A Machine Learning Generalization of LSI-OR

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

By
Rahim Oraji

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 Level of Service Inventory-Ontario Revision (LSI-OR) is used as a risk/need assessment tool to classify, manage, and treat the offender population so that they receive supportive services consistent with their custodial needs. This thesis adopts a machine learning approach employing the Naïve Bayes technique as an alternative to the LSI-OR.

The study was conducted on a group of (72725) offenders with different races and includes males (82.62%) and females (17.38%). Participants were monitored for two years to collect recidivism information. A basic analysis of the dataset revealed that 1) 83.18% of population used a unique pattern to answer 43 LSI-OR items, 2) the total LSI-OR scores in the entire population and also in male and female population followed two beta distribution functions, one for each recidivism class, and 3) the recidivism rate was approximated by a normal distribution function.

It was shown that the Naïve Bayes classifier can be considered as an extended LSI-OR classifier that accepts multiple continuous and discrete features as input. In other words, the Naïve Bayes classifier provides a simple framework for studying the effect of distinct features on classification efficiency and accuracy.

The results of running the Naïve Bayes classifier with various input features revealed that the Naïve Bayes classifier presented better performance than the LSI-OR. However, there was no obvious trend in the accuracies predicted by both models to indicate the superiority of one model over the other.

The only feature whose value could be treated as a continuous variable was the LSI-OR score. Many models were created based on continuous and discrete LSI-OR scores producing either the same performance and mean accuracy or slightly better.

The dataset contained many features that are never used by the LSI-OR assessment for instance, the offence severity. A model was built at each index of offence severity based on LSI-OR scores and 43 LSI-OR items as input features. The results of running the experiment indicate that considering 43 LSI-OR items gives more stable results in terms of accuracy than the LSI-OR scores.

Acknowledgements

I would like to express my sincere gratitude to all those who have directly or indirectly contributed to this work.
I would like to express my deepest gratitude to my supervisor Dr. Raymond J. Spiteri, Dr. Mahshid Atapour, and Dr. Daniel Anvari for help, guidance and financial support.
I would like to thank the department of computer science for financial support.
My appreciation also goes to Dr Mehdi Ghasmi for his unconditional help in different situations.

I dedicate this thesis to my parents who are sadly no more.

Contents

List of Tables

List of Figures

List of Abbreviations

Adj	Adjusted
AUC	the Area Under the Curve
CLT	Central Limit Theorem
E(X)	Expectation value of X
F	F-measure
FN	False Negative
FP	False Positive
FPR	False Positive Rate
FNR	False Negative Rate
H	High
L	Low
LM	Levenberg–Marquardt
LSI-OR	Level of Service Inventory Ontario Revision
M	Medium
MAP	Maximum A Posteriori Estimation
NB	Naïve Bayes
PDF	Probability Density Function
P	Precision
R	Recidivism status
R	Recall
ROC	Receiver Operating Characteristic
Std	Standard deviation
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TNR	True Negative Rate
UP	Unique Pattern
VH	Very High
VL	Very Low
$\delta_{\max}$	Maximum probability difference
$\mathcal{A}$	Accuracy
$\hat{\mathcal{B}}$	Beta distribution function
$\mathcal{E}$	Error rate
$\bar{\mathcal{E}}$	Average error
Γ	Gamma function

Chapter 1

Introduction

In Canada, the federal and provincial governments share the administration of adult correctional services, including secure and safe housing, effective correctional settings that properly meet offenders programming needs, and so on; e.g., 139337 adult offenders have been supervised in provincial/territorial or federal correctional services in 2013/2014 [1]. The federal government has jurisdiction the sentencing of over adult offenders (18 years and older) to a term of two years or more, whereas the jurisdiction over other offenders receiving shorter sentences is given to the provincial/territorial government [2]. The justice system in Canada supervises adult offenders from diverse races, genders, and backgrounds. Hence, proposing a correctional program or treatment that considers the individual differences among offenders and protects the general public after their release is not an easy task.

To this end, a variety of assessment tools need to be used by correctional services to classify, manage, and treat the offenders population so that they receive supportive services consistent with their custodial needs. One of the assessment tools in use and empirically tested today is the Level of Service Inventory (LSI). The LSI system has been employed by many correctional agencies, and many studies have been done to estimate the predictive and preventive validity of LSI across various offenders groups [3, 4, 5, 6]. The LSI tool is a risk assessment instrument based on social learning and was developed by two Canadian psychologists Don Andrews and James Bonta in early 1980s [7]. In 1990, a revised version of the LSI called LSI-R was released. Roughly speaking, there have been several generations of risk assessment [8]. The first-generation assessment relied mostly on clinical judgment. In the second generation of risk assessment, a static risk factor played the main role in the determination of offender risk. The third generation of assessment tools considers both static and dynamic risk factors; for instance, the LSI is an example of third-generation assessment tool. A reduced version of the LSI, known as the LSI Ontario revision (LSI-OR), that has fewer items than the original LSI, is currently used in Ontario provincial corrections [9].

The original LSI-R assessment has 54 items scored as either Yes or No on a scale 0 to 3 and grouped into 10 domains as follows (the number of items in each domain is given in parentheses): criminal history (10), education/employment (10), financial (2), family/marital (4), accommodation (3), leisure/recreation (2), companions (5), alcohol/drug problems (9), emotional/personal (5), and attitudes/orientation (4) [10].

In the LSI-OR version, the number of items is reduced from 54 to 43, and each item is scored by 0 or 1 corresponding to No or Yes answers, respectively. In this version, the lowest achievable score is 0 and the highest is 43. In the general risk/need factors section of the LSI-OR system, 43 components referring to the background and characteristics of an

offender are considered. The 43 items are classified in 8 categories as follows: criminal history (8), education/employment (9), family/marital (4), leisure/recreation (2), companions (4), procriminal attitude/orientation (4), substance abuse (8), and antisocial pattern (4). Finally, the total score, which is the sum of the scores of each item, is organized in five risk levels as given in Table 1.1.

Table 1.1: Five risk levels in LSI-OR.

	Very low (VL)	**Low (L)**	**Medium (M)**	**High (H)**	**Very high (VH)**
Score range	$0-4$	$5-10$	$11-19$	$20-29$	$30-43$

According to the ministry of community safety and correctional services policy, an LSI-OR assessment is required for all adult inmates who received sentences of at least one month in custody [11]. The result of the LSI-OR assessment (or LSI-OR score) is employed for prediction of recidivism regardless of race or gender and is used for obtaining information relating to offender needs. In fact, the LSI-OR score is the only factor considered in the LSI-OR assessment.

There are a number of factors (or features), such as the 43 LSI-OR items, age, race, and gender, that could be considered in the LSI-OR prediction method. Hence, a new version of LSI-OR or a new classification method other than LSI-OR would be needed to support multiple features. Moreover, it may be desirable that the new method produce the same results as LSI-OR when the number of features is cut down to one.[1] It is proven in Chapter 2 that the Naïve Bayes classifier (NB classifier) as a machine learning classifier satisfies above conditions. Therefore, this study is aimed at presenting and testing the NB classifier as an alternative to the LSI-OR, employing various continuous and discrete input features. Additionally, depending on input features, various classification models based on the NB classifier with distinct accuracies were built. The outline of this work is as follows.

Chapter 2 includes a summary of NB classifier and performance measures. In Section 2.1, the Bayes rule is illustrated. In Section 2.2, the theory of NB classifier is briefly discussed, and then it is shown that under certain circumstances the LSI-OR method and NB classifier give the same results. In Section 2.3, two methods of testing are introduced that are applied to evaluate the accuracy and performance of the NB and LSI-OR models and the effectiveness of input features. In Section 2.4, the definition of expectation value for both continuous and discrete random variables is given. In Section 2.5, the confusion matrix and other performance measures such as sensitivity, specificity, recall, precision, F-measure, and receiver operating characteristic curve are briefly presented. In Section 2.6, a summary of input features used in the NB classifier and extracted from the dataset is given.

Chapter 3 contains the results and discussion. This chapter is divided into three sections. In the introductory section of this chapter, basic information about the dataset is provided, including the determination of 8-sub patterns assigned to each offender. In Section 3.1, density distribution functions of the dataset and the dataset grouped into females and males are given. In Section 3.2, the results of employing the NB classifier are presented, and in some cases the results of NB classifier are compared with the results of LSI-OR method.

Chapter 4 provides conclusions and future possible research directions. Appendices A, B,

[1] This is the main reason why the the Naïve Bayes classifier is prefered over other classification techniques such as neural networks, support vector machines, and so on.

C, D, E, and F contain tables, figures, supporting materials, and other ancillary information.

Chapter 2

Naïve Bayes Classifier

There are many classification techniques such as, support vector machines (SVM), k-nearest neighbor (kNN), NB classifier, and so on that can be applied to classify a dataset into distinct groups. In this work, the NB classifier as a classification algorithm is employed to analyze the dataset. The NB classifier is fast, simple, and easy to implement, and is used in many different fields [12, 13].

It is shown later in this chapter that the LSI-OR method is a NB classifier only accepting, the total score classified in five risk levels (VL, L, M, H, and VH)[1] as an input feature. Indeed, there are a number of possible features that can affect the accuracy of the LSI-OR method. However, according to the LSI-OR assessment as a prediction tool, a prediction is performed based on employing one discrete feature. Hence, it is possible to support multiple features in the LSI-OR if one uses an extended version of the LSI-OR or the NB classifier.

In the first part of this chapter, the main focus is on the concept of the NB classifier and performance measures. In the second part of this chapter, a summary of input features used in the NB classifier is provided.

2.1 The Bayes rule

According to the Bayes rule, the joint probability of two events A and B ($P(A, B)$) is given by

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A), \tag{2.1}$$

where $P(A|B)$ and $P(B|A)$ are the conditional probability of A given B and the conditional probability of B given A, respectively. A more familiar form of equation (2.1) is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{2.2}$$

where $P(A|B)$ is the posterior probability, $P(B|A)$ is the likelihood, and $P(A)$ is the prior probability.

[1] It is possible to use the total scores $0, 1, \ldots,$ and 43 without considering five risk levels.

2.2 Naïve Bayes classifier

NB classifier is a probabilistic classifier employing Bayes rule (2.1) and assuming that for a given class attribute (c), the features ($\mathbf{F}$=$(f_1, \ldots, f_k)$) are independent of each other, or in other words, they have independent distributions[2]. Mathematically,

$$P(\mathbf{F}|c) = P(f_1|c) \ldots P(f_k|c). \tag{2.3}$$

Let $\mathcal{D} = [\mathcal{D}_1, \mathcal{D}_2, \ldots \mathcal{D}_k]$ be a vector of features and $C = \{c_1, c_2, \ldots, c_m\}$ be the target class with m discrete values ($m > 1$). The probability of certain features being in a certain class say $C = \{c_i\}$ is given by $P(C = c_i|\mathcal{D})$, and it can be calculated from the (2.1) as follows

$$P(c_i|\mathcal{D}) = \frac{P(\mathcal{D}|c_i)}{P(\mathcal{D})} P(c_i), \tag{2.4}$$

where $P(\mathcal{D}|c_i) = P(\mathcal{D}_1|c_i)P(\mathcal{D}_2|c_i) \ldots P(\mathcal{D}_k|c_i) = \prod_{l=1}^{k} P(\mathcal{D}_l|c_i)$. $P(\mathcal{D})$ is a scaling factor and can be ignored. Applying the maximum a posteriori estimation principle (MAP)[15] to (2.4) gives[3]

$$\hat{\mathcal{Y}} = \underset{i \in \{1,\ldots,m\}}{\operatorname{argmax}} \prod_{l=1}^{k} P(\mathcal{D}_l|c_i) P(c_i), \tag{2.5}$$

where $\hat{\mathcal{Y}} \in C = \{c_1, \ldots c_m\}$.

In a continuous situation[4], when one or more features follow a distribution function like $\mathcal{Q}(\mu, \sigma^2, X)$[5], the conditional probability becomes

$$P(\mathcal{D}_0|C = \{c\}) = \mathcal{Q}(\mu_c, \sigma_c^2, X = x_0), \tag{2.6}$$

where $x_0 \in \mathcal{D}_0$ and (μ_c, σ_c^2) are respectively the average and variance of X values, associated with class c.

It can be shown that in a simple case where just one feature and one class exist, the NB classifier and LSI-OR method produce the same results. To show this, let us assume that the number of records in the dataset is $\mathcal{N}_0$, and the dataset has two fields. The feature field ($\mathcal{D}$) is a vector with the length of $\mathcal{N}_0$, where each element is one of five distinct values given by $\{d_1$=VL, d_2=L, d_3=M, d_4=H, d_5=VH$\}$[6]. The class field (C) is a vector with $\mathcal{N}_0$ components

[2]The correlation among LSI-OR features is discussed in [14]. We note that LSI is an assessment tool, as opposed to a survey, so there is no need to have redundancy between questions.

[3]e.g., argmax $y(x) = 4 - (x-1)^2$ is $x = 1$, because $y_{\max}|_{x=1} = 4$.

[4]Some features are random variables (e.g., X).

[5]This form is used for a normal distribution function. For a beta distribution function with two shape parameters a and b, the functional form becomes $\mathcal{Q}(a, b, X)$.

[6]Five risk levels.

and has two unique values of $\{c_1, c_2\}$. Starting from equation (2.4) and ignoring $P(\mathcal{D})$ gives

$$\begin{aligned} P(c_i|d_j) &\sim P(d_j|c_i).P(c_i), \\ &= \frac{N_{d_j c_i}}{N_{c_i}}.\frac{N_{c_i}}{\mathcal{N}_0}, \\ &= \frac{N_{d_j c_i}}{N_{d_j}}.\frac{N_{d_j}}{\mathcal{N}_0}, \end{aligned} \tag{2.7}$$

where $N_{d_j c_i}$ is the total number of d_j being in class c_i , N_{c_i} is the total number of c_i , and N_{d_j} is the total number of d_j.
Now, the problem can be simplified by setting $i = 1, 2$ and considering a special case $j = 1$. Hence, (2.7) becomes

$$P(c_1|d_1) = \underbrace{\frac{N_{d_1 c_1}}{N_{d_1}}}_{p}.\frac{N_{d_1}}{\mathcal{N}_0}, \tag{2.8}$$

$$P(c_2|d_1) = \underbrace{\frac{N_{d_1 c_2}}{N_{d_1}}}_{q}.\frac{N_{d_1}}{\mathcal{N}_0}, \tag{2.9}$$

where $p + q = 1$. According to the NB classifier, the d_1 is in class c_1[7] if

$$P(c_1|d_1) > P(c_2|d_1) \rightarrow p > q \rightarrow p > \frac{1}{2}. \tag{2.10}$$

Recall that in the LSI-OR assessment, offenders are classified in five risk levels {VL, L, M, H, VH}, and an offender in a particular risk level most likely recidivates if the recidivism rate defined by[8]

$$p = \frac{\text{the number of offenders in particular risk level who recidivate}}{\text{total number of offenders in particular risk level}} = \frac{N_{d_1 c_1}}{N_{d_1}}, \tag{2.11}$$

is higher than 50% ($p > \frac{1}{2}$). Thus, by comparing (2.10) and (2.11), one can see that both NB and LSI-OR give same results.

2.3 Model evaluation

In general, the performance and accuracy of a model can be checked in various ways. In the following section, two methods of simple split and k-fold validation employed in this work as evaluation methods are briefly reviewed.

[7] A similar argument can be used for class c_2.
[8] A general definition for the recidivism rate is given in Section 2.6.

2.3.1 Simple split

In this case, the dataset is split into two groups of training and test sets. Usually, around 66% of dataset is randomly chosen and is devoted for training the classifier, and the rest 34% is considered for testing purposes [16]. The error rate of model is calculated by comparing the estimated value $\hat{\mathcal{Y}}_{\text{predict}}$ and the test value $\mathcal{Y}_{\text{test}}$. Hence, the error rate becomes

$$\begin{aligned} \mathcal{E}(\text{error rate}) &= \frac{\text{the number of misclassifications}}{\text{the total number of test cases}}, \\ &= \frac{\sum_{i=1}^{N} \mathcal{I}(\hat{\mathcal{Y}}_{\text{predict}}, \mathcal{Y}_{\text{test}})}{N}, \end{aligned} \tag{2.12}$$

where $\mathcal{I}(\hat{\mathcal{Y}}_{\text{predict}}, \mathcal{Y}_{\text{test}}) = 1$ if $\hat{\mathcal{Y}}_{\text{predict}} \neq \mathcal{Y}_{\text{test}}$ otherwise $\mathcal{I}(\hat{\mathcal{Y}}_{\text{predict}}, \mathcal{Y}_{\text{test}}) = 0$.

2.3.2 k-fold validation

In this method [17], the original dataset is grouped into k equal sized subgroups, of which $k-1$ groups are used for training and 1 group is used for testing. The k-fold process is repeated k times, and then, the average of error rates is calculated employing the equation (2.12) as follows,

$$\bar{\mathcal{E}}(\text{average error}) = \frac{\sum_{i=1}^{k} \mathcal{E}_i}{k}. \tag{2.13}$$

The accuracy of a model ($\mathcal{A}$) can be obtained by subtracting the error rate from 1 or $\mathcal{A} = 1 - \mathcal{E}$.

2.4 Expectation value

The expectation value of a discrete random variable $X = (x_1, \ldots, x_M)$ with M distinct values is defined by

$$\mathbf{E}(X) = \sum_{i=1}^{M} x_i f_i, \tag{2.14}$$

where n_i is the number of x_i, $f_i = \dfrac{n_i}{N}$ is the frequency of x_i, and $N = \sum_{k=1}^{M} n_k$. For a continuous random variable X, the expectation value becomes

$$\mathbf{E}(X) = \int_{-\infty}^{+\infty} x\mathcal{Q}(x)dx, \tag{2.15}$$

where $\mathcal{Q}(x)$ is the probability density function.

2.5 Confusion matrix

The performance of a classifier can be visualized and examined by a matrix known as the confusion matrix [18, 17]. The rows and columns of the confusion matrix present observed

and predicted class labels, respectively. The confusion matrix for a class feature with two unique values of 1 and 0 is

		Predicted	
		1	**0**
Observed	**1**	True Positive	False Negative
	0	False Positive	True Negative

where **1** $\equiv$ Positive, **0** $\equiv$ Negative, true positive (**TP**) is the number of testing instances being in class 1 that are correctly predicted to be in class 1, false negative (**FN**) is the number of testing instances being in class 1 that are incorrectly predicted to be in class 0, true negative (**TN**) is the number of testing instances being in class 0 that are correctly predicted to be in class 0, and false positive (**FP**) is the number of testing instances being in class 0 that are incorrectly predicted to be in class 1. Using the confusion matrix, one can easily obtain the accuracy and some other performance measures such as, sensitivity, specificity, recall, precision, F-measure, and receiver operating characteristic curve in terms of the entities of the confusion matrix as follows.

2.5.1 Accuracy ($\mathcal{A}$)

The accuracy of a model in terms of confusion matrix elements is defined by the following formula

$$\mathcal{A} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{2.16}$$

where the numerator is the total number of correctly predicted 0s and 1s, and the denominator is the total number of instances.

2.5.2 Sensitivity and specificity

The sensitivity or true positive rate (**TPR**) (or recall) and specificity or true negative rate (**TNR**) are defined by

$$\textbf{TPR} = \frac{TP}{TP + FN}, \ \textbf{TNR} = \frac{TN}{TN + FP}. \tag{2.17}$$

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

$$\textbf{FPR} = 1 - \text{specificity}, \ \textbf{FNR} = 1 - \text{sensitivity}. \tag{2.18}$$

2.5.3 Precision

Precision (**P**) or the percentage of correctly predicated 1s is defined by the following formula,

$$\textbf{P} = \frac{TP}{TP + FP} \ . \tag{2.19}$$

2.5.4 F-Measure

The F-Measure or in other words the harmonic mean of precision (**P**) and recall (**R**)[19] is given by

$$\mathbf{F} = (1 + \beta^2)\ \frac{\mathbf{P} \cdot \mathbf{R}}{\beta^2 \mathbf{P} + \mathbf{R}}, \tag{2.20}$$

where $\beta \in (0, +\infty]$ is applied to control the relative weight designated to **R** and **P**. In this study it was assumed that $\beta = 1$.

2.5.5 Receiver operating characteristics curve

A receiver operating characteristics curve (ROC curve) is a two-dimensional curve with two coordinates (FPR, TPR). It has TPR (sensitivity) as its vertical axis and FPR (1-specificity) as its horizontal axis [20, 21]. Both axes range from 0 to 1. Some special points in a typical ROC curve are as follows. The origin (FPR=0, TPR=0) $\Rightarrow$ TP = 0 and FP = 0 : all negative instances (0s) are predicted, the top left corner (FPR=0, TPR=1) $\Rightarrow$ FP = 0 and FN = 0 : all instances are correctly predicted, the bottom right corner (FPR=1, TPR=0) $\Rightarrow$ TP = 0 and TN = 0 : all instances are wrongly predicted, and (FPR=1, TPR=1) $\Rightarrow$ FN = 0 and TN = 0 : all positives instances (1s) are predicted.

