Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.11 n.54, p. 67-82. July/August., 2025.
DOI: https://doi.org/10.5935/jetia. v11i54.1718

RESEARCH ARTICLE OPEN ACCESS

THE EFFECT OF DATA CONVERSION METHODS (NAIVE BAYES, C5.0 & SUPPORT VECTOR MACHINE) ON THE PERFORMANCE OF CLASSIFICATION ALGORITHMS IN DATA MINING

Hussein Ali Attallah [1], Ahmed Al-Asadi [2] and Sadeer Sadeq [3]

[1,2] Communication Engineering Department, University of Technology, Baghdad, Iraq.
[3] University of Information Technology and Communications Baghdad, Iraq

[1]https://orcid.org/0000-0002-3443-3547, [2]https://orcid.org/0000-0002-5052-1969, [3]https://orcid.org/0009-0000-8415-8056

E-mail: hussein.a.attallah@uotechnology.edu.iq, ahmed.a.hussain@uotechnology.edu.iq, sadeer.abduljabbar@uoitc.edu.iq

ARTICLE INFO

Article History
Received: March 16, 2025
Revised: April 20, 2025
Accepted: June 15, 2025
Published: July 31, 2025

Keywords:
Comparative study,
Conversation methods,
Data transformation,
NB, SVM, C5.0,
Normalization,
Statistical analysis.

ABSTRACT

In the study, sample distributions (Normal, Chi-square, F), number of observations (100, 500, 1000, 10000) and class distribution rates (0.1, 0.2, 0.3, 0.4, 0.5) were evaluated. It was aimed to examine the effects of data transformation on naive Bayes (NB), C5.0 and support vector machines (SVM) by applying minimum-maximum and z-score normalisation and equal width and equal frequency spacing discrimination methods to different types of data produced by simulation. In this research, the minimum-maximum and z-score normalisation of the data produced by simulation from a normal distribution, chi-square distribution and F distribution according to four different numbers of observations and five different equilibrium distribution ratios of classes, and spacing discrimination transformations of equal girth (width) (EG) and equal frequency (EF). The results and comparative study showed that both normalisation and discrimination methods were influential in the performance of SVM and contributed to better results. According to the classification success achieved with SVM, normalisation methods are more effective in average, and chi-square distribution among both approaches, and EF unsupervised discrimination method is more effective in F-distribution.

I. INTRODUCTION

Making the right decision is of great importance in order to obtain positive results in medical practices. Although the developments in the field of health and technology are pleasing, some uncertainties encountered in medicine make it difficult for clinicians to make decisions [1]. Since 1975, the rapidly increasing number of diagnostic and treatment options and the increasing importance of cost aspects have made decisions difficult for healthcare providers [2]. Data mining, which includes many disciplines, has become one of the auxiliary tools used to develop new clinical principles and to make scientifically high-quality decisions with the information obtained. Data mining is generally defined as a process consisting of steps such as data collection, preparation for analysis, and modelling to obtain valid and valuable new information in the field of research [3],[4]. The main objective of the current research work is classification, which is one of the data mining techniques, and naive Bayes (NB), decision trees, C5.0 algorithm, and support vector machines (SVM) are the main methods used.

The naive Bayes algorithm is a simple classification method based on the Bayes theorem. The naive Bayes algorithm has strong assumptions that are rarely provided. Despite this, it can give good results against many problematic data [5]. In classification problems, class estimation uses the probabilities of the classes in the dependent variable and the class conditional probabilities for a new observation [6]. Decision trees are a supervised classification method with a hierarchical structure that considers a particular order of importance for the variables in the data set [7],[8]. Here, the order of the variables is determined with variable selection methods, and a decision is made for the classes of the new observations by following a path from top to bottom. The developed version of the ID3 decision tree algorithm,

C5.0, has the advantages of modelling continuous data, automatically selecting features in multidimensional data, and not requiring a priori determination of the model. Another method used in the study is SVM [9]. Its popularity has increased recently due to its good performance in nonlinear and multidimensional data. It performs classification by drawing a boundary that maximises the area between observations in one class and the other class in a multidimensional space. Classification methods have advantages and disadvantages [10]. Therefore, their performance superiority may vary depending on the sample size and data type. In order to obtain more robust and balanced results, several methods can be applied to the data where the algorithms are disadvantaged before classification, thus increasing the performance success. One of the available options for this purpose is to apply transformation to the data. In data mining, minimum-maximum and z-score normalisation are frequently used for data transformation. For a similar purpose, unsupervised discretisation applications, also known as binning, are another approach that uses equal width and equal frequency spacing methods [11]. In this study, it was aimed to investigate the effects of data transformation on naive Bayes, C5.0 and SVM by applying minimum-maximum and z-score normalisation and equal width and equal frequency spacing discretisation methods to different types of data generated by simulation according to sample distributions (Normal, Chi-square, F), observation numbers (100, 500, 1000, 10000) and class distribution ratios (0.1, 0.2, 0.3, 0.4, 0.5).

II. CLASSIFICATION DEFINITION AND PROCESS

In classification applications, each category in the dependent variable is called "class". In medical research, examples of these classes can be given, such as three different types of treatment and two different disease diagnoses [12]. NB, C5.0 and SVM are frequently used classifiers for classification problems in data mining. In data mining, reaching all class conditional intersection probabilities becomes more difficult as the number of independent variables increases, making the solution more complex [13]. In the NB method, to simplify the operations, it is assumed that the variables in each class, that is, class conditional variables, are independent of each other. This assumption is referred to as “conditional independence” in many sources [14].

II.1 NAIVE BAYES (NB):

In the study [15], Naïve Bayes Algorithm uses probability and statistics to solve classification problems. This method performs classification by calculating the value of the probability P($x \mid y$) by knowing the probability of class X. Determination of the class in the classification is done by selecting the max value of P($x \mid y$) Based on probability. The advantage of classification is that it requires a relatively small amount of training data to estimate the parameters needed for classification. Based on the Naïve Bayes algorithm, the following is the equation for calculating the value of $P(x \mid y)$ [16]:

$$P(X \mid Y) = \frac{P(Y|X).(X)}{P(Y)} \quad (1)$$

where:
P(X|Y) = Posterior probability, namely the probability value of X based on condition Y

P(Y|X) = probability of Y determined by X is true
P(X) = Probability of evidence of disease X
P(Y) = Probability of the value of Y

Successful results are achieved with NB in genetic and drug development studies. For [17] developed an approach based on the naive Bayesian algorithm using RNA sequencing data, which could help diagnose tissue-derived hepatobiliary or pancreatic cancer in synchronous tumours. While more than a 95% success rate was achieved with the 10-fold cross-validation method with the created model, they reported correctly classifying 17 (94.4%) of 18 clinical cancer tissue samples (six negative controls) used for external validity.

II.2 DECISION TREES:

In medical data analysis, it is essential to communicate data mining results to people in an understandable way [18]. Two algorithms form the basis of decision trees researchers developed without knowing each other.

