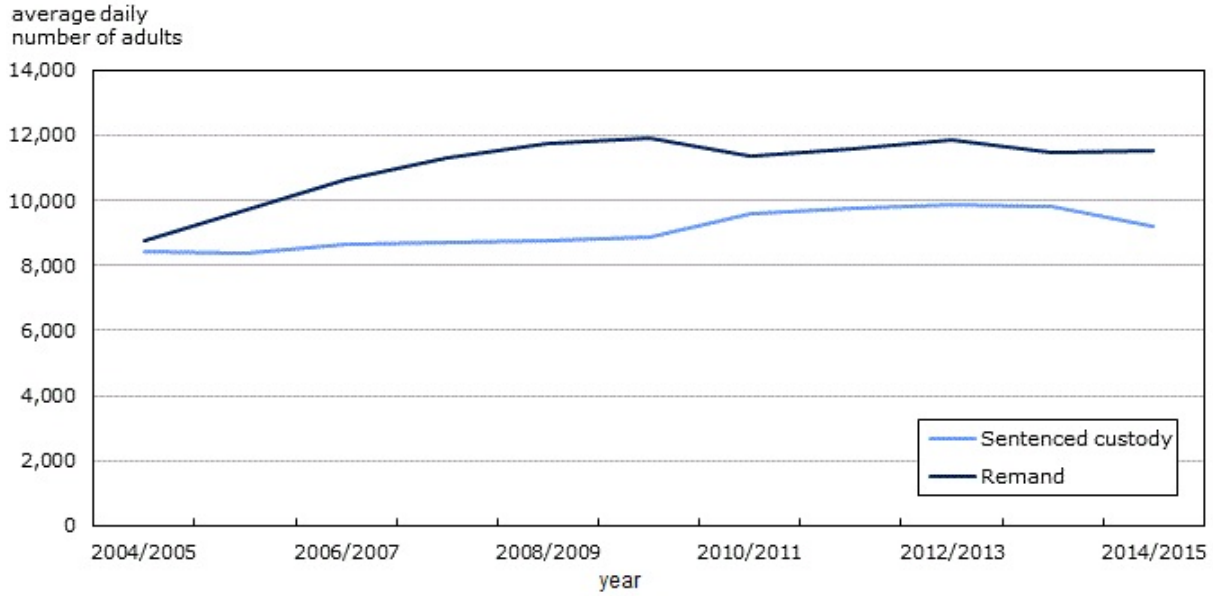

Figure 1.1: Trends in the average daily number of sentenced custody and remanded adult offenders in territorial and provincial custody from 2004/2005 to 2014/2015 (*Correctional Services Program*, 2017).

ulation increases. A higher pretrial population requires more security and concentrated supervision (*Ozkan*, 2017). Also, it might impose a higher cost on the correctional system due to providing required space for an unpredictable remand time duration (*Johnson*, 2003). Studies have shown that many offenders in pretrial custody are held in facilities with maximum security where they are kept in small cells with two or three other offenders. Generally, the offenders do not have any access to recreational facilities. Importantly, there is a high uncertainty for the length of incarceration time (*John Howard Society of Ontario*, 2007). The other serious consequences for offenders held in remand include that they might lose their jobs and houses, become depressed due to the separation from the family, be imposed higher living costs for finding emergency child care, and in some cases they might miss medical treatments (*Canadian Civil Liberties Association and Education Trust*, 2014).

Recidivism has been studied for many years ranging from general recidivism to specific types of crimes to find the factors that play a key role in re-offending (*Fu et al.*, 2011; *Schmidt and Witte*, 1988; *Andrews et al.*, 2008; *Thomson et al.*, 2018). However, there is little research measuring the effectiveness of these risk factors. Recently, artificial intelligence has drawn the attention of researchers to address the issue of recidivism and predict the possibility of re-offending by examining the risk factors (*Alufaisan et al.*, 2017).

In this study, a remand risk assessment tool (RRAT) is proposed that can potentially reduce the pretrial population by lowering the number of unnecessary detentions. This tool determines which offenders are at risk to commit a new crime during their trial using machine learning (ML) techniques. Using the RRAT could lead to reduction in expenses related to the criminal justice system and offenders. This study uses the

Saskatchewan primary risk assessment (SPRA) as an assessment measure for general recidivism in both male and female adult offenders under the jurisdiction of the Ministry of Corrections and Policing. The SPRA, comprised of offenders information in the form of 15 questions, is considered as the input of the proposed RRAT.

To find the best ML model for the RRAT, five ML classifiers known as the support vector classifier (SVC), decision tree classifier (DTC), random forest classifier (RFC), naive Bayesian classifier (NBC), and extreme learning classifier (ELC) are compared in terms of their performance. Then, the best models from the aforementioned models are used to select the most important features affecting the pretrial recidivism.

According to the results, the ELC and the RFC outperformed all other models in the comparison by providing the lowest false positive rate and the highest accuracy, precision, specificity, and area under the receiver operating characteristic curve. Then, the RFC and the ELC combined with binary particle swarm optimization (BPSO) are used for feature selection. The results show that the ELC–BPSO achieved higher accuracy compared to the RFC by selecting seven features from the original 15 features of the SPRA. The ELC–BPSO is able to achieve an accuracy of around 74% using the SPRA data from 15,117 defendants.

1.1 Ethics

Some of the data in this thesis are highly sensitive, and great care was taken to uphold a high ethical standard. The results given in this thesis do not divulge aspects of the data that could be seen as sensitive or proprietary. The University of Saskatchewan ethics file that contains the necessary permissions to work with the data is BEH# 16-166.

1.2 Outline

The thesis contains the research methods used and the results obtained when using the five ML models SVC, DTC, RFC, NBC, and ELC on the SPRA dataset. This thesis is organized as follows. Chapter 2 gives a summary of other studies that consist of important information to comprehend aspects of this study. Chapter 3 gives a description of the machine learning methods used. Chapter 4 gives a detailed description of the data sets used, the methodology, results, and discussion. Chapter 5 gives a summary of the final results and future possible research directions.

Chapter 2

Literature Review

This section introduces concepts and background information associated with each of the main topics permeating this thesis. Section 2.1 discusses machine learning topics as a whole and how they have been used in prediction. Section 2.2 discusses feature selection and some of its applications. Section 2.3 provides general information about recidivism risk assessment tools. Specifically, this chapter gives information about the Saskatchewan primary risk assessment and how it has been enhanced.

2.1 Machine Learning

The term machine learning (ML) was coined by Arthur Samuel in 1959 (*Samuel*, 1959). ML can be defined as the extraction of knowledge from data (*Muller and Guido*, 2017). Nowadays, ML has become useful in many daily-life applications, from automatic recommendations of what to cook (*Jayaraman et al.*, 2017) to which music to listen to (*Kim et al.*, 2007). Apart from its daily-life applications, ML has had a huge impact on research that includes data.

One such example is using ML to detect and predict lung cancer (*Alam et al.*, 2018). In this study, an algorithm is proposed using the support vector classifier (SVC) to detect and predict lung cancer. The algorithm is able to detect cancer-affected cells from computed tomography images and the related stage of the cancer, such as initial, middle, or final. If the algorithm cannot find any cancer-affected cells in the input images, it calculates the probability of lung cancer. For predicting lung cancer, each image is converted into a high contrast picture that contains only black and white pixels. Then, the quantity of white pixels in each image is checked against a specific threshold to check for ordinary or unusual lungs. The proposed algorithm obtained a 97% accuracy for detecting cancer in 126 lung cancer-affected images among 130 images. Moreover, it obtained an 87% accuracy in predicting lung cancer.

In *Shabtai et al.* (2010), eight different ML classifiers are applied to static features that are extracted from Android .apk files. Android .apk files are comprised of valuable information, such as requested permissions, framework methods, and user interface widgets (*Shabtai et al.*, 2010). The goal of the research is to find an optimal combination of classification and feature selection techniques as well as number of features to classify Android applications into games and tools. The other goal of the research is to detect malicious applications. The following eight ML classifiers are evaluated: decision tree classifier (DTC), naive Bayes classifier (NBC),

Bayesian networks, SVC, boosted Bayesian networks (BBN), boosted decision tree classifier, random forest classifier (RFC), and voting feature intervals. Also, three feature selection techniques, chi-squared (*F. Imam et al.*, 2002), Fisher score (*Golub et al.*, 1999), and information gain (IG) (*Shannon*, 1948) are compared for selecting the best features. The combination of BBN and 800 features selected by IG proved to have the highest accuracy of 91% and the lowest false positive rate of 17%.

In *Sönmez et al.* (2018), features of phishing websites are defined, and the extreme learning classifier (ELC) is performed to detect phishing websites. Phishing is a kind of cybercrime that aims at stealing information used by organizations and people to conduct transactions. Phishing websites have special features within their content that can be used in detecting them (*Sönmez et al.*, 2018). In the research, data from 11,000 websites in the UC Irvine Machine Learning Repository database are examined, and 30 features, including features related to phishing websites, are extracted. Then, the ELC is used to detect phishing websites. To evaluate the performance of the ELC, six different activation functions within the ELC are compared, and the sigmoid activation function was determined to have the best performance by achieving the highest accuracy. Then, the ELC is compared with the SVC and the NBC. The results show that the ELC outperformed the two other techniques with an accuracy of 95%.

In *Zawbaa et al.* (2014), an automatic fruit classification system is developed using the RFC. The proposed system includes three stages: pre-processing, feature extraction, and classification. In the pre-processing stage, the fruit images are resized. Then, in the feature extraction stage, images are analyzed, and fruit features, such as shape, color characteristics, and scale invariant feature transform, are extracted. Finally, the classification is done using the RFC, and the result is compared with k-nearest neighbors and the SVC. Based on the results, the RFC obtained the best accuracy of 96%.

It is worth mentioning that predicting human behavior is generally more difficult compared to predicting abovementioned targets and might not lead to ML models with high accuracy levels (i.e., above 80%) (*Cziko*, 1989). One reason for this difficulty is individual differences (*Cronbach*, 1975). For instance, a teaching method used for a student might not work for another student. Another reason that makes predicting human behavior difficult is the constant change in the human behavior over time in such a way that, for example, using a teaching method for a student with specific characteristics might not work after a period of time for the same type of student (*Cronbach*, 1975).

2.2 Feature Selection

In recent years, the dimensionality of data has increased dramatically (*Tang et al.*, 2014). Usually, datasets have a large number of features including relevant, irrelevant, and redundant features (*Xue et al.*, 2013). Irrelevant and redundant features can cause serious problems for existing ML methods, e.g., reducing the model performance (*Gheyas and Smith*, 2010). Moreover, a large number of features can cause a model to overfit the data (*Tang et al.*, 2014). Overfitting happens when a model is fit closely to the features and

variables of the training data and obtains good accuracy on the training data, but it achieves poor results on new data (*Muller and Guido*, 2017). Feature selection is a popular technique for reducing the dimensionality of data (*Tang et al.*, 2014). Using a selection criterion, feature selection selects a subset of relevant features from the original data (*Tang et al.*, 2014). Reducing the number of features could decrease the computation time, simplify the ML model, and increase the model performance (*Dash and Liu*, 1997; *Ünler and Murat*, 2010). Nowadays, feature selection has many applications, such as genetics, diagnosing diseases, and text mining (*Salas-Gonzalez et al.*, 2010; *Li et al.*, 2004; *Mugunthadevi et al.*, 2011; *Chandrashekar and Sahin*, 2014).

In a study done by Kumar and Shaikh, the RFC is applied on a dataset to detect heart disease (*Kumar and Shaikh*, 2017). In this study, the performance of four feature selection techniques in conjunction with the RFC is evaluated. The applied feature selection techniques are relief feature selection, random forest selector, recursive feature elimination, and Boruta feature selection (*Guyon et al.*, 2002; *Kira and Rendell*, 1992; *Kursa and Rudnicki*, 2010). First, the RFC is implemented using the default 13 features of the dataset, 500 trees, and a maximum of three variables at each split and had an 80% accuracy. Then, each of the four feature selection techniques is applied to the data within the RFC. It is found that relief feature selection and the random forest selector produce the same accuracy of 82% by selecting 8 and 9 features, respectively. Moreover, recursive feature elimination and Boruta feature selection choose the same number of features (9) with 84% and 98% accuracy, respectively. The results show that feature selection techniques are able to increase the classifier accuracy. Also, Boruta feature selection selects the set of features that gives the highest accuracy for this problem.

Cao et al. focused on selecting representative speech emotion features and used the SVC to recognize emotions (*Cao et al.*, 2017). First, emotional speeches of two men and two women are recorded. Each person speaks the same 300 emotional short words using the following six emotions: surprise, happy, sad, angry, fear, and neutral. Second, random forest feature selection (RFFS) is used to remove the redundant features that have high correlations with each other. To evaluate the performance of RFFS, another feature selection technique called Spearman correlation analysis is used (*Narayan et al.*, 2012). Then, the features selected by each aforementioned method are used separately within the SVC to perform emotion speech recognition. Emotional speech of three people among four people is used to train the SVC. Also, the emotional speech of the fourth person is used for testing the model performance. The results present that the RFFS method reduces the number of features from 384 to 242. Moreover, the RFFS method outperforms Spearman correlation analysis by achieving 2.2% higher recognition accuracy.

2.3 Recidivism Risk Assessment Tools

Recidivism is defined as "a return to criminal activity, usually measured by arrest, after being convicted of a criminal offense" (*Bartol and Bartol*, 2008). Recidivism has a long history of research. There are several

factors that are related to recidivism and have been assessed in different domains, such as prior criminality, demographic factors, employment, education, and antisocial tendencies (*Blumstein et al.*, 1986; *Gendreau et al.*, 1996; *Gottfredson and Jarjoura*, 1996; *Greenwood and Abrahamse*, 1982; *Pratt and Cullen*, 2000; *U.S. Sentencing Commission*, 2004, 2016).

Jurisdictional policies also consider the historical level of recidivism in different states to estimate the likelihood of recidivism. For instance, compared to the states with high-risk offenders, the likelihood of experiencing recidivism is low in the states with low-risk offenders (*Urahn*, 2011).

Offense type, as another determining factor, can help to estimate the likelihood of recidivism. For example, compared to fraud, larceny, or drug trafficking offenders, the recidivism rates among robbery and firearms offenders is higher (*U.S. Sentencing Commission*, 2004). A study evaluated 336 murderers in 2007, and it is found that no offenders committed another murder, but new violent or drug offenders, who led to homicide while committing a felony, had the highest recidivism rate of 27% (*Roberts et al.*, 2007).

Unlike the offense types, such as domestic violence and accident-related homicide that had shown lower recidivism rates, property crimes, followed by drug offense, public order offense, and violent offense, are the most likely to experience recidivism (*D. Cooper et al.*, 2014). On the other hand, although there is a low probability of recidivism for sex offenders, diverse types of sex offences show different recidivism rates (*Sample and Bray*, 2003, 2006).

The following sections describe the effects of basic demographic, criminal history, and some other individual factors as well as a promising intervention method, risk-need-responsivity (RNR), that incorporates those factors.

1. **Demographic Factors**. Langan and Levin argue that race plays a key role in recidivism (*Langan and Levin*, 2002). In the USA, African-Americans are more likely to recidivate compared to Whites (*Langan and Levin*, 2002). Based on the Bureau of Justice Statistics report, African-Americans recidivated at a higher rate in comparison to Whites within a five-year period (*D. Cooper et al.*, 2014).

Due to ethical and practical issues, using race as one of the predictors of future recidivism is controversial. From the ethical point of view, one may argue that race should not be a factor for predicting future recidivism. Berk underscores the fact that by excluding race and gender one may be sparing some African-American men substantial time in prisons, but at the cost of the deaths of other young African-American men (*Berk*, 2012). From the practical point of view, including race can increase prediction accuracy. Taxman et al. argue that "demographic-neutral" models that exclude age, gender, ethnic, or racial factors lack crucial information in predicting recidivism (*Taxman et al.*, 2013). Berk argues that removing demographic factors of race, gender, and age will reduce the performance of predicting recidivism (*Berk*, 2012).

In terms of gender, the recidivism rate of male offenders was found to always be higher than female offenders (*D. Cooper et al.*, 2014; *Langan and Levin*, 2002). For drug offenders, males are more likely to reoffend compared to females (*Stahler et al.*, 2013). Among violent offenders, Piquero et al. found that the recidivism rate of males is considerably higher than that of females (*Piquero et al.*, 2015).

Stahler introduces age as a predictor for both general and crime-specific recidivism (*Stahler et al.*, 2013). Studies show that as the offender gets older, the probability of recidivism decreases (*Langan and Levin*, 2002). For example, for the ages in the range 25 to 39 years old, the predicted number is 69.7% whereas for the ages 40 years old or older it is 60.3% (*D. Cooper et al.*, 2014). More analyses show that federal offenders released before age 21 had the highest recidivism rate of 67.6%, but for offenders older than sixty at the time of release, a recidivism rate of 16.0% is reported (*U.S. Sentencing Commission*, 2016). Generally, young male offenders have higher risk for violent recidivism (*Piquero et al.*, 2015).