The ROC curve is used to compare the performance of various classifiers by considering the area under the curve (AUC) or more precisely the area between the curve and the FPR-axis. For non-intersecting curves, the AUC is proportional to the performance of classifiers. For instance, in Figure 2.1 the ROC curves are plotted for four classifiers. The best and worst performances belong to D and A, respectively ($AUC_D > AUC_C > AUC_B > AUC_A$).

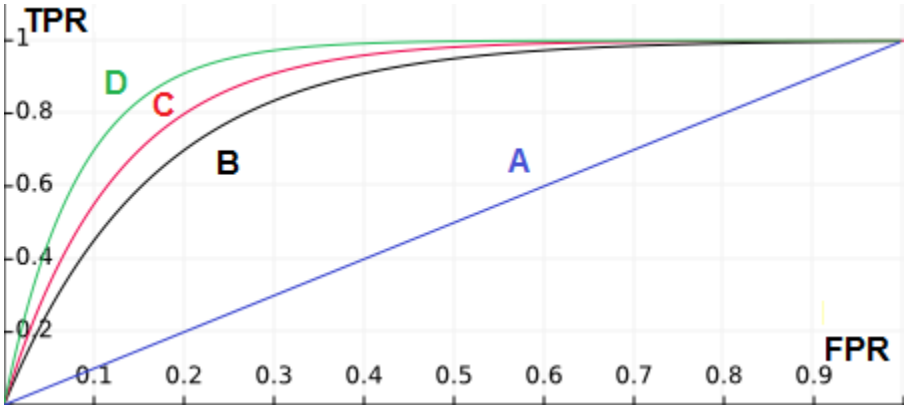

Figure 2.1: ROC curves for various classifiers. A higher AUC results a higher performance.

In this study, the confusion matrix and other performance measures are used as comparison tools to evaluate the performance of NB and LSI-OR classifiers over the dataset.

2.6 Dataset

It was mentioned in Chapter 1 that the LSI-OR assessment is required for all adult inmates, and it also was stated that the result of LSI-OR assessment serves as a prediction tool to obtain information relating to offender needs. In the prediction part, the data analysis is mainly focused on 48 extracted fields from the dataset and utilized in this work as follows (the number of fields is given in parentheses).

- **LSI-OR items** (43)
  43 fields, the LSI-OR items $(A_1, A_2, \dots A_{43})$, are used as input features for NB classifier. Indeed, the NB classifier performs a prediction based on considering the effect of each LSI-OR item.
- **Total LSI-OR score** (1)
  As an input feature of the NB classifier, a total LSI-OR score (or simply total score) is assigned to each offender given by equation (2.21), and the total score is calculated based on the score of 43 items in the LSI-OR assessment. Recalling that the total score is the only feature utilized by the LSI-OR to predict recidivism.

$$\text{LSI-OR}_{\text{ score}} = \sum_{q=1}^{43} A_q \in \Big\{0, 1, \dots, 43\Big\}. \tag{2.21}$$

The minimum and maximum values of total score are 0 and 43, respectively, and each element $A_1, A_2, \dots A_{43}$ in the dataset is scored as either 0 or 1. Basically, the total score is an integer; however, in some cases, such as obtaining the distribution function of dataset, the total score is treated as a continuous variable changing from 0 to 43.
It was stated in Chapter 1 that for each question in the LSI-OR items there are two possibilities of either the answer is Yes (1) or No (0). Hence, there are 2^{43} ways to answer all 43 questions. In general, the number of ways of answering N questions $Q = \{Q_1, Q_2, \dots Q_N\}$ having a certain total score k is given by

$$\mathcal{N}_k = \binom{N}{k} = \frac{N!}{(N-k)!k!}\ , \tag{2.22}$$

where $\sum_{i=1}^{N} Q_i = k$ and $\sum_{k=1}^{N} \mathcal{N}_k = 2^N$. For instance, for $N = 43$ the $\mathcal{N}_k$ is shown in Figure 2.2.

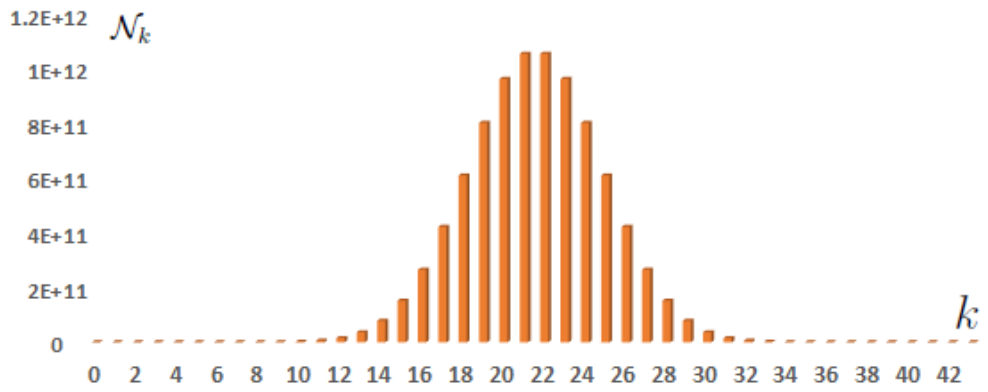

Figure 2.2: The number of ways of answering 43 questions ($\mathcal{N}_k$) with a total score of $k \in \{0, 1, 2, \ldots, 43\}$. The maximum values occur at $k = 21$ and $k = 22$.

Mathematically speaking, the total score can be treated either as a whole pattern of 0s or 1s or as 8-sub patterns of 0s or 1s (Figure 2.3).[9] In each case, a certain number of participants answers the LSI-OR items in a unique way or, in other words, use a unique pattern (UP). The remaining offenders just repeat the UPs.

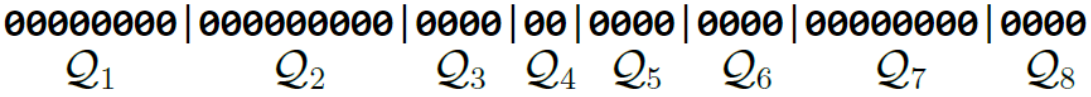

Figure 2.3: The 8-sub patterns of total LSI-OR score.

In addition, there is a possibility of grouping the dataset in five risk levels as well rather than considering each total score. In this case, the total scores satisfy the following condition

$$\text{LSI-OR}_{\text{ score}} = \sum_{q=1}^{43} A_q \in \Big\{\text{VL, L, M, H, VH}\Big\}. \tag{2.23}$$

Furthermore, it is possible to consider the total score as a continuous random variable varying between 0 and 43. Thus, two distribution functions, one for each class attribute, are needed to obtain the probability of getting a certain total score.

- **Status of recidivism (1 field)**
  Recidivism refers to a person's regression into criminal activity, after the person has served time for a previous crime. The status of recidivism is recorded for each offender in the dataset with a given values of 0 (not occurred or $R = 0$) or 1 (occurred or $R = 1$). This is the only class feature used in the NB classifier to classify inmates. The rate of recidivism (or simply recidivism rate) for a population of offenders is defined by[10]

$$\text{recidivism rate} = \frac{\text{the number of offenders who recidivate } (R = 1)}{\text{population size}}. \tag{2.24}$$

[9] These are not the only ways that one can define the pattern of 0s or 1s.
[10] Sometimes the percentage of recidivism rate is used.

Equation (2.24) can be considered the average rate of recidivism. Thus, according to the central limit theorem (CLT), the recidivism rate can be approximated by a normal distribution function given by $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$, where σ is the known standard deviation of population, n is the sample size, and μ is the unknown mean of population. The confidence interval for the mean of population (μ_n) in terms of sample mean ($\bar{x}_n$) and sample size (n) can simply be obtained from the following equation,

$$\mu_n = \bar{x_n} \pm z^* \frac{\sigma}{\sqrt{n}}, \tag{2.25}$$

where z^* is the critical value for a certain confidence level.

- **Gender** (1)
  One field in the dataset is dedicated to the gender of the offenders, stored as either F(female) or M(male). This feature is applied to the NB model to show whether the prediction of recidivism is affected by gender or not.
- **Offense severity** (1)
  For each offender, the number between 1 (unknown) to 26 (homicide) is designated that shows the category of crime committed by that particular case (Appendix A). This is one of features that is never used in the LSI-OR, but it is applied to the NB classifier.
- **Race** (1)
  A field with two nominal values of 0 and 1 corresponding to No or Yes is assigned to the race of offenders as either they are aboriginal Canadian or not. This feature is applied to the NB model to investigate the effect of offender's race on the prediction of recidivism among offenders.

The data are analyzed using the data mining open-source or commercial software packages, R-3.2.2 [22], Weka-3.6.13 [23], TableCurve2Dv5.01 [24], Tanagra-1.4.50 [25], and Microsoft Excel 2013.

Chapter 3

Results

The results of data analysis including the estimation of the density distribution function and the NB classifier are discussed in this chapter. In this work, the original dataset under study is provided by the Ontario Provincial Police[1], and it contains 90781 criminal records relating to 686 types of offences, categorized into 29 groups (Appendix A on page 42). After removing repeated records, the data are reduced to 72725 records, one record for each offender. Table 3.1 contains some basic information about the dataset.

Table 3.1: A basic information about the dataset.

Total	UP[1](%)	Race[2]	Male (%)	Female (%)	Min BthYr[3]	Max BthYr[4]
72725	60491(83.18)	726(1.00)	60086(82.62)	12639(17.38)	1926	1993

(1) unique patterns used by offenders in LSI-OR assessment (43 items),
(2) aboriginal Canadian, (3) minimum birth year , (4) maximum birth year

The offenders can be classified in five risk levels according to the range of total scores, provided in Table 1.1. In Figure 3.1, the percentage of offenders at each risk level for the current dataset is given.

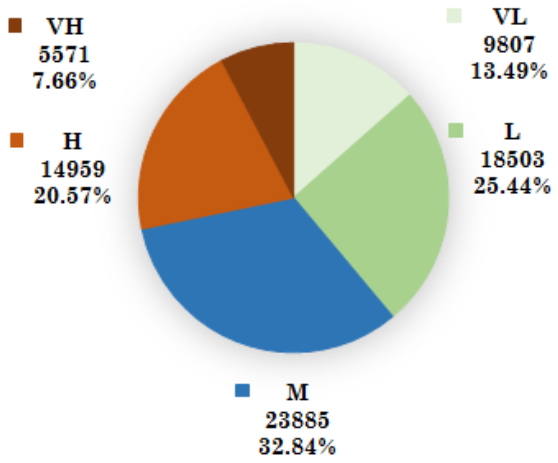

Figure 3.1: Percentage of offenders at various risk levels.

[1]Permission to use this dataset for reasearch purpose was granted by the Forensic Centre of the University of Saskatchewan.

The percentage of recidivism rate at each risk level is presented in Figure 3.2 (red) and that of the whole population is shown in Figure 3.3 (green).

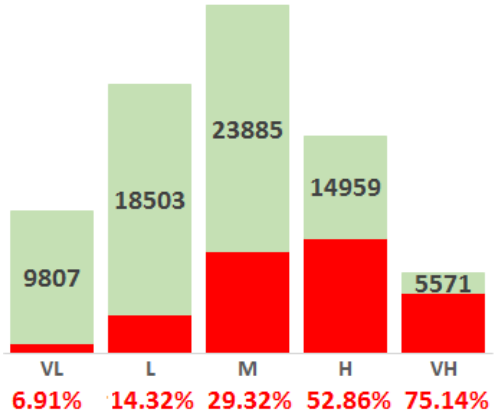

Figure 3.2: Recidivism rate (%) at various risk levels (red).

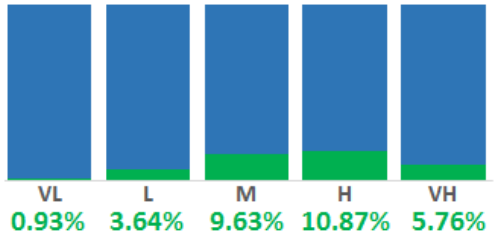

Figure 3.3: Recidivism rate (%) of each risk level in total population or 72725 cases (green). The total recidivism rate is 30.83% (0.93%+3.64%+9.63%+10.87%+5.76%).

It can be observed from Figure 3.2 that moving from VL (6.91%) to VH (75.14%) shows an increase in the rate of recidivism occurrence in each risk level. The same trend is found in the dataset grouped by the total scores (Figure 3.4).

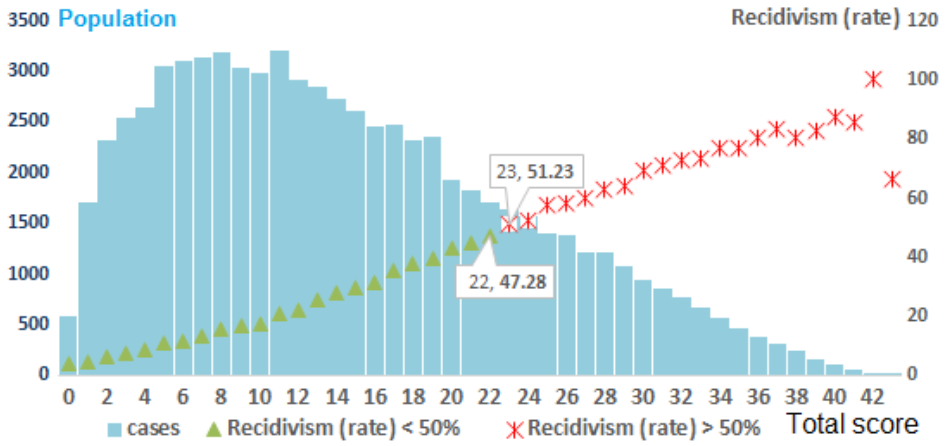

Figure 3.4: The recidivism rate (%) at each total score.

In Table 3.1 on page 13, the second column contains the percentage of offenders who used the UPs in the LSI-OR assessment. In the same manner, the number of UPs was individually extracted from the dataset for each 8-sub patterns (Figure 3.5). As given in Figure 3.5, the greatest number of UPs are observed in $\mathcal{Q}_1$ (criminal history), $\mathcal{Q}_2$ (education/employment), and $\mathcal{Q}_7$ (substance abuse). These results suggest that in a NB-based model based on the 8-sub patterns as input features, only three features that create diversity among offenders are most effective.

One can also consider total scores and obtain the number of UPs for each single total score. The number of UPs for each score is displayed in Figure 3.6. According to the Figure 3.6, $\mathcal{Q}_1$, $\mathcal{Q}_2$, and $\mathcal{Q}_7$ receive the highest UP values between total scores of 4 and 38.

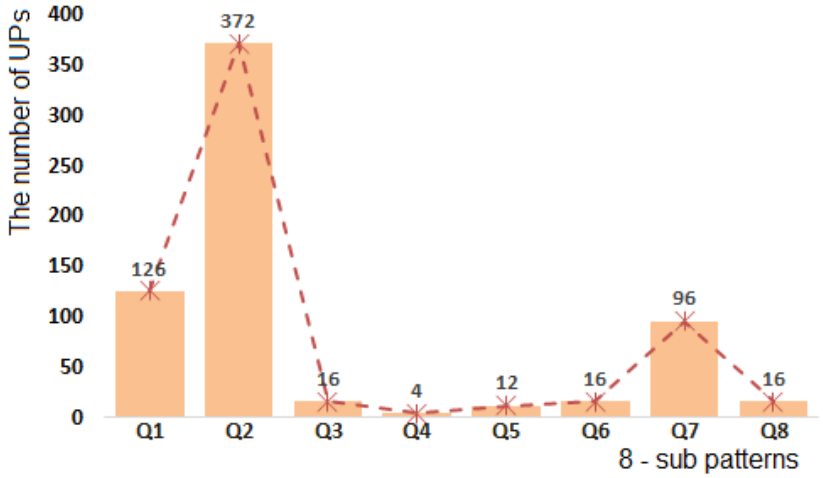

Figure 3.5: The number of UPs for the dataset.

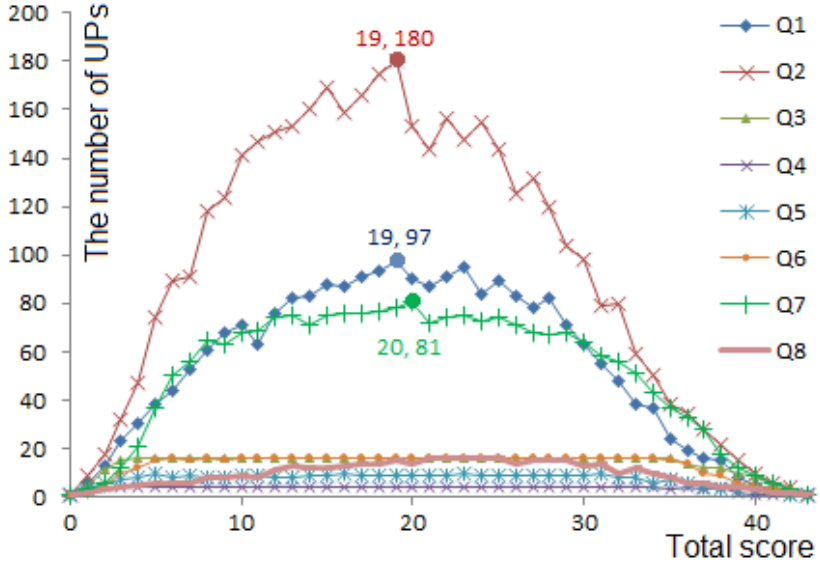

Figure 3.6: The number of UPs for each single total score.

It was claimed in Chapter 2 that a normal distribution function could describe the recidivism rate if the standard deviation of population was known. Therefore, for the purpose of obtaining the distribution function directly from the dataset (illustrated in Figure 3.7), $100,000$ samples with the sample size of 600 are randomly chosen. Figure 3.7 displays the density function that is centered around 0.3083 with the standard deviation of $\sigma = 0.0188$.

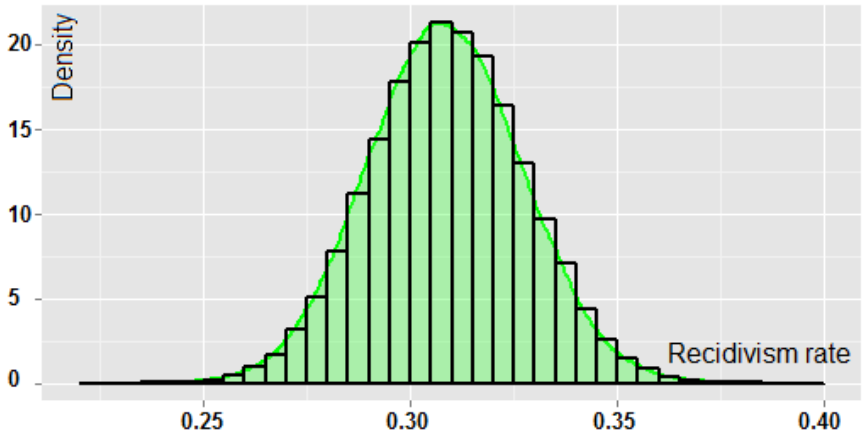

Figure 3.7: The density function of recidivism rate obtained from $100,000$ randomly selected samples with a sample size of 600.

For a 95% confidence interval ($z^* = 1.96$), the equation (2.25) given in Chapter 1 becomes

$$\mu_n|_{\text{Recidivism rate}} = \bar{x}_n \pm 1.96\frac{0.4618}{\sqrt{n}}, \tag{3.1}$$

where 0.4618 is the standard deviation of population, n is the sample size, and $\bar{x}_n$ is the sample mean. For instance, for a randomly chosen sample of $n = 600$ and $\bar{x}_{600} = 0.315$, the μ_{600} is

$$\mu_{600}|_{\text{Recidivism rate}} = 0.315 \pm 1.96\frac{0.4618}{\sqrt{600}}$$

or

$$0.278 \leq \mu_{600}|_{\text{Recidivism rate}} \leq 0.352. \tag{3.2}$$

According to equation (3.2), the mean of recidivism rate falls between 0.278 and 0.352. This can be verified by plugging $\mu_{600} \approx 0.308$ (Figure 3.7) into (3.2).

3.1 Probability density function estimation by different techniques

The probability density function (PDF) estimation of the dataset is performed employing the total scores and using two distinct methods of curve fitting. In fact, one can categorize the total scores in different ways, considering the gender of offenders and the status of recidivism. The results show that in all cases studied in this section, the data as a subset of the original dataset followed a beta distribution function with two unknown shape parameters of a and b.