II.3 STRUCTURE OF DECISION TREES:

The structure of a decision tree consists of nodes and branches. As shown in (Figure 1), the topmost starting node is called "root", and the last node containing the classes of the dependent variable is called "leaf". In decision trees, all nodes except leaves represent an independent variable. These nodes are also called "decision nodes" because they decide which class to reach with independent variable questions (e.g., "What is the gender?" and "What is the age?" etc.). The branches emerging from each decision node contain all the answers to the variable (For example, “female-male”, “65 years old and under - 65 years old, etc.). In other words, it represents their categories. These branches, formed as many as the number of categories belonging to the relevant variable, are connected to a leaf or a new decision node. The leaves reached at the last stage express the decisions regarding the classes.

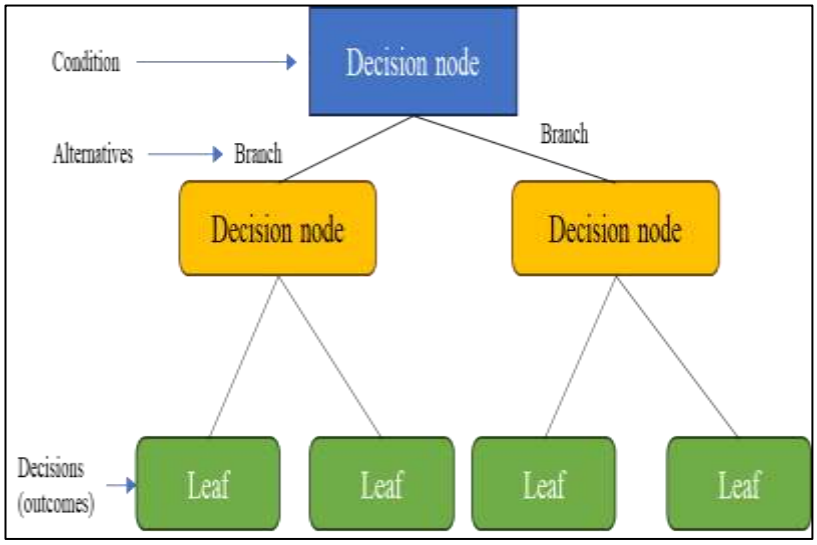

Figure.1: The structure of a decision tree consists of nodes and branches.
Source: [19].

II.4 C5.0 DECISION TREE ALGORITHM:

The C5.0 algorithm is a development or improvement of the C4.5 algorithm where the development is carried out on memory usage, attribute separation, and acceleration, according to Arif, Muhammad. (2018) [20],[21], the C4.5 algorithm is a development or improvement of the ID3 algorithm; improvements made such as being able to handle discrete or continuous attributes, being able to handle empty attributes or missing values, and being able to summarise or prune decision trees.

The C5.0 algorithm is a data mining classification algorithm applied explicitly to decision tree techniques. In selecting attributes for solving objects in several classes, the attribute that produces the most significant information gain must be selected [22]. Entropy measures uncertainty associated with a random variable, while gain is the expected value. If the entropy value is "0", it is easier to classify the observations. In data mining, the terms entropy and information gain are related. Entropy is used to select the variables that provide the highest information gain for the nodes to determine the most appropriate variable order in the classification. By selecting the variables that provide the highest information gain, the information required for the classification, namely entropy, is reduced to a minimum level. Suppose the entropy value is found to be "0", no further information is needed to classify the observations and that all observations belong to the same class. The formula for the entropy index is given below [23]:

$$Entropy\ (S)\ = \sum_{i=1}^{m} -\ p_i\ log_2(p_i) \tag{2}$$

Where;
$\boldsymbol{S}$: $Case\ value$
$\boldsymbol{M}$: $Number\ of\ classes\ in\ the\ variable$
$\boldsymbol{p_i}$: $Proportion\ of\ S\ and\ S_i$

In the analysis using the C5.0 Algorithm, there are several steps, namely, inputting the studied data. Next, selecting the root node begins with calculating the entropy value. Then, the process continues by finding the value gained. After that, look for the gain ratio value.

II.5 SUPPORT VECTOR MACHINE ALGORITHM (SVM):

According to research [24], SVM is a method or algorithm for classifying and predicting. The working principle of this method is to find the most optimal separation space for a dataset in different classes.

$$f(x_d) = \sum_{i=1}^{ns} a_i y_i \overrightarrow{x_i} \overrightarrow{x_d} + b \tag{3}$$

Where:
$\boldsymbol{ns}$ = Number of support vectors
$\boldsymbol{a_i}$ = Weight value of each data point
$\boldsymbol{y_i}$ = Data class
$\overrightarrow{\boldsymbol{x_i}}$= Support vector variable
$\overrightarrow{\boldsymbol{x_d}}$ = Data to be classified
$\boldsymbol{b}$ = Error or bias value

The SVM algorithm places all observations according to their values in a dimensional space equal to the number of variables. Then, it finds the hyperplane that provides the best separation between the observations according to their classes and classifies new observations according to this hyperplane [25].

A hyperplane is a p-1-dimensional subspace plane in p-dimensional space. The following equation makes its mathematical definition:

$$\beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \quad (4)$$

When this equation is applied to a point in space,

$$\beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p < 0 \quad (4\text{-}1)$$

$$\beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p > 0 \quad (4\text{-}2)$$

Results can be obtained. If the result is negative, the point is in one class; otherwise, if it is positive, it is in the other. It can be easily found on which side the observations are in the space divided into two, according to the sign of the hyperplane equation result. According to [26], identifying acute exacerbations in chronic obstructive pulmonary disease is crucial to reducing mortality and financial burden. He developed different classification models for this disease and conducted a comparison study to find the best model. From the results obtained, SVM showed the highest success performance.

Based on the above points, in the study, sample distributions (Normal, Chi-square, F-distribution), number of observations (100, 500, 1000, 10000) and class distribution rates (0.1, 0.2, 0.3, 0.4, 0.5) were evaluated. It aimed to examine the effects of data transformation on NB, C5.0 and SVM by applying minimum-maximum and z-score normalisation and equal width and frequency spacing discrimination methods to different types of data produced by simulation.

III. MATERIALS AND METHODS

Not all obtained data can be processed for classification from the data collection. Data analysis first needs to be done because there are attributes that do not need to go through various stages of data analysis in order to get quality data.

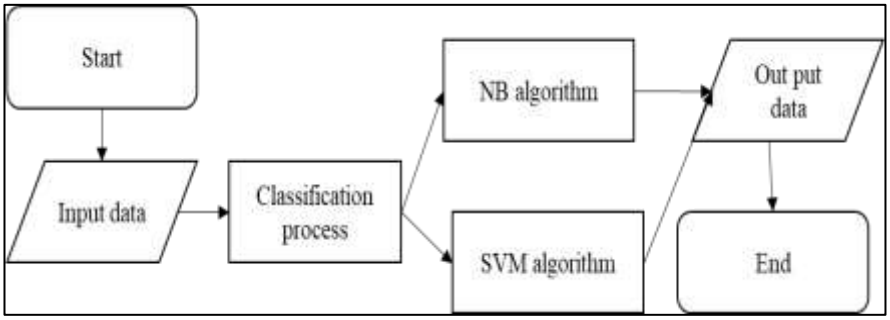

Figure 2: Research Flow.
Source: Authors, (2025).