2. **Criminal History**. As frequently mentioned in many research papers and government reports, the most important recidivism determining factor is the offender prior criminal record (*Greenwood and Abrahamse*, 1982; *Blumstein et al.*, 1988; *Piquero et al.*, 2003). The released statistics in Bureau of Justice Statistics report show that around 50% of offenders with three priors were likely to recidivate, whereas this rate is reported to be about 80% for offenders with more than 10 priors in three years (*Langan and Levin*, 2002). The released report in 2014 indicated that 26.4% of offenders with four or fewer priors experienced recidivism, while 56.1% of offenders with 10 or more arrests were rearrested over a year after their release (*D. Cooper et al.*, 2014). The historical criminal record not only can be a quantitative number of prior arrests but also might include other components of prior criminality, such as frequency, seriousness, and recency (*Hoffman*, 1983).

3. **Other individual-level factors**. The recidivism rate might change in proportion to antisocial attitudes or personalities (*Serin et al.*, 2013). In fact, individuals who have fewer social bonds are more probable to recidivate. In contrast, the offenders with different personality disorders are less likely not to be sentenced (*Dejong*, 1997; *Yang et al.*, 2010). A comprehensive review has reported criminogenic needs, history of antisocial behavior, social achievement, age, gender, race, and family factors as the most related features for predicting adult recidivism and intellectual functioning, personal distress factors, and socioeconomic status as the least reliable features (*Gendreau et al.*, 1996).

Hanson and Harris examined 208 sexual offence recidivists and 201 non-recidivists in a study (*Hanson and Harris*, 2000). They found that recidivists generally had poor social supports, poor sexual attitudes, poor self-control strategies, antisocial lifestyles, and increased anger and subjective distress prior to reoffending. In a similar study, Stalans et al. presented the existence of generalized aggression before reoffending as the strongest predictor of violent recidivism (*Stalans et al.*, 2004). Peer delinquency as another feature of future crimes can also affect recidivism (*Matsueda*, 1989; *Warr and Stafford*, 1991; *Benda*, 2003; *Warr*, 1998).

Substance abuse as an important fact among offenders can increase the recidivism rate. Previous research has shown that compared with the general population, substance abuse is more spread among offender populations (*Lurigio et al.*, 2003; *Taxman et al.*, 2007). The importance of understanding the different kinds of substance use, such as lifetime use (ever used), regular use (abuse), and daily life use (dependence) for recidivism is shown by Taxman et al. (*Taxman et al.*, 2013). According to Taxman et al., substance use can increase the recidivism rate among aggressive delinquents. Also, the prior research by the U.S.

Sentencing Commission demonstrated that the probability of recidivism is highly related to education level, marriage, employment, illegal substance use, and the type of punishment (*U.S. Sentencing Commission*, 2004). Offender thinking patterns are also of great importance among other features. The prior research found that most offenders have errors in thinking, especially dominance, entitlement, self-justification, displacing blame, optimistic perceptions of realities, and blaming society (*Taxman et al.*, 2013; *Yochelson and Samenow*, 1976). This issue is related to specific needs of offenders and addressed in the following sections.

4. **The risk-need-responsivity (RNR) model and recidivism**. The RNR model, proposed by Andrews and Bonta, aims at the reducing recidivism rate through a selective focus on prisoners (*Andrews and Bonta*, 2010). The level of service is matched to the level of risk (whom to treat) through the risk principle. In order to distinguish between criminogenic needs and non-criminogenic needs for reducing recidivism, the need principle, which concentrates on the criminogenic needs, is used (what to treat). Finally, the responsivity principle shows how a treatment can be delivered (how to treat) (*Andrews and Bonta*, 2010).

In order to achieve the three main parts of correctional goals, i.e., recidivism reduction, least restrictive sanctioning, and cost effectiveness, the RNR model focuses on the importance of classification in risk level and in treatment (*Taxman et al.*, 2013). The RNR model matches offenders to appropriate supervision levels and services based on their static risks (with no or little change over time) and dynamic criminogenic needs (with change over time) (*Taxman et al.*, 2013). Static risk factors include the age of the first arrest, the number of prior arrests, the number of prior convictions, the number of escapes or infractions in prison, the number of probation violations, and the number of incarcerations, while dynamic risk factors indicate factors that can change. Dynamic risk factors include employment, peer association or substance use and are categorized as criminogenic needs when they relate to recidivism (*Andrews and Bonta*, 2010; *Taxman et al.*, 2013). Some reports indicate that the most predictive items for future re-offending are static or historical items (*Coid et al.*, 2007). Other studies argue that dynamic predictors are as useful as the static predictors in predicting recidivism (*Gendreau et al.*, 1996).

Recidivism has been studied for many years, ranging from general recidivism to specific types of crimes to find the factors that play a key role in re-offending (*Fu et al.*, 2011; *Schmidt and Witte*, 1988; *Andrews et al.*, 2008; *Thomson et al.*, 2018). Recently, ML has drawn the attention of researchers to address the issue of predicting the possibility of re-offending by examining a set of risk factors (*Wang et al.*, 2010; *Ozkan*, 2017).

Wang et al. used the three ML techniques of SVC, logistic regression (LR), and neural networks (NNs) to predict the probability of recidivism by taking nine variables as the inputs (*Wang et al.*, 2010). The studied variables are marital status, age, ethnicity, history of serious alcohol problem, status of past hard drugs usage, status of being convicted for a crime against property, gender, the number of previous incarcerations, and the number of months served for the sample sentence. Using 10-fold cross validation, the optimal number of hidden nodes is estimated to be 10 from a range of four to 25 hidden nodes. Also, the sigmoid function is selected as the activation function for the hidden nodes. The results show that the SVC outperformed the two other techniques with higher accuracy and specificity.

Palocsay et al. focused on the use and comparison of NNs and LR to classify offenders from two datasets with nine features into recidivists and non-recidivists (*Palocsay et al.*, 2000). The nine features are ethnicity, history of alcohol use, history of drug use, gender, the number of previous incarcerations, the time served in prison, age, convictions for, and whether the offender was a felony of misdemeanor. In order to tune the number of hidden nodes and find the best NNs model, the number of hidden nodes are varied from five to 50, and the training and test accuracy for each created network is analyzed. The model with 26 hidden nodes performed better than the other models with 69% accuracy. The result indicates that NN outperforms LR with presenting higher accuracy.

Ozkan studied the performance of five different classification methods for predicting recidivism (*Ozkan*, 2017). In the first step, 80 features that create a good predictive model are extracted from the data by removing the highly correlated features and then applying the Least Absolute Shrinkage and Selection Operator methods (*Tibshirani*, 1996). The five used classifiers are LR, RFC, SVC, XGBoost, and NNs. According to the results, XGBoost had the best accuracy (78%) and area under the curve (AUC). The SVC obtained highest sensitivity and the lowest false negative rate. Finally, the LR classifier scored highest in precision, specificity, and false positive rate. The findings show that depending on the problem, LR can outperform other ML classifiers. But, because ultimately all the classifiers performed closely, it can be concluded that from one dataset to another, the best classifier may change.

2.3.1 Saskatchewan Primary Risk Assessment

The Saskatchewan primary risk assessment (SPRA) is a questionnaire that consists of 15 questions that was created based on the RNR model (*Andrews et al.*, 1990). The SPRA was modeled to measure general recidivism in adult offenders under the jurisdiction of the Ministry of Corrections and Policing (*Patrick et al.*, 2013) in Saskatchewan. The SPRA was developed from two previous risk assessment tools, the Wisconsin case classification system (WCCS) and the primary risk assessment (PRA).

In 1979, the WCCS was created in Wisconsin containing 21 questions to predict general recidivism in adult offenders (*Heinz et al.*, 1979). Later, the Manitoba Community and Youth Services adopted and revised the WCCS and created the PRA with only 15 questions of the WCCS. After the use of the PRA by the Manitoba Community and Youth Services, it was found that the performance of the PRA in predicting recidivism is better than the original WCCS (*Bonta et al.*, 2011).

In the 1990s, adult corrections in Province of Saskatchewan adopted the PRA. The PRA was established as being an effective predictive model (*Patrick et al.*, 2013). Later, some questions were removed from the PRA due to lack of theoretical or statistical basis, and instead, other questions were added to the assessment (*Patrick et al.*, 2013). Consequently, the revised version with 15 questions, called the SPRA, was implemented in Community Corrections in 2007.

There are five purposes for which assessors complete the SPRA: court reports, probation supervision, conditional sentencing, bail, and jail (*Patrick et al.*, 2013). The number of possible responses for questions

varies from 2 to 4. Also, the maximum score for each question varies from 1 to 3 (see Appendix A for more details). The total score of the test, which is the sum of the scores of each question, varies from 0 to 22. Four risk levels of low, medium, high, and very high are assigned to each offender based on their total score on the SPRA (*Patrick et al.*, 2013). The low risk level contains total scores from zero to five. The medium risk level contains total scores from six to 11. The high risk level contains total scores from 12 to 16, and the very high risk level contains total risk scores from 17 to 22.

Chapter 3

Machine Learning

ML can be defined as a set of techniques used to automatically learn patterns in data. Then, the techniques use those patterns to predict the future data (*Murphy*, 2012). ML has four types: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning (*Nawrocka et al.*, 2018). In the supervised learning technique, the predictive model learns from a data set that is already labeled (*Muller and Guido*, 2017). The main types of supervised learning problems include regression and classification (*Muller and Guido*, 2017). In the regression and classification problems, the goal is to predict continuous and discrete values, respectively. This chapter details the ML techniques that are applied to the data of this study to give a better understanding of how they are built and used. The ML techniques support vector machine, decision tree, random forests, naive Bayesian, and extreme learning machine are covered in this study. All the techniques mentioned here can be used for both regression and classification problems. Because this study aims at predicting discrete values, it only focuses on the classification aspect of each technique. So, this chapter presents support vector classifier (SVC), decision tree classifier (DTC), random forest classifier (RFC), naive Bayesian classifier (NBC), and extreme learning classifier (ELC) in details.

3.1 Support Vector Classifier

The SVC was developed in the 1990s to solve classification problems (*Cortes and Vapnik*, 1995). Basically, the idea behind the SVC is to find a hyperplane that best divides the data into two or more existing classes. A hyperplane is an n-variable linear polynomial that separates and classifies a set of data, where $n \geq 1$. The SVC is a natural approach for classifying a linearly separable dataset in a finite dimensional space (*James et al.*, 2013). However, some datasets are not linearly separable. In this case, the SVC applies another approach, which is known as the *kernel trick* (*Hofmann et al.*, 2008). In the kernel trick approach, the SVC maps the vectors (data points) into a higher-dimensional space and then uses a linear classifier in the new space.

Figure 3.1 illustrates an SVC trained with a two-class dataset. Assume this two-class dataset has two features, x_1 and x_2. The goal of the SVC is to design a hyperplane that classifies all the vectors into two classes. In Figure 3.1, two hyperplanes that can correctly classify all the vectors are shown with the red and blue solid lines. The best hyperplane is the one that creates the maximum margin from both classes, where

the margin is the distance between the hyperplane and the nearest vectors from either class. The nearest vectors are known as *support vectors*. The support vectors are the only vectors such that their movement directly affects the maximal margin hyperplane. The movement of other vectors has no effect on the maximal margin hyperplane.

In Figure 3.1, the margins for the red and blue hyperplanes are indicated by z_1 and z_2, respectively. As can be seen from the figure, the value of z_2 is greater than z_1 so the best choice to classify this dataset is the blue hyperplane.

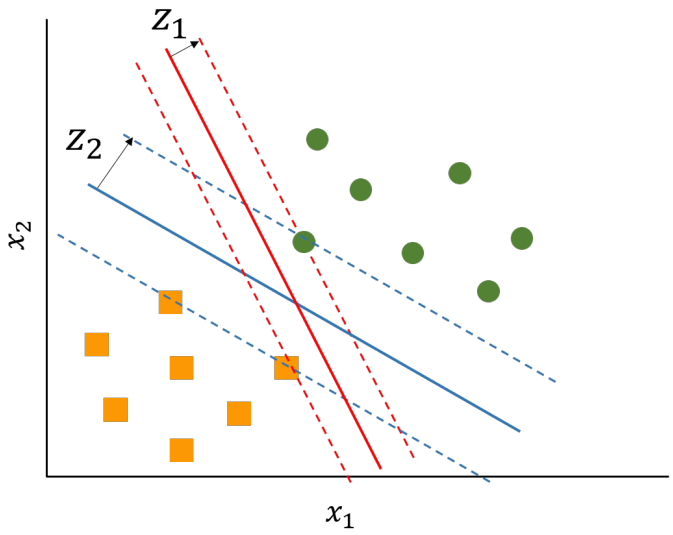

Figure 3.1: An SVC trained by a dataset with two classes

Figure 3.2 illustrates an SVC applied on dataset with two classes. The optimal hyperplane and the margins are visually described in the figure with solid and dashed lines, respectively. Vectors on the margins are the support vectors.

To find the optimal hyperplane for a linearly separable dataset as shown in Figure 3.2, there are a few steps that should be followed:

1. Define the hyperplane H_0 such that

$$H_0 : \boldsymbol{\omega} x + b = 0,$$

where $\boldsymbol{\omega}$ is the vector of weights for each feature, and x is the input vector.

2. By considering one class labeled as positive and the other class labeled as negative, two parallel hyperplanes H_1 and H_2 that define the margins are described as

$$H_1 : \boldsymbol{\omega} x + b = 1,$$

$$H_2 : \boldsymbol{\omega} x + b = -1,$$

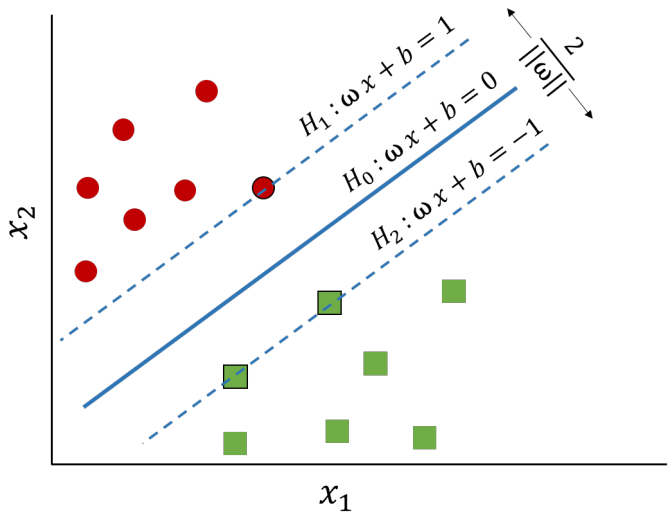

Figure 3.2: Optimal hyperplane with maximum margin for an SVC trained with a two-class dataset

where any vector on or above H_1 is related to the positive class, and any vector on or below H_2 is related to the negative class. The vectors can be represented as

$$\begin{aligned} \boldsymbol{\omega} x_i + b \geq +1 \quad \text{when} \quad y_i = +1, \\ \boldsymbol{\omega} x_i + b \leq -1 \quad \text{when} \quad y_i = -1. \end{aligned} \tag{3.1}$$

Equations (3.1) can be combined into one equation as:

$$y_i(\boldsymbol{\omega} x_i + b) \geq 1$$

3. By recalling the distance between a point(x_0, y_0) and a line $Ax + By + C = 0$ that is

$$\frac{|Ax_0 + By_0 + c|}{\text{sqrt}(A^2 + B^2)},$$

the distance between H_1 and H_0 leads to

$$\frac{|\boldsymbol{\omega} x + b|}{||\boldsymbol{\omega}||} = \frac{1}{||\boldsymbol{\omega}||}.$$

As a result, the margin, which is the total distance between H_1 and H_2, can be represented as

$$\frac{2}{||\boldsymbol{\omega}||}. \tag{3.2}$$

In order to maximize the margin, $||\boldsymbol{\omega}||$ should be minimized. Using Lagrange multipliers, equation (3.2) can be re-expressed in the equivalent form of $\min \frac{1}{2}||\boldsymbol{\omega}||^2$ (*Burges*, 1998).

3.2 Decision Tree Classifier

The DTC is an algorithm that solves a classification problem by learning a hierarchy of if/else questions and answers and creates a tree representation (*Muller and Guido*, 2017). The goal of the DTC is to get the

right classification result by asking the least number of if/else questions. Figure 3.3 shows an example of a DTC.