The parameter estimation is carried out in two steps. In the first step, the moment matching estimation (MME) method (Appendix B on page 43) is applied to dataset to obtain the shape parameters (a and b), and in the second step, the parameters are re-estimated, employing a nonlinear least-square fitting technique (Levenberg–Marquardt algorithm).
The main purpose of using the LM method is to make a comparison between the estimated values and to determine the goodness of fit. It is shown later in this section that in all cases, the good performance of MME method in estimating parameters is guaranteed by the result of the goodness-of-fit test and the closeness of estimated parameters by both fitting methods. Finally, the goodness-of-fit test is determined by calculating the index of correlation or r^2 and the adjusted value of r^2 that accounts for the degrees of freedom Adj (r^2). Generally, r^2 and Adj r^2 close to one indicate a good fit. The procedure of PDF estimation in each population is illustrated in Figure 3.8.

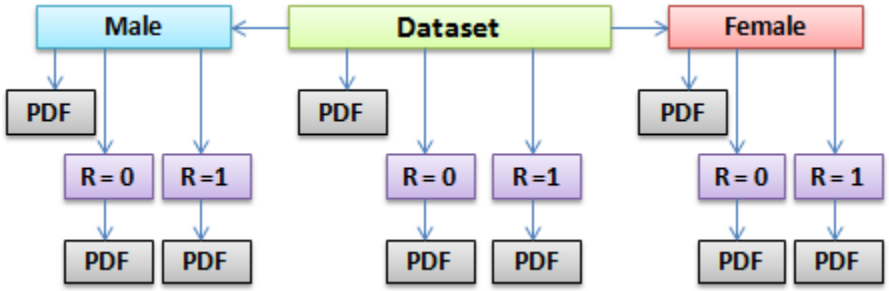

Figure 3.8: The PDF estimation of various population emerged from the original dataset. R stands for the recidivism status.

- **Dataset**
  In this case, the only considered factor for the study is the total score associated with each offender. Other factors such as gender and recidivism status are ignored. According to the observed results, given in Table 3.2, the data points are scattered widely around the mean value due to a considerably high standard deviation, and the distribution function of the dataset is positively skewed and follows a platykurtic distribution pattern because of negative kurtosis (Appendix B on page 43). Using the skewness and kurtosis of population given in Table 3.2 and plotting a Cullen–Frey graph help us to estimate the type of a PDF for the dataset.

Table 3.2: Statistical parameters, extracted from the dataset.

Number of cases	Min	Max	Std	Mean	Mode	Median	Skewness	Kurtosis
72725	0	43	9.12	14.58	7.48	13	0.55	−0.47

According to this graph (Figure 3.9), the best option is a beta distribution function with two estimated values of $a_{\mathrm{MME}} = 1.35$ and $b_{\mathrm{MME}} = 2.63$ (Figure 3.10 and Appendix C on page 45). In this case, the MME method is applied to determine a_{MME} and b_{MME}. The LM technique implemented in the TableCurve software [24] is employed to re-estimate the parameters a and b. In this case, both r^2 and Adj r^2 are around 0.995 that indicates a good model fit (Appendix D, Listing D.1 on page 47).

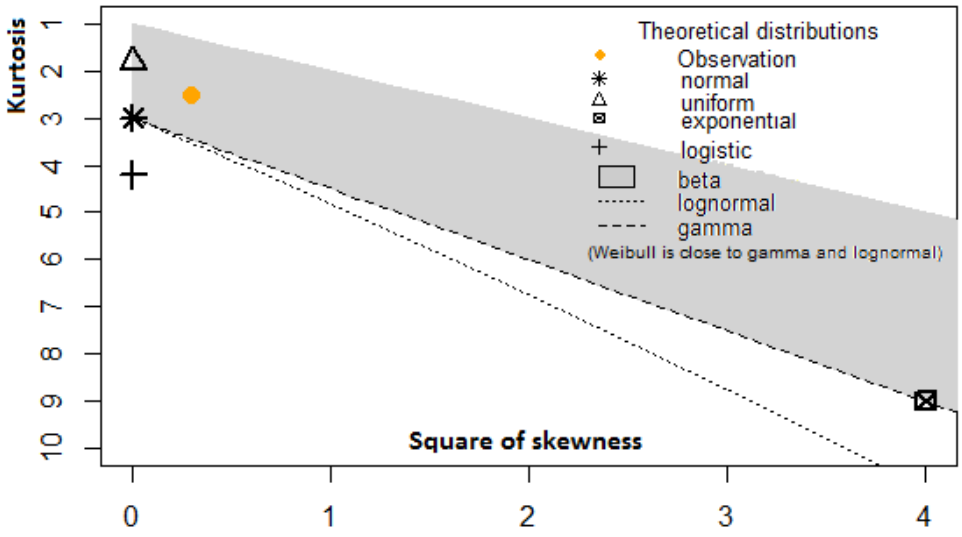

Figure 3.9: Cullen–Frey graph (kurtosis versus square of skewness). The dataset (observation), indicated in orange, is located in the grey area covered by the beta distribution function.

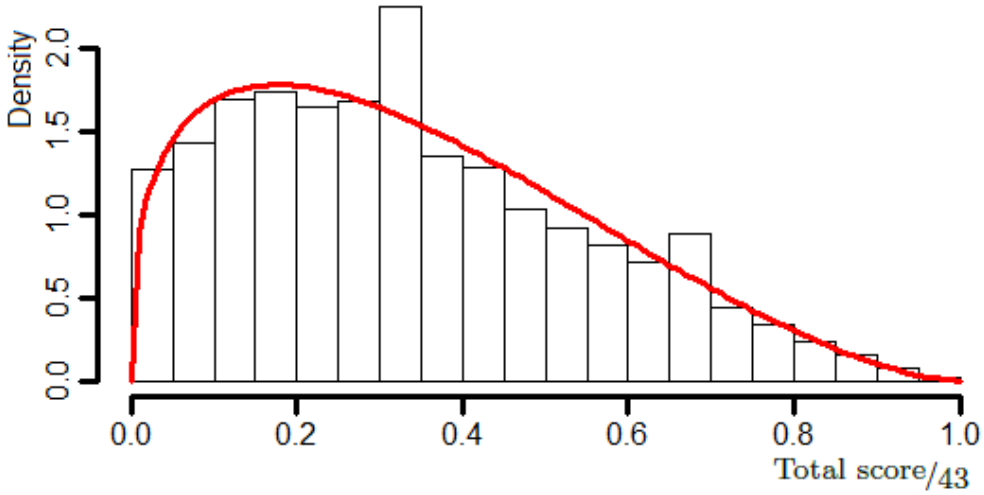

Figure 3.10: The histogram of dataset and the beta distribution function with two parameters of $a_{\mathrm{MME}} = 1.35$ and $b_{\mathrm{MME}} = 2.63$ (red).

It can be observed from the MME results ($a_{\mathrm{MME}} = 1.35$ and $b_{\mathrm{MME}} = 2.63$) and the LM values ($a_{\mathrm{LM}} = 1.43$ and $b_{\mathrm{LM}} = 2.82$) that a small difference in both values of a and b is detected (Figure 3.11) or in other words, both methods estimated almost the same values. This can be verified by the following calculation,

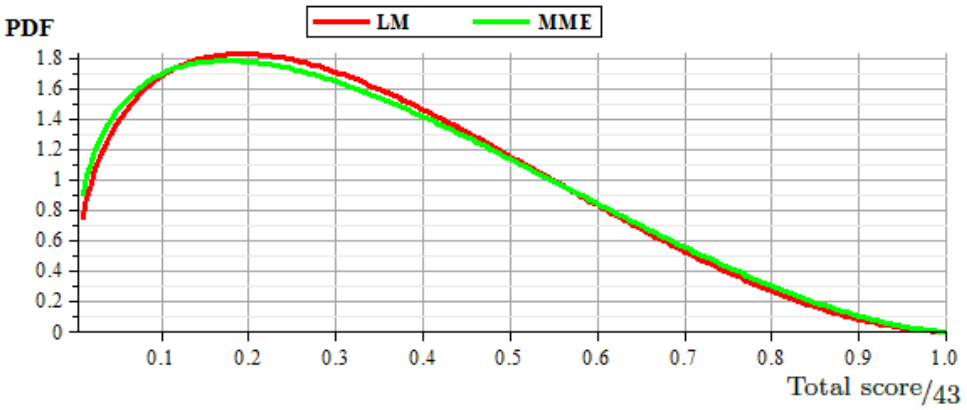

Figure 3.11: A small variation is observed between MME and LM methods (Δa =0.08 and Δb =0.19).

$$\delta_{\max} = \int_0^1 |\hat{\mathcal{B}}_{\mathrm{MME}} - \hat{\mathcal{B}}_{\mathrm{LM}}|\, dx \approx 0.035, \tag{3.3}$$

where $\delta_{\max}$ gives the maximum probability difference between PDFs (the grey area shown in Figure 3.12).

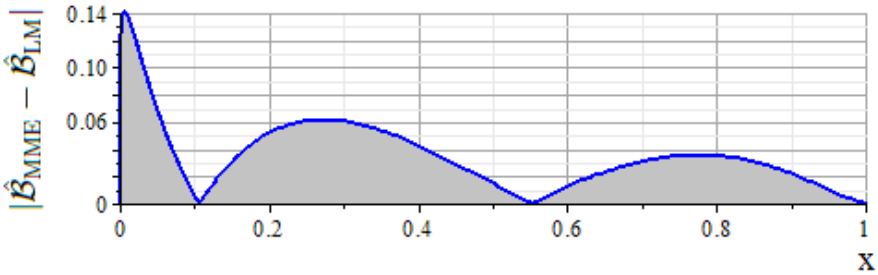

Figure 3.12: The difference values between PDFs.

The $\delta_{\max}$ also defines an upper bound to $\delta E(X)$ as follows,

$$\begin{aligned} \delta E(X) &= \int_0^1 X|\hat{\mathcal{B}}_{\mathrm{MME}} - \hat{\mathcal{B}}_{\mathrm{LM}}|\, dx, \\ &< \int_0^1 |\hat{\mathcal{B}}_{\mathrm{MME}} - \hat{\mathcal{B}}_{\mathrm{LM}}|\, dx, \\ &< \delta_{\max}\ , \end{aligned}$$

where $0 \leq X \leq 1$.

- **Dataset with two subpopulation groups**
  The population is divided into two subclasses of either relapsing into criminal activity has occurred ($R = 1$) or not ($R = 0$). The results show that each case of $R = 1$ and $R = 0$ follows different PDF. Similar to the previous section, obtaining the Cullen–Frey graph for each case of $R = 1$ and $R = 0$ (Figures 3.13 and 3.14) suggests two beta functions.

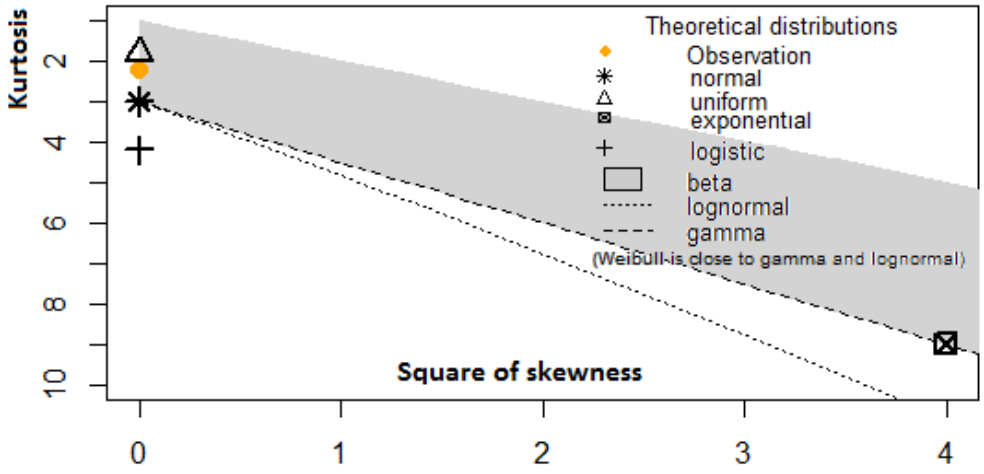

Figure 3.13: Cullen–Frey graph (kurtosis versus square of skewness) for which the recidivism status is $R = 1$. The dataset (observation), indicated in orange, is located in the grey area covered by the beta distribution function.

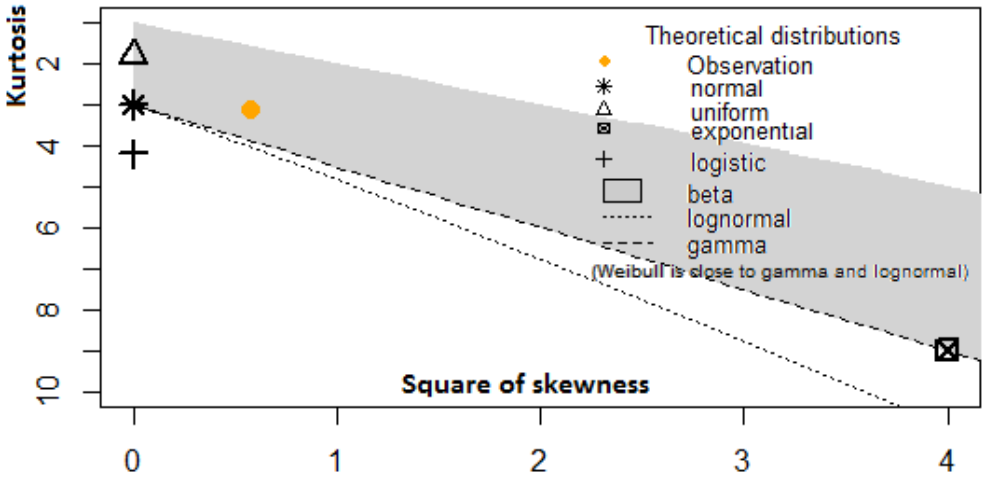

Figure 3.14: Cullen–Frey graph (kurtosis versus square of skewness) for which the recidivism status is $R = 0$. The dataset (observation), indicated in orange, is located in the grey area covered by the beta distribution function.

The statistical parameters including mean, mode, median, the number of cases (counts), standard deviation, kurtosis, and skewness for each case are given in Table 3.3. According

to this table, one expects a symmetric platykurtic distribution for the case of $R = 1$ by virtue of zero skewness and negative kurtosis. The graph of PDF with two estimated shape parameters of $a_{MME} = 2.25$ and $b_{MME} = 2.44$ is plotted in Figure 3.15. In the case of $R = 0$, both skewness and kurtosis have positive values. The PDF function is asymmetric and leptokurtic and is positively skewed. The estimated shape parameters are $a_{MME} = 1.41$ and $b_{MME} = 3.70$, and the PDF receives its maximum value around 5.7 (Figure 3.16).

Table 3.3: Statistical parameters, extracted from the dataset for $R = 1$ and $R = 0$.

R	Count	Std	Mode	Median	Mean	Skewness	Kurtosis	Min	Max
1 (Yes)	22424 (33.83%)	9.00	18.54	21	20.64	0.00	−0.76	0	43
0 (No)	50301 (66.17%)	7.78	6.39	11	11.87	0.76	0.11	0	43

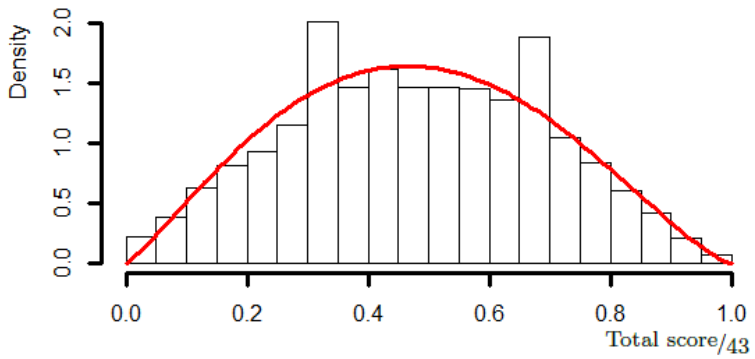

Figure 3.15: Histogram and PDF in the case of $R = 1$. The skewness of dataset is almost zero which guarantees the symmetry of left and right tails.

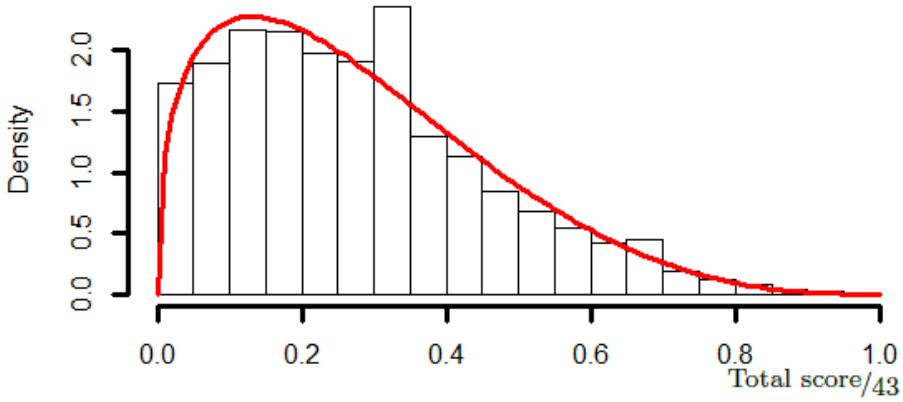

Figure 3.16: Histogram and PDF in the case of $R = 0$. The estimation of MME method is a asymmetric beta distribution function $\hat{\mathcal{B}}(1.41, 3.70, x)$. The skewness of dataset is positive, indicating that the distribution is positively skewed.

Similar to the previous section, the LM technique also is employed to dataset to determine the two parameters a and b in both cases of $R = 1$ and $R = 0$. In both cases, r^2 and Adj

r^2 are very close to 1 (Appendix D, Listing D.2 on page 48 and Listing D.3 on page 49). The details of LM and MME methods are given in Table 3.4.

Table 3.4: Parameter determination using MME and LM techniques.

R	a_{MME}	b_{MME}	a_{LM}	b_{LM}	Δ a	Δ b	δ_{max}
1 (Yes)	2.25	2.44	2.29	2.44	0.04	0	0.017
0 (No)	1.41	3.70	1.52	4.02	0.11	0.32	0.042

The PDFs from both methods are given in Figures 3.17 and 3.18. It can be observed from these figures that in practice, both curves for both cases $R = 1$ and $R = 0$ are indistinguishable.

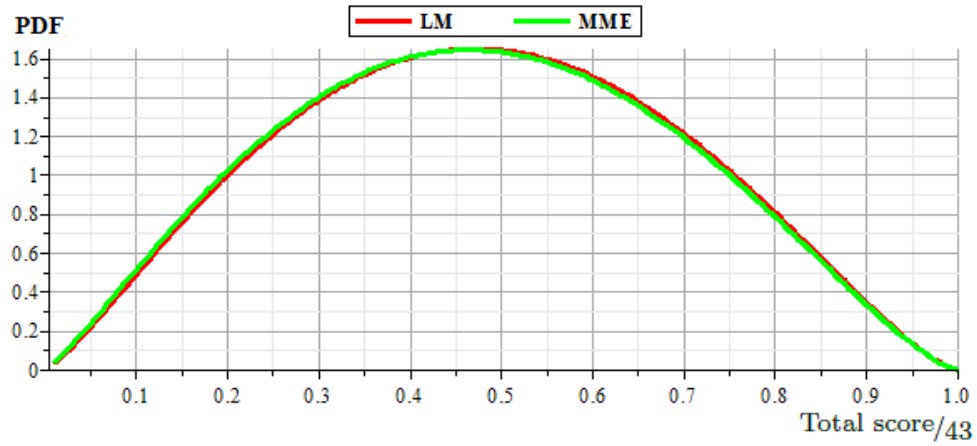

Figure 3.17: $R = 1$, a small variation is observed between MME and LM methods (Δa =0.04 and Δb =0).