III.1 RESEARCH FLOW DIAGRAM

The effects of normalisation, z-score, and unsupervised discrimination methods applied in the data pre-processing process on classification algorithms' performance were examined using simulation data sets. Data sets were created in a structure with 11 variables, including 10 independent variables of quantitative type and a dependent variable of 2 categories.

Table 1: Sampling distribution parameters.

Variable	Group	Normal distribution		Chi-square distribution	F distribution	
		Mean	SD	df1	df1	df2
Variable-1	1	1.8	0.2	30	4	4
	0	2	0.2	32	5	5
Variable-2	1	20	4	10	50	50
	0	22	5	8	10	10
Variable-3	1	4	0.3	200	100	10
	0	4.2	0.2	195	100	50
Variable-4	1	10	1	100	10	1
	0	9	0.8	95	5	15
Variable-5	1	30	4	2	100	10
	0	28.5	3	4	100	5
Variable-6	1	150	40	5	30	1
	0	180	50	7	10	5
Variable-7	1	40	5	15	100	5
	0	41	4	17	100	50
Variable-8	1	100	15	500	100	10
	0	95	20	475	10	10
Variable-9	1	5	0.8	200	1	5
	0	4.5	0.5	210	10	5
Variable-10	1	50	10	1000	10	5
	0	52	10	1050	10	3

Source: Authors, (2025).

sd: standard deviation;
df1: 1st degrees of freedom;
df2: 2nd degrees of freedom.

Methods were applied before classification to examine the effects of data transformation, and NB, C5.0 decision tree, and SVM algorithms were used. Classification results were first obtained with raw data without pre-processing to examine the effects of data transformation methods on the algorithms. Then, the classification analysis was repeated by applying equal width spacing (EW) and equal frequency spacing (EF) methods to the generated data, such as minimum-maximum normalisaton, $z - score$ normalisation and unsupervised discrimination methods (Table 2).

Table 2: Data transformation methods.

Minimum-Maximum normalisation	$\frac{X_{min} - x}{X_{max} - X_{min}}$	X_{min} ; minimum value X_{max} ; maximum value
Z-Score normalisation	$\frac{x - \bar{X}}{\sigma}$	$\bar{X}$; variable mean σ ; standard deviation
Equal width spacing discretisation (EW)	$a = \frac{X_{max} - X_{min}}{k}$, $s = X_{min} + (i \times a)$	k; number of intervals a; interval width i ;1,2,…,k-1 s ; limits
Equal frequency spacing discretisation (EF)	$f = n/k$	k; number of intervals n; number of observations f; interval frequency

Source: Authors, (2025).

The values in the raw data are respectively applied by minimum-maximum normalisation and z-score normalisation; It was brought to the range of "0" to "1" and "-3" to "3". For these operations, functions are written in the R −Program according to the formulas of both methods were used. The k −Value in the formula of the EW and EF discrimination methods was determined as "10" for both methods. The data was divided according to this value and converted into categorical form. All applications in the simulation study were repeated 1000 times, and the findings were summarised using mean and standard deviation values.

As a result of the study, "Overall Accuracy", "Area Under the Curve (AUC)", "Sensitivity", "Specificity", and "Positive Predictive Value" were used to evaluate the performance of classification algorithms", "Negative Predictive Value" measurements were used. In the unbalanced distribution of classes in the dependent variable, the "balanced accuracy" criterion was used instead of the "General Accuracy Rate". The definition and calculation formulas for these criteria used to evaluate performance are given in (Table 3).

Table 3: Performance evaluation criteria.

Accuracy	The ratio of actually "Positive" and "Negative" to the total	$\frac{(\mathrm{TP} + \mathrm{TN})}{(TP + TP + FP + FN)}$
Balanced Accuracy rate	Average of sensitivity and selectivity rates	$\frac{(TP/(TP + FN) + TN/(TN + FP))}{2}$
Sensitivity rate	Ratio of predicted "Positive" to actually "Positive."	$TP/(TP + FN)$
Selectivity rate	Ratio of predicted "Negative" to actually "Negative"	$TN/(TN + FP)$
Positive predictive value	Ratio of actually "Positive" to those predicted to be "Positive"	$TP/(TP + FP)$
Negative predictive value	Ratio of predicted "Negative" to actually "Negative"	$TN/(TN + FN)$
AUC	Area under the curve with sensitivity on the vertical axis and 1-selectivity on the horizontal axis	

Source: Authors, (2025).

TP: *True positive;* ***TN:*** *True negative;*
FP: *False positive;* ***FN:*** *False negative.*

IV. RESULTS AND DISCUSSION

IV.1 RESULTS

The algorithms were compared by taking the highest accuracy rate, AUC value, sensitivity and selectivity rate, and the classification algorithms' positive and negative predictive value in the raw data or after data transformation.

As a result of the study, the findings regarding the highest accuracy rates achieved in the three algorithms. While there was an increase in all combinations of data generated from the normal distribution in SVM with data transformation, there was no increase in the NB algorithm. The NB algorithm achieved higher overall accuracy rates than other algorithms with the Gaussian probability density function, regardless of the number of observations and class distribution rates. With the increase in the number of observations for data produced from normal distribution in SVM, the change in data transformations draws attention. As the number of observations increases, the superiority of z-score normalisation over minimum and maximum increases when moving from balanced to unbalanced class

distribution. With the C5.0 algorithm, there is an increase in accuracy rates with z-score normalisation only in the 0.4 and 0.5 class distributions in 500 observations and the 0.2 class distribution in 10000 observations. Still, according to the findings in Table 4, these increases are less than 0.01%.

When examined for chi-square distribution, an increase in the overall accuracy rate was achieved with data transformations in all data types of SVM and with 100 observations and a class distribution ratio of 0.1, NB with the kernel probability density function and the SVM algorithm achieved the highest accuracy rates in all remaining combinations. While the highest accuracy rates are achieved with z-score normalisation in SVM for data more prominent than 100 observations and a fully balanced class distribution (0.5), the superiority of minimum-maximum normalisation is noteworthy when the classes are not distributed balanced. In 100 observations, the best results were found with the data using minimum-maximum normalisation in all distributions except the 0.1 class distribution in SVM.