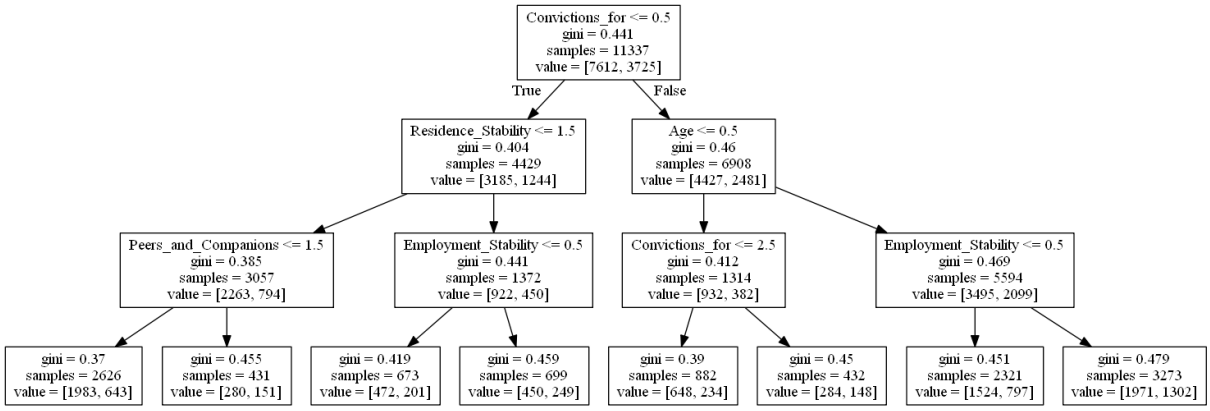

Figure 3.3: A DTC trained by a dataset with two classes

In the DTC, the top node is called the *root*. Also, each node contains either a question, which is called *test*, or a classification result, which is called *leaf*. Moreover, the answer of a test is connected to the next test through *edges*. Furthermore, the length of the longest path from the root to a leaf is called the *depth*.

In the first step for creating a decision tree, the algorithm searches for the most informative feature about the target and creates a test based on that feature in the root. In the second step, the algorithm splits the data into subsets based on the test in the root. For each subset, the two steps are repeated until all the edges lead to a leaf.

The DTC can use different criteria to decide the ordering of features based on finding the most informative feature and then splitting the data. In this study, the two most popular criteria of Entropy and Gini index are used separately. These two criteria measure how much information a feature gives us about a class (*Raileanu and Stoffel*, 2004). In the following sections, the two splitting criteria are discussed in detail.

3.2.1 Splitting Criteria

3.2.1.1 Entropy

Entropy is a common way of measuring the level of impurity in a set of features. For a set of features, X, the Entropy calculation for a given feature, X_m, is

$$S(X_m) = -\sum_{i=1}^{J} p_{mj} \log_b p_{mj},$$

where J is the number of classes in feature X_m, and p_{mj} is the proportion of instances belonging to class j considering feature X_m. Also, the value of b is commonly one of the three numbers, 2, Euler's number e, or 10. In this problem, the classification value is binary, and because of this, b is 2.

For a binary classification problem, if all examples are from only one class, then entropy yields 0. If half of the records are of one class and half are of the other class, then entropy yields 1.

3.2.1.2 Gini Index

The Gini index is a metric to measure how often a randomly chosen element would be incorrectly identified. This means a feature with lower Gini index should be preferred. For a set of features, X, the Gini index for a given feature, X_m, is

$$\begin{aligned} G(X_m) &= \sum_{j=1}^{J} p_{mj}(1 - p_{mj}) \\ &= 1 - \sum_{j=1}^{J} {p_{mj}}^2, \end{aligned}$$

where J is the number of classes in feature X_m and p_{mj} is the proportion of instances of feature X_m that belong to class j.

Although the DTC is fast and easy to interpret, it has high variance (*Murphy*, 2012). This means a small change to the input can cause a large change in the structure of the tree. Because of the hierarchical structure of the tree, a small change at the top affects the rest of the tree. The next section discusses a solution to this problem.

3.3 Random Forests Classifier

The RFC is known as an ensemble machine learning method. In this method, a group of weak learners is used to form a strong learner by which the performance of the model is improved by reducing the model variance (*Muller and Guido*, 2017). The RFC is an algorithm that uses multiple different DTCs to create a powerful algorithm and averages the results of all the DTCs. Consequently, an RFC can generally outperform a single DTC by decreasing the problem of high variance among DTCs (*Muller and Guido*, 2017; *Murphy*, 2012).

The RFC gets its name from randomly choosing data for each tree to make sure the trees are different and independent. In the RFC method, the data are randomly divided into some number of subsets with equal sizes. For each subset a DTC is built using the subset data. Finally, to use the constructed RFC for prediction on new data, the RFC first predicts the target using each DTC in the forest. Then, it uses the majority vote of all the DTCs prediction and assigns the target with the highest probability to the new data.

In the RFC, there are different parameters that affect the accuracy of the model, such as the number of DTCs, depth, and how random the data are chosen for the DTCs (*Muller and Guido*, 2017). Using a larger number of DTCs creates a more robust model by reducing overfitting. However, having more DTCs in the forest requires more time and memory to train the model.

3.4 Naïve Bayesian Classifier

The (NBC) is a simple and fast algorithm that is built based on the Bayes rule and the assumption of strong independence between variables (*Huang and Li*, 2011). The Bayes rule for two events of A and B is defined as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where $P(A|B)$ is the posterior probability, $P(B|A)$ is the likelihood, $P(A)$ is the prior, and $P(B)$ is the evidence. For example, in this study, the prior might be the probability of recidivism being "True" for a given offender, and the evidence might be the probability of the offender to be married. In this case, the posterior probability would be the probability that recidivism is "True" given that the offender is married. Also, the likelihood would be the probability that the offender is married given that recidivism is "True".

Let $\mathbf{X} = \{x_1, x_2, ..., x_M\}$ be a vector of features and $\mathbf{C} = \{c_1, c_2, ..., c_J\}$ a vector of classes. In NBC, the probability of a certain feature being in a certain class can be calculated using the Bayes rule as

$$P(c_j|\mathbf{X}) = \frac{P(\mathbf{X}|c_j)P(c_j)}{P(\mathbf{X})}, \tag{3.3}$$

where $P(c_j|\mathbf{X})$ is the posterior probability of class c_j given a set of features $\mathbf{X}$, $P(\mathbf{X}|c_j) = \prod_{m=1}^{M} P(x_m|c_j), j = 1, 2, \ldots, J$, and is the likelihood of the features given the class, $P(c_j)$ is the prior probability of the class, and $P(\mathbf{X})$ is the prior probability of the features. Applying the maximum a posteriori decision rule (*Murphy*, 2012) to (3.3) gives

$$\hat{\mathcal{Y}} = \underset{j \in \{1,\ldots,J\}}{\operatorname{argmax}} P(c_j) \prod_{m=1}^{M} P(x_m|c_j),$$

where $\hat{\mathcal{Y}} \in \mathbf{C} = \{c_1, c_2, \ldots c_J\}$. In this study, an example of a feature vector and corresponding class variable can be $\mathbf{X}$ = {Marital Status= Married, Age= 40 or over, Gender= Male} and $\mathbf{C}$ = {recidivism= True}, respectively. Here, $P(x_1|c_1)$ means the probability that the offender is "married" given that recidivism is "True".

3.5 Extreme Learning Classifier

The ELC is an efficient algorithm for training single-hidden-layer feed-forward neural networks (SLFNs) with shorter training time and better performance without iteration compared to the conventional algorithms, such as the gradient-based learning algorithms (*Huang et al.*, 2006). In the ELC approach, by randomly selecting the input weights and biases, SLFNs can be viewed as a linear system with the output weights analytically determined using a generalized inverse. Figure 3.4 illustrates the structure of SLFN.

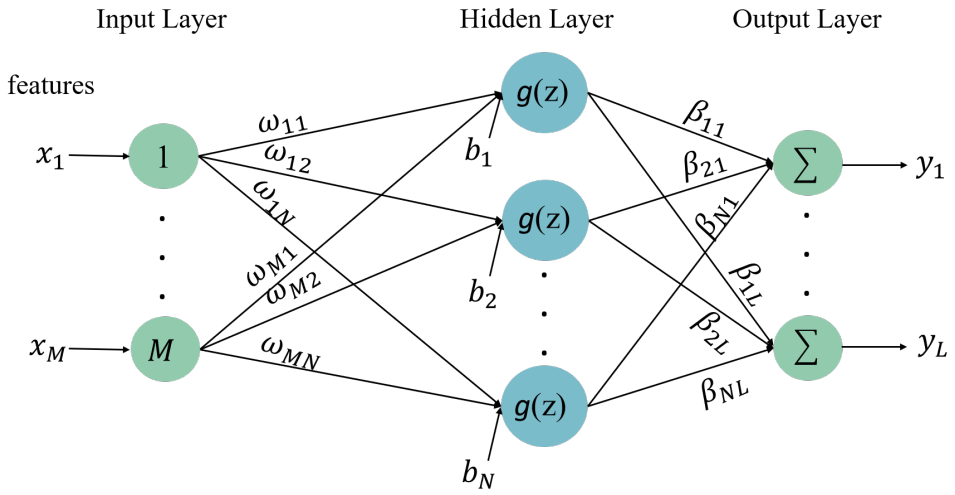

Figure 3.4: SLFN structure

The mathematical form of an SLFN with K training sets and M features can be simplified to $\mathbf{HB} = \mathbf{Y}$, where $\mathbf{H}$, $\mathbf{B}$, and $\mathbf{Y}$ are defined as,

$$\mathbf{H} = \begin{bmatrix} g(\boldsymbol{X}_1\boldsymbol{\omega}_1 + b_1) & \cdots & g(\boldsymbol{X}_1\boldsymbol{\omega}_N + b_N) \\ \vdots & \vdots & \vdots \\ g(\boldsymbol{X}_K\boldsymbol{\omega}_1 + b_1) & \cdots & g(\boldsymbol{X}_K\boldsymbol{\omega}_N + b_N) \end{bmatrix}_{K\times N},$$

$$\boldsymbol{X} = \begin{bmatrix} x_{1M}, & \cdots & , x_{KM} \end{bmatrix}, \boldsymbol{\omega}_n = \begin{bmatrix} \omega_{M1}, & \cdots & , \omega_{Mn} \end{bmatrix}^T,$$
$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta}_1, & \cdots & , \boldsymbol{\beta}_N \end{bmatrix}, \boldsymbol{\beta}_j = \begin{bmatrix} \beta_{j1}, & \cdots & , \beta_{jL} \end{bmatrix}^T,$$
$$\mathbf{Y} = \begin{bmatrix} \boldsymbol{Y}_1, & \cdots & , \boldsymbol{Y}_K \end{bmatrix}, \boldsymbol{Y}_i = \begin{bmatrix} Y_{i1}, & \cdots & , Y_L \end{bmatrix}^T$$

where N is the number of hidden nodes and L is the number of outputs. Also b is the bias, ω and β are the input and output weights, respectively, and g is the activation function. There are many activation functions that can be used, but in this research g is chosen to be the Sigmoid function

$$g(z) = \frac{1}{1 + \exp(-z)}.$$

After calculation of $\mathbf{H}$, the output weights can be determined using $\mathbf{B} = \mathbf{H}^{\dagger}\mathbf{Y}$, where $\mathbf{H}^{\dagger}$ is the Moore–Penrose (generalized) inverse of matrix $\mathbf{H}$; i.e., $\mathbf{B} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{Y}$ (*Huang et al.*, 2006).

3.6 Feature Selection

Feature selection refers to selecting a subset of features $\boldsymbol{A}$ from a set of features $\boldsymbol{B}$ where $\boldsymbol{B} \supseteq \boldsymbol{A}$ using specific feature selection techniques or optimization techniques (*Brezočnik*, 2017). Optimization techniques have various applications, such as NNs training, function optimization, and feature selection (*Chuang et al.*,

2008). In the context of feature selection, an optimization technique can be used to create an optimal model with the fewest number of features, better classification accuracy, and lower complexity (*Chuang et al.*, 2008). There are different feature selection techniques that can be used to reduce the number of features, such as Pearson's Correlation, genetic algorithm, and principal component analysis (*James et al.*, 2013). The following sections explain the optimization techniques called particle swarm optimization (PSO) and binary PSO (BPSO) used for feature selection in this thesis. BPSO was used in this study because it is capable of searching large number of features, has low computation time, is easy to implement, and has few tunable parameters (*Ahmad*, 2015).

3.6.1 Particle Swarm Optimization

Inspired by the natural behaviors of bird flocking and fish schooling, PSO is an evolutionary optimization algorithm to optimize an objective function. PSO was initially proposed by Kennedy and Eberhart in 1995 (*Kennedy and Eberhart*, 1995).

PSO is meant to give the globally optimal solution for an optimization problem using a number of non-optimal initial solutions. Each solution is considered as a particle in a swarm. For each particle i, two vectors are defined: a position vector represented as $\boldsymbol{X}_i = (x_{i1}, x_{i2}, ..., x_{iM})$ and a velocity vector represented by $\boldsymbol{V}_i = (v_{i1}, v_{i2}, ..., v_{iM})$, where M is the number of features. Using the position and velocity, each particle can move in the search space to reach the optimal solution.

At each iteration, each particle updates its position and velocity according to its own experience and that of its neighbors. The best previous position of the particle is recorded as the personal best $P_{best,i}$, and the best position obtained by the population thus far is called G_{best}. Based on $P_{best,i}$ and G_{best}, PSO searches for the optimal solutions by updating the velocity and the position of each particle according to the following equations:

$$\boldsymbol{V}_i^{k+1} = \omega \boldsymbol{V}_i^k + c_1 r_1 (P_{best,i}^k - \boldsymbol{X}_i^k) + c_2 r_2 (G_{best}^k - \boldsymbol{X}_i^k), \tag{3.4}$$

$$\boldsymbol{X}_i^{k+1} = \boldsymbol{X}_i^k + \boldsymbol{V}_i^{k+1}$$

In these equations, $\boldsymbol{V}_i^{k+1}$ and $\boldsymbol{X}_i^{k+1}$ are the updated velocity and position vectors of the particle i in the iteration $(k+1)$. Moreover, r_1 and r_2 are random numbers in the range $[0, 1]$, c_1 and c_2 are the learning factors, and ω is the inertia weight factor. In principle, the iteration continues until G_{best} converges to a constant value. In practice, however, a maximum number of allowable function evaluations is also used as a stopping criterion.

3.6.2 Binary Particle Swarm Optimization

In 1997, Kennedy and Eberhart introduced a new version of the PSO algorithm called BPSO (*Kennedy and Eberhart*, 1997). The PSO and BPSO algorithms have two main differences (*Ünler and Murat*, 2010). First, in the PSO algorithm, the position vector contains continuous values, whereas in the BPSO algorithm the position vector contains binary values. Second, in the BPSO, the velocity vector of each particle is a probability vector, where each element determines the probability of the associated binary value in the position vector taking the value one.

The BPSO can be used for feature selection within a predictive model (*Xue et al.*, 2013). Each particle shows a possible solution, which is a binary pattern of features. For instance, 50 particles means 50 possible solutions or 50 binary patterns from the existing features. The position vectors of the particles represent the features of the studied data, and binary values are assigned to the position vectors to show which features are selected for building the model. For example, the particle 110000 represents a solution where only the first two features are selected for constructing the model.

At each iteration of the BPSO, the position of each particle is calculated using the equation (3.4), and the velocity vector is transformed to a probability vector using the Sigmoid function

$$S(\boldsymbol{V}_i^{k+1}) = \frac{1}{1+\exp^{(-\boldsymbol{V}_i^{k+1})}}, \tag{3.5}$$

where $S(\boldsymbol{V}_i^{k+1})$ represents the probability that the position j in $\boldsymbol{X}_i^k$ is 1. Then, the BPSO updates the positions of the new particles using

$$x_{ij}^{k+1} = \begin{cases} 1 & \text{if } \delta < S(\boldsymbol{V}_i^{k+1}) \\ 0 & \text{otherwise} \end{cases}, \tag{3.6}$$

where δ is a uniformly random number in the range [0,1]. A BPSO can be used within a predictive model by the steps listed in Algorithm 1.

Algorithm 1 BPSO

procedure BPSO FEATURE SELECTION (data, I=number of iterations, R=number of particles)
    Create R particles using all the features in the data.
    for $k = 1$ to I iterations **do**
        for $i = 1$ to R particles **do**
            Construct a model using the features selected in the particle i.
        Compare the accuracy of all the constructed models and select the particle of the model with the highest accuracy as $P_{best,i}$.
        Select the best $P_{best,i}$ as G_{best}.
        Using the equations in (3.4), (3.5), and (3.6), create R new particles by updating the velocity and the position of the previous particles.