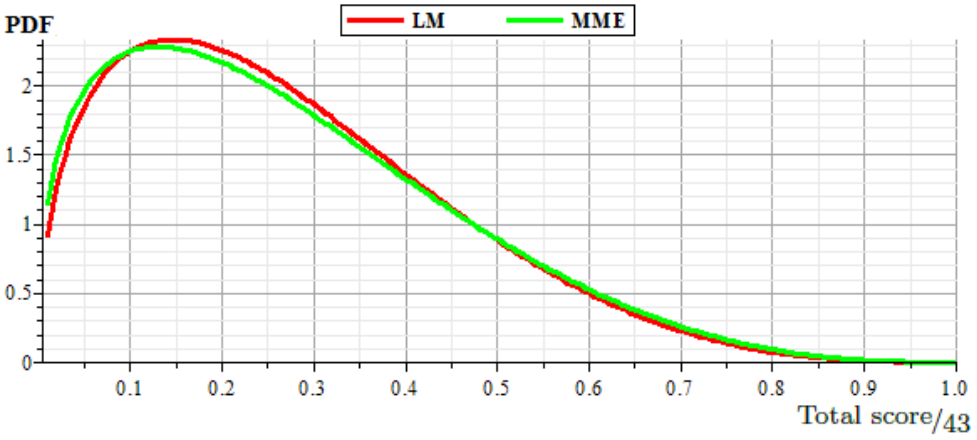

Figure 3.18: $R = 0$, a small variation is observed between MME and LM methods (Δa =0.11 and Δb =0.32).

- **Dataset with two subpopulation groups of females and males**
  In this case, based on the gender of the offenders, the dataset is partitioned into two groups, females and males, and the PDF in each case is derived independently. Furthermore, for each group, the PDF is obtained in two different situations, with and without considering the status of recidivism. Lastly, a single table is used to present the results of both groups. Based on the basic information about the dataset given in the Table 3.1 on page 13, about 17.38% of the total population are females (hence males 82.62%), and the recidivism rate among females is around 25.75% (31.90% among males). For the sake of comparison, the recidivism rate among the female and male populations at each total score is presented in Figure 3.19.

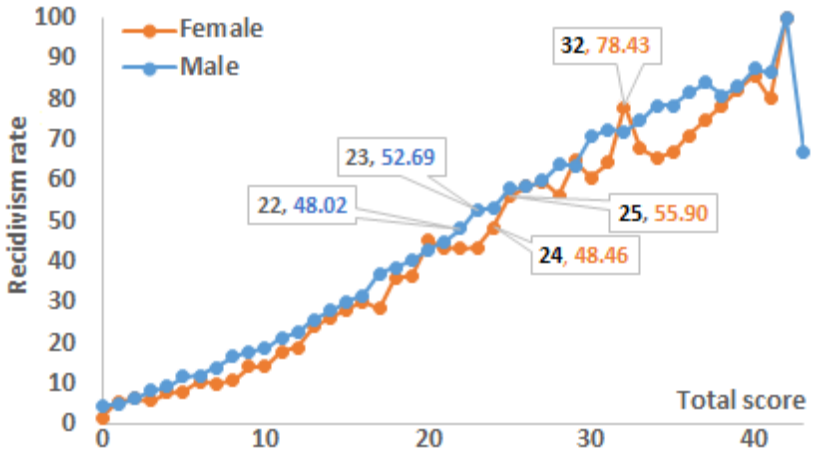

Figure 3.19: The recidivism rate among female and male population at each total score.

The same procedure used in previous sections for obtaining the distribution functions is employed here to obtain the PDFs. For the total population without considering recidivism, the histogram and the Cullen–Frey graph are given in Appendix E on page 50, Figures E.1 and E.2 (for males Appendix F on page 56, Figures F.1 and F.2). Using this information and the information regarding to the measures of central tendency, skewness, and kurtosis given in Table 3.5 suggests a beta distribution function with two parameters of a_{MME} and b_{MME} (Table 3.6).

Table 3.5: Statistical parameters, extracted from the dataset for females (F) and males (M).

	population	Std	Mode	Median	Mean	Skewness	Kurtosis	Min	Max
F	12639	8.57	7.72	12	13.47	0.61	−0.31	0	42
M	60086	9.22	7.31	13	14.81	0.53	−0.50	0	43

Table 3.6 also contains two more estimated parameters a_{LM} and b_{LM} that are derived by the LM method. The results of the goodness-of-fit test, including r^2 and Adj r^2 are given in Appendix E, Listing E.1 on page 50 and Appendix F, Listing F.1 on page 56.

Table 3.6: Parameter determination, using MME and LM techniques, females (F) and males (M).

	$\mathbf{a}_{\mathrm{MME}}$	$\mathbf{b}_{\mathrm{MME}}$	$\mathbf{a}_{\mathrm{LM}}$	$\mathbf{b}_{\mathrm{LM}}$	Δ **a**	Δ **b**	$\delta_{\max}$
F	1.36	2.87	1.46	3.11	0.10	0.24	0.041
M	1.35	2.57	1.43	2.75	0.08	0.18	0.034

The difference between two estimated parameters (Δa and Δb) are given in the last two columns of Table 3.6. For the females, the difference values are $\Delta a = 0.1$ and $\Delta b = 0.24$ (for males $\Delta a = 0.08$ and $\Delta b = 0.18$).
Considering the status of recidivism among females and males results two PDFs for each gender. According to the information provided in Table 3.7 and Appendix E on page 50, Figures E.4, E.5, E.7, and E.8, a symmetric density function for $R = 1$ and an asymmetric density function for $R = 0$ are suggested (for male Appendix F on page 56, Figures F.4, F.5, F.7, and F.8).

Table 3.7: Statistical parameters, extracted from the dataset for $R = 1$ and $R = 0$, females (F) and males (M).

	R	**Count**	**Std**	**Mode**	**Median**	**Mean**	**Skewness**	**Kurtosis**	**Min-Max**
F	1	3255 (25.75%)	8.55	19.05	19	19.62	0.02	-0.65	$0-42$
F	0	9384 (74.25%)	7.48	7.26	10	11.33	0.79	0.22	$0-41$
M	1	19169 (31.90%)	9.07	18.43	21	20.82	-0.01	-0.78	$0-43$
M	0	40917 (68.01%)	7.84	6.22	11	11.99	0.75	0.09	$0-43$

The results of LM method including r^2 and Adj r^2 are given in Appendix E on page 50, Listing E.2 and Listing E.3 (for males Appendix F on page 56, Listing F.2 and Listing F.3). More details about the estimated values and the shape of PDFs are provided in Table 3.8 and Appendix E on page 50, Figures E.6 and E.9 (for males Appendix F on page 56, Figures F.6 and F.9).

Table 3.8: Parameter determination, using MME and LM techniques for $R = 1$ and $R = 0$, and females (F) and males (M).

	R	$\mathbf{a}_{\mathrm{MME}}$	$\mathbf{b}_{\mathrm{MME}}$	$\mathbf{a}_{\mathrm{LM}}$	$\mathbf{b}_{\mathrm{LM}}$	Δ **a**	Δ **b**	$\delta_{\max}$
F	1	2.34	2.67	2.44	2.73	0.10	0.06	0.024
F	0	1.38	3.62	1.55	4.20	0.17	0.58	0.072
M	1	2.24	2.38	2.26	2.37	0.02	0.01	0.012
M	0	1.41	3.65	1.52	3.95	0.11	0.30	0.041

- **Summary**
  A comparison between distribution functions, with and without considering the status of recidivism, is made among different groups (Tables 3.9, 3.10 and 3.11). In all cases with similar conditions, almost the same distribution function with same parameters and same measures of central tendency (mean, mode, and median) are observed (Figures 3.20, 3.21, and 3.22). It can be observed from figures and tables that a small deviation between the original dataset and the population of females and males is discovered. For the females, the differences between estimated values are greater than those for the males.

Table 3.9: Comparison between various population without considering the status of recidivism. A small variation is discovered in each column.

Population	a_{MME}	b_{MME}	Mean	Mod	Median
Entire dataset	1.35	2.63	14.58	7.48	13
Female	1.36	2.87	13.47	7.72	12
Male	1.35	2.57	14.81	7.31	13

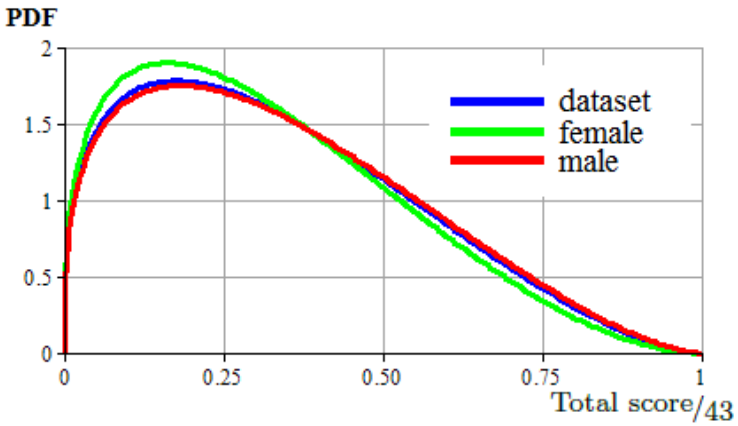

Figure 3.20: Comparison between density functions without considering the status of recidivism. The dataset, female, and male are shown in blue, green, and red, respectively.

Table 3.10: Comparison between various population with $R = 0$. A small variation is discovered in each column.

Subpopulation groups	a_{MME}	b_{MME}	Mean	Median	Mod
Entire dataset	1.41	3.70	11.87	11	6.39
Female	1.38	3.62	11.33	10	7.26
Male	1.41	3.65	11.99	11	6.22

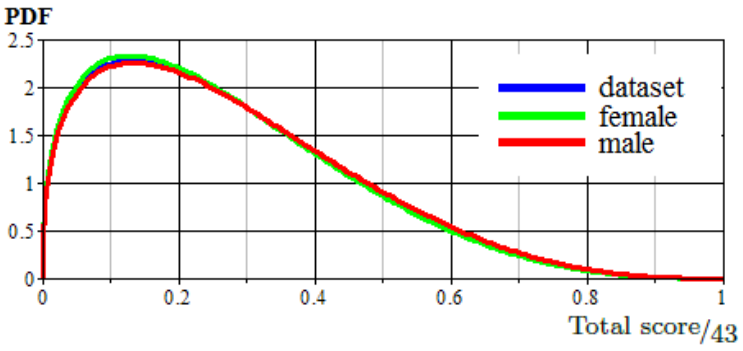

Figure 3.21: Comparison between density functions with $R = 0$. The dataset, female, and male are shown in blue, green, and red, respectively.

Table 3.11: Comparison between various population with $R = 1$. A small variation is discovered in each column.

Subpopulation groups	a_{MME}	b_{MME}	Mean	Median	Mod
Entire dataset	2.25	2.44	20.64	21	18.54
Female	2.34	2.67	19.62	19	19.05
Male	2.24	2.38	20.82	21	18.43

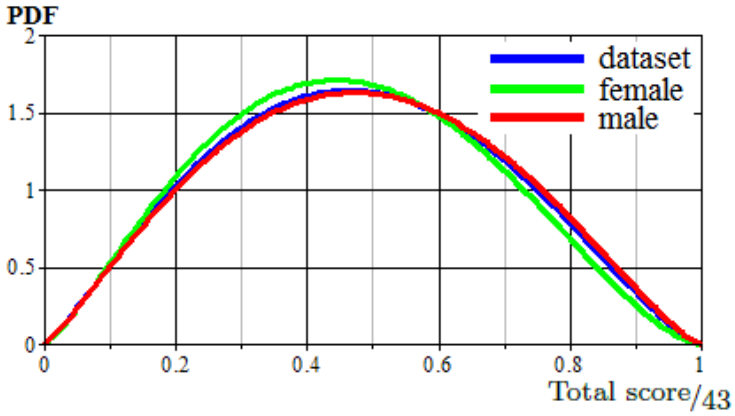

Figure 3.22: Comparison between density functions with $R = 1$. The dataset, female, and male are shown in blue, green, and red, respectively.

These results suggest that sufficiently large subset of the original dataset, with or without considering the status of recidivism, most likely follows a beta distribution function with almost the same estimated shape parameters derived in this section.

One can use the PDFs to estimate the mean, mode, median, variance, standard deviation, skewness, kurtosis, and the percentage of total scores that falls within a certain interval without having access to the dataset and to generate the total scores among offenders in a computer-based simulation program (for example, agent-based modeling). Moreover, the PDFs as a continuous feature can be applied to an NB classifier.

3.2 Naïve Bayes results

In this section, the results of NB classifier accepting a combination of 47 features as input and the status of recidivism ($R = 1$ or $R = 0$) as a class feature are discussed. Basically, the dataset contains nominal attributes such as total score, race, 43 LSI-OR items, genders, and the status of recidivism. However, in some cases the total score is treated as a continuous variable following a beta distribution function.

3.2.1 Individual total scores

It was mentioned previously in Chapter 2 that a classification of offender population, employing the LSI-OR method, was possible considering individual total scores. The NB classifier as a machine learning tool is able to do the same task, using the single total score as an input feature. Moreover, the NB classifier is able to consider the effect of each LSI-OR item on the final prediction by accepting 43 LSI-OR items as input features at each total score. The results of running the NB model, constructed by 43 LSI-OR items, are given in Figure 3.23. The horizontal axis in Figure 3.23 presents the total score calculated from equation (2.21) on page 10. The vertical axis on the left-hand side (light blue) is associated to the bar chart and shows the population size in each score displayed on the horizontal axis. The accuracies of LSI-OR and NB methods (the vertical axis on the right-hand side) are illustrated by two red and green curves, respectively. Two straight lines given in Figure 3.23 show the expectation value of accuracy in each case.

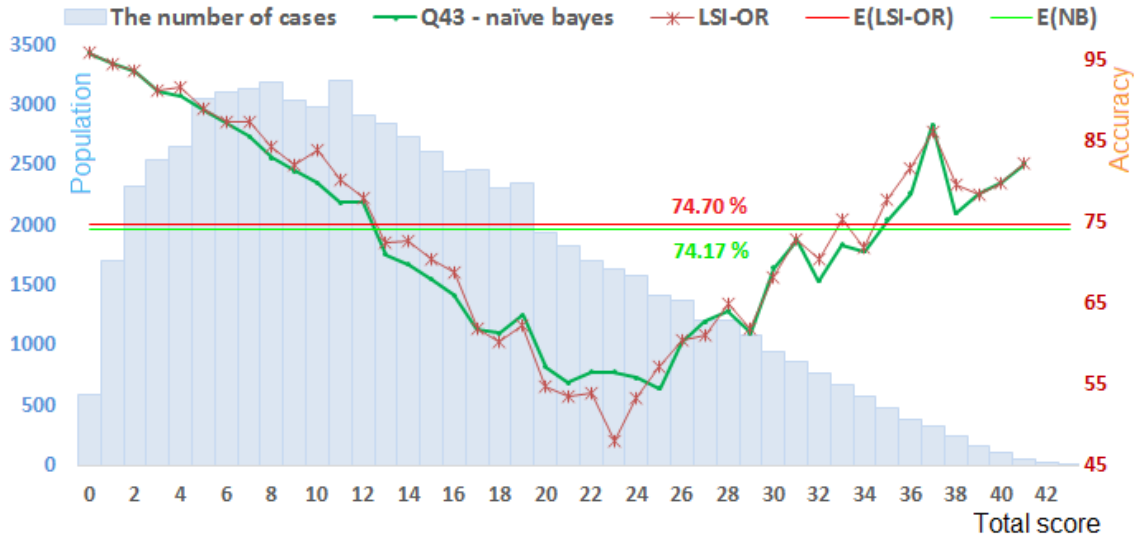

Figure 3.23: A comparison between LSI-OR and NB predictions. The accuracy in each case is calculated employing simple split 66% and 34% devoted for training and testing, respectively. E(NB) and E(LSI-OR) display the expectation values of each method.

The difference between accuracies (red and green curves in Figure 3.23) is significant starting from the lowest value of 0.09% to the highest value of 8.65%. The highest value occurs at

the total score of 23, where the recidivism rate is about 51.23%. (Figure 3.4). However, a slight variation about 0.53% is observed among the expectation value of accuracies for each method, derived from equation (2.14). It seems that taking into account 43 items for building a model employing the NB classifier considerably changes the accuracy of model at some scores. This might be explained by extracting and studying the pattern of 43 LSI-OR items associated with the offenders (Figure 3.24); e.g., five UPs belonging to different cases are shown in Figure 3.25. In Figure 3.24, the red graph shows the number of UPs at each total score. The plot does not present a symmetric behaviour over the total scores 0 to 43 as was predicted in Figure 2.2 given in Chapter 2. The blue plot shows the number of cases at each score, and the green plot is derived by dividing the number of UPs by the number of cases.

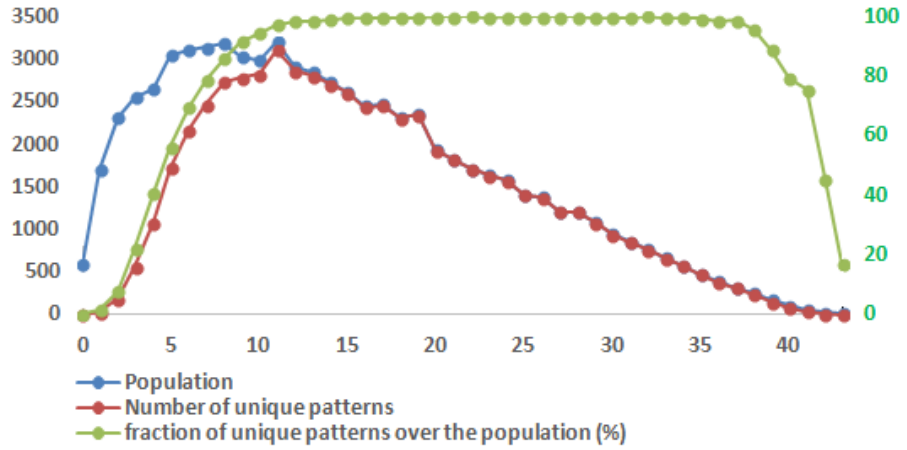

Figure 3.24: A comparison between the number of UPs related to cases (offenders) and the number of population at each individual total score.

As discussed before in Chapter 2, the equation (2.3) on page 5 was used to obtain the probability of each pattern; e.g., for the pattern $11\ldots10$, the probability becomes $P(11\ldots10|c) = P(1|c)P(1|c)\ldots P(1|c)P(0|c)$. In turn, the formula also was applied to build a prediction model. It can be seen that as the number of UPs increases (the red plot in Figure 3.24) the number of independent probabilities affecting all elements of the confusion matrix increases as well. Whereas having just a few UPs makes a few independent predictions that do not affect the confusion matrix much.

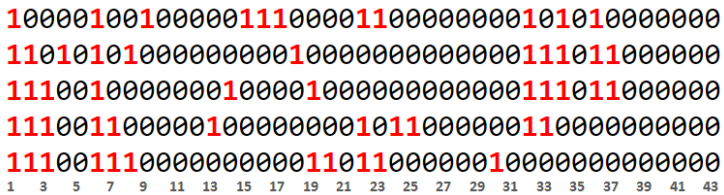

Figure 3.25: The first five UPs with a total score of 11. The population is 3198 and in total 3109 UPs are discovered.

To make a comparison between two methods using the confusion matrix, let us focus on the

total score of 11. The results are given in the following two Figures 3.26 and 3.27. It can be seen from Figure 3.26 that the confusion matrix in the LSI-OR method obtains zero values at class $R = 1$, and all instances in this class are misclassified causing TPR, FPR, precision, recall, and F-measure to become zero (yellow highlighted components).

```
=== Summary ===
Correctly Classified Instances         873               80.239  %
Incorrectly Classified Instances       215               19.761  %
Kappa statistic                          0
Mean absolute error                      0.3187
Root mean squared error                  0.3982
Total Number of Instances             1088
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1          0.802     1         0.89       0.5      0
                 0         0          0         0         0          0.5      1
Weighted Avg.    0.802     0.802      0.644     0.802     0.714      0.5
=== Confusion Matrix ===
   R
  0    1   <-- classified as
 873   0 | 0
 215   0 | 1 R
```

Figure 3.26: The LSI-OR method, the confusion matrix, and other performance measures obtained for the total score 11.

In general, the confusion matrix derived from the LSI-OR method always has a column with zero values either in class $R = 0$ or $R = 1$, resulting an ROC area value of 0.5.

```
=== Summary ===
Correctly Classified Instances         841               77.2978 %
Incorrectly Classified Instances       247               22.7022 %
Kappa statistic                          0.0665
Mean absolute error                      0.3168
Root mean squared error                  0.4085
Total Number of Instances             1088

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.935     0.884      0.811     0.935     0.869      0.602    0
                 0.116     0.065      0.305     0.116     0.168      0.602    1
Weighted Avg.    0.773     0.722      0.711     0.773     0.73       0.602
=== Confusion Matrix ===
   R
  0    1   <-- classified as
 816  57 | 0
 190  25 | 1 R
```

Figure 3.27: The NB classifier, the confusion matrix, and other performance measures obtained for the total score 11.

In comparison to the LSI-OR method, the NB classifier shows better performance (Figure 3.28), and the confusion matrix contains both classified (yellow highlighted components) and

mis-classified instances (green highlighted components) in most of the cases.