Table 4: Average accuracy rates (%) obtained with C5.0 for the Continuation of Normal Distribution.

n	Ratio	Raw data	Minimum-Maximum (n)	Z-Score (n)	EW	EF
100	0.1	65.33±17.18	65.33±17.18	65.98±17.35	51.41±5.44	51±4.03
	0.2	71.98±13.33	71.98±13.33	72.69±13.46	53.59±8.18	53.94±8.77
	0.3	74.25±11.72	74.25±11.72	74.99±11.83	58.61±11.33	57.25±11.26
	0.4	75.64±10.5	75.64±10.5	76.39±10.61	65.62±11.26	65.83±11.49
	0.5	76.12±9.97	76.12±9.97	76.88±10.07	67.17±11.48	65.94±11.22
500	0.1	70.51±8.43	70.51±8.43	71.2±8.52	61.26±8.3	57.36±7.81
	0.2	76.71±6.1	76.71±6.1	77.47±6.17	71.02±6.94	70.83±6.93
	0.3	79.33±4.79	79.32±4.79	80.1±4.83	75.68±5.51	75.64±5.33
	0.4	80.46±4.42	80.46±4.42	81.26±4.48	77.67±4.65	77.41±4.84
	0.5	80.68±4.14	80.68±4.14	81.48±4.18	78.52±4.39	77.97±4.59
1000	0.1	72.8±5.88	72.8±5.88	73.52±5.94	65.91±6.5	60.88±6.37
	0.2	78.11±4.27	78.11±4.27	78.89±4.31	74.43±4.57	72.84±4.85
	0.3	80.62±3.43	80.62±3.43	81.41±3.47	78.38±3.82	78.22±4.02
	0.4	81.68±3	81.68±3	82.49±3.03	80.38±3.16	80.01±3.27
	0.5	81.89±2.91	81.89±2.91	82.7±2.94	80.71±3.01	80.24±3.19
10000	0.1	77.02±2.03	77.01±2.03	77.78±2.06	74.74±2.02	72.59±2.17
	0.2	81.74±1.4	81.74±1.4	82.55±1.42	80.56±1.38	79.56±1.47
	0.3	83.95±1.09	83.95±1.09	84.78±1.1	83.31±1.09	82.73±1.04
	0.4	85.03±0.91	85.03±0.91	85.88±0.92	84.7±0.89	84.29±0.94
	0.5	85.41±0.83	85.41±0.83	86.25±0.84	85.19±0.8	84.93±0.88

Source: Authors, (2025).

For the 0.1 class distribution rate in 1000 and 10000 observations in the data produced from the F distribution, the C5.0 algorithm is used. In all remaining combinations with raw data and z-score transformation, it was seen that the NB algorithm gave the highest results with EF discrimination. EF transformation in SVM increased accuracy more effectively than other transformation methods.

In the study, the findings regarding the highest AUC values obtained from all simulation scenarios with NB, C5.0 and SVM algorithms and how they were obtained are included in (Table 5).

There was no increase in the data produced from a normal distribution compared to the results obtained from the raw data after data transformation in NB and C5.0 algorithms. It was observed that NB achieved a higher AUC value than other algorithms, regardless of the number of observations and class distribution rates. Notably, in the SVM algorithm, the highest AUC values in data that are more significant than 100 observations are most frequently achieved with z-score normalisation.

There was no increase in the data generated from the chi-square distribution after data transformation in NB. The highest AUC values were achieved with the SVM algorithm in all observation numbers and class distribution ratios. However, NB and SVM gave the same high AUC values in 0.2-0.3-0.4-0.5 class distributions in 10000 observations. At the highest values obtained with SVM, it is noteworthy that z-score normalisation is superior in fully balanced distributions of classes, and minimum-maximum normalisation is superior in unbalanced distributions.

The data generated from the F distribution showed that NB and SVM gave the highest AUC values after EF discrimination. The NB algorithm reached the highest values without varying depending on the number of observations or class distribution.

In the C5.0 algorithm, AUC values generally did not increase after the transformations on the data generated from all distributions.

In the study, the findings regarding the highest sensitivity rates obtained with NB, C5.0 and SVM algorithms and how they were obtained are given in (Table 6).

There was no increase in the data produced from normal distribution compared to the sensitivity rates obtained from the raw data after data transformation in NB and C5.0 algorithms. NB has the highest sensitivity rates with the Gaussian probability density function applied to the raw data in all data types except 100 observations and 0.4 class distribution ratio. Looking at the SVM findings, it can be seen that the superiority of z-score normalisation over minimum and maximum increases as the number of observations increases and the balance ratio in classes decreases.

In the data generated from the chi-square distribution, there was no increase in sensitivity rates in the C5.0 algorithm after data transformation. The highest sensitivity rates were achieved with the SVM algorithm with 100 observations and 0.1-0.2 class distribution rates. The superiority of minimum-maximum normalisation draws attention to the highest values obtained with SVM.

In the data produced from the F distribution, there was no increase in sensitivity rates in the C5.0 algorithm after data transformation. It was observed that NB and DVM, each individually, gave the highest sensitivity rates most frequently after EF discrimination. Regardless of the number of observations or class distribution, the best results in sensitivity rates were achieved with the NB algorithm.

In the study, the findings regarding the highest selectivity rates obtained with NB, C5.0 and SVM algorithms and how they were obtained are given in Table 7.

In the data produced from normal distribution, there was no increase in the selectivity rates obtained from the raw data after data transformation in NB algorithms. The highest sensitivity rates were achieved with NB (Gaussian) in the data where the classes were distributed precisely balanced (0.5) and with SVM in all unbalanced distributions. Both algorithms achieved their highest values with the most frequent raw data. In the C5.0 algorithm, the effect of EG discrimination in increasing the selectivity rate, especially in unbalanced class distributions larger than 100 observations, is noteworthy.

In the data generated from the chi-square distribution, there was no increase in NB after data transformation. While C5.0 (EF) provides the highest selectivity rates in 0.1-0.2 class distributions per 100 observations and NB (Gaussian) in 0.4-0.5 class distributions per 10000 observations, the highest selectivity rates are achieved with SVM in all remaining data types was done. In SVM, the highest values were obtained with raw data, most frequently in data larger than 100 observations and with unbalanced class distributions, and with z-score normalisation in a fully balanced (0.5) distribution. In the C5.0 algorithm, while the superiority of EF discrimination was noted in unbalanced distributions in 100 observations, high selectivity rates were achieved more frequently with EG discrimination in more extensive observations.

In the data produced from the F distribution, it was seen that NB showed the highest selectivity success in fully balanced distributions for all observation numbers. While NB showed the highest success in the 0.4 class distribution out of 100 observations, the highest selectivity rates were achieved with SVM in all the remaining unbalanced class distributions. SVM reached its highest rate with EF in fully balanced distributions and minimum-maximum normalisation in unbalanced distributions. EG discrimination seems most effective on the C5.0 algorithm, increasing the selectivity rate, especially in unbalanced class distributions.

The study includes the findings regarding the highest positive predictive values (PKV) obtained from all simulation scenarios with NB, C5.0 and SVM algorithms and how they were obtained in (Table 8).