3.7 Model Validation

In general, a predictive model can be validated in various ways. This study uses two methods, simple split and cross-validation (CV), to validate the implemented predictive models. In the following sections, the two aforementioned methods are briefly reviewed.

3.7.1 Simple Split

In the simple split method, the original data are randomized and split into two sets called training and testing sets. In this method, the samples are selected with uniform distribution, e.g., each sample has the same probability for being selected (*Reitermanov*, 2010). The prediction error of model is calculated by comparing the predicted value $\hat{\mathcal{Y}}_{\text{predict}}$ and the actual value $\mathcal{Y}_{\text{actual}}$. Hence, the prediction error becomes

$$\begin{aligned}\mathcal{E} &= \frac{\text{the number of misclassifications}}{\text{the total number of test cases}} \\ &= \frac{\sum_{i=1}^{N} \mathcal{I}(\hat{\mathcal{Y}}_{\text{predict}}, \mathcal{Y}_{\text{actual}})}{N},\end{aligned}$$

where the misclassification indicator function $\mathcal{I}$ is given by $\mathcal{I}(\hat{\mathcal{Y}}_{\text{predict}}, \mathcal{Y}_{\text{actual}}) = 1$ if $\hat{\mathcal{Y}}_{\text{predict}} \neq \mathcal{Y}_{\text{actual}}$ and $\mathcal{I}(\hat{\mathcal{Y}}_{\text{predict}}, \mathcal{Y}_{\text{actual}}) = 0$ otherwise.

3.7.2 Cross-Validation

CV is a statistical method that aims at minimizing the probability of overfitting and creating a more unbiased model (*Kononenko and Kukar*, 2007; *Refaeilzadeh et al.*, 2009). Another goal of CV is to select the best model produced by different training algorithms (*Reitermanov*, 2010). Also, CV can be used to find the optimal parameters of models with different levels of complexity (*Reed and Marks*, 1998), such as RFCs with different depths or number of DTCs in the forest. CV has many methods, but k-fold CV is the most

popular one (*Murphy*, 2012). The reason for the popularity of k-fold CV is that all observations are used for both training and validation. In addition, each observation is used exactly once for validation.

In the k-fold CV method, the original data are divided into k subsets with equal sizes. In this case, $k-1$ subsets form the training set, and the other subset forms the validation set. The model is trained based on the training set, and the prediction error is estimated using the validation set. This process is repeated for each subset (for a total of k times), and then, the average prediction error is determined using,

$$\bar{\mathcal{E}} = \frac{\sum_{i=1}^{k} \mathcal{E}_i}{k}. \tag{3.7}$$

The main disadvantage of k-fold CV is that depending on the value of k this method can be computationally expensive or it can overfit if too many models are validated (*Muller and Guido*, 2017). Choosing the number of folds depends on the computation time and the number of samples in the data. With a larger value of k, the error tends to be smaller, but the process can be computationally expensive. Usually, researchers choose $k = 10$, but for a large number of samples it is better to choose a smaller value for k in order to decrease the computation time (*Reitermanov*, 2010).

To increase the validation level of a model, a combination of both methods of simple split and k-fold CV can be used by the steps listed in Algorithm 2 (*Ozkan*, 2017; *Hastie et al.*, 2008).

Algorithm 2 Combination of simple split and cross validation

procedure SIMPLE SPLIT–CROSS VALIDATION (data, K=number of folds, P=number of possible values for the model parameter)
    Split the randomized original data into training and test sets.
    Create a list of P possible values for the model parameter.
    for $i = 1$ to P parameter values **do**
        Split the training data into K folds.
        for $j = 1$ to K folds **do**
            Train a model using K-1 folds and the parameter value i
            Calculate the prediction error of the model using the fold j.
        Using the equation in (3.7) determine the average prediction error of all the calculated prediction errors.
    Compare the average prediction error of all the trained models and select the parameter value of the model with the lowest average prediction error.
    Using the selected parameter value and the entire training set, train a new model.
    Evaluate the model performance using the test set.

3.8 Model Performance Evaluation

After training and validation of the model, the next step is to find out how effective the model is in terms of its performance on a test set. Different metrics are used to evaluate the performance of a model (*Muller and Guido*, 2017). The following sections focus on the metrics used in this study to estimate the performance of the constructed models.

3.8.1 Confusion Matrix

A confusion matrix, also known as an error matrix, is a table used to summarize the performance of a classification model on the test data (*Kononenko and Kukar*, 2007). A confusion matrix has two dimensions, labeled as actual and predicted, with a set of classes on both dimensions. Table 3.1 presents an example of a confusion matrix for a binary classification model with the classes of positive and negative. In the table, values on the diagonal of the matrix show the number of correct predictions for classes positive and negative, whereas off-diagonal values show the number of misclassifications for those classes. A confusion matrix cannot measure the performance of a classification model, but the values inside the matrix can be used to create some performance measures, such as accuracy, precision, sensitivity, specificity, F1-score, and the receiver operating characteristic (ROC) curve.

Table 3.1: Confusion matrix for a binary classification model

		Predicted	
		Positive	**Negative**
Actual	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

3.8.2 Accuracy

Accuracy is the number of correct predictions made by the model over the total number of predictions (*Muller and Guido*, 2017). Using the confusion matrix, accuracy is defined by,

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

3.8.3 Precision

Precision or positive predictive value shows the number of correct predictions among the samples that are predicted to be positive and is calculated by (*Muller and Guido*, 2017),

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

3.8.4 Sensitivity

Sensitivity or true positive rate (TPR) shows the number of correct predictions in class positive and is calculated by (*Muller and Guido*, 2017),

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

3.8.5 Specificity

Specificity or true negative rate shows the proportion of correct predictions in class negative and is defined by (*Muller and Guido*, 2017),

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}.$$

Having sensitivity and specificity, false negative rate (FNR) and false positive rate (FPR) become

$$\text{FNR} = 1 - \text{Sensitivity}, \ \text{FPR} = 1 - \text{Specificity}.$$

3.8.6 F1-Score

The F1-Score is a middle ground between precision and sensitivity (*Muller and Guido*, 2017). It combines precision and sensitivity by taking the harmonic mean of precision and sensitivity. The F1-Score is calculated by

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}.$$

3.8.7 Receiver Operating Characteristic Curve

The receiver operating characteristic (ROC) curve is commonly used to illustrate the performance of a binary classifier (*Fawcett*, 2006). The ROC curve is a two-dimensional curve representing TPR and FPR on its vertical and horizontal axes, respectively. The performance of a ROC curve is evaluated by a single number that defines the area under the curve (AUC) or the area between the curve and the FPR axis. The AUC can be used to compare the performance of multiple classifiers. A classifier with higher AUC has a higher performance. Figure 3.5 depicts the ROC curves for three classifiers. Classifiers A and C have the best and worst performances, respectively ($\text{AUC}_\text{A} > \text{AUC}_\text{B} > \text{AUC}_\text{C}$).

It should be mentioned there are other metrics for evaluating the performance of ML classifiers, such as Precision-Recall curve and learning curves. A Precision-Recall curve is used when the number of data points in one class of the dataset is higher than the number of data points in the other class (*Saito and Rehmsmeier*, 2015). Also, a learning curves shows how error of a model changes as the training set size increases (*Perlich*, 2010). Because the aforementioned curves are beyond the scope of this study, they are not used to compare the performance of the five studied classifiers.

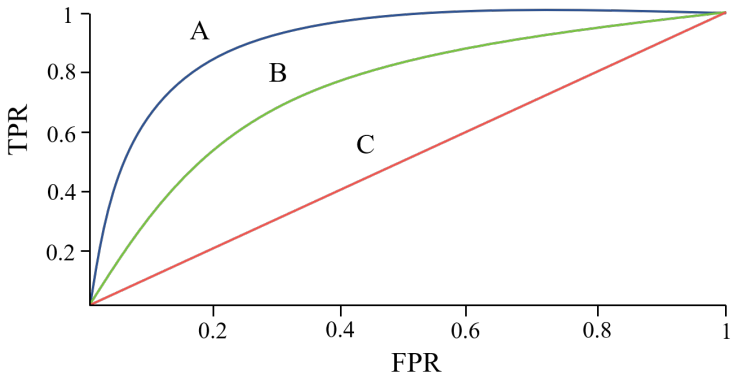

Figure 3.5: ROC curve for three classifiers. A higher AUC results in a higher performance.

Chapter 4

Methodology and Results

The purpose of this chapter is to describe the structure of the provided data files and the method used to clean, validate, and organize them. Moreover, this chapter presents the result of data analysis of two datasets, namely Court_Offence_Appearance and SPRA. Furthermore, the performance of the five ML methods SVC, NBC, DTC, RFC, and ELC are compared to see whether any can be used to increase the predictive performance in the criminal justice system within the context of pretrial recidivism. Finally, the results of feature selection techniques using the best models from the five ML methods are discussed. In this study, the original datasets are provided by the Ministry of Corrections and Policing.

4.1 Remand Data Structure

The remand data are comprised of five data files including two sets of data, namely court and corrections. The court data files are comprised of the records of those offenders who committed crimes and went through their trial processes. The records in the court data files include demographic features of the offender, crime type, court dates, and court decisions. The corrections data files consist of the records of those offenders who are released from a correctional facility. The records in the corrections data files include demographic features, crime type, and corrections duration. It is worth mentioning that because not all the offenders are given a sentence during their trial and sent to correctional facilities, the corrections data files have fewer data than the court data files. Basic information about the five provided data files including court and corrections is given in Table 4.1. As can be seen from this table, most of the files contain multiple records for some offenders. The number of offenders who are in common among the five provided data files is calculated and shown in Figure 4.1.

According to Figure 4.1, there are 30,726 offenders in common between the two corrections data files Correction_Risk_Assessment and RealSubject. Moreover, there are 100,591 offenders in common between the two court data files Court_Offence_Appearance and Court_RealSubject. Furthermore, there are 22,120 offenders in common among the aforementioned data files and RealSubject_Match, which contains records of 53,021 offenders.

Table 4.1: Basic information about each data file

	File name	# of records	# of offenders
1	Court_Offence_Appearance	4,114,630	100,630
2	Court_RealSubject	100,641	100,641
3	Correction_Risk_Assessment	1,493,371	31,818
4	RealSubject	42,494	42,486
5	RealSubject_Match	375,039	53,021

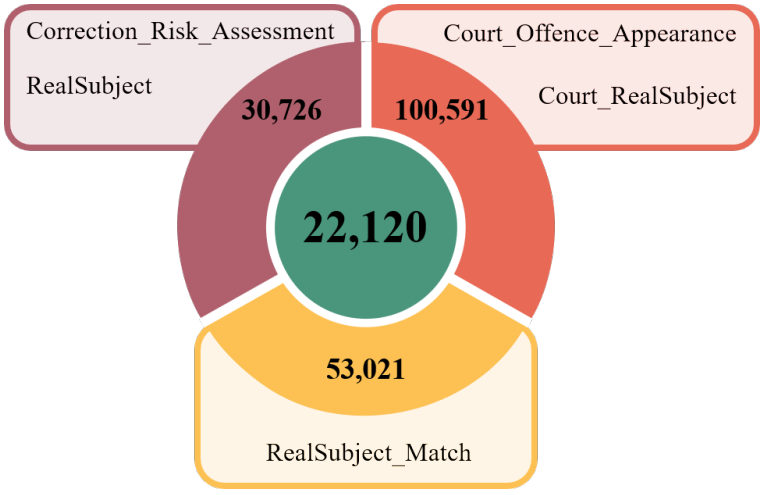

Figure 4.1: Number of offenders in common among all provided data files

4.1.1 Cleaning and Organizing the Remand Data

Originally, all the data files were in a comma-separated values (CSV) format, but for the purpose of data mining, all of the CSV files are converted to the SQLite format. SQLite is a software library that provides a relational database management system. The lite in SQLite means light weight in terms of setup, database administration, and required resources (*Bhosale et al.*, 2015). Data cleaning can be challenging sometimes. Here is a list of challenges regarding the remand data conversion in this research:

1. Each table was divided into multiple CSV files that needed to be merged.

2. Some tables had multiple columns with the same header, making direct data extraction impossible. In such cases, distinct names were assigned to the columns that had the same header.

3. There were many missing values and columns that had to be filled either by calculating or by requesting the Ministry of Corrections and Policing to provide them. For instance, in the Corrections_Risk_Assessment database, there are 158 missing values in the RISKSCORE column related to the SPRA test. They are calculated by summation of the score of each question in the SPRA test.

4. Some columns were removed because of being either empty or irrelevant for this study.

4.1.2 Basic Statistical Analysis of the Remand Data

In this section, a basic statistical analysis for recidivism is performed using the Court_Offence_Appearance database. The original dataset contains 417,432 criminal records. After removing records with no final decision or first appearance, the dataset is reduced to 385,586 records. Due to the fact that the status of offenders while waiting for their trial is not identified in the dataset, the entire data are used to find a base rate for recidivism without considering whether the defendant was remanded or not for trial.

In this study, recidivism is considered to have occurred if at least one new crime is observed before the final decision date of the previous crime. Table 4.2 shows the results of some basic statistical analyses for recidivism using the Court_Offence_Appearance database. It should be mentioned that in this study, each criminal record is considered as a case.

4.1.3 Analysis of the Case Period

The case period for an offender is considered as the period of time the offender was in the trial process. The case period for each offender is calculated by subtracting the first trial date from the final date. Understanding the case period gives a general idea of the costs for the offenders and the correctional facilities. A longer case period results in more costs as the remand duration or the supervision duration increases. Also, a longer case period means that the offender had more time to recidivate if they were not remanded. Table 4.3 shows some basic statistical analysis of the case periods related to those offenders who were considered guilty in their cases.

Looking at Table 4.3, although the minimum and maximum case periods are almost the same between

Table 4.2: Basic statistical analysis of the remand data

Court_Offence_Appearance	
Total offenders	100,630
Male Offenders	72,239
Female Offenders	26,752
Undefined-gender	1,639
Time duration	2008 – 2015
Guilty offenders	63,433
Guilty offenders recidivated at least once	21,520
Total Cases	385,586
Total Guilty Cases	189,233
Recidivism Rate	33.9%

Table 4.3: Basic statistical analysis of the cases period

	Minimum	Maximum	Average
Total Offenders	0 days	6 Years and 6 months and 21 days	4 Months and 29 days
Recidivists	0 days	6 Years and 6 months and 21 days	9 Months and 12 days
Non-recidivists	0 days	6 Years and 6 months and 5 days	2 Months and 27 days

recidivists and non-recidivists, the average case periods of recidivists is around six months more than non-recidivists.

Table 4.4: Frequency of the case period for all offenders

Case Period	Frequency of offenders
Up to 1 year	60,169
Between 1 to 2 years	11,736
Between 2 to 3 years	2,732
Between 3 to 4 years	749
Between 4 to 5 years	257
Between 5 to 6 years	78
Between 6 to 7 years	14

As can be observed from Table 4.4, the number of offenders decreases (from 60,169 to 14) as the case period increases (from less than one year to more than six years). Based on this table, because the number of offenders whose case period is less than one year is by far larger than the other groups, this number (60,169) is broken down by month and shown in Figure 4.2.

As can be seen from Figure 4.2, as the case period increases, the number of offenders declines. In Figure 4.2, the frequency of offenders whose case period is up to one month is higher than the other groups. The data related to the first column of Figure 4.2 are broken down by day and are shown in Figure 4.3. As can be seen in Figure 4.3, for most of the offenders (23,984 offenders) the case period took less than a day. A more comprehensive analysis can be done to determine the case period depending on the type of the offences that the offenders committed.

4.1.4 Analysis of the Survival Period

The survival period is considered to be the period from the time that offenders committed a crime to the time that they recidivated while were waiting for their trial. The survival period for each offender is calculated by subtracting the first date of the current case from the first date of the next case. Table 4.5 shows some basic statistical analysis of the survival period related to those offenders who are considered guilty in their cases and have recidivated. Based on the table, some offenders recidivated within a day, and for others it took more than six years to recidivate.