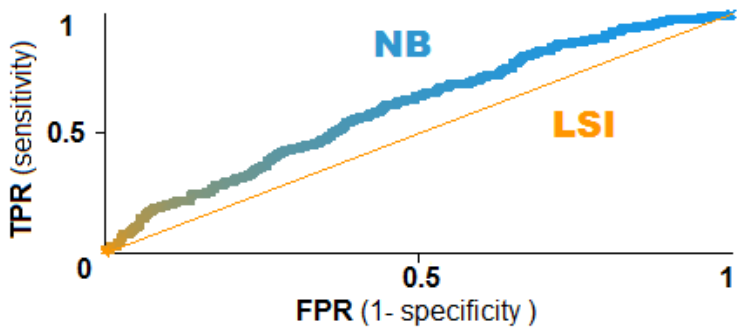

Figure 3.28: The NB classifier shows better performance than the LSI-OR method ($AUC_{NB} > AUC_{LSI}$).

The performance measures for the LSI-OR method are given in Figures 3.29 and 3.30. According to these figures, the performance measures take zero or non-zero values in two regions defined by the total scores as follows:
1) The total score is equal or less than 22 where the recidivism rate is lower than 50%. the second column of confusion matrix becomes zero causing the performance measures in $R = 1$ to gain zero values (except the ROC-area value).
2) The total score is greater than 22 where the rate of recidivism is higher than 50%. The first column of confusion matrix becomes zero resulting the performance measures in $R = 0$ to obtain zero values (except the ROC-area value).

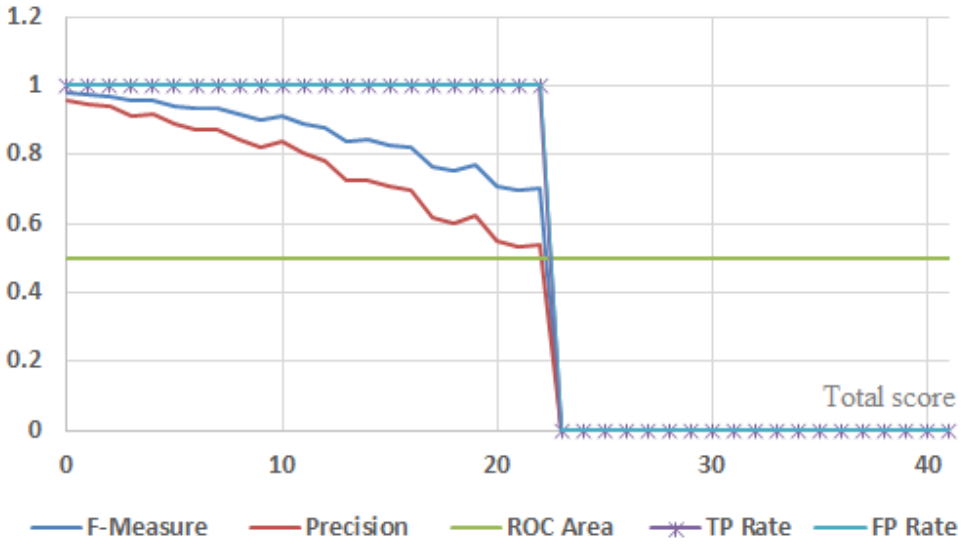

Figure 3.29: LSI-OR performance measures for $R = 0$. A jump in some performance measures from non-zero to zero values is observed moving from the total scores of 22 to 23.

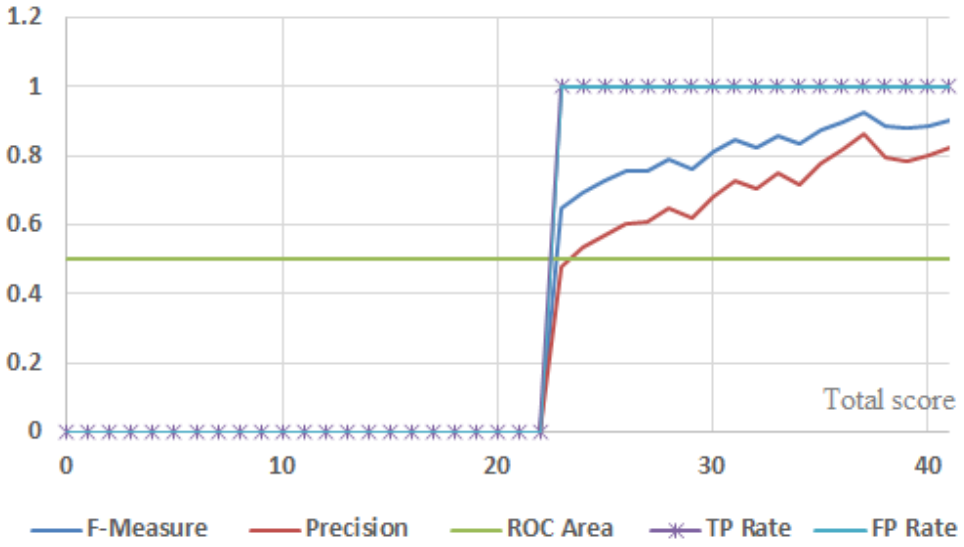

Figure 3.30: LSI-OR performance measures for $R = 1$. A jump in some performance measures from zero to non-zero values is observed moving from the total scores of 22 to 23.

Moreover, it can be noticed from the Figures 3.29 and 3.30 that the performance of the LSI-OR models built for each total score is always constant, and, in terms of the ROC area, receives the value of 0.5. Figures 3.31 and 3.32 present the performance measures of the NB classifier. The performance of NB models (ROC-area values) at each score varies around 0.6, presenting a better performance than the LSI-OR models. In comparison to the LSI-OR models, the performance measures in the NB models are gradually changing from zero to one or one to zero. During the gradual change, the testing instances are treated differently and no information is lost.

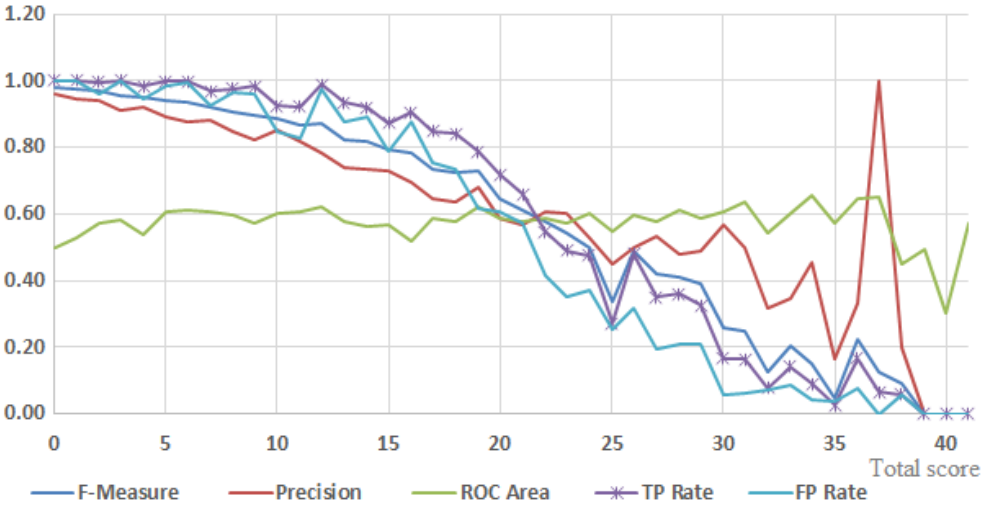

Figure 3.31: NB performance measures for $R = 0$. In some performance measures, a gradual transition from one to zero is observed.

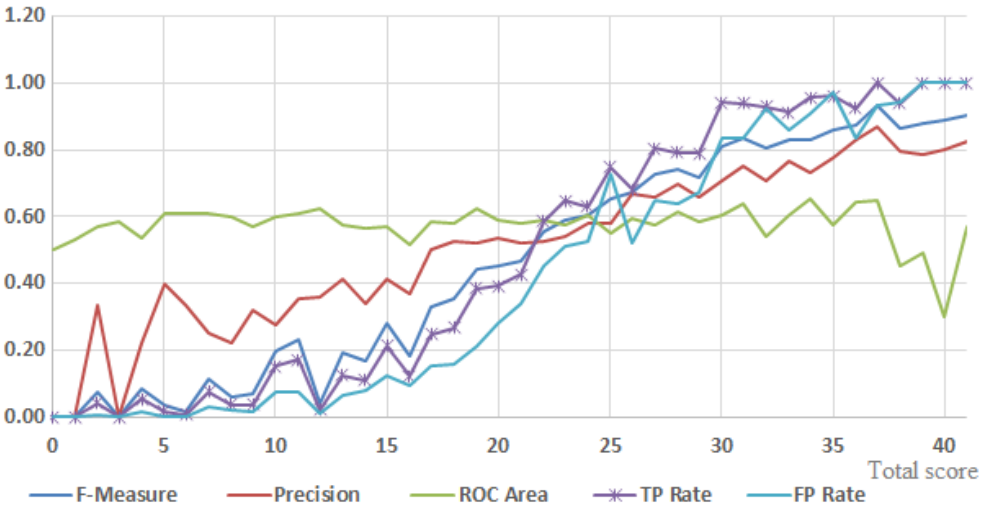

Figure 3.32: NB performance measures for $R = 1$. In some performance measures, a gradual transition from zero to one is observed.

3.2.2 Discrete risk levels

In this experiment, the total scores are grouped in five risk levels. Similar to Section 3.2.1, the NB classifier adopts 43 LSI-OR items. The models are evaluated with 34% of the dataset at each risk level (Figure 3.33). The results of running both the LSI-OR and NB methods on the dataset show that in spite of having lower performance, at most risk levels the LSI-OR method presented better accuracy than the NB classifier.

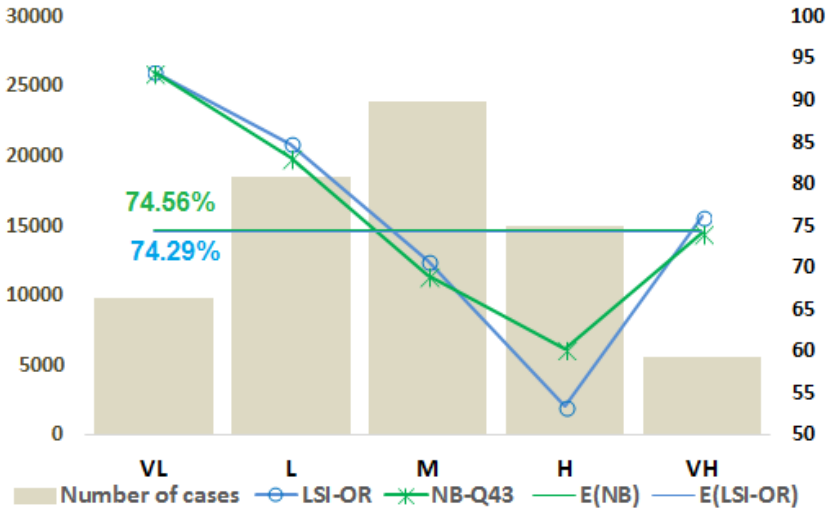

Figure 3.33: A comparison between LSI-OR and NB predictions. The accuracy at each risk level is calculated employing simple split 66% and 34% devoted for training and testing, respectively. E(NB) and E(LSI-OR) display the expectation values of each method.

The number of cases and UPs is given in Table 3.12 . The UPs achieve the highest value at the risk level of H where the accuracy of the LSI-OR method falls below the NB classifier.

Table 3.12: The number of cases, the number of UPs, and the percentage of UPs at each risk level.

Class levels	Number of cases	Number of UPs	(%)
VL	9807	1843	18.79
L	18503	14682	79.35
M	23885	23646	99.00
H	14959	14941	99.88
VH	5571	5391	96.77

More details about the H risk level, including the confusion matrices and accuracy measures, are provided in Figures 3.34, 3.35, and 3.36. The LSI-OR method, shown in Figure 3.34, does not perform any classification of instances at the class of $R = 0$ (yellow highlighted components in the Figure 3.34), and the performance never exceeds 0.5 (Figure 3.36). In fact, the LSI-OR method prediction depends on the recidivism status in the training dataset. For instance, at this particular risk level, the recidivism rate is 52.86% (Figure 3.2); thus, all testing instances are labeled as being in class $R = 1$ and hence have an accuracy value of $\frac{2716}{2371} \approx 0.53391$.

The NB classifier in this case shows better performance, higher accuracy, and slightly better expectation value (E(NB) in Figure 3.33). The confusion matrix includes both classified and misclassified instances in both classes of $R = 0$ and $R = 1$ (yellow and green highlighted components in Figure 3.35). Because the NB classifier considers the 43 LSI-OR items as input features, the accuracy and performance of NB models could be affected by the number of UPs at each risk level.

```
=== Summary ===
Correctly Classified Instances        2716               53.391  %
Incorrectly Classified Instances      2371               46.609  %
Kappa statistic                          0
Mean absolute error                      0.4982
Root mean squared error                  0.4989
Total Number of Instances             5087
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.5      0
                 1         1          0.534     1         0.696      0.5      1
Weighted Avg.    0.534     0.534      0.285     0.534     0.372      0.5
=== Confusion Matrix ===
   0   1   <-- classified as
 0 2371 | 0
 0 2716 | 1
```

Figure 3.34: A comparison between LSI-OR and NB predictions. No prediction is observed at the class of $R = 0$.

```
=== Summary ===
Correctly Classified Instances        3071               60.3696 %
Incorrectly Classified Instances      2016               39.6304 %
Kappa statistic                          0.1991
Mean absolute error                      0.4571
Root mean squared error                  0.4888
Total Number of Instances             5087

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.529     0.331      0.582     0.529      0.555      0.635    0
                 0.669     0.471      0.619     0.669      0.643      0.635    1
Weighted Avg.    0.604     0.406      0.602     0.604      0.602      0.635

=== Confusion Matrix ===
   a    b   <-- classified as
1255 1116 | a = 0
 900 1816 | b = 1
```

Figure 3.35: A comparison between LSI-OR and NB predictions. The confusion matrix contained both classified and misclassified instances.

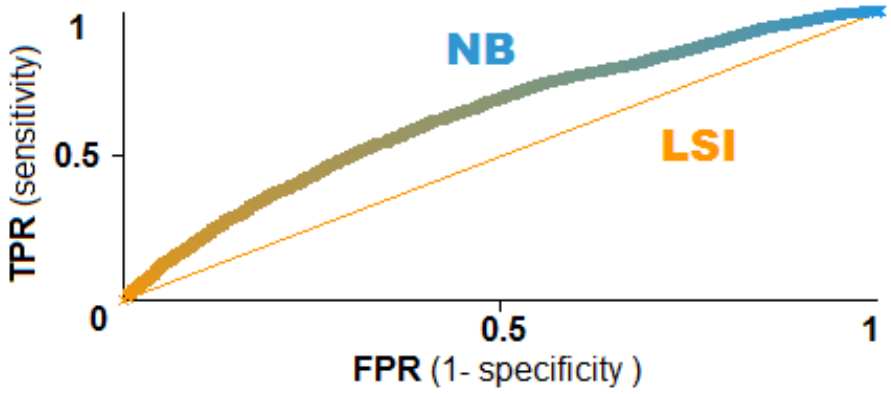

Figure 3.36: NB classifier (NB) shows better performance than LSI-OR method ($AUC_{NB} > AUC_{LSI}$).

3.2.3 Discrete offence severity indices

It was mentioned in the start of this chapter that there were about 686 types of offenses that could be categorized into 29 offence severity indices with various rates of recidivism (Figure 3.37). Thus, for each offence severity index, it is possible to build a model with an option of having either total scores or 43 LSI-OR items as input features. The results of model prediction are illustrated in Figure 3.38. As the figure shows, in comparison with the 43 lSI-OR items, the prediction of model built based on the total score fluctuates between 17.05% and 87.63% (for the 43 LSI-OR items is between 65.37% and 90.40%). In 15 cases (65.22%), the total scores display better accuracy than the 43 LSI-OR items. Furthermore, the expectation value of the total scores is higher than that of the 43 LSI-OR features.

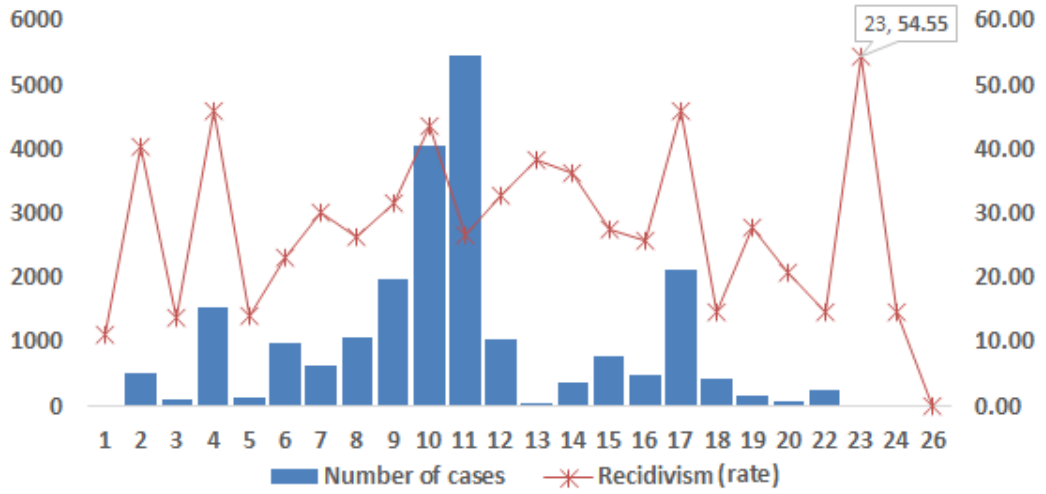

Figure 3.37: The rate of recidivism among 29 groups. The highest recidivism rate (54.55%) is observed at the offence severity index of 23 (or break & enter & related offences).

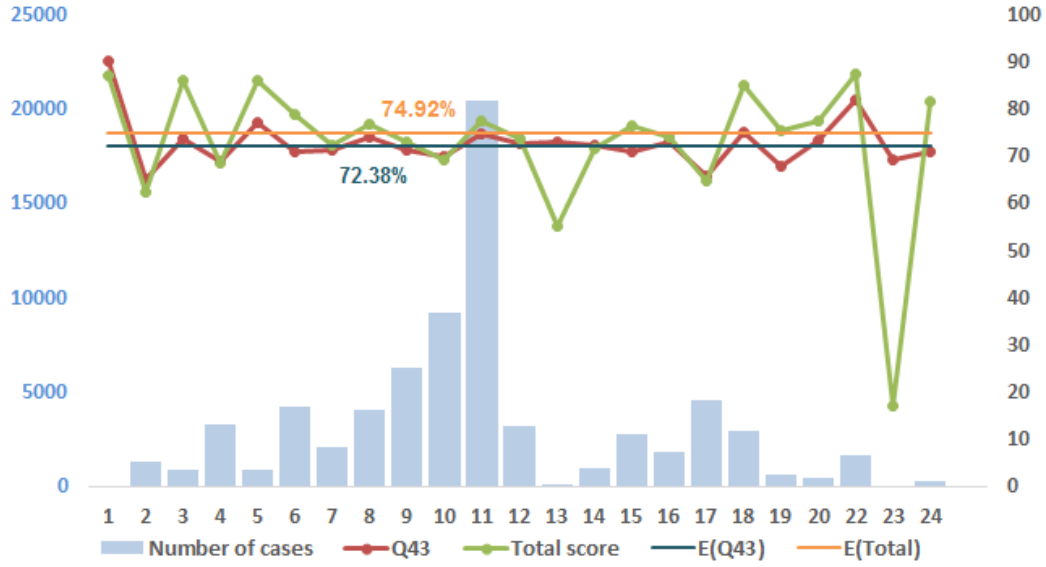

Figure 3.38: A comparison between LSI-OR and NB predictions employing simple split 66% and 10-fold validation testing techniques. The accuracy in each class level is calculated by averaging the results of two testing methods. The highest number of cases (20427) is occurred at the offence severity index of 11 (criminal code traffic offences).

3.2.4 Various discrete features

In order to see the effect of various features on the prediction of recidivism, a combination of features is applied to build a model each time. The accuracy of models are evaluated applying two techniques of testing shown in Table 3.13 and 3.14. The last column of the

table is dedicated to the average of two tests. The ROC-area value for each testing method is obtained for $R = 1$.
As given in Table 3.13, employing just the single feature of race results around 70% mean accuracy. Roughly speaking, this can be explained by the following argument. An NB model can be built by considering 99% of the dataset (Table 3.1 on page 13). In fact, the rest of the population (1%) only slightly affects the prediction and accuracy of the classifier and can be ignored. Hence, similar to the discussion given in sub-section 3.2.2 on page 32, one can claim that the prediction of the NB model depends on the recidivism status in the training set which, is around 30%, and resulting all instances in the testing set to be in class $R = 0$ (or giving about 70% mean accuracy).
The table reveals that the worst performance and accuracy are reported by the model constructed by the input feature of race, whereas, the best accuracy is seen in the input features of total score and race. The ROC areas have almost same values in all cases except in the input feature of race.