In the data produced from the normal distribution, there was no increase compared to the PKD obtained from the raw data after the data transformation in the NB algorithm. The highest PKD was obtained with SVM (M) in 100 observations and 0.1 class distribution, and NB (Gaussian) in all remaining data types. In the C5.0 algorithm, the effect of EG discrimination is noteworthy, especially in cases with unbalanced class distributions larger than 100 observations.

In the data generated from the chi-square distribution, the data transformation did not produce any increase in NB. SVM for PKD showed the highest success in all data types except 1000 observations 0.2 class distribution, and 10000 observations. The highest PKD was reached with NB (Gaussian) in 10000 observations. The highest positive predictive values were most frequently obtained in SVM with z-score transformation. While the C5.0 algorithm reaches the highest selectivity rates with raw data at 100 observations, the superiority of EG over other transformations or raw data stands out, especially for unbalanced class distributions at 1000 observations.

In the data generated from the F distribution, it was seen that NB reached the highest PKD in the 0.4-0.5 class distributions for observation numbers 100 to 500 and in the 0.3-0.4-0.5 class distributions for 1000 to 10000 observation numbers. In all remaining unbalanced class distributions, SVM showed the highest success. While the highest PKD was achieved with raw data in 100 observations in NB and EF in other observations, the minimum, maximum normalisation success rate in SVM is noteworthy. In the C5.0 algorithm, it is seen that PKD success is most often achieved with raw data.

The study includes the findings regarding the highest negative predictive values (NCV) obtained with NB, C5.0 and SVM algorithms and how they were obtained in Table 9.

In the data produced from the normal distribution, there was no increase compared to the NKV obtained from the raw data after data transformation in the NB and C5.0 algorithms. The highest NKD in all data types was obtained with NB (Gaussian). Looking at the SVM findings, it can be seen that the superiority of z-score normalisation over minimum and maximum increases as the number of observations increases and the balance ratio in classes decreases.

There was no increase in the data produced from the chi-square distribution compared to the NVC obtained from the raw data after the data transformation in the C5.0 algorithm. The highest NCD was obtained with SVM in 100 observations and all data types except 0.1 class distribution. The minimum-maximum normalisation effect was high in reaching the highest values in SVM. In NB, high values were most frequently reached with raw data.

In the data generated from the F distribution, the data transformation in the C5.0 algorithm did not have an increasing effect on NCD. It was observed that the highest NCD was achieved with C5.0 in the 0.1 class distribution for observation numbers between 1000 and 10000, and NB (EF) was achieved for all the remaining data types. The effect of EF discrimination is noteworthy in SVM and NB.

IV.2 DISCUSSION

In the current study, results were obtained on the raw data produced by simulation from different sample distributions, at different numbers of observations and different equilibrium distribution ratios of classes, as well as on the data obtained with four different transformation methods using NB, C5.0 and support vector machine classification algorithms. The performances of the three algorithms after data transformation and the raw data were compared with six different performance evaluation criteria [27].

In the study, when NB, C5.0 and SVM algorithms were compared in terms of general performance, it was seen that NB gave more successful results in data produced from a normal distribution and F distribution, and SVM gave more successful results in data produced from the chi-square distribution. When the results obtained from data generated from three sample distributions and without any pre-processing were compared with the results obtained after normalisation, it was seen that normalisation methods generally did not increase the classification performance of the NB algorithm. In the C5.0 decision tree algorithm, normalisation generally did not make any

difference compared to the results obtained from the raw data. In SVM, normalisation transformations generally increased all performance measures except the selectivity rate [28].

Kumar et al. (2019) [29], [30] investigated the all the roles of the linear discriminant analysis, k-nearest neighbour, NB, SVM, decision trees, random forest, and multilayer perceptron algorithms in performing minimum-maximum normalisation on Alzheimer's data. In their study, NB gave better results in the overall performance comparison of the algorithms. He reported that minimum-maximum normalisation did not affect the performance of decision trees and NB algorithms much, while the SVM algorithm significantly improved its performance.

Siledar and Chaudhary (2017) [31] investigated the effect of z-score normalisation in the decision tree (C4.5) and NB on the 24-variable "Credit" data with 30000 observations obtained from the UCI data warehouse. They reported no change in the decision tree (C4.5) after Z-score normalisation, but there was an increase in NB. In our study, there was no increase in NB after z-score normalisation. It is thought that the reason for this difference is that we prefer the "Kernel" probability density function instead of "Gaussian" for data that is not normally distributed.

The current study showed that the z-score transformation was more effective in increasing the performance, especially with the increase in the number of observations in the data produced from the normal distribution in the SVM algorithm. In data generated from chi-square distribution, z-score normalisation gave higher results in a fully balanced class distribution, and minimum-maximum normalisation gave higher results in data with an unbalanced distribution of classes. While minimum-maximum normalisation caused a decrease in performance in data produced from the F distribution, more successful results were obtained with z-score normalisation in data larger than 100 observations.

In our study, selectivity rates differed from other performance measures in SVM regarding conversion results. While other performance criteria in the SVM algorithm generally increased with normalisation, there was an increase in selectivity rates only in fully balanced class distributions. There was generally no increase in normally distributed data in unbalanced class distributions, but an increase was achieved in chi-square distribution data with 100 observations. Higher selectivity rates were achieved with minimum-maximum normalisation in the F distribution without changing according to the number of observations in unbalanced class distributions. There was no increase in C5.0 and NB as in other performance criteria.

Suma et al. (2016) [32] compared normalisation methods on the “Breast” data obtained from the UCI data warehouse regarding general accuracy rate, sensitivity and selectivity rates in linear discriminant analysis, SVM, NB, artificial neural networks and decision tree algorithms. In their study, they achieved the best performance with NB. While there was no increase in performance values with minimum-maximum normalisation in NB, it was shown that there was an increase in general accuracy rates with z-score normalisation. They showed that both methods increased the accuracy rates equally for SVM, while the minimum-maximum normalisation increased the sensitivity and selectivity rates. When Suma et al.'s study is compared with the NB findings in our study, it is thought that this difference is due to our preference for the "Kernel" probability density function instead of "Gaussian" for data that is not normally distributed.

In our study, it was observed that there was a decrease in the performance of NB in the data produced from normal distribution and chi-square distribution with unsupervised discrimination methods. More successful results were obtained in the NB classification, in which the data generated from the F distribution and converted to categorical form by the equal frequency spacing (EF) method, one of the unsupervised discrimination methods, was compared to the results obtained from the raw data.

Wang et al. (2023) [33] compared the error rates of different discrimination methods in the classification result of the NB algorithm on 35 data sets obtained from the UCI data warehouse. In their comparison, it was seen that EF provided lower error rates in NB compared to EG.

The current study observed that unsupervised discrimination methods caused a decrease in the accuracy rate, AUC value, sensitivity rate and negative predictive value performances of the C5.0 algorithm. When classification performances were considered in terms of selectivity rate, it was seen that EG gave better results, especially with the increase in the number of observations in data where classes were unevenly distributed. When the effect of data transformations on performance in terms of positive predictive value is examined, the superiority of EG draws attention to data produced from normal and chi-square distribution, in which classes are unevenly distributed and with several observations of 1000 and above.