Table 4.5: Basic statistical analysis of the survival period for the recidivist offenders

	Minimum	Maximum	Average
Survival period	0 days	6 Years and 3 months and 14 days	3 Months and 15 days

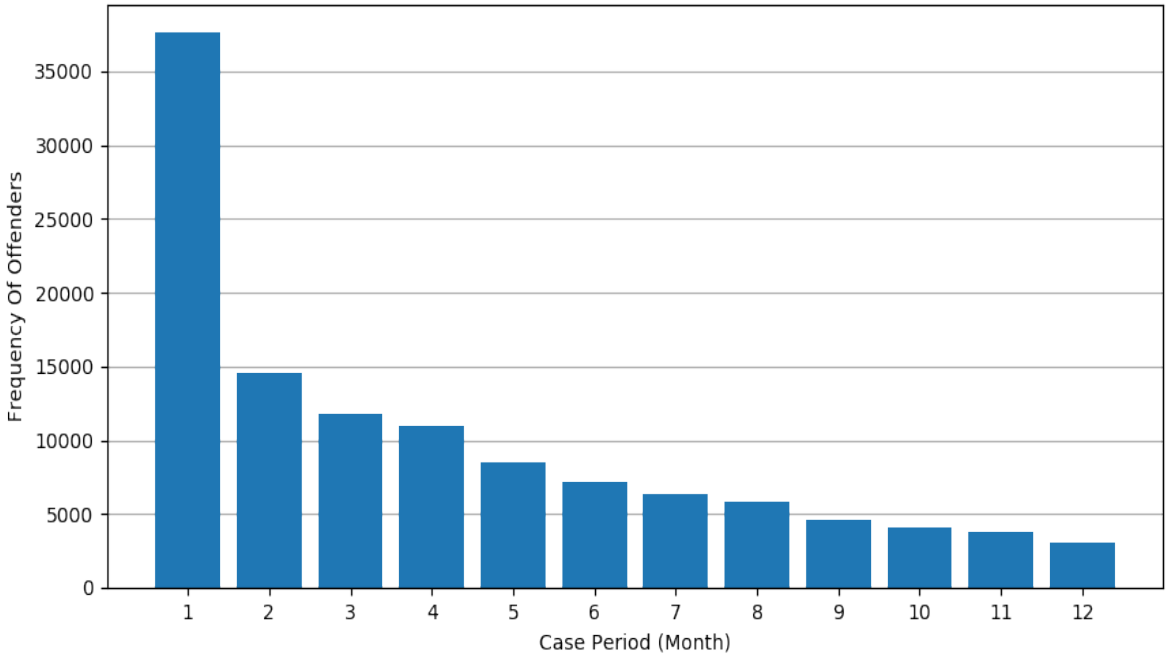

Figure 4.2: Frequency of the offenders whose case period is less than one year

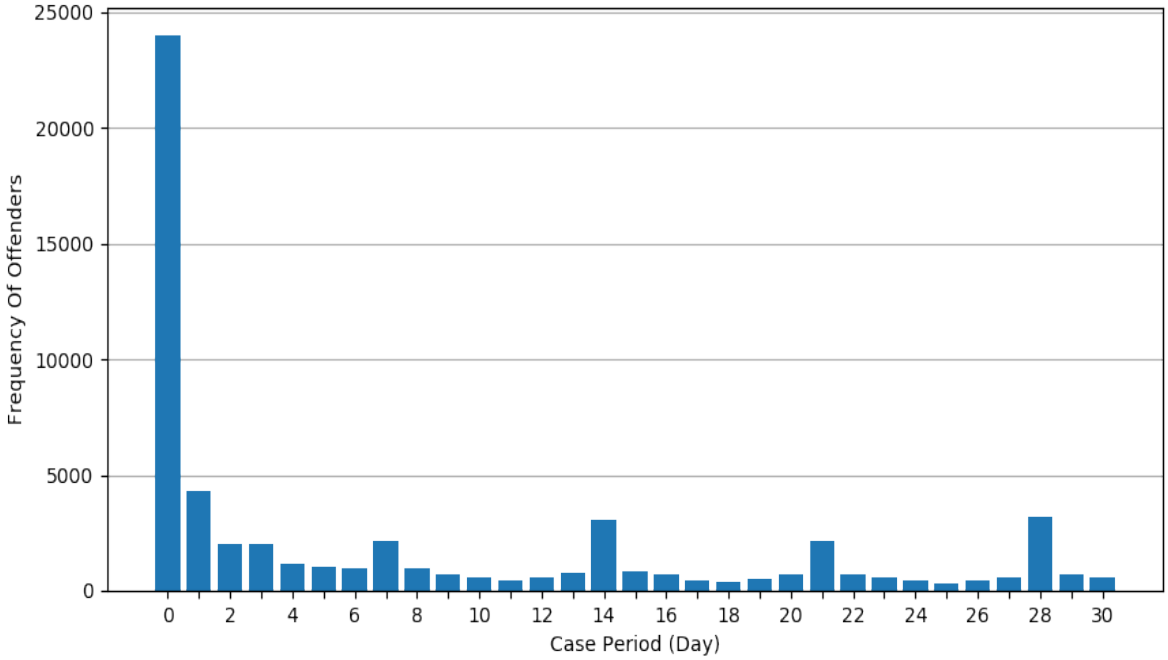

Figure 4.3: Frequency of the offenders whose case period is less than one month

Table 4.6: Frequency of the survival period for the recidivists

Survival Period	Frequency
Up to 1 year	20,437
Between 1 to 2 years	2,711
Between 2 to 3 years	518
Between 3 to 4 years	122
Between 4 to 5 years	56
Between 5 to 6 years	15
Between 6 to 7 years	2

In Table 4.6, because the number of recidivists whose survival period is less than one year is by far larger than the other groups, this number (20,437) is broken down by month and is shown in Figure 4.4. In Figure 4.4, the frequency of offenders whose survival period is up to one month is higher than the other groups (around 10,549). Also, as the survival period increases, the number of offenders drops. The data related to the first column of Figure 4.4 are broken down by day and are shown in Figure 4.5. As can be seen in Figure 4.5, the number of offenders fluctuates over the survival period, and the most common survival period was 28 days.

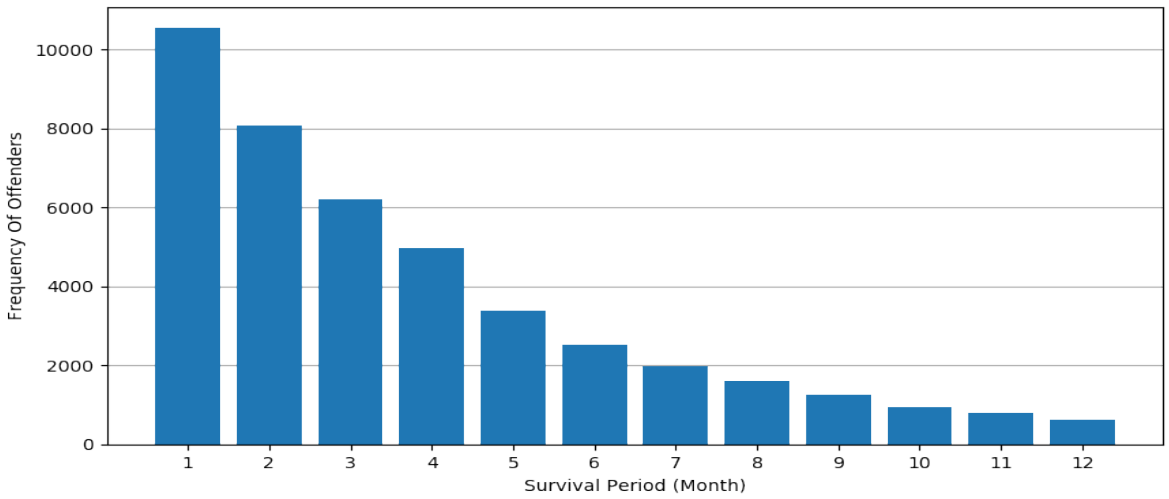

Figure 4.4: Frequency of recidivist offenders whose survival period is less than one year

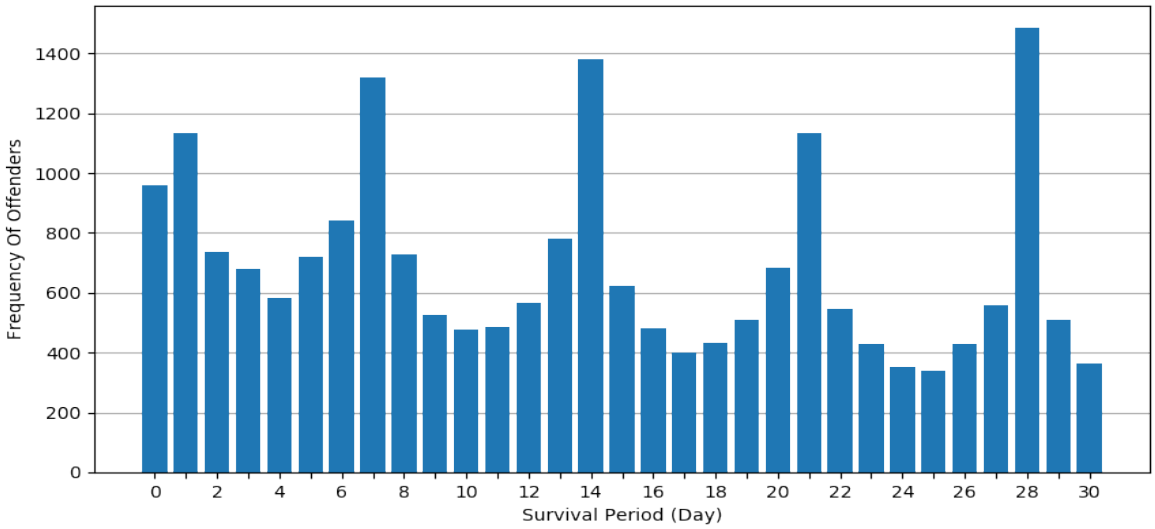

Figure 4.5: Frequency of recidivist offenders whose survival period is less than one month

4.2 Matching Data Between Court and Corrections Databases

Generally, different IDs are allocated to the offenders in court and corrections databases. To link data related to the court and corrections databases and combine the data files, there is a need to match offenders between these two databases. Currently, the RealSubject_Match table provides the connection between offenders in the court and corrections databases by matching the IDs. But this table contains the data only up to 2015. This research aims at expanding the above mentioned matching and updating the RealSubject_Match table by adding the existing data up to 2017. This section discusses the method used to achieve this goal.

To match offenders between the court and corrections databases, at least two databases should be chosen that have some columns in common. Here, among the five provided databases, the RealSubject and Court_RealSubject databases are chosen from court and corrections databases, respectively, because of having three columns in common.

By considering the three columns NAMESORT, DATEOFBIRTH, and GENDER, 36,375 offenders between the RealSubject and Court_RealSubject databases are matched using the Levenshtein distance (LD) algorithm (*Haldar and Mukhopadhyay*, 2011). The LD algorithm is a fuzzy logic technique that finds the number of different characters between two words. In other words, it counts the minimum number of changes namely insertions, deletions, and substitutions needed to convert one word into the other. For instance, the result of one indicates the two words are different in just one character. At the completion of this process, the matched offenders are stored in a new database called New_Match. Figure 4.6 shows the number of offenders in each database.

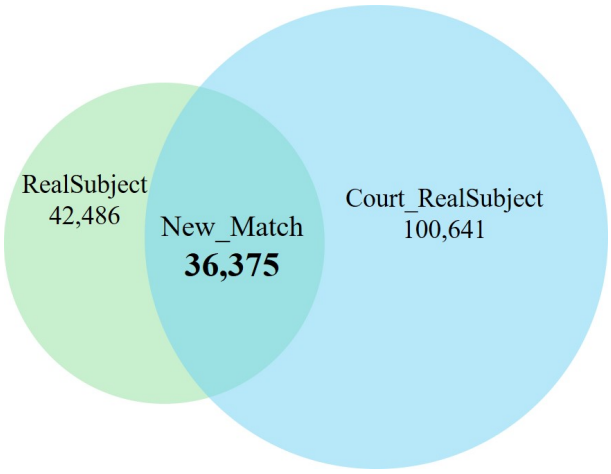

Figure 4.6: Number of offenders in each database

To achieve the above results and construct the New_Match database, the steps listed below are followed.

Step 1: Find offenders between RealSubject and Court_RealSubject databases with exactly the same NAMESORT, DATEOFBIRTH, and GENDER. Then, store these offenders in the New_Match database and remove them from the original databases.

Step 2: By considering the rest of offenders in RealSubject and Court_RealSubject, store offenders with the same DATEOFBIRTH and GENDER in a database called Diff_Name. In this case, each offender in one database can be matched to more than one offender in the other database because of having the same DATEOFBIRTH and GENDER. Table 4.7 shows a possible sample of the Diff_Name database.

Table 4.7: A sample of the Diff_Name database

	Court_RealSubject			**RealSubject**	
NAMESORT	**GENDER**	**DATEOFBIRTH**	**NAMESORT**	**GENDER**	**DATEOFBIRTH**
TO AN	M	9/18/1980	AL CR	M	9/18/1980
TO AN	M	9/18/1980	TO BR	M	9/18/1980
TO AN	M	9/18/1980	TO AM	M	9/18/1980

Step 3: Apply the LD algorithm to all rows of the Diff_Name database and find all the names with one or two differences in characters.

For instance, by applying *Step 3* to the data in Table 4.7, the LD algorithm finds the last row as the matched data with only one difference in characters of the last names. However, there may be more than one offender in each database with the same DATEOFBIRTH, GENDER, and a highly similar NAMESORT. In this case, the LD algorithm cannot decide how to match them. So, for now, those data are removed from the database. Figure 4.7 shows an example of this problem.

It should be mentioned that in this study, the threshold of the LD algorithm is considered to be less than

Court_RealSubject			
NAMESORT	GENDER	DATEOFBIRTH	ID
LO MA	F	9/18/1980	232211

RealSubject			
NAMESORT	GENDER	DATEOFBIRTH	ID
LO MA G.	F	9/18/1980	200145
LO MA GI	F	9/18/1980	200578

Figure 4.7: A sample problem in matching

three due to the fact that names usually have more than two characters and having a threshold greater than two may give poor results. For example, an LD algorithm with threshold less than four considers the two names Joe and Sam as matched names with three differences. In this example, although the LD algorithm correctly identified the number of differences, the two names do not match.

After matching, the New_Match database is checked with the provided matched database RealSubject_Match, and 22,159 offenders between those two databases are found to be in common. Merging the New_Match database with the RealSubject_Match database results in a new database called Total_Match with 67,237 offenders. The numbers of offenders in each of the aforementioned databases are shown in Figure 4.8.

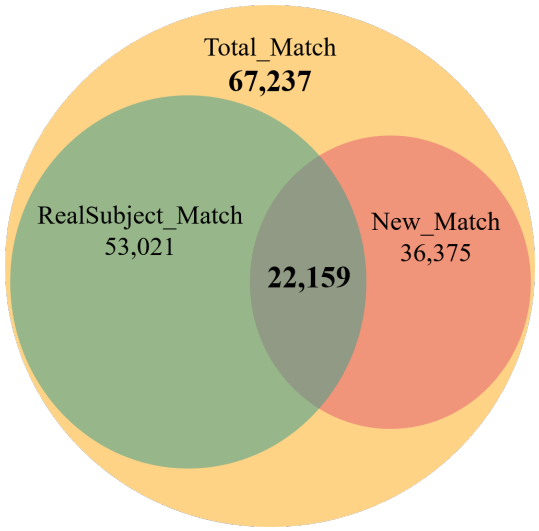

Figure 4.8: Number of offenders in databases containing matched data

In addition, the number of offenders who are in common among the five provided databases and the constructed Total_Match database is calculated and shown in Figure 4.9. As can be seen, after creating the Total_Match database, the number of offenders that are in common among all databases increased from 22,120 (see Figure 4.1) to 28,456.

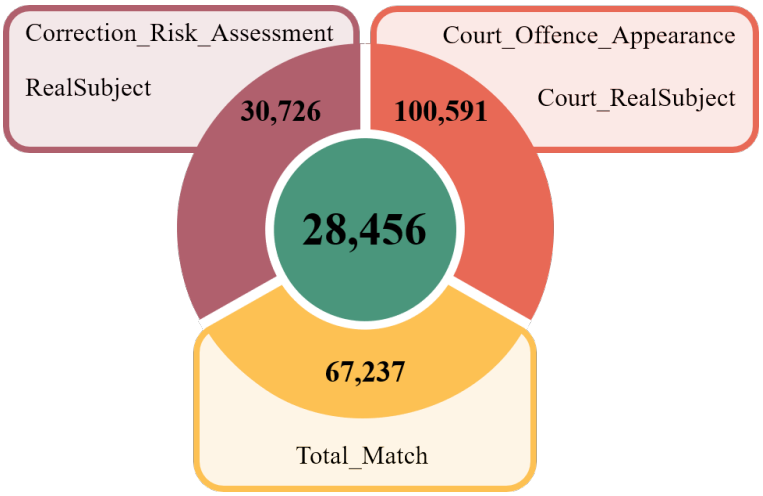

Figure 4.9: Number of offenders in common among all databases

4.3 The Saskatchewan Primary Risk Assessment (SPRA)

The province of Saskatchewan has designed a questionnaire called the Saskatchewan Primary Risk Assessment (SPRA) as a test containing 15 questions to measure the rate of recidivism among adult offenders in the province. In the SPRA, questions have 2 to 4 possible choices. Also, the maximum score for each question varies from 1 to 3 (see Appendix A for more details). The risk score of the test, which is the sum of the scores of each question, varies from 0 to 22. Table 4.8 shows different ranges of risk score organized in four risk levels of low, medium, high, and very high proposed by Patrick et al., (*Patrick et al.*, 2013). The 15 questions of the SPRA are given in Appendix A.