Table 3.13: The accuracy of models, built based on various discrete fetures.

Features	Simple split(66%)	ROC	10-Fold(%)	ROC	Mean(%)
Total score	75.01	0.77	74.95	0.77	74.98
Total score and gender	75.09	0.77	75.01	0.77	75.05
Total score and race	75.13	0.77	75.01	0.77	75.07
43-items	72.68	0.78	72.71	0.78	72.69
43-items and gender	72.68	0.78	72.73	0.78	72.70
43-items and race	72.72	0.78	72.79	0.78	72.76
Race	70.30	0.52	70.16	0.52	70.23

Recalling that the LSI-OR method is a NB classifier having a single feature as input feature, the first and last rows of Table 3.13 and the second and fourth rows of Table 3.14 can be considered as results of LSI-OR method prediction, as well. Categorizing data into two groups of females and males and applying two different features of the total score and the 43 LSI-OR items present interesting results (Table 3.14). The highest accuracy is seen among female offenders with the input feature of the total score. The ROC area value is not significantly changed in each case.

Table 3.14: The accuracy and performance of models, built based on the discrete total score and 43 LSI-OR items. The dataset is grouped by gender.

Features	Simple split(66%)	ROC	10-Fold(%)	ROC	Mean(%)
Male (43-items)	72.35	0.78	72.46	0.77	72.40
Male (total score)	74.52	0.77	74.49	0.77	74.50
Female (43-items)	74.42	0.78	73.75	0.77	74.09
Female (total score)	77.38	0.77	77.48	0.76	77.43

These results bring up two issues that are not considered in the LSI-OR. The first issue is that a combination of features may give results better than the features used individually, and the second issue is that dividing the dataset into smaller groups may improve the results in some subgroups.

3.2.5 Continuous features

It is possible to treat the total scores as a continuous variable, varying between 0 and 43 (first row in Table 3.15). This approach is also applicable to the dataset, classified by the gender of offenders (second and third rows in Table 3.15). In this section, the results of continuous cases are discussed employing two beta PDF functions for $R = 0$ and $R = 1$ obtained in Section 3.1.
However, there is always an overlapping between two PDF functions of $P(x \in \mathcal{D}|R = 0) = \hat{\mathcal{B}}(1.41, 3.7, x)$ and $P(x \in \mathcal{D}|R = 1) = \hat{\mathcal{B}}(2.25, 2.44, x)$ (Figure 3.39) that can affect the prediction and accuracy of the classifier, but according to the results given in Tables 3.13, 3.14, and 3.15, a small improvement in the classification of dataset based on gender is observed. No change is discovered in the entire dataset having the discrete total score as a feature input (first row of Tables 3.13 and 3.15).

Table 3.15: The accuracy of models, built based on the continuous total score. The ROC area values are calculated for $R = 1$.

Features	Simple split(66%)	ROC	10-Fold(%)	ROC	Mean(%)
Entire dataset (total score)	75.01	0.77	74.95	0.77	74.98
Female (total score)	77.47	0.78	77.41	0.77	77.44
Male (total score)	74.52	0.77	74.49	0.77	74.50

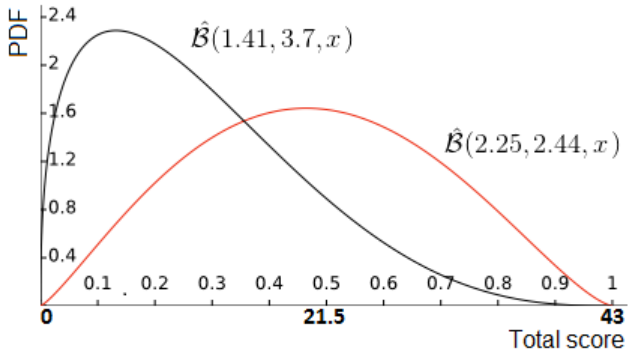

Figure 3.39: The beta functions $P(x \in \mathcal{D}|R = 0) = \hat{\mathcal{B}}(1.41, 3.7, x)$ and $P(x \in \mathcal{D}|R = 1) = \hat{\mathcal{B}}(2.25, 2.44, x)$ for the dataset. The functions are given in Section 3.1.

In practice, implementing a PDF function for building a NB model is much easier than using 44 total scores, and also the determination of $2 \cdot 44$ conditional probabilities[2] is not required. This can speed up the computation process.

[2]As needed in the discrete version.

Chapter 4

Conclusions and Future Work

The main goal of this thesis was to introduce the NB classifier and its advantages as a replacement for the LSI-OR that presents more predictive power, includes the LSI-OR results in a special case, and supports multiple discrete and continuous features. A series of experiments was designed and conducted to study the dataset provided by the Ontario Provincial Police and to evaluate the performance of the NB classifier.

In the first part of this work, a basic data analysis of the dataset, including obtaining 8-sub patterns for the total score and determining the PDF of recidivism rate, was performed. According to the analysis, the greatest number of UPs were observed in three categories: criminal history, education/employment, and substance abuse, and the results suggested that in a NB-based model based on the 8-sub patterns as input features, only three features were significantly effective. Taking advantage of the equality of the recidivism rate and the average rate of recidivism showed that the recidivism rate could be approximated by a normal distribution function. In addition, it was shown that when one has access to a small sample of offenders, the normal distribution defines an acceptable interval for the recidivism rate for the offenders.

As continuous features for the NB classifier, the PDF functions of the total score for various populations including the dataset, females, and males were obtained employing MME and LM techniques. The MME method guaranteed the existence of well-known PDFs, and the LM method provided the goodness-of-fit test. It was observed that the total scores in the dataset and in the female and male populations followed an asymmetric beta distribution function regardless of the status of recidivism. Taking into account the status of recidivism resulted in two beta distribution functions for the dataset and female and male populations.The PDFs could be used to generate the total scores among offenders in a computer-based simulation program (for example, agent-based modeling).

In the second part, an evaluation of the NB classifier was carried out. Many NB and LSI-OR models based on various features were built and tested. The total score, 43 LSI-OR items, and five risk levels as input features were separately applied to the NB classifier, and the results were compared in terms of performance measures. The effect of UPs on the confusion matrix was discussed as well. It was seen that unlike the LSI-OR, the NB classifier could easily consider the effect of each input feature that resulted both classification and misclassification in the confusion matrix and showed gradual changes in performance measures. In the models built by five risk levels and individual total scores, the NB classifier always showed better performance than the LSI-OR. On the other hand, there was no obvious trend in the accuracies predicted by both models to indicate the superiority of one model over the

other. A combination of various discrete features was applied to the NB classifier to obtain a set of features with the highest accuracies. It was realized that the minimum and maximum mean accuracies were achieved in the models built by a single input feature of race and a combination of two input features of total score and race, respectively. The result indicated that the total score was not the only feature that gave the highest accuracy. Dividing the dataset into two groups of females and males produced different results in terms of the total score and 43 LSI-OR items. The highest accuracy was seen among female offenders with the input feature of the total score.

In the last part, the continuous version of the total score as an input feature was applied to the NB classifier, and the results were compared with the discrete version. Considering the total score as a continuous variable in the dataset did not improve the performance and mean accuracy of model, and the results were similar to the results of the model built by the discrete version. However, a small improvement was observed among the female and male populations. Treating the total score as a continuous variable could improve the speed of computation, and could make the implementation of NB classifier easier.

In general, it was seen that the NB classifier was capable of accepting many discrete and continuous features, and the NB classifier was also capable of providing an effective tool for considering and studying the effect of multiple features on the output, whereas in the LSI-OR approach, only one discrete feature total LSI score is utilized.

In addition, the NB classifier showed better performance than the LSI-OR and provided a simple framework for studying the effect of individual features on the accuracies and performance measures.

The following list of suggestions is proposed for future possible research directions.

1) In a model built with 43 LSI-OR items as input features, it was suggested that the accuracy of a model could be affected by the number of UPs. Additional studies are needed to understand the effectiveness of UPs on the accuracy and other performance measures of the model.

2) A central issue in a classification model is to choose a subset of features with low intercorrelation but that are strongly correlated with the class feature. This suggests removing noisy, irrelevant, and unnecessary features from the feature set. A study is recommended to discover and select the best features out of 43 LSI-OR items or other features.

3) All statistical parameters related to a certain population such as mean, mode, median, and so on can be easily obtained from PDFs. In turn, the PDFs are defined as functions of shape parameters. Hence, the shape parameters of PDFs obtained in this work may change over time because of growing the size of the dataset. Therefore, it is necessary to monitor the variation of parameters in the future.

Bibliography

[1] "Statistics canada: Adult correctional statistics in canada, 2013/2014." http://www.statcan.gc.ca/pub/85-002-x/2015001/article/14163-eng.htm, 2013-14.

[2] S. M. Hogg, "The level of service inventory (ontario revision) scale validation for gender and ethnicity: addressing reliability and predictive validity," Master's thesis, University of Saskatchewan, 2011.

[3] C. T. Lowenkamp, B. Lovins, and E. J. Latessa, "Validating the level of service inventory revised and the level of service inventory: Screening version with a sample of probationers," *The Prison Journal*, vol. 89, no. 2, pp. 192–204, 2009.

[4] M. Ostermann and B. A. Herrschaft, "Validating the level of service inventory revised: A gendered perspective," *The Prison Journal*, vol. 93, no. 3, pp. 291–312, 2013.

[5] L. Girard and J. S. Wormith, "The predictive validity of the level of service inventory-ontario revision on general and violent recidivism among various offender groups," *Criminal Justice and Behavior*, vol. 31, no. 2, pp. 150–181, 2004.

[6] A. W. Flores, C. T. Lowenkamp, A. M. Holsinger, and E. J. Latessa, "Predicting outcome with the level of service inventory-revised: The importance of implementation integrity," *Journal of Criminal Justice*, vol. 34, pp. 523–529, 2006.

[7] B. Vose, F. T. Cullen, and P. Smith, "The empirical status of the level of service inventory," *Federal Probation*, vol. 72, no. 3, 2008.

[8] D. A. Andrews, J. Bonta, and J. S. Wormith, "The recent past and near future of risk and/or need assessment," *Crime and Delinquency*, vol. 52, no. 1, pp. 7–27, 2006.

[9] D. A. Andrews, J. Bonta, and J. S. Wormith, "The LSI-OR: Interview and scoring guide. Toronto, ON: Ontario ministry of solicitor general and correctional services," 1995.

[10] D. A. Andrews and J. Bonta, "LSI-R: The level of service inventory-revised user's manual," *Multi-Health Systems Inc.*, 1995.

[11] J. S. Wormith, "Research to practice: Applying risk/needs assessment to offender classification," *Forum on Corrections Research*, vol. 9, pp. 26–31, 1997.

[12] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk e-mail," *Learning for text categorization: papers from the 1998 workshop. AAAI Technical Report WS-98-05*, vol. 62, pp. 98–105, 1998.

[13] J. Kazmierska and J. Malicki., "Application of the nave bayesian classifier to optimize treatment decisions," *Radiotherapy and Oncology*, vol. 86, no. 2, pp. 211–216, 2008.

[14] M. Ghasemi, S. Wormith, D. Anvari, M. Atapour, and R.J. Spiteri, "LSI, machine learning, and intervention," *Unpublished manuscript*, 2016.

[15] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[16] C. Ziegler, *Mining for Strategic Competitive Intelligence: Foundations and Applications*. Springer, 2012.

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

[18] Y. Zhao and Y. Cen, *Data Mining Applications with R*. Academic Press, 2013.

[19] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," *Proceeding SIGIR '94 Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 3–12, 1994.

[20] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[21] P. Cichosz, *Data Mining Algorithms : Explained Using R*. John Wiley & Sons, Ltd, 2015.

[22] "The R project for statistical computing." https://www.r-project.org/, 2015.

[23] "Weka 3. Free data mining software in java." http://www.cs.waikato.ac.nz/ml/weka/, 2015.

[24] "TableCurve 2D. Regression and curve fitting software." http://www.sigmaplot.co.uk/products/tablecurve2d/tablecurve2d.php, 2002.

[25] "Tanagra project. Free data mining software." http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html, 2003.

Appendix A

Offence Severity Index

Table A.1: Offence severity index.

Offence Severity	Offence Type
1	Unknown
2	Municipal Bylaw Offences
3	Other Provincial Offences
4	Liquor Licence Act Offences
5	Highway Traffic Act Offences
6	Parole Violations
7	Other Federal Statute Offences
8	Misc. Offences against Public Order
9	Drinking & Driving Offences
10	Breach of Court Order/Escape
11	Criminal Code Traffic Offences
12	Drug Possession Offences
13	Obstruction of Justice Offences
14	Morals & Gaming Offences
15	Arson/Property Damage Offences
16	Assault & Related Offences
17	Theft/Possession Offences
18	Misc. Offences against the Person
19	Fraud & Related Offences
20	Weapons Offences
21	Traffic/Import Drug Offences
22	Non-Violent Sexual Offences
23	Break & Enter & Related Offences
24	Violent Sexual Offences
25	Serious Violent Offences
26	Homicide & Related Offences

Appendix B

Skewness and Kurtosis

The skewness of a random variable X is given by the following formula,

$$\text{Skewness}(X) = \frac{m_3}{{m_2}^{\frac{3}{2}}}, \tag{B.1}$$

where α-th central moment given in (B.2) can be employed to obtain m_2 and m_3.

$$m_\alpha = \frac{\sum_{i=1}^{N} (X_i - X_{\text{mean}})^\alpha}{N}. \tag{B.2}$$

It can be seen from (B.1), the skewness of X determines the degree of asymmetry of a distribution around its mean(X_{mean}) e.g., in a symmetric situation like normal distribution, the skewness becomes zero and ($X_{\text{mean}} = X_{\text{median}}$). The skewness value can accepts a zero, a negative, or a positive value (figure B.1). A positive value indicates that the right tail of distribution is longer than the left tail and is skewed toward positive direction($X_{\text{mean}} > X_{\text{median}}$), a negative value indicates that the left tail of distribution is longer than the right tail and is skewed toward negative direction($X_{\text{mean}} < X_{\text{median}}$), and finally zero value means that the distribution is totaly symmetric. ($X_{\text{mean}} = X_{\text{median}}$).

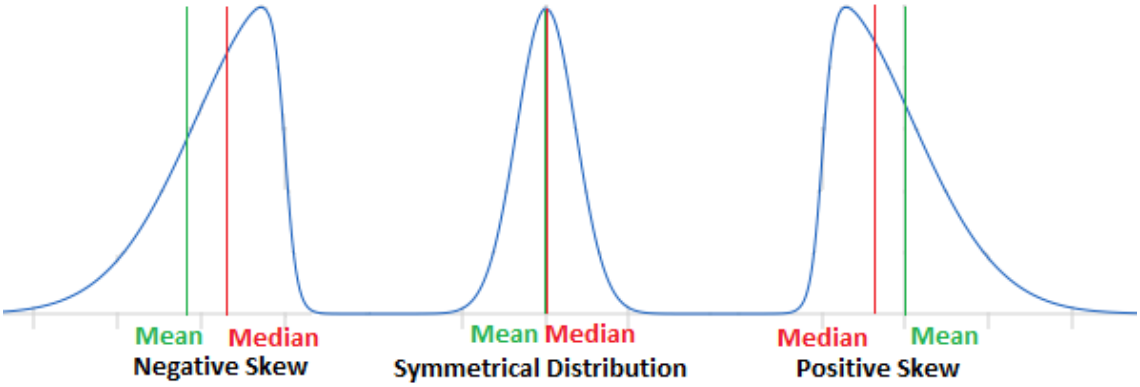

Figure B.1: A data distribution with various skewnesses.

The kurtosis of random variable X in terms of second and forth central moments (m_2 and m_4) is determined by

$$\text{Kurtosis}(X) = \frac{m_4}{{m_2}^2} - 3. \tag{B.3}$$

Intuitively, the kurtosis is a quantity that measures the shape (peakedness) of a data distribution. For example, in a normal distribution $m_4 = 3\sigma^4$, $m_2 = \sigma^2$, and therefore according to (B.3) the kurtosis becomes zero. A data distribution is said to be a peaked distribution (leptokurtic) or a normal distribution (mesokurtic) or a flat distribution (platykurtic) if the kurtosis becomes positive or zero or negative, respectively (figure B.2) .

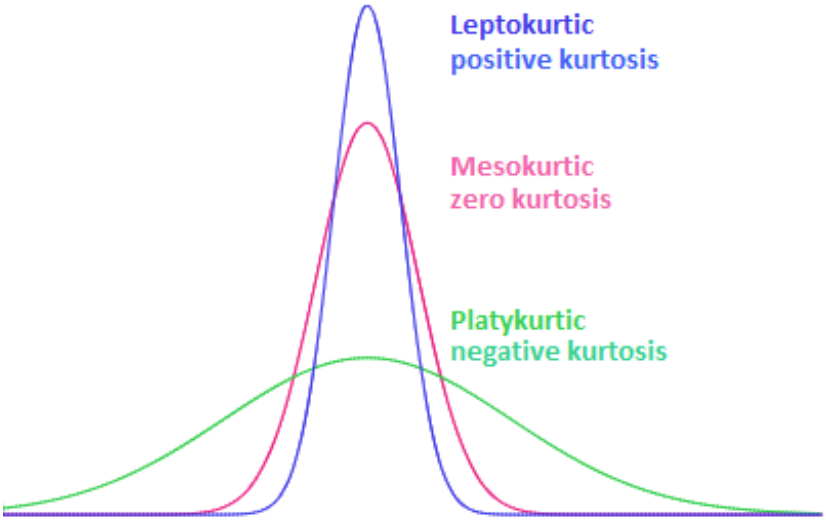

Figure B.2: Data distribution with various kurtosises.

In practice, many statisticians use the first term of equation (B.3) and usually the number 3 is eliminated. Finally, using (B.1) and (B.3) one can easily show that the skewness and kurtosis of a data set are invariant under scale transformation ($X \to X_{\max} Y$).
In general, the shape of most continuous distribution functions can be summarized in mean, variance[1], skewness[2], and kurtosis[3]. If a known distribution function and the observed data have same central moments, the distribution function of observed data can be usually approximated by the given distribution function. The approach of matching moments is called moment matching estimation (MME).

[1] second central moment
[2] third central moment
[3] fourth central moment

Appendix C

Beta distribution

The beta distribution function contains two free parameters (a, b), and it is defined by following formula,

$$\hat{\mathcal{B}}(a,b,x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \quad a>0\ ,\ b>0, \tag{C.1}$$

where

$$\begin{aligned} \Gamma(1) &= 1, \\ \Gamma(q+1) &= q\Gamma(q) \qquad q \in \mathbb{R}^+, \\ \Gamma(n+1) &= 1.2.3\cdots(n-1) = (n-1)! \qquad n \in \mathbb{Z}^+. \end{aligned}$$

In general one can show that

$$\begin{aligned} M_s &= \int_0^1 \hat{\mathcal{B}}(a,b,x) x^s dx, \\ &= \frac{\Gamma(a+b)\Gamma(a+s)}{\Gamma(a+b+s)\Gamma(a)}. \end{aligned} \tag{C.2}$$

The result in (C.2) can be used to calculate mean(μ), standard deviation (σ), skewness, and kurtosis:

$$\begin{aligned} \mu &= \frac{a}{a+b}, \\ \sigma &= \frac{1}{(a+b)}\sqrt{\frac{ab}{a+b+1}}, \\ \text{Skewness} &= \frac{2(b-a)}{(a+b+2)}\sqrt{\frac{a+b+1}{ab}}, \\ \text{Kurtosis} &= \frac{6[a^3+a^2(1-2b)+b^2(1+b)-2ab(2+b)]}{ab(a+b+2)(a+b+3)}. \end{aligned} \tag{C.3}$$

The maximum value of the beta distribution function (or the mode) can be derived by setting

$$\frac{d\hat{\mathcal{B}}(a,b,x)}{dx} = 0,$$

which results

$$\text{Mode} = x_{\max} = \frac{a-1}{a+b-2}, \tag{C.4}$$

and

$$x_{\max} = \mu.$$

if $(a = b)$. Finally the median is given by

$$\text{Median} = \frac{a - \frac{1}{3}}{a + b - \frac{2}{3}},$$

where $a, b > 1$.