Jiang et al. (2018) [17] used data from the UCI data warehouse to compare the results before discrimination in different classifiers with the overall accuracy rates obtained after applying different discrimination methods. This study showed that while there was a decrease in general accuracy rates in decision trees (C4.5) with EG, one of the discrimination methods, compared to before discrimination, there was an increase in NB. Our study's overall accuracy rates decreased with EG in the C5.0 decision tree algorithm and NB. The reason for the contradiction in the NB findings is thought to be the use of the "Kernel" probability density function instead of "Gaussian" for data that is not normally distributed.

Our study achieved an increase in SVM performance with unsupervised discrimination methods. EF discrimination was more effective than EG in SVM performance in data generated from normal distribution and F distribution. Generally, higher performances were obtained with EG in the discrimination of data generated from the chi-square distribution. According to the performance values obtained after discrimination, NB showed higher success in data produced from a normal distribution and F distribution, and SVM showed higher success in data produced from chi-square.

In the literature review [34], no study examining normalisation and discrimination methods together was found. When the normalisation and discrimination methods are examined in our study, the superiority of the transformation methods over each other differs in the classification algorithms of NB, C5.0 and SVM. In the C5.0 algorithm, no increase in performance was generally achieved with either method. Better results were obtained with the EF discrimination method in the data produced from the F distribution in NB. In SVM, a change was observed in the superiority of normalisation and discrimination methods over each other according to the data distribution. In SVM, z-score helped to achieve the best results in the majority of data produced from a normal distribution, and minimum-

maximum normalisation helped to achieve the best results in the majority of data produced from the chi-square distribution. In the data generated from the F distribution, EF discrimination was more effective in increasing the performance of SVM.

V. CONCLUSIONS

In our study, the algorithm most affected by the normalisation and discrimination methods applied before classification was SVM, while the C5.0 decision tree algorithm was determined to be the least affected algorithm. The C5.0 algorithm gave more robust results regarding data transformations in the face of data that varied in terms of data distribution, number of observations and class distribution ratio. It was seen that both normalisation and discrimination methods were influential in the performance of SVM and contributed to better results. The advantages of normalisation methods over each other varied depending on the distribution of the data, the number of observations, and whether the classes were evenly distributed. As the number of observations increased in normally distributed data, the effect of z-score transformation in improving performance also increased. In data generated from chi-square distribution, z-score normalisation gave better results in fully balanced class distributions, and minimum-maximum normalisation gave better results in unbalanced class distributions. In the data generated from the F distribution, more successful results were obtained in the data transformed with z-score normalisation compared to the minimum-maximum normalisation. In the transformations made with discrimination, there was a change in the superiority of the two methods over each other according to the data distribution patterns. EF discrimination was more effective in SVM performance than EG for data generated from normal distribution and F distribution. Generally, higher performances were obtained with EG in the discrimination of data generated from the chi-square distribution. According to the classification success achieved with SVM, among both approaches, normalisation methods are more effective in average and chi-square distribution, and EF unsupervised discrimination method is more effective in F distribution. In the NB algorithm, there was generally no increase in performance values with normalisation methods. The effect of unsupervised discrimination methods was seen in the classification made with the data produced from the F distribution. Better results were obtained from the data converted with the EF method compared to the raw data not subjected to pre-processing. When the advantages of the classification algorithms used in the study are examined, the data produced from the normal distribution and F distribution are, respectively. It was observed that normalisation methods and SVM gave more successful results in the data produced from raw data and EF and NB, chi-square distribution.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Hussein Ali Attallah, Ahmed Al-Asadi.
Methodology: Hussein Ali Attallah, Sadeer Sadeq.
Investigation: Hussein Ali Attallah, Ahmed Al-Asadi.
Discussion of results: All Authors.
Writing – Original Draft: Hussein Ali Attallah, Sadeer Sadeq.
Writing – Review and Editing: Ahmed Al-Asadi, Sadeer Sadeq.
Supervision: Hussein Ali Attallah, Ahmed Al-Asadi.
Approval of the final text: Ahmed Al-Asadi, Sadeer Sadeq.

VII. REFERENCES

[1] I. Masic, "Medical Decision Making - an Overview," Acta Inform Med, vol. 30, pp. 230-235, Sep 2022. doi:https://doi.org/10.5455/aim.2022.30.230-235.

[2] K. Kim and Y. M. Lee, "Understanding uncertainty in medicine: concepts and implications in medical education," Korean J Med Educ, vol. 30, pp. 181-188, Sep 2018. doi:https://doi.org/10.3946/kjme.2018.92.

[3] W.-T. Wu, Y.-J. Li, A.-Z. Feng, L. Li, T. Huang, A.-D. Xu, et al., "Data mining in clinical big data: the frequently used databases, steps, and methodological models," Military Medical Research, vol. 8, p. 44, 2021-08-11 2021. doi:https://doi.org/10.1186/s40779-021-00338-z.

[4] Y. A. Enaya, M. Jawad Mohammed, and G. A. Bilal, "Password-free Authentication for Smartphone Touchscreen Based on Finger Size Pattern," International Journal of Interactive Mobile Technologies (iJIM), vol. 14, pp. pp. 163-179, 11/24 2020. doi:https://doi.org/10.3991/ijim.v14i19.17239.

[5] S. S. Bafjaish, "Comparative analysis of Naive Bayesian techniques in health-related for classification task," Journal of Soft Computing and Data Mining, vol. 1, pp. 1-10, 2020. doi:https://doi.org/10.30880/jscdm.2020.01.02.001.

[6] S. Sucipto, D. D. Prasetya, and T. Widiyaningtyas, "A review questions classification based on Bloom taxonomy using a data mining approach," ITEGAM-JETIA, vol. 10, pp. 161-170, 2024. doi:https://doi.org/10.5935/jetia.v10i48.1204.

[7] Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," Shanghai Arch Psychiatry, vol. 27, pp. 130-5, Apr 25 2015. doi:https://doi.org/10.11919/j.issn.1002-0829.215044.

[8] L. V. Vaz, M. C. Gonçalves, I. C. P. Dias, and E. O. B. Nara, "Application of a production planning model based on linear programming and machine learning techniques," ITEGAM-JETIA, vol. 10, pp. 17-29, 2024. doi:https://doi.org/10.5935/jetia.v10i45.920.

[9] E. E. Ogheneovo and P. A. Nlerum, "Iterative Dichotomizer 3 (ID3) Decision Tree: A Machine Learning Algorithm for Data Classification and Predictive Analysis," International Journal of Advanced Engineering Research and Science, vol. 7, pp. 514-521, 2020 2020. doi:https://doi.org/10.22161/ijaers.74.60.