Table 4.8: Four risk levels in SPRA

	Low (L)	**Medium (M)**	**High (H)**	**Very high (VH)**
Risk score range	0 – 5	6 – 11	12 – 16	17 – 22

4.3.1 Experimental Data Preparation: SPRA

A primary goal for this study is to use data from the SPRA to design an intelligent model to predict recidivism among offenders who committed certain crimes and are waiting to be summoned. To this end, first the data related to the SPRA are extracted from the Correction_Risk_Assessment database. The sample initially contained 26,243 offenders, but 47 offenders are removed due to incomplete SPRAs. The COMPLETEDFLAG column in the Correction_Risk_Assessment database shows whether the SPRA is complete or not by values of 'Y' and 'N', respectively. The rest of the offenders are stored in a database called SPRA_Data. In this step, this database contains 26,196 offenders.

For each offender, there are 15 rows in the database that show the answers to the questions of the SPRA.

However, for doing prediction using a ML model with standard libraries in Python, all the data for each offender should be stored in one row with separate columns to be used as the input of the model. Therefore, in the second step, all rows are converted to columns to have only one row for each offender, and the answers to the questions are stored in separate columns. Tables 4.9 and 4.10 present an example of the original form of the SPRA_Data and its converted version, respectively.

Table 4.9: An example of the original version of the SPRA_Data

ID	RISKSCORE	QUESTION	ANSWER
221130	18	Age	0
221130	18	Gender	1
221130	18	Marital Status	0
221130	18	Attitude	0
...	...	...	...

Table 4.10: An example of the converted version of the SPRA_Data

ID	RISKSCORE	Age	Gender	Marital Status	Attitude	...
221130	18	0	1	0	0	...

Third, using the Total_Match database, the offenders who are in both the COURT_OFFENCE _APPEARANCE and the SPRA database, and have at least one guilty case are preserved in the SPRA_Data database, and the rest are removed. In this case, the number of offenders declined to 22,729. The initial number of offenders in the aforementioned databases is shown in Figure 4.10.

Fourth, for each offender, using the Court_Offence_Appearance database, only the cases that happened after the SPRA dates are added to the SPRA_Data database. Due to the fact that for many offenders there are no records after their SPRA date, they are removed from the SPRA_Data database. In this case, the number of offenders is reduced to 11,661.

In the last step, a new column is added to the SPRA_Data database to show recidivism status in a binary format. If an offender recidivated during the case that happened after the SPRA date, the recidivism status is equal to 1, otherwise it is equal to 0. Table 4.11 represents an example of the ultimate SPRA_Data database. Also, the flowchart in Figure 4.11 shows how the number of offenders reduced through the steps. Looking at Figure 4.11, the number of offenders dropped from 26,243 to 11,661 during the SPRA data preparation.

4.3.2 Basic Statistical Analysis of the SPRA Data

In this section, some basic statistical analyses are provided for the SPRA data. The data were collected by correctional staff during 2007–2015 from offenders who had been given a sentence. The original data

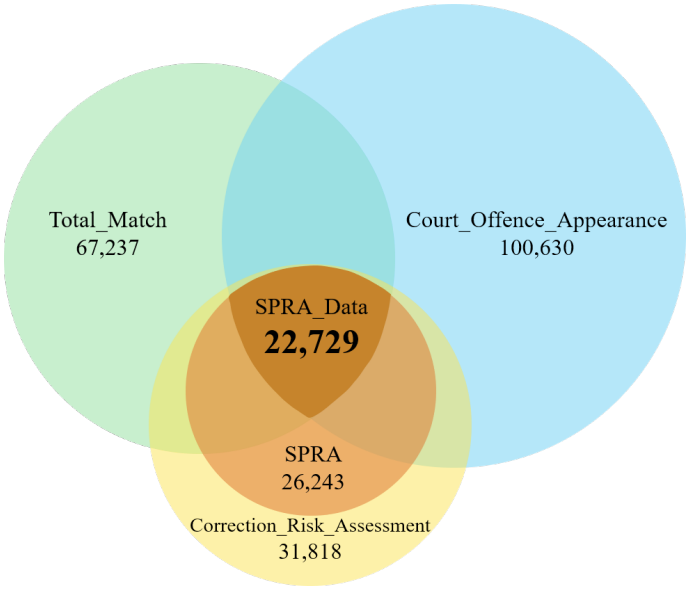

Figure 4.10: The initial numbers of offenders in each database

Table 4.11: An example of the final version of the SPRA_Data database

ID	Recidivism_Status	RISKSCORE	Age	Gender	Marital Status	...
221130	1	18	0	1	1	...
221354	0	9	0	0	1	...

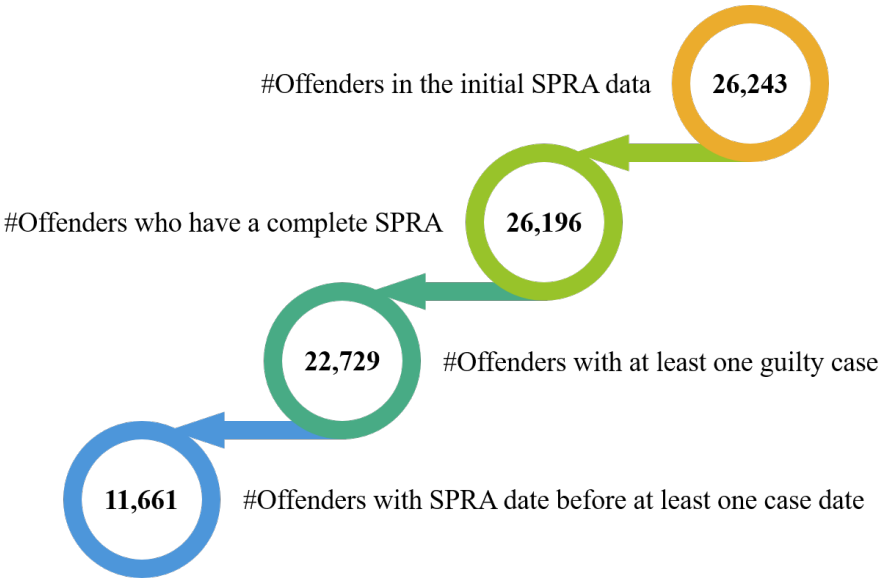

Figure 4.11: Total number of offenders after each step of the SPRA data preparation

contain 32,280 cases. After removing incomplete cases (cases with no final decision from court), the data are reduced to 15,117 cases. Tables 4.12 summarizes the result of basic statistical analysis of the SPRA data (see Appendix B for more detailed results).

Table 4.12: Basic statistical analysis of the SPRA

SPRA_Data	
Total Offenders	11,661
Total Cases	15,117
Male Offenders	9,162 (78.5%)
Female Offenders	2,499 (21.5%)
Min Completed Date	2007-05-01
Max Completed Date	2015-03-14
Min Risk Score	0
Max Risk Score	22
Total Questions	15
Recidivism Rate	59%

Based on Table 4.12, the total number of offenders is less than the total number of records. This means that some offenders have more than one record in the data. Also, the data are mostly comprised of male offenders (78.5% male offenders versus 21.5% female offenders). Furthermore, Table 4.12 shows that 59% of the offenders recidivated at least once during their pretrial. Figure 4.12 illustrates the distribution of cases for recidivists and non-recidivists over all the SPRA risk scores.

As can be seen from Figure 4.12, non-recidivists have the higher number of cases in each risk score compared to recidivists. This leads to the majority of cases being related to non-recidivists over all the SPRA risk scores (10,128 cases for non-recidivists versus 4,989 cases for recidivists). Moreover, most of recidivists (574 offenders) and non-recidivists (1,159 offenders) have the SPRA score of 11. Furthermore, only one offender scored 22 in the SPRA, and this person was a non-recidivist.

Figure 4.13 represents the numbers and the percentages of offenders in each of the risk levels low, medium, high, and very high as proposed by Patrick et al., (*Patrick et al.*, 2013). Based on the figure, the majority of offenders (81%) have a medium or high risk of recidivism. Only 7% of offenders (1,122) have very high risk of recidivism, and 12% of offenders (1,708) have low risk of recidivism. The total population of non-recidivists (green) and recidivists (red) is presented in Figure 4.14. In this figure, the percentage of recidivists at each risk level is shown in white. Also, the total number of offenders at each risk level is shown above each bar.

It can be observed from Figure 4.14 that the number of non-recidivists at each risk level is higher than the number of recidivists. Also, moving from low (21.7%) to very high (38.9%) shows an increase in the

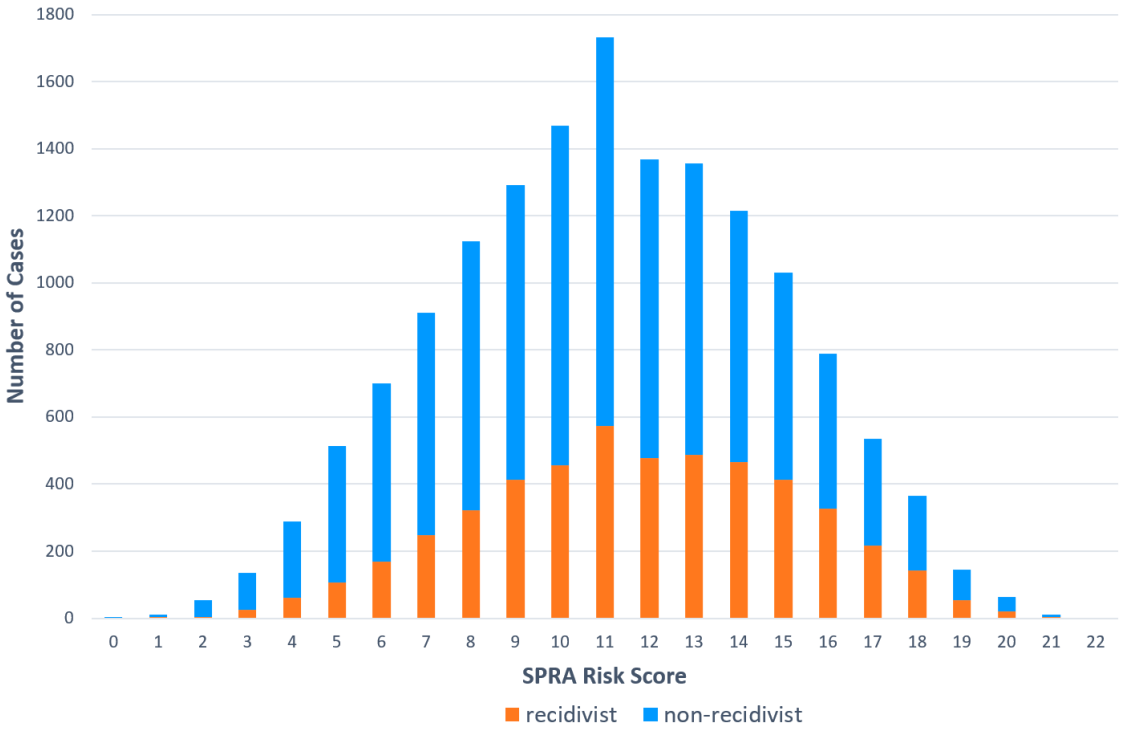

Figure 4.12: SPRA risk score distribution

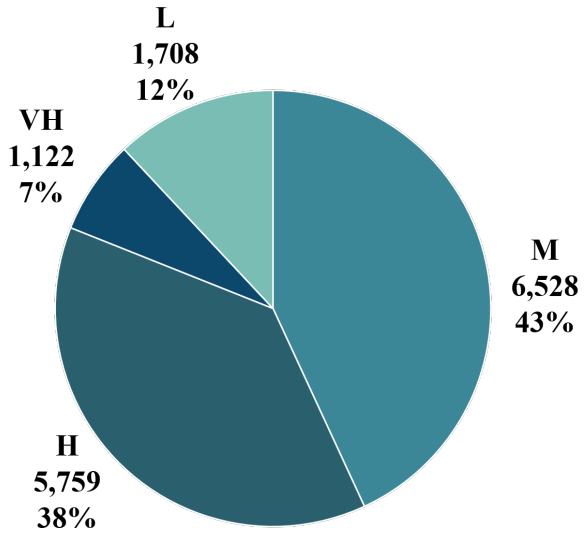

Figure 4.13: Numbers and percentages of offenders at various risk levels.

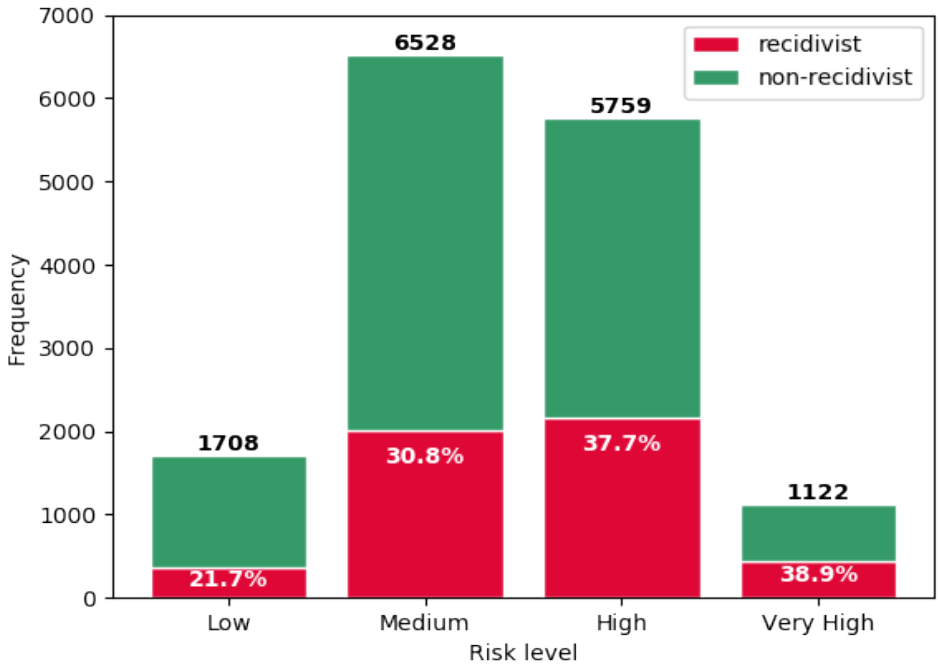

Figure 4.14: Recidivism rate at each risk level. The numbers above each bar show the total number of offenders at the associated risk level

rate of recidivism occurrence in each risk level. Figure 4.15 shows the ROC curve and the AUC score for the SPRA data. According to the figure, the AUC score is around 0.57, which means that the probability of a recidivist scoring higher than a non-recidivist in the SPRA is 57%. Figure 4.15 shows poor correlation between the SPRA risk scores and the recidivism rate based on AUC (0.57).

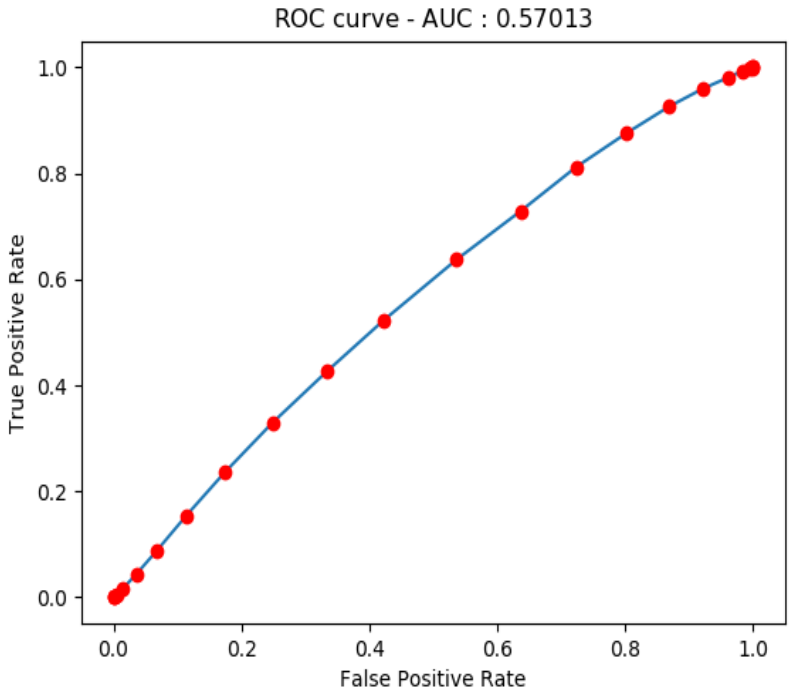

Figure 4.15: The SPRA ROC curve

4.4 Machine Learning Results

The goal of building a model on the given data is to be able to provide court professionals with a model that is able to help predict recidivists. It is also beneficial to provide a list of features that have the most impact on recidivism to the court professionals. The court professionals are looking for a model that uses the fewest number of features. Detecting the most influential features in recidivism may help to validate the SPRA features, make changes to the SPRA, or create a new form of the SPRA. It should be mentioned that in this study, the Python programming language version 3.7 on a Corei7-7660U personal computer with 16GB RAM is used.