Appendix D

Density functions-dataset

Listing D.1: The goodness-of-fit test result for total dataset.

```
                  F(x,a,b,c) = a x^{b-1}(1-x)^{c-1}

r^2 Coef Det   DF Adj r^2    Fit Std Err   F-value
0.9945423662  0.9945061429  0.0457550700  41275.001834

Parm  Value       Std Error   t-value     95% Confidence Limits   P>|t|
a    5.445154293 0.058605502 92.91199759 5.329981906 5.560326680 0.00000
b    1.433927178 0.004499885 318.6586530 1.425083939 1.442770416 0.00000
c    2.822155381 0.012110962 233.0248788 2.798354743 2.845956020 0.00000

Area Xmin-Xmax Area Precision
0.9865578695  -7.75808e+2868
Function min  X-Value       Function max  X-Value
4.114369e-06  0.9995630330  1.8042827028  0.1923366579
1st Deriv min X-Value       1st Deriv max X-Value
-3.197266045  0.5440079647  187.94627778  0.0004369679
2nd Deriv min X-Value       2nd Deriv max X-Value
4887.0222118  0.0503929725  13180.386773  0.9995630330

Procedure     Minimization  Iterations
LevMarqdt     Least Squares 8
r^2 Coef Det   DF Adj r^2    Fit Std Err   Max Abs Err
0.9945423662  0.9945061429  0.0457550700  0.2653776578
Source  Sum of Squares  DF     Mean Square      F Statistic     P>F
Regr    172.82061       2      86.410307        41275           0.00000
Error   0.94836747      453    0.0020935264
Total   173.76898       455
```

D.1 Dataset (R=0)

Listing D.2: The goodness-of-fit test result for total dataset (R=0).

```
F(x,a,b,c) = a x^{b-1}(1-x)^{c-1}

r^2 Coef Det   DF Adj r^2    Fit Std Err   F-value
0.9978750061  0.9978610258  0.0383528763  107301.21840

Parm  Value       Std Error   t-value     95% Confidence Limits   P>|t|
a    10.11444516 0.100082091 101.0614888 9.917766986 10.31112333 0.00000
b    1.523653346 0.003735504 407.8842745 1.516312451 1.530994241 0.00000
c    4.016911407 0.013994702 287.0308706 3.989409460 4.044413353 0.00000

Area Xmin-Xmax Area Precision
0.9840693809  -7.75808e+2868
Function min  X-Value       Function max  X-Value
1.598903e-09  0.9994359520  2.2939783497  0.1479010835
1st Deriv min X-Value       1st Deriv max X-Value
-5.029632232  0.3706242855  185.92247264  0.0005640488
2nd Deriv min X-Value       2nd Deriv max X-Value
7233.0674792  0.0505074592  15.359501845  0.5802618423

Procedure     Minimization  Iterations
LevMarqdt     Least Squares 8
r^2 Coef Det   DF Adj r^2    Fit Std Err   Max Abs Err
0.9978750061  0.9978610258  0.0383528763  0.4162284738
Source  Sum of Squares  DF     Mean Square      F Statistic     P>F
Regr    315.66798       2      157.83399        107301          0.00000
Error   0.67222101      457    0.0014709431
Total   316.3402        459
```

D.2 Dataset (R=1)

Listing D.3: The goodness-of-fit test result for total dataset (R=1).

```
                 F(x,a,b,c) = a x^{b-1}(1-x)^{c-1}

r^2 Coef Det   DF Adj r^2    Fit Std Err   F-value
0.9950887038  0.9950552177  0.0373040772  44675.998365

Parm  Value       Std Error   t-value     95% Confidence Limits   P>|t|
a    10.80541331 0.129211377 83.62586592 10.55146672 11.05935990 0.00000
b    2.287177624 0.007668757 298.2462097 2.272105773 2.302249475 0.00000
c    2.437528791 0.008358961 291.6066569 2.421100441 2.453957141 0.00000

Area Xmin-Xmax Area Precision
0.9963256137  5.510314e-10
Function min  X-Value       Function max  X-Value
0.0001408579  0.9996000000  1.6414418287  0.4724096285
1st Deriv min X-Value       1st Deriv max X-Value
-4.626659168  0.8525556285  5.4112507290  0.0922638870
2nd Deriv min X-Value       2nd Deriv max X-Value
-8.857488655  0.3708314133  -572.0314265  0.0503591616

Procedure     Minimization  Iterations
LevMarqdt     Least Squares 6
r^2 Coef Det   DF Adj r^2    Fit Std Err   Max Abs Err
0.9950887038  0.9950552177  0.0373040772  0.0828673748
Source  Sum of Squares  DF     Mean Square      F Statistic     P>F
Regr    124.34172       2      62.170859        44676           0.00000
Error   0.61369303      441    0.0013915942
Total   124.95541       443
```

Appendix E

Density functions-females

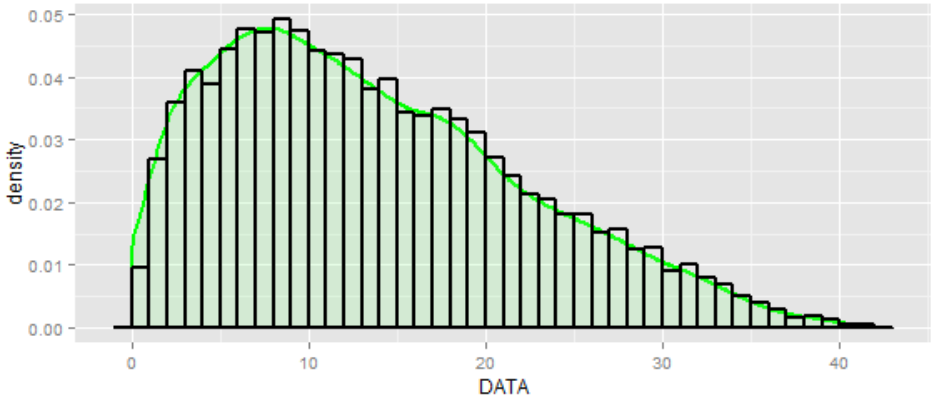

Figure E.1: Histogram of female offenders.

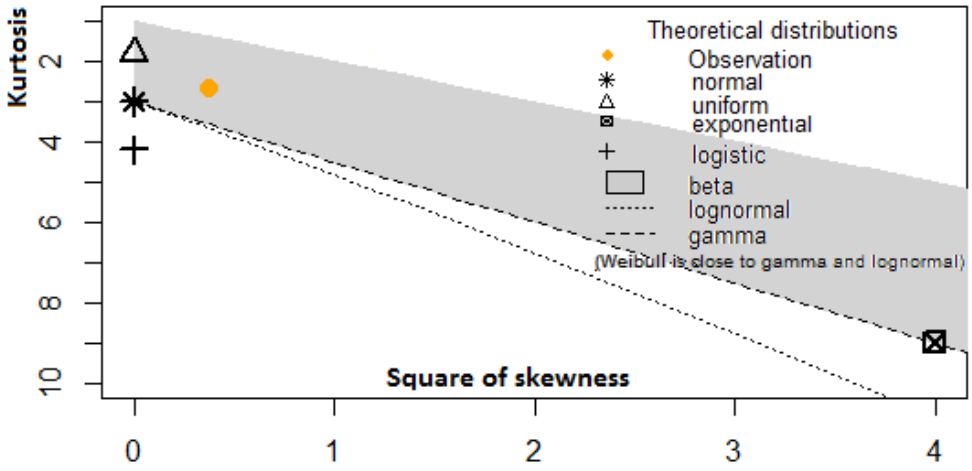

Figure E.2: Cullen–Frey graph (kurtosis versus square of skewness). The dataset (observation), indicated in orange, is located in the grey area covered by the beta distribution function.

Listing E.1: The goodness-of-fit test result for females.

```
                    F(x,a,b,c) = a x^{b-1}(1-x)^{c-1}

r2 Coef Det      DF Adj r2        Fit Std Err      F-value
```

```
0.9949753737    0.9949406412    0.0475889101    43069.301019

Parm    Value          Std Error    t-value       95% Confidence Limits    P>|t|
a     6.416908221   0.076036791   84.39214952   6.267463047   6.566353396   0.00000
b     1.461948596   0.004820968   303.2479633   1.452473310   1.471423882   0.00000
c     3.108415277   0.014182251   219.1764457   3.080541022   3.136289532   0.00000

Area Xmin-Xmax  Area Precision
0.9808245671    -7.75808e+2868
Function min    X-Value         Function max    X-Value
3.84857e-06     0.9988804010    1.9124236719    0.1797210907
1st Deriv min   X-Value         1st Deriv max   X-Value
-3.561807682    0.4861149994    113.87328831    0.0011196014
2nd Deriv min   X-Value         2nd Deriv max   X-Value
5214.1139438    0.0510076391    8.8472662087    0.8793302009

Procedure       Minimization    Iterations
LevMarqdt       Least Squares   8
r2 Coef Det     DF Adj r2       Fit Std Err     Max Abs Err
0.9949753737    0.9949406412    0.0475889101    0.2913484276
Source     Sum of Squares     DF      Mean Square        F Statistic     P>F
Regr       195.07847          2       97.539234          43069.3         0.00000
Error      0.9851464          435     0.0022647044
Total      196.06361          437
```

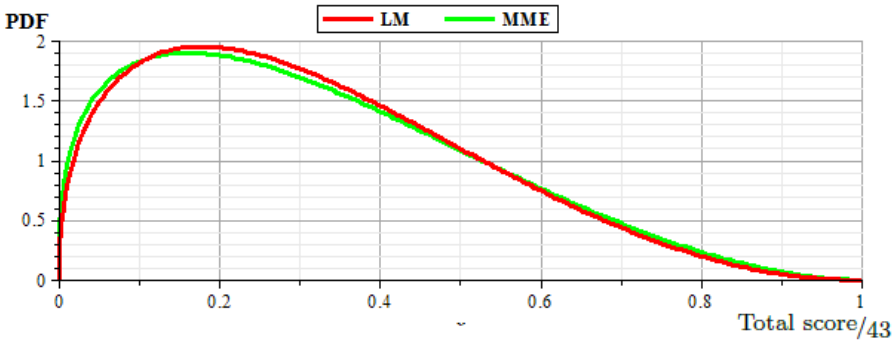

Figure E.3: A small variation was observed between MME and the LM method (Δa =0.10 and Δb =0.24).

E.1 Female offenders (R=0)

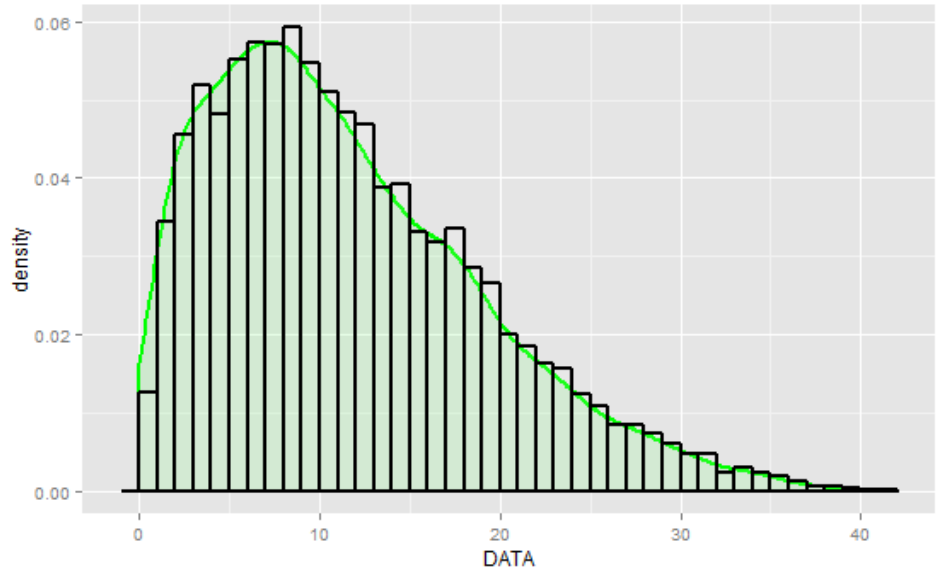

Figure E.4: The histogram of female offenders, case R=0.

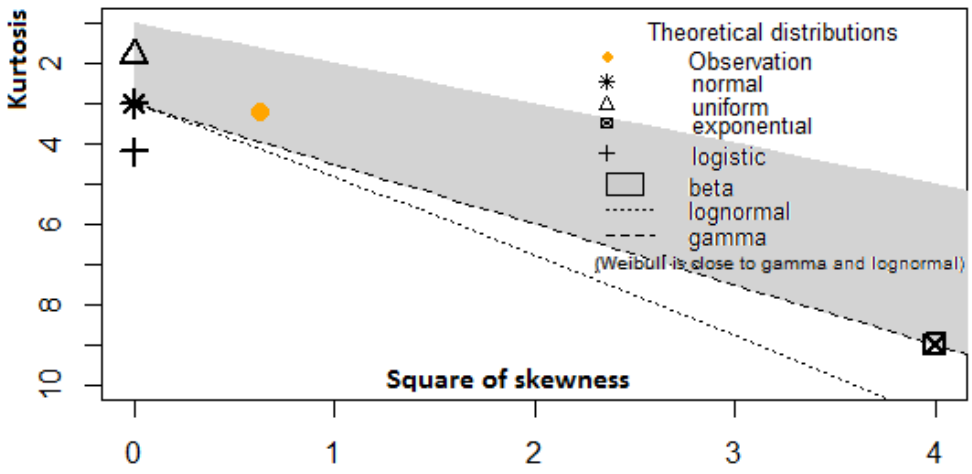

Figure E.5: Cullen–Frey graph (kurtosis versus square of skewness) for which the recidivism status is 0 (R=0). The dataset (observation), indicated in orange, is located in the grey area covered by the beta distribution function.

Listing E.2: The goodness-of-fit test result for females (R=0).

```
1 F(x,a,b,c) = a x^{b-1}(1-x)^{c-1}
2
```

```
r^2 Coef Det   DF Adj r^2    Fit Std Err   F-value
0.9957758254  0.9957475385  0.0551752343  52921.977207

Parm  Value       Std Error   t-value     95% Confidence Limits   P>|t|
a    11.13616350 0.164055424 67.88049562 10.81375170 11.45857531 0.00000
b    1.548463780 0.005528498 280.0875968 1.537598836 1.559328725 0.00000
c    4.197522413 0.021235107 197.6690002 4.155789875 4.239254951 0.00000

Area Xmin-Xmax Area Precision
0.9777290413  -7.75808e+2868
Function min  X-Value       Function max  X-Value
3.029111e-08  0.9979040140  2.3401814273  0.1464137284
1st Deriv min X-Value       1st Deriv max X-Value
-5.300258877  0.3597503110  207.32407926  0.0004039799
2nd Deriv min X-Value       2nd Deriv max X-Value
7441.6777175  0.0502789808  16.201277807  0.5593426973

Procedure     Minimization  Iterations
LevMarqdt     Least Squares 8
r^2 Coef Det   DF Adj r^2    Fit Std Err   Max Abs Err
0.9957758254  0.9957475385  0.0551752343  0.5453657743
Source  Sum of Squares  DF     Mean Square      F Statistic     P>F
Regr    322.22144       2      161.11072        52922           0.00000
Error   1.3668936       449    0.0030443065
Total   323.58833       451
```

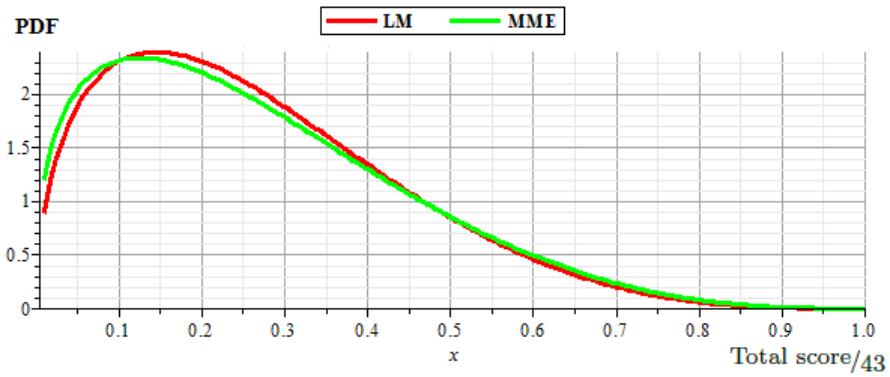

Figure E.6: R=0, a small variation was observed between MME and LM methods (Δa =0.17 and Δb =0.58).

E.2 Female offenders (R=1)

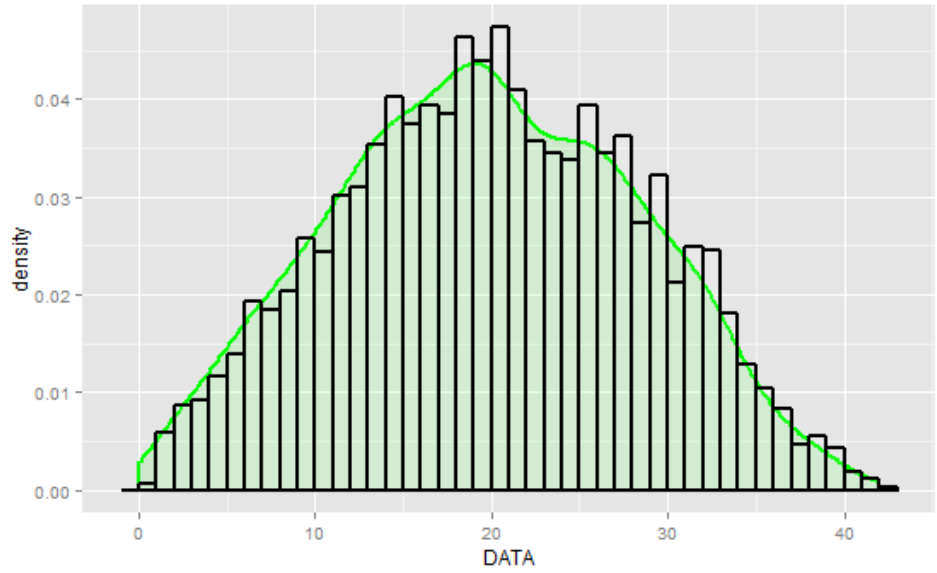

Figure E.7: The histogram of female offenders, case R=1.

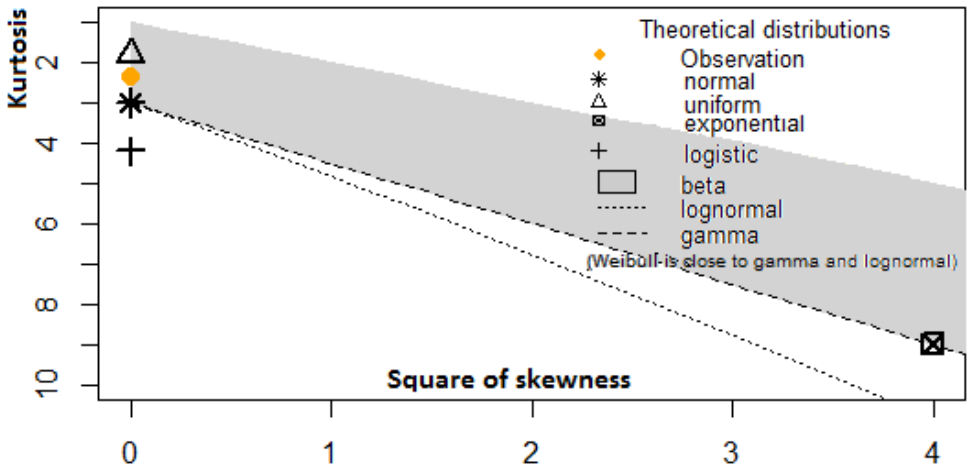

Figure E.8: Cullen–Frey graph (kurtosis versus square of skewness) for which the recidivism status is 1 (R=1). The dataset (observation), indicated in orange, is located in the grey area covered by the beta distribution function.