[10] Z. Zhang, Z. Zhao, D.-S. Yeom, and Z. Lv, "Decision Tree Algorithm-Based Model and Computer Simulation for Evaluating the Effectiveness of Physical Education in Universities," Complexity, vol. 2020, pp. 1-11, 2020-10-28 2020. doi:https://doi.org/10.1155/2020/8868793.

[11] U. Stańczyk, B. Zielosko, and G. Baron, "Significance of Single-Interval Discrete Attributes: Case Study on Two-Level Discretisation," Applied Sciences, vol. 14, p. 4088, 2024-05-11 2024. doi:https://doi.org/10.3390/app14104088.

[12] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review," Healthcare (Basel), vol. 10, p. 541, Mar 15 2022. doi:https://doi.org/10.3390/healthcare10030541.

[13] Y. Zhang, Y. Xin, Q. Li, J. Ma, S. Li, X. Lv, et al., "Empirical study of seven data mining algorithms on different characteristics of datasets for biomedical classification applications," BioMedical Engineering OnLine, vol. 16, p. 125, 12/2017 2017. doi:https://doi.org/10.1186/s12938-017-0416-x.

[14] L. Wang, Y. Xie, M. Pang, and J. Wei, "Alleviating the attribute conditional independence and I.I.D. assumptions of averaged one-dependence estimator by double weighting," Knowledge-Based Systems, vol. 250, p. 109078, 08/2022 2022. doi:https://doi.org/10.1016/j.knosys.2022.109078.

[15] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," Journal of Information Science, vol. 44, pp. 48-59, 02/2018 2018. doi:https://doi.org/10.1177/0165551516677946.

[16] D. Berrar, "Bayes' Theorem and Naive Bayes Classifier," in Encyclopedia of Bioinformatics and Computational Biology, ed: Elsevier, 2019, pp. 403-412. doi:https://doi.org/10.1016/B978-0-12-809633-8.20473-1.

[17] W. Jiang, Y. Shen, Y. Ding, C. Ye, Y. Zheng, P. Zhao, et al., "A naive Bayes algorithm for tissue origin diagnosis (TOD-Bayes) of synchronous multifocal tumors in the hepatobiliary and pancreatic system," Int J Cancer, vol. 142, pp. 357-368, Jan 15 2018. doi:https://doi.org/10.1002/ijc.31054.

[18] F. Hajjej, M. A. Alohali, M. Badr, and M. A. Rahman, "A Comparison of Decision Tree Algorithms in the Assessment of Biomedical Data," Biomed Res Int, vol. 2022, p. 9449497, 2022-7-7 2022. doi:https://doi.org/10.1155/2022/9449497.

[19] A. M. Ahmed, A. Rizaner, and A. H. Ulusoy, "A Decision Tree Algorithm Combined with Linear Regression for Data Classification," in 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), 2018, pp. 1-5. doi:https://doi.org/10.1109/ICCCEEE.2018.8515759.

[20] M. Arif, "Decision Tree Algorithms C4.5 and C5.0 in Data Mining: A Review," International Journal of Database Theory and Application, vol. 11, pp. 1-8, 2018-03-31 2018. doi:https://doi.org/10.14257/ijdta.2018.11.1.01.

[21] O. B. Ali, S. Hammami, M. Hasni, F. H'Mida, and A. N. S. Moh, "Using Machine Learning to evaluate Industry 4.0 Maturity: A comprehensive analysis highlighting Lean's impact on Digital Transformation," ITEGAM-JETIA, vol. 10, pp. 156-167, 2024. doi:https://doi.org/10.5935/jetia.v10i50.1262.

[22] R. Revathy and R. Lawrance, "Comparative analysis of C4. 5 and C5. 0 algorithms on crop pest data," International Journal of Innovative Research in Computer and Communication Engineering, vol. 5, pp. 50-58, 2017. https://www.researchgate.net/profile/Revathy-Rathinasamy/publication/332036647.

[23] M. Marzuki, M. Iqbal, A. Nivada, H. Sofyan, T. Usman, N. Nazaruddin, et al., "Implementation of decision tree using C5.0 algorithm in preference and electability survey results on regional head election in Aceh," Journal of Physics: Conference Series, vol. 1882, p. 012132, 2021-05-01 2021. doi:https://doi.org/10.1088/1742-6596/1882/1/012132.

[24] T. P. Bagchi, "Support Vector Machines--An Overview," January 27-30, 2022 https://www.researchgate.net/publication/358021073_Support_Vector_Machines--An_Overview.

[25] B. Said, L. Mazouz, T. T. NAAS, Ö. Yildirim, and R. D. Mohammedi, "Broken magnets fault detection in pmsm using a convolutional neural network and SVM," ITEGAM-JETIA, vol. 10, pp. 55-62, 2024. doi:https://doi.org/10.5935/jetia.v10i48.1185.

[26] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Med Inform Decis Mak, vol. 19, p. 281, Dec 21 2019. doi:https://doi.org/10.1186/s12911-019-1004-8.

[27] M. A. Al-Hashem, A. M. Alqudah, and Q. Qananwah, "Performance Evaluation of Different Machine Learning Classification Algorithms for Disease Diagnosis," International Journal of E-Health and Medical Communications, vol. 12, pp. 1-28, 2021-8-11 2021. doi:https://doi.org/10.4018/IJEHMC.20211101.oa5.

[28] V. Vapnik and R. Izmailov, "Reinforced SVM method and memorization mechanisms," Pattern Recognition, vol. 119, p. 108018, 11/2021 2021. doi:https://doi.org/10.1016/j.patcog.2021.108018.

[29] N. Kumar, J. Manhas, and V. Sharma, "A comparative analysis to visualize the behavior of different machine learning algorithms for normalized and un-normalized data in predicting Alzheimer's disease," Journal of Computational and Theoretical Nanoscience, vol. 16, pp. 3840-3848, 2019. doi:https://doi.org/10.1166/jctn.2019.8259.

[30] Y. A. Enaya, A. A. Karim, S. M. Saleh, and S. W. Shneen, "Adapting Wired TCP for Wireless Ad-hoc Networks Using Fuzzy Logic Control," Journal Européen des Systèmes Automatisés, vol. 57, p. 1377, 2024. doi:https://doi.org/10.18280/jesa.570513.

[31] S. Siledar and S. Chaudhary, "Comparative analysis of naive bays classifier and decision tree c4. 5 on credit payment data set," International Journal of Research in Engineering and Technology, vol. 6, pp. 43-44, 2017. https://www.researchgate.net/profile/Seema-Siledar/publication/328654658.

[32] V. R. Suma, S. Renjith, S. Ashok, and M. V. Judy, "Analytical Study of Selected Classification Algorithms for Clinical Dataset," Indian Journal of Science and Technology, vol. 9, 2016-03-22 2016. doi:https://doi.org/10.17485/ijst/2016/v9i11/67151.