4.4.1 Tuning ML Classifiers

In this section, the result of tuning the parameters of each ML classifier is discussed. The data are split up into two groups, 80% training and 20% testing. Using the 10-fold CV method on the training set, the optimal parameters for the studied ML classifiers are found. Table 4.13 shows the optimal values of the parameters selected for each ML classifier using the 10-fold CV method.

Table 4.13: Optimal values of the parameters selected using the 10-fold CV method for each ML classifier

Method	Optimal values
SVC	Kernel = Sigmoid
NBC	Type = Bernoulli
DTC	max_depth = 7, criterion = gini
RFC	max_depth = 15, n_estimator = 35, criterion = gini
ELC	activation_function = sigmoid, hidden_neurons = 20

4.4.2 Comparison of the ML Classifiers

In this section, the results of ML classifiers accepting a combination of 15 features as input and the status of recidivism (1 or 0) as target are discussed. The data are split up into two groups, 80% training and 20% testing. This puts 12,093 data points in the training set and 3,024 in the testing set. The results are obtained by applying the SVC, NBC, DTC, RFC, and ELC on these split data sets. Table 4.14 shows the validity scores for the five studied classifiers. The best values are shown in bold-face.

Table 4.14: Comparison of the five ML classifiers using various performance metrics

	SVC	NBC	DTC	RFC	ELC
Training Accuracy (%)	59	65	68	70	73
Test Accuracy (%)	59	65	67	70	**72**
Precision (%)	70	68	68	74	**75**
Sensitivity (%)	70	89	**97**	64	68
Specificity (%)	38	45	25	67	**77**
AUC	0.54	0.53	0.50	0.60	**0.62**
FPR (%)	62	55	75	33	**23**
FNR (%)	30	11	**3**	36	32
F1-score (%)	70	77	**80**	69	71

The results from Table 4.14 show that the ELC obtained higher accuracy, precision, specificity, and AUC compared to the other classifiers. The RFC followed the ELC and scored second in the aforementioned performance metrics. It would have been more ideal to do a deeper analysis and explore the characteristics of the observations that are correctly classified by all the five ML classifiers; however, limited access to data precluded this possibility.

4.4.3 Feature Selection Results

In this study, because the two methods of the ELC and the RFC beat the other classifiers in many performance metrics (see Table 4.14), they are used to detect the most influential features in recidivism. Table 4.15 shows the number of features selected using two feature selection techniques, the accuracy and the changes in accuracy compared to using all 15 features. The BPSO method with 200 particles and 100 iterations is used within the ELC for feature selection. It should be mentioned that the number of particles and iterations are selected using the 10-fold CV method. This method selected seven features from the original 15 features of the SPRA. For the purpose of comparison, the top seven features used by the RFC are chosen.

Based on Table 4.15, first, the ELC outperforms the RFC by 2%. Then, using the RFC with the seven most important features, the accuracy goes up by 2%, but using the ELC–BPSO is still 2% better. Table 4.16 is a summary of the seven features selected using the ELC–BPSO and the RFC. If a feature is selected by a method, the associated box is filled with a check mark.

Table 4.15: The accuracy using various number of features

Method	# features Selected	Accuracy (%)	Δ Accuracy (%)
ELC–BPSO	7	74	+2
ELC	15	72	-
RFC	7	72	+2
RFC	15	70	-

Table 4.16: The overlay of features selected using ELC–BPSO and RFC

	ELM–BPSO	RFC
Age	✓	✓
Convictions for	✓	
Employment stability	✓	✓
Peers and companions	✓	
Number of Prior Criminal Code Convictions	✓	✓
Gender	✓	✓
Drug and alcohol use	✓	✓
Attitude		✓
Academic and Vocational Skills		✓

4.5 Discussion

Looking at Table 4.14, the ELC had the highest accuracy, precision, specificity, and AUC. Also, the ELC had the lowest FPR. In order to detect future recidivists, a model with high sensitivity should be considered, which refers to percentage of recidivists who were correctly identified as recidivists. As far as the sensitivity is concerned, the DTC with 97% is the clear winner followed by the NBC. Similarly, in order to detect non-recidivists, a model with high specificity should be considered. The ELC with the highest specificity, correctly classified 77% of the actual non-recidivists.

In order to have a balanced predictive power that is good for detecting positive cases but also careful in not incorrectly labeling cases as positive, the F1-score can be used as a general metric of the predictive performances. Doing so, the DTC provided the best performance (80%), with the NBC following at 77%. Also, despite the fact that the classifiers used in this study are different in their working mechanisms, the RFC and the ELC yielded similar results. These two techniques performed better than the other classifiers in terms of accuracy with 72% and 70%, respectively. The SVC with 59% provided the lowest accuracy among all the classifiers.

It should be mentioned that in this study, the status of offenders while waiting for their trial is not identified. As a result, the data include records of the offenders that were remanded prior to their trial and hence could not recidivate. If the status of offenders was identified, those offenders who were remanded would be removed from the datasets, and that could result in an increase in the performance of the predictive models studied.

Due to the fact that the ELC and the RFC obtained better results in most of the performance metrics compared to the other classifiers, only these two methods are used for finding the most influential features in recidivism. As can be seen in Table 4.15, the ELC selected seven features from the 15 features of the SPRA using the BPSO, technique and the same number of features is selected using the RFC. The ELC–BPSO beat the RFC with accuracy of 74% when using the seven selected features. As shown in Table 4.16, the two classifiers agreed on the five features of Age, Employment stability, Number of prior criminal code convictions, Gender, and Drug and alcohol use to be the most influential features in recidivism. The features selected by both methods could provide more predictive information than the other features in the data set. The results of this analysis can help criminal psychologists to improve the current risk assessment by considering only the selected features or by allocating more weight to the selected features.

Chapter 5

Conclusion and Suggestions for Future Work

5.1 Conclusion

This research studies the effectiveness of machine learning (ML) models within the specific context of pretrial recidivism in order to construct a tool called the remand risk assessment tool to be available for the criminal justice system in Saskatchewan. A large volume of information provided by the Ministry of Corrections and Policing in Saskatchewan was cleaned and organized to create a dataset called SPRA to be used for the purpose of this research. A number of ML models are implemented and compared in terms of their performance at predicting recidivism. The ELC and the RFC are chosen as the best models by providing the lowest FPR and the highest accuracy, precision, specificity, and AUC. However, other models could be used in other studies depending on their goals. For instance, in our study, may prefer to have a model with high F1-score rather than high overall accuracy. In this case, the DTC is preferred.

As a result, this study concludes that choosing the best algorithm for constructing a model depends on the properties of the available data, such as the number of features, the number of observations, and the type of input values, whereas choosing the best model from the constructed ones depends on the desired goal. Generally, it is a good idea to start with a simple model, such as the NBC, the linear SVC, or the DTC, and assess the results. After understanding more about the data, one can use more complex models, such as the RFC, the kernelized SVC, or the ELC and focus on improving the performance of the model by tuning its parameters.

This research also aims at helping court professionals to decide whether to remand an offender using an efficient model with the fewest number of features. To this end, the most efficient models, namely the ELC and the RFC, are chosen from the explored models and are used to find the most important features that affect the pretrial recidivism. Due to the fact that the ELC is not able to select the most important features by itself, the BPSO is used within the ELC to do feature selection. The results show that the ELC–BPSO and the RFC chose seven features as the most important features from the original 15 features of the SPRA. The two methods agreed on the five features of Age, Employment stability, Number of prior criminal code convictions, Gender, and Drug and alcohol use to be the most influential features in pretrial recidivism. Providing court professionals with a set of features that are more influential in recidivism than the others can help in providing a minimum requirement of data features that should be collected to aid in constructing

and enhancing the future predictive models. Also, feature selection can help criminal psychologist to validate the features in the SPRA and improve it or make alternate forms of the SPRA.

5.2 Future Work

For future extension of this study, the following research directions are suggested:

- In a more realistic case, if a comprehensive data set is provided that includes the status of offenders while waiting for their trial, this study could be applied to only the data of those offenders who had not been remanded during their pretrial. In this case, the results of the predictive models can be expected to be closer to reality.
- It is suggested to study the impact of features affecting the pretrial recidivism for female and male offenders separately. Dividing the data of female and male offenders and implementing an ML model on each dataset may result in a more accurate model for each gender type.
- An offense-type recidivism classification can be done to specifically predict the type of crime that an offender may commit. For instance, the output of the predictive model can be violent recidivism, cybercrime recidivism, or property crime recidivism. Technically, this is a multi-class classification, and it is suggested to use ML models that are capable of handling these kinds of classification tasks. An offence-type recidivism classification could help the court professionals to decide whether to remand or monitor an offender based on the severity of his predicted crime. In this way, the pretrial population and the cost of detention decrease.
- It is suggested to extend the analysis to predict the survival period for each offender whose recidivism status is predicted as positive, e.g., predicting how many days will take for an offender to recidivate. To do so, an ML model can be trained with the data of the offenders with priors. Predicting the survival period could potentially aid the criminal justice system to decrease the cost of detention by not remanding the offenders from the time they were arrested. Instead, if the constructed model predicts that an offender will recidivate after a certain amount of time, court professionals can use this information for customizing the frequency of monitoring.

5.3 Closing Remarks

As expressed by Kuhn, there is a natural fundamental change in using traditional assumptions and approaches in any research area when a new technique, or generally a particular way of thinking, emerges in the world of science and technology (*Kuhn*, 1962). Similarly, in statistical science, a slow but growing change is being experienced in traditional statistical analyses towards novel estimation techniques. Also, it is of high importance to note that, today, data take many diverse forms and structures that are now available for

any kinds of study, such as forecasting, estimation, classification, and clustering. Therefore, this availability along with the emerging novel techniques and tools enable scientists to study and alleviate many critical social challenges. In this context, data-driven techniques like forecasting and estimation are increasingly drawing more attention in social sciences. As Bushway opined, state-of-the-art tools should be used in criminology to keep pace with ongoing science (*Bushway*, 2013). Hopefully, this study is a step towards proposing new strategies in criminology.

Bibliography

Ahmad, I. (2015), Feature selection using particle swarm optimization in intrusion detection, *International Journal of Distributed Sensor Networks*, *11*(10), 806,954, doi:10.1155/2015/806954.

Alam, J., S. Alam, and A. Hossan3 (2018), Multi-stage lung cancer detection and prediction using multi-class svm classifier, *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, doi:10.1109/IC4ME2.2018.8465593.

Alufaisan, Y., Y. Zhou, M. Kantarcioglu, and B. Thuraisingham (2017), From myths to norms: Demystifying data mining models with instancebased transparency, in *IEEE 3rd International Conference on Collaboration and Internet Computing (CIC)*, IEEE.

Andrews, D. A., and J. Bonta (2010), *The psychology of criminal conduct*, LexisNexis Group, New Providence, NJ.

Andrews, D. A., J. Bonta, and R. D. Hoge (1990), Classification for effective rehabilitation: Rediscovering psychology, *Criminal Justice and Behavior*, *17*(1), 19–52.

Andrews, D. A., J. Bonta, and R. D. Hoge (2008), Classification for effective rehabilitation: Rediscovering psychology, *Criminal Justice and Behavior*, *17*(1), 19–52.

Bartol, C. R., and A. M. Bartol (2008), *Criminal Behavior: A Psychosocial Approach*, Pearson.

Benda, B. B. (2003), Survival analysis of criminal recidivism of boot camp graduates using elements from general and developmental explanatory models, *International Journal of Offender Therapy and Comparative Criminology*, *47*(1), 89–110, doi:10.1177/0306624X02239277.

Berk, R. A. (2012), *Criminal Justice Forecasts of Risk: A Machine Learning Approach*, Springer, New York, NY.

Bhosale, S., T. Patil, and P. Patil (2015), Sqlite: Light database system, *International Journal of Computer Science and Mobile Computing*, *4*(4), 882–885.

Blumstein, A., J. Cohen, J. A. Roth, and C. A. Visher (1986), *Criminal Careers and "Career Criminals"*, vol. 1, The National Academies Press, Washington, DC.

Blumstein, A., J. Cohen, and D. P. Farrington (1988), Criminal career research: Its value for criminology, *Criminology*, *26*(1), 1–35, doi:10.1111/j.1745-9125.1988.tb00829.x.

Bonta, J., T. Rugge, B. Sedo, and R. Coles (2011), *Case Management in Manitoba Probation 2004-01*, Public Safety and Emergency Preparedness Canada.

Brezočnik, L. (2017), Feature selection for classification using particle swarm optimization, in *IEEE EUROCON 2017 -17th International Conference on Smart Technologies*, pp. 966–971.

Burges, C. J. (1998), A tutorial on support vector machines for pattern recognition, *Data mining and knowledge discovery*, *2*(2), 121–167, doi:10.1023/A:1009715923555.

Bushway, S. (2013), Is there any logic to using logit finding the right tool for the increasingly important job of risk prediction, *Criminology and Public Policy*, *12*, 563–567, doi:10.1111/1745-9133.12059.

Canadian Civil Liberties Association and Education Trust (2014), Set up to fail: Bail and the revolving door of pre-trial detention.

Cao, W.-H., J.-P. Xu, and Z.-T. Liu (2017), Speaker-independent speech emotion recognition based on random forest feature selection algorithm, in *2017 36th Chinese Control Conference (CCC)*, pp. 10,995–10,998, doi:10.23919/ChiCC.2017.8029112.

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

Chuang, L.-Y., H.-W. Chang, C.-J. Tu, and C.-H. Yang (2008), Improved binary PSO for feature selection using gene expression data, *Computational Biology and Chemistry*, *32*(1), 29–38.

Coid, J., M. Yang, S. Ullrich, T. Zhang, A. Roberts, C. Roberts, R. Rogers, and D. Farrington (2007), Predicting and understanding risk of re-offending: the prisoner cohort study, *UK Ministry of Justice*.

Correctional Services Program (2017), Trends in the use of remand in canada, 2004/2005 to 2014/2015, *Juristat*.

Cortes, C., and V. Vapnik (1995), Support-vector networks, *Machine Learning*, *20*(3), 273–297, doi:10.1007/BF00994018.

Cronbach, L. J. (1975), Beyond the two disciplines of scientific psychology, *American Psychologist*, *30*(2), 116–127, doi:10.1037/h0076829.

Cziko, G. A. (1989), Unpredictability and indeterminism in human behavior: Arguments and implications for educational research, *Educational Researcher*, *18*(3), 17–25.

D. Cooper, A., M. R. Durose, and H. N Snyder (2014), *Recidivism of prisoners released in 30 states in 2005: Patterns from 2005 to 2010*, Bureau of Justice Statistics, Washington, DC.

Dash, M., and H. Liu (1997), Feature selection for classification, *Intelligent Data Analysis*, *1*(1–4), 131–156.

Dejong, C. (1997), Survival analysis and specific deterrence: Integrating theoretical and empirical models of recidivism, *Criminology*, *35*(4), 561–576.

F. Imam, I., R. S. Michalski, and L. Kerschberg (2002), Discovering attribute dependence in databases by integrating symbolic learning and statistical analysis techniques.

Fawcett, T. (2006), An introduction to ROC analysis, *Pattern Recognition Letters*, *27*, 861–874.

Fu, A. C., R. A. Galione, A. J. Keith, M. G. Miller, C. A. Paxton, C. M. Vaccarello, M. C. Smith, and K. P. White (2011), Albemarle-Charlottesville regional jail overcrowding systems analysis, *IEEE Systems and Information Engineering Design Symposium*.

Gendreau, P., T. Little, and C. Goggin (1996), Meta-analysis of the predictors of adult offender recidivism: What works!, *Criminology*, *34*(4), 575–607.

Gheyas, I. A., and L. S. Smith (2010), Feature subset selection in large dimensionality domains, *Pattern Recognition*, *43*(1), 5–13.

Golub, T., D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. D Bloomfield, and E. S Lander (1999), Molecular classification of cancer: class discovery and class prediction by gene monitoring, *Science (New York, N.Y.)*, *286*, 531–537.

Gottfredson, S. D., and G. R. Jarjoura (1996), Race, gender, and guidelines-based decision making, *Journal of Research in Crime and Delinquency*, *33*(1), 49–69.