Listing E.3: The goodness-of-fit test result for females (R=1).

```
F(x,a,b,c) = a x^{b-1}(1-x)^{c-1}

```

```
r^2 Coef Det   DF Adj r^2     Fit Std Err   F-value
0.9873766673  0.9872856337  0.0637733929  16308.532835

Parm  Value       Std Error   t-value     95% Confidence Limits   P>|t|
a    15.30374630 0.353168768 43.33267179 14.60953335 15.99795925 0.00000
b    2.444593381 0.014504926 168.5353961 2.416081496 2.473105266 0.00000
c    2.731509921 0.016782989 162.7546737 2.698520117 2.764499725 0.00000

Area Xmin-Xmax Area Precision
0.9881251219  1.095988e-10
Function min  X-Value       Function max  X-Value
3.944198e-05  0.9994077720  1.7152652593  0.4548319103
1st Deriv min X-Value       1st Deriv max X-Value
-4.684573672  0.7923919675  5.7754408859  0.1172720340
2nd Deriv min X-Value       2nd Deriv max X-Value
-18.82350530  0.3505665773  -753.6508511  0.0505324729

Procedure     Minimization  Iterations
LevMarqdt     Least Squares 6
r^2 Coef Det   DF Adj r^2     Fit Std Err   Max Abs Err
0.9873766673  0.9872856337  0.0637733929  0.1484481742
Source  Sum of Squares  DF     Mean Square      F Statistic     P>F
Regr    132.65509       2      66.327547        16308.5         0.00000
Error   1.695958        417    0.0040670456
Total   134.35105       419
```

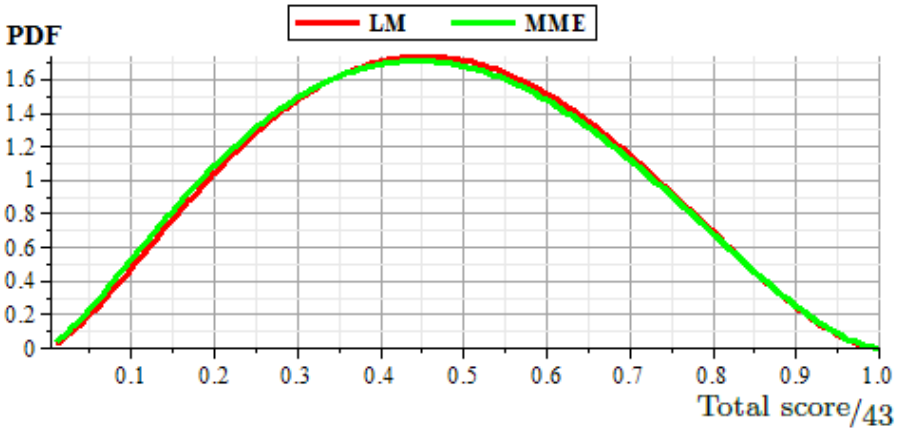

Figure E.9: R=1, a small variation was observed between MME and LM methods (Δa =0.10 and Δb =0.06).

Appendix F

Density functions-males

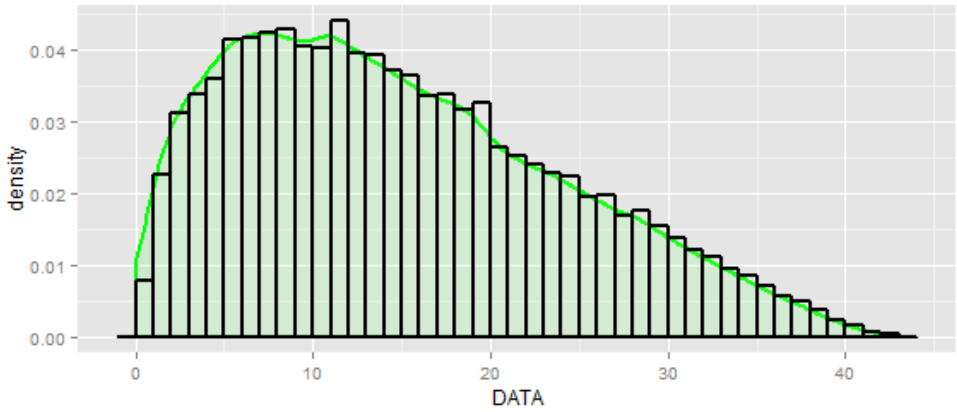

Figure F.1: The histogram of male offenders.

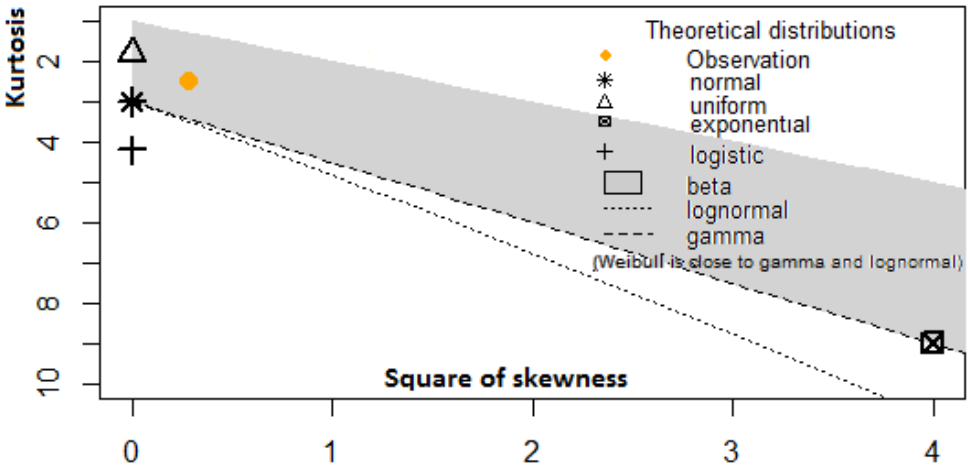

Figure F.2: Cullen–Frey graph (kurtosis versus square of skewness). The dataset (observation), indicated in orange, is located in the grey area covered by the beta distribution function.

Listing F.1: The goodness-of-fit test result for males.

```
                F(x,a,b,c) = a x^{b-1}(1-x)^{c-1}

r^2 Coef Det   DF Adj r^2     Fit Std Err   F-value
0.9944068366  0.9943693824  0.0450330542  39913.787751
```

```
Parm  Value       Std Error   t-value     95% Confidence Limits   P>|t|
a    5.263522376 0.055864773 94.21898723 5.153733489 5.373311262 0.00000
b    1.433705872 0.004488830 319.3941392 1.424884148 1.442527595 0.00000
c    2.752887126 0.011689774 235.4953178 2.729913663 2.775860589 0.00000

Area Xmin-Xmax Area Precision
0.9858206138  7.317129e-10
Function min  X-Value       Function max  X-Value
0.0001111819  0.9978468970  1.7711889013  0.1983478449
1st Deriv min X-Value       1st Deriv max X-Value
-3.115560537  0.5644109511  72.995515790  0.0021531069
2nd Deriv min X-Value       2nd Deriv max X-Value
4370.0493974  0.0519377927  3673.9497298  0.9978468970

Procedure     Minimization  Iterations
LevMarqdt     Least Squares 7
r^2 Coef Det   DF Adj r^2     Fit Std Err   Max Abs Err
0.9944068366  0.9943693824  0.0450330542  0.1118001929
Source  Sum of Squares  DF     Mean Square       F Statistic     P>F
Regr    161.8884        2      80.944202         39913.8         0.00000
Error   0.91056121      449    0.002027976
Total   162.79897       451
```

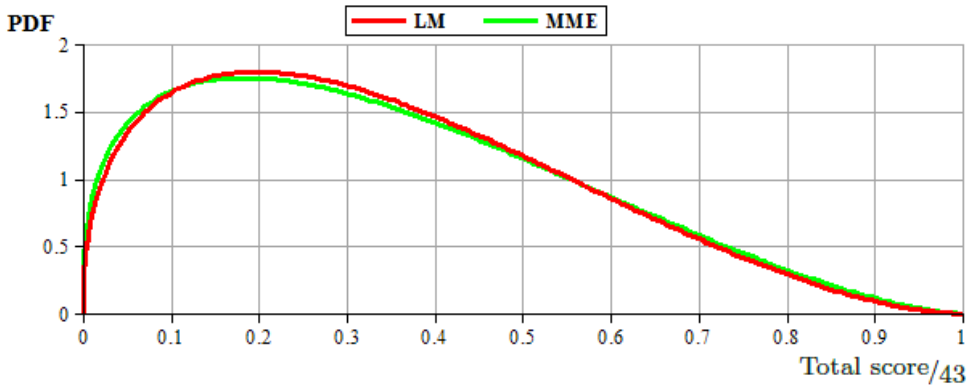

Figure F.3: A small variation was observed between MME and the LM method (Δa =0.08 and Δb =0.18).

F.1 Male offenders (R=0)

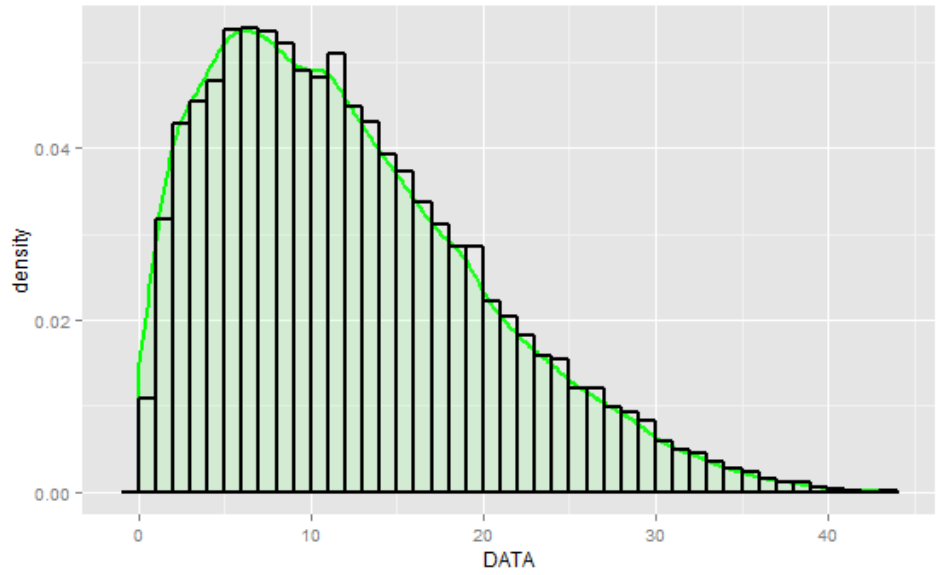

Figure F.4: The histogram of male offenders, case R=0.

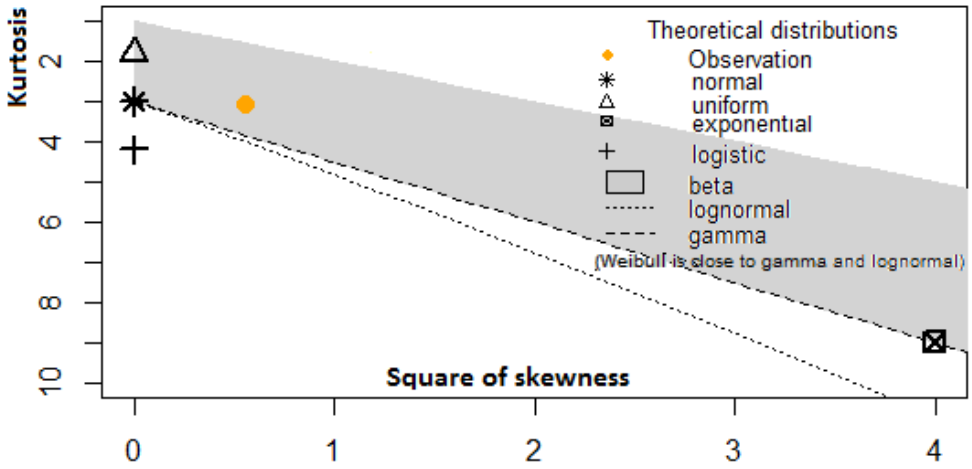

Figure F.5: Cullen–Frey graph (kurtosis versus square of skewness) for which the recidivism status is 0 (R=0). The dataset (observation), indicated in orange, is located in the grey area covered by the beta distribution function.

Listing F.2: The goodness-of-fit test result for females (R=0).

```
F(x,a,b,c) = a x^{b-1}(1-x)^{c-1}
```

```
r^2 Coef Det   DF Adj r^2    Fit Std Err   F-value
0.9975626268  0.9975465208  0.0405380792  93110.689896

Parm  Value       Std Error   t-value     95% Confidence Limits   P>|t|
a    9.845727641 0.103105319 95.49194691 9.643105952 10.04834933 0.00000
b    1.521562697 0.003975378 382.7466636 1.513750318 1.529375076 0.00000
c    3.951833394 0.014636490 269.9987112 3.923069890 3.980596899 0.00000

Area Xmin-Xmax Area Precision
0.9839348038  -7.75808e+2868
Function min  X-Value       Function max  X-Value
1.560133e-10  0.9997805960  2.2656057102  0.1501593134
1st Deriv min X-Value       1st Deriv max X-Value
-4.912386713  0.3773015156  288.54856761  0.0002194046
2nd Deriv min X-Value       2nd Deriv max X-Value
7163.8598352  0.0501974636  12.796871859  0.6000790738

Procedure     Minimization  Iterations
LevMarqdt     Least Squares 8
r^2 Coef Det   DF Adj r^2    Fit Std Err   Max Abs Err
0.9975626268  0.9975465208  0.0405380792  0.4821007497
Source  Sum of Squares  DF     Mean Square      F Statistic     P>F
Regr    306.02427       2      153.01214        93110.7         0.00000
Error   0.74771782      455    0.0016433359
Total   306.77199       457
```

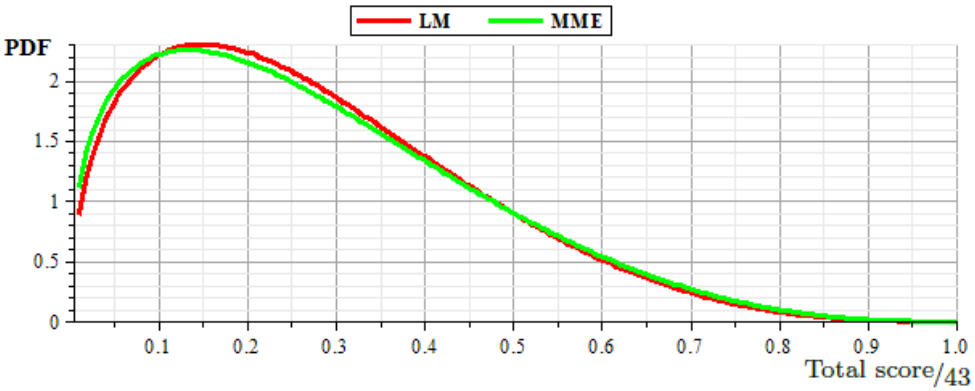

Figure F.6: A small variation was observed between MME and the LM method (Δa =0.11 and Δb =0.30).

F.2 Male offenders (R=1)

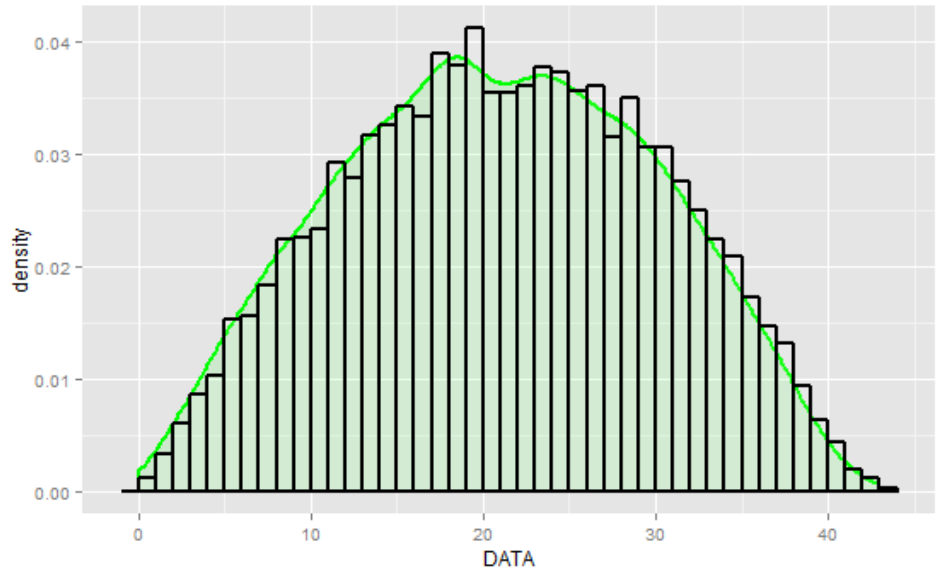

Figure F.7: The histogram of male offenders, case R=1.

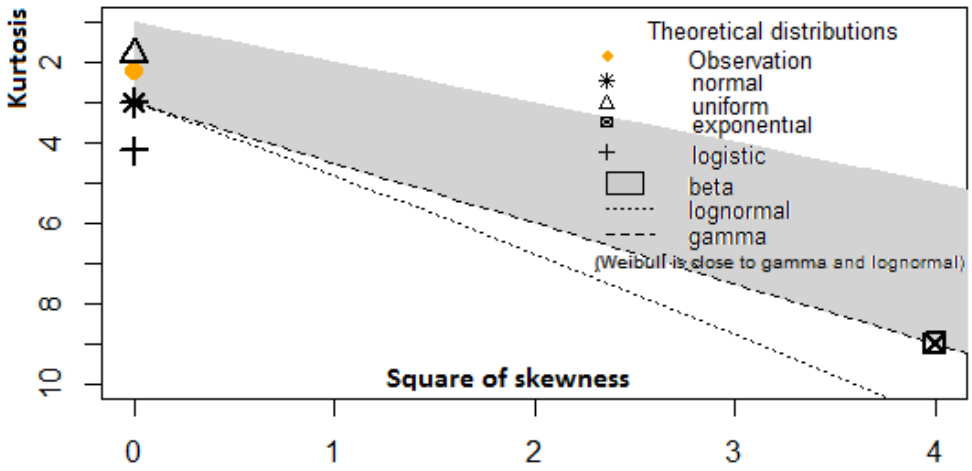

Figure F.8: Cullen–Frey graph (kurtosis versus square of skewness) for which the recidivism status is 1 (R=1). The dataset (observation), indicated in orange, is located in the grey area covered by the beta distribution function.

Listing F.3: The goodness-of-fit test result for males (R=1).

```
F(x,a,b,c) = a x^{b-1}(1-x)^{c-1}

```

```
r2 Coef Det      DF Adj r2        Fit Std Err      F-value
0.9956407910     0.9956109334     0.0345366300     50133.671375

Parm     Value          Std Error     t-value       95% Confidence Limits    P>|t|
a     10.00726959   0.108837260   91.94709241   9.793362752   10.22117644   0.00000
b     2.258225082   0.007029982   321.2277052   2.244408478   2.272041686   0.00000
c     2.369458517   0.007506243   315.6650473   2.354705879   2.384211156   0.00000

Area Xmin-Xmax   Area Precision
0.9969972583     2.393475e-09
Function min     X-Value          Function max     X-Value
4.818208e-05     0.9998690270     1.6230309295     0.4788343081
1st Deriv min    X-Value          1st Deriv max    X-Value
-4.669512652     0.8703919515     5.3032815543     0.0872767754
2nd Deriv min    X-Value          2nd Deriv max    X-Value
-7.568256016     0.3813375489     -524.6391704     0.0501178757

Procedure        Minimization     Iterations
LevMarqdt        Least Squares    6
r2 Coef Det      DF Adj r2        Fit Std Err      Max Abs Err
0.9956407910     0.9956109334     0.0345366300     0.0769087902
Source      Sum of Squares      DF      Mean Square         F Statistic      P>F
Regr        119.59676           2       59.798381           50133.7          0.00000
Error       0.5236299           439     0.0011927788
Total       120.12039           441
```

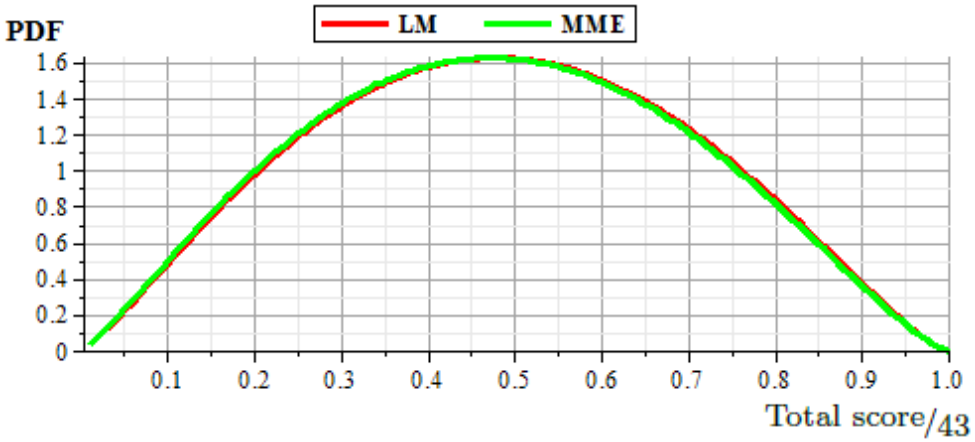

Figure F.9: A small variation was observed between MME and the LM method (Δa =0.02 and Δb =0.01).