Greenwood, P. W., and A. Abrahamse (1982), *Selective Incapacitation*, Rand Corporation, Santa Monica, CA.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik (2002), Gene selection for cancer classification usingsupport vector machines, *Machine Learning*, *46*, 389–422.

Haldar, R., and D. Mukhopadhyay (2011), Levenshtein distance technique in dictionary lookup methods: An improved approach, *Computing Research Repository - CORR*.

Hanson, R. K., and A. J. R. Harris (2000), Where should we intervene? Dynamic predictors of sexual offense recidivism, *Criminal Justice and Behavior*, *27*(1), 6–35.

Hastie, T., R. Tibshirani, and J. Friedman (2008), *The Elements of Statistical Learning*, Springer.

Heinz, R., B. Bemus, and C. C. D. Project (1979), *The Wisconsin Case Classification/Staff Deployment Project: A Two Year Follow-up Report*, Project report, Wisconsin Division of Corrections.

Hoffman, P. B. (1983), Screening for risk: A revised salient factor score (sfs 81), *Journal of Criminal Justice*, *11*(6), 539–547.

Hofmann, T., B. Scholkopf, and A. J. Smola (2008), Kernel methods in machine learning, *The Annals of Statistics*, *36*(3), 1171–1220.

Huang, G.-B., Q.-Y. Zhu, and C.-K. Siew (2006), Extreme learning machine: Theory and applications, *Neurocomputing*, *70*(1-3), 489–501.

Huang, Y., and L. Li (2011), Naive bayes classification algorithm based on small sample set, in *2011 IEEE International Conference on Cloud Computing and Intelligence Systems*, pp. 31–36.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013), *An Introduction to Statistical Learning*, Springer.

Jayaraman, S., T. Choudhury, and P. Kumar (2017), Analysis of classification models based on cuisine prediction using machine learning, *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, doi:10.1109/SmartTechCon.2017.8358611.

John Howard Society of Ontario (2007), Remand in Ontario. Second report to the board, standing committee on prison conditions in Ontario.

Johnson, S. (2003), Custodial remand in Canada, 1986/1987 to 2000/2001, *Juristat*, *23*.

Kennedy, J., and R. Eberhart (1995), Particle swarm optimization, in *IEEE International Conference on Neural Networks*, pp. 1942–1948, IEEE.

Kennedy, J., and R. C. Eberhart (1997), A discrete binary version of the particle swarm algorithm, in *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, vol. 5, pp. 4104–4108, doi:10.1109/ICSMC.1997.637339.

Kim, D., K. su Kim, K.-H. Park, J.-H. Lee, and K. M. Lee (2007), A music recommendation system with a dynamic k-means clustering algorithm, *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, doi:10.1109/ICMLA.2007.97.

Kira, K., and L. A. Rendell (1992), A practical approach to feature selection, *Machine Learning Proceedings 1992*, pp. 249–256.

Kononenko, I., and M. Kukar (2007), *Machine Learning and Data Mining: Introduction to Principles and Algorithms*, Horwood Publishing.

Kuhn, T. S. (1962), *The Structure of Scientific Revolutions*, The University of Chicago Press.

Kumar, S. S., and T. Shaikh (2017), Empirical evaluation of the performance of feature selection approaches on random forest, in *2017 International Conference on Computer and Applications (ICCA)*, pp. 227–231, doi:10.1109/COMAPP.2017.8079769.

Kursa, M. B., and W. R. Rudnicki (2010), Feature selection with the boruta package, *Journal of Statistical Software*, *36*(11), 1–13.

Langan, P. A., and D. J. Levin (2002), Recidivism of prisoners released in 1994, *Federal Sentencing Reporter*, *15*(1), 58–65.

Li, T., C. Zhang, and M. Ogihara (2004), A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, *Bioinformatics*, *20*(15), 2429–2437.

Lurigio, A. J., Y. I. Cho, J. A. Swartz, T. P. Johnson, I. Graf, and L. Pickup (2003), Standardized assessment of substance-related, other psychiatric, and comorbid disorders among probationers, *International Journal of Offender Therapy and Comparative Criminology*, *47*(6), 630–652, doi:10.1177/0306624X03257710.

Matsueda, R. L. (1989), The dynamics of moral beliefs and minor deviance, *Social Forces*, *68*(2), 428–457, doi:10.2307/2579255.

Mugunthadevi, K., S. Punitha, M. Punithavalli, and K. Mugunthadevi (2011), Survey on feature selection in document clustering, *International Journal on Computer Science and Engineering*, *3*(3), 1240–1241.

Muller, A. C., and S. Guido (2017), *Introduction to Machine Learning with Python*, O'Reilly Media, Inc.

Murphy, K. P. (2012), *Machine Learning : A Probabilistic Perspective*, Massachusetts Institute of Technology.

Narayan, A. J., J. E. Herbers, E. J. Plowman, A. H. Gewirtz, and A. S. Masten (2012), Expressed emotion in homeless families: A methodological study of the five-minute speech sample, *journal of the Division of Family Psychology of the American Psychological Association (Division 43)*, *26*(4), 648–653.

Nawrocka, A., A. Kot, and M. Nawrocki (2018), Application of machine learning in recommendation systems, *2018 19th International Carpathian Control Conference (ICCC)*, doi:10.1109/CarpathianCC.2018.8399650.

Ünler, A., and A. E. Murat (2010), A discrete particle swarm optimization method for feature selection in binary classification problems, *European Journal of Operational Research*, *206*(3), 528–539.

Ozkan, T. (2017), Predicting recidivism through machine learning, Ph.D. thesis, University of Texas at Dallas.

Palocsay, S., P. Wang, and R. Brookshire (2000), Predicting criminal recidivism using neural networks, *Socio-Economic Planning Sciences*, *34*, 271–284.

Patrick, G., L. Orton, and J. S. Wormith (2013), The predictive validity of the Saskatchewan primary risk assessment (SPRA), *Centre for Forensic Behavioural Science and Justice Studies*.

Perlich, C. (2010), *Learning Curves in Machine Learning*, pp. 577–580, Springer US, Boston, MA, doi: 10.1007/978-0-387-30164-8_452.

Piquero, A. R., D. P. Farrington, and A. Blumstein (2003), The criminal career paradigm, *Crime and Justice*, *30*, 359–506.

Piquero, A. R., W. G. Jennings, B. Diamond, and J. M. Reingle (2015), A systematic review of age, sex, ethnicity, and race as predictors of violent recidivism, *International Journal of Offender Therapy and Comparative Criminology*, *59*(1), 5–26.

Pratt, T. C., and F. T. Cullen (2000), The empirical status of gottfredson and Hirschi's general theory of crime: A meta-analysis, *Criminology*, *38*(3), 931–964.

Raileanu, L. E., and K. Stoffel (2004), Theoretical comparison between the gini index and information gain criteria, *Annals of Mathematics and Artificial Intelligence*, *41*(1), 77–93, doi:10.1023/B:AMAI.0000018580.96245.c6.

Reed, R. D., and R. J. Marks (1998), *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*, Massachusetts Institute of Technology, The MIT Press, Cambridge, Massachusetts, London, England.

Refaeilzadeh, P., L. Tang, and H. Liu (2009), Cross-validation, *Encyclopedia of Database Systems*, pp. 532–538, doi:10.1007/978-0-387-39940-9.

Reitermanov, Z. (2010), Data splitting, in *WDS'10 Proceedings of Contributed Papers*, pp. 31–36.

Roberts, A. R., K. M. Zgoba, and S. M. Shahidullah (2007), Recidivism among four types of homicide offenders: An exploratory analysis of 336 homicide offenders in New Jersey, *Aggression and Violent Behavior*, *12*(5), 493–507.

Saito, T., and M. Rehmsmeier (2015), The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets, *PLOS ONE*, *10*(3), 1–21, doi:10.1371/journal.pone.0118432.

Salas-Gonzalez, D., J. M. Górriz, J. Ramírez, M. López, I. Álvarez, F. Segovia, and C. G. Puntonet (2010), Computer-aided diagnosis of Alzheimer's disease using support vector machines and classification trees, *Physics in Medicine & Biology*, *55*(10), 418–425.

Sample, L. L., and T. M. Bray (2003), Are sex offenders dangerous?, *Criminology & Public Policy*, *3*(1), 59–82.

Sample, L. L., and T. M. Bray (2006), Are sex offenders different? an examination of rearrest patterns, *Criminal Justice Policy Review*, *17*(1), 83–102.

Samuel, A. (1959), Some studies in machine learning using the game of checkers, *IBM Journal of Research and Development*, *3*(3), 210–229, doi:10.1147/rd.33.0210.

Schmidt, P., and A. D. Witte (1988), *Predicting Recidivism Using Survival Models*, Springer.

Serin, R. C., C. D. Lloyd, L. Helmus, D. M. Derkzen, and D. Luong (2013), Does intraindividual change predict offender recidivism? Searching for the holy grail in assessing offender change, *Aggression and Violent Behavior*, *18*(1), 32–53.

Shabtai, A., Y. Fledel, and Y. Elovici (2010), Automated static code analysis for classifying android applications using machine learning, *2010 International Conference on Computational Intelligence and Security (2010)*, pp. 329–333, doi:10.1109/CIS.2010.77.

Shannon, C. E. (1948), A mathematical theory of communication, *The Bell System Technical Journal*, *27*, 379–423.

Sönmez, Y., T. Tuncer, H. Gökal, and E. Avcı (2018), Phishing web sites features classification based on extreme learning machine, in *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*, IEEE, doi:10.1109/ISDFS.2018.8355342.

Stahler, J., J. Mennis, S. Belenko, and W. Welsh (2013), Predicting recidivism for released state prison offenders, *Criminal Justice and Behavior*, *40*(6), 690–711.

Stalans, L. J., P. Yarnold, M. Seng, D. E. Olson, and M. Repp (2004), Identifying three types of violent offenders and predicting violent recidivism while on probation: A classification tree analysis, *Law and Human Behavior*, *28*(3), 253–271, doi:10.1023/B:LAHU.0000029138.92866.af.

Tang, J., S. Alelyani, and H. Liu (2014), Feature selection for classification: A review, in *Data Classification: Algorithms and Applications*.

Taxman, F. S., M. L. Perdoni, and L. D. Harrison (2007), Drug treatment services for adult offenders: The state of the state, *Journal of Substance Abuse Treatmen*, *32*(3), 239–254, doi:10.1016/j.jsat.2006.12.019.

Taxman, F. S., A. Pattavina, M. S. Caudy, J. Byrne, and J. Durso (2013), *Simulation Strategies to Reduce Recidivism*, 73–111 pp., Springer, New York, NY.

The Government of British Columbia (2019), Custody sentences, https://www2.gov.bc.ca/gov/content/justice, [Online; accessed 6-September-2019].

Thomson, A., J. Tiihonen, J. Miettunen, M. Virkkunen, and N. Lindberg (2018), Firesetting and general criminal recidivism among a consecutive sample of finnish pretrial male firesetters: A register-based follow-up study, *Psychiatry Research*, *259*, 377–384.

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Urahn, S. K. (2011), *State of Recidivism: The Revolving Door of America's Prisons*, The PEW Center on the States, Washington, DC.

U.S. Sentencing Commission (2004), *Measuring recidivism: The criminal history computation of the federal sentencing guidelines*, Washington, DC.

U.S. Sentencing Commission (2016), *Recidivism among federal offenders: A comprehensive overview*, Washington, DC.

Wang, P., R. Mathieu, J. Ke, and H. J. Cai (2010), Predicting criminal recidivism with support vector machine, in *IEEE International Conference on Management and Service Science*, IEEE.

Warr, M. (1998), Life-course transitions and desistance from crime, *Criminology*, *36*(2), 183–216, doi:10.1111/j.1745-9125.1998.tb01246.x.

Warr, M., and M. C. Stafford (1991), The influence of delinquent peers: What they think or what they do?, *Criminology*, *29*(4), 851–866, doi:10.1111/j.1745-9125.1991.tb01090.x.

Xue, B., M. Zhang, and W. N. Browne (2013), Particle swarm optimization for feature selection in classification: A multi-objective approach, *IEEE TRANSACTIONS ON CYBERNETICS*, *43*(6), 1656–1671.

Yang, M., Y. Liu, and J. Coid (2010), Applying neural networks and other statistical models to the classification of serious offenders and the prediction of recidivism, *UK Ministry of Justice Research Series 6/10*, *35*(4), 561–576.

Yochelson, S., and S. E. Samenow (1976), *The Criminal Personality: A Profile for Change*, vol. 1, Jason Aronson, New York, NY.

Zawbaa, H. M., M. Hazman, M. Abbass, and A. E. Hassanien (2014), Automatic fruit classification using random forest algorithm, in *14th International Conference on Hybrid Intelligent Systems*, IEEE, doi:10.1109/HIS.2014.7086191.

Appendix A

The SPRA Questions

Table A.1: SPRA questions

Question	Score
1. Academic and Vocational Skills	
Completed Grade 10 or marketable skill	0
Has Less Than Grade 10 and no marketable skill	1
2. Age	
40 or over	0
39 or less	1
3. Antisocial Behavior	
No evidence of a pattern of antisocial behaviour	0
Evidence of a pattern of antisocial behaviour	1
4. Attitude	
Attitudes Pro-social and supportive of justice system	0
Either pro-criminal attitudes or not supportive of justice system	1
Pro-criminal attitude and not supportive of justice system	2
5. Convictions for	
Not Applicable	0
Fraud, Forgery, Worthless Cheques	1
Theft, Break and Enter, Robbery	2
Convictions for both 1 and 2	3
6. Drug and Alcohol Use	
Evidence of impact in one area	0
No Evidence of impact	1
Evidence of impact in two or more areas	2
7. Employment Stability	
Employed 50% or more over last 12 months	0
Unemployed 50% or more over last 12 months	1
8. Family/Marital Relationships	
Pro-social support	0
Antisocial support/lack of pro-social support	1
9. Financial Situation	
No Serious Problems	0
Evidence of Serious Problems	1
10. Gender	
Female	0
Male	1
11. Number of Prior Criminal Code Convictions	
No Priors	0
1 Conviction	1
2 or More	2

12. Peers and Companions	
No Known Problems With Peers	0
Some Association With Negative Peers	1
Associates Mainly With Negative Peers	2
13. Residence Stability	
None	0
One	1
Two or More	2
14. Self Management	
Good insight and strategies	0
Lack of insight and/or strategies	1
15. Unemployed at time of offence	
Employed at time of offence	0
Unemployed at time of offence	1
Maximum Total Score	22

Appendix B

The SPRA Analysis

Tables B.1–B.9 show the results of basic statistical analysis of the SPRA data.

Table B.1: Frequencies of all offenders categorized by gender

	Frequency	Percent
Male	11,739	77.6%
Female	3,378	22.4%
Total	15,117	100%

Table B.2: Frequencies of all offenders categorized by age group

	Frequency	Percent
40 or over	2,585	17.1%
39 or less	12,532	82.9%
Total	15,117	100%

Table B.3: Frequencies of male and female offenders categorized by age group

	Males	Females
40 or over	2,035 (17.3%)	550 (16.3%)
39 or less	9,704 (82.7%)	2,828 (83.7%)
Total	11,739	3,378

Table B.4: Frequencies of recidivists and non-recidivists categorized by gender

	Males	Females	Total
Recidivists	3,823	1,166	4,989
Non-Recidivists	7,916	2,212	10,128
Total	11,739	3,378	15,117

Table B.5: Frequencies of all offenders categorized by level of education

	Frequency	Percent
Completed Grade 10 or marketable skill	11,231	74.3%
Grade 10 and no marketable skill	3,886	25.7%
Total	15,117	100%

Table B.6: Frequencies of male and female offenders categorized by level of education

	Males	Females
Completed Grade 10 or marketable skill	9,000 (76.6%)	2,231 (66%)
Grade 10 and no marketable skill	2,739 (23.4%)	1,147 (34%)
Total	11,739	3,378

Table B.7: Frequencies of offenders categorized by SPRA risk levels

	Frequency	Percent	Cumulative Percent
Low (0-6)	1,708	11%	11%
Medium (7-11)	6,528	43%	54%
High (12-16)	5,759	38%	92%
Very High (17-22)	1,122	8%	100%
Total	15,117	100%	-

Table B.8: Frequencies of male and female offenders categorized by SPRA risk levels

	Males	Females
Low (0-6)	1,221	478
Medium (7-11)	5,142	1,386
High (12-16)	4,513	1,246
Very High (17-22)	863	259
Total	11,739	3,378

Table B.9: Frequencies of recidivists and non-recidivists by SPRA risk levels

	recidivists	non-recidivists	Total
Low (0-6)	371	1,337	1,708
Medium (7-11)	2,012	4,516	6,528
High (12-16)	2,169	3,590	5,759
Very High (17-22)	437	685	1,122
Total	4,989	10,128	15,117