[33] S. Wang, J. Ren, and R. Bai, "A semi-supervised adaptive discriminative discretization method improving discrimination power of regularized naive Bayes," Expert Systems with Applications, vol. 225, p. 120094, 09/2023 2023. doi:https://doi.org/10.1016/j.eswa.2023.120094.

[34] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," Information, vol. 15, p. 235, 2024-04-19 2024. doi:https://doi.org/10.3390/info15040235.

APPENDIX

APPENDIX A

Table 5. The Highest AUC Values Obtained in The Study.

AUC values												
Dispersed	**Ratio**	$N = 100$			$N = 500$			$N = 1000$			$N = 10000$		
		NB	$C5.0$	SVM	NB	NB	$C5.0$	SVM	NB	NB	$C5.0$	SVM	NB
Normal	0.1	0.899	0.681	0.891	0.953	0.749	0.931	0.956	0.777	0.934	0.96	0.862	0.95
		*,#,!,	*,#,!,	*,#,!,	*,#,!,	*,#,!,	#	*,#,!,	#	!,	*,#,!,	!,	df
	0.2	0.927	0.748	0.913	0.956	0.812	0.935	0.958	0.831	0.94	0.959	0.883	0.951
		*,#,!,	*,#,!,	*,#,!,	*,#,!,	*,#,!,	#	*,#,!,	*,#,!,	!,	*,#,!,	*,#,!,	df
	0.3	0.94	0.77	0.92	0.957	0.838	0.936	0.957	0.851	0.943	0.959	0.895	0.951
		*,#,!,	*,#,!,	*,#,!,	*,#,!,	*,#,!,	!,	*,#,!,	*,#,!,	!,	*,#,!,	*,#,!,	!,
	0.4	0.944	0.783	0.926	0.958	0.85	0.939	0.958	0.862	0.945	0.959	0.901	0.953
		*,#,!,	*,#,!,	*,#,!,	*,#,!,	*,#,!,	!,	*,#,!,	*,#,!,	!,	*,#,!,	*,#,!,	!,
	0.5	0.947	0.791	0.925	0.958	0.852	0.94	0.958	0.865	0.944	0.96	0.904	0.954
		*,#,!,	*,#,!,	*,#,!,	*,#,!,	*,#,!,	!,	*,#,!,	*,#,!,	!,	*,#,!,	*,#,!,	!,
χ^2	0.1	0.791	0.644	0.854	0.881	0.723	0.909	0.898	0.72	0.917	0.919	0.806	0.921
		a.#	*,#,!,	!,	*,#,!,	*,#,!,	#	*,#,!,	*,#,!,	#	*,#,!,	*,#,!,	#
	0.2	0.836	0.698	0.888	0.896	0.757	0.916	0.905	0.773	0.916	0.921	0.833	0.921
		*,#,!,	*,#,!,	#	*,#,!,	*,#,!,	#	*,#,!,	*,#,!,	#	*,#,!,	*,#,!,	#
	0.3	0.854	0.731	0.898	0.9	0.783	0.915	0.908	0.799	0.918	0.921	0.844	0.921
		*,#,!,	*,#,!,	#	*,#,!,	*,#,!,	#	*,#,!,	*,#,!,	#	*,#,!,	*,#,!,	#
	0.4	0.855	0.738	0.901	0.902	0.796	0.916	0.911	0.811	0.92	0.922	0.851	0.922
		*,#,!,	*,#,!,	#	*,#,!,	*,#,!,	#	*,#,!,	*,#,!,	#	*,#,!,	*,#,!,	!,.$
	0.5	0.86	0.744	0.896	0.906	0.801	0.914	0.911	0.813	0.917	0.922	0.853	0.922
		*,#,!,	*,#,!,	#	*,#,!,	*,#,!,	!,	*,#,!,	*,#,!,	!,	*,#,!,	*,#,!,	!,.$
$F - distribution$	0.1	0.851	0.694	0.763	0.936	0.808	0.898	0.954	0.859	0.925	0.97	0.902	0.964
		$	*,#,!,	$	$	*,#,!,	$	$	*,#,!,	$	$	*,#,!,	$
	0.2	0.898	0.768	0.837	0.959	0.881	0.926	0.967	0.907	0.944	0.975	0.948	0.971
		$	*,#,!,	$	$	*,#,!,	$	$	*,#,!,	$	$	a.#	$
	0.3	0.922	0.813	0.857	0.967	0.906	0.934	0.973	0.929	0.952	0.977	0.959	0.974
		*,#,!,	$	$	*,#,!,	$	$	*,#,!,	$	$	*,#,!,	$	*,#,!,
	0.4	0.921	0.822	0.857	0.969	0.915	0.938	0.974	0.934	0.953	0.978	0.962	0.975
		*,#,!,	$	$	*,#,!,	$	$	*,#,!,	$	$	*,#,!,	$	$
	0.5	0.924	0.834	0.857	0.97	0.921	0.925	0.975	0.937	0.954	0.978	0.962	0.975
		$	*,#,!,	$	$	*,#,!,	$	$	*,#,!,	$	$	*,#,!,	$

Note ∗: Raw data; #: Minimum-Maximum Normalization;
!: Z-Score Normalization; df: Equal Width Spacing Discrimination;
$: Equal Frequency Spacing Discrimination.

Source: Authors, (2025).

APPENDIX B

Table 6. The Highest Sensitivity Rates Obtained in The Study.

Sensitivity Ratios													
Distribution	**Ratio**	$N = 100$			$N = 500$			$N = 1000$			$N = 10000$		
		NB	$C5.0$	SVM	NB	NB	$C5.0$	SVM	NB	NB	$C5.0$	SVM	NB
Normal	0.1	0.475	0.349	0.331	0.606	0.435	0.573	0.622	0.477	0.612	0.636	0.555	0.623
		*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	!
	0.2	0.67	0.528	0.618	0.725	0.594	0.729	0.732	0.614	0.717	0.738	0.669	0.73
		*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	!
	0.3	0.762	0.613	0.742	0.793	0.683	0.773	0.796	0.699	0.782	0.801	0.74	0.796
		*,#,!	*,#,!	#	*,#,!	*	#	*,#,!	*,#,!	!	*,#,!	*,#,!	!
	0.4	0.823	0.69	0.827	0.843	0.743	0.822	0.841	0.758	0.832	0.845	0.792	0.842
		*,#,!	*,#,!	#	*,#,!	*,#,!	!	*,#,!	*,#,!	!	*,#,!	*,#,!	!
	0.5	0.866	0.745	0.83	0.879	0.795	0.866	0.879	0.803	0.869	0.881	0.834	0.88
		*,#,!	*,#,!	!	*,#,!	*,#,!	!	*,#,!	*,#,!	!	*,#,!	*,#,!	!
χ^2	0.1	0.248	0.264	0.172	0.369	0.336	0.408	0.39	0.36	0.468	0.444	0.399	0.522
		*,#,!	*,#,!	!	*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!.$	*,#,!	#
	0.2	0.474	0.437	0.468	0.557	0.497	0.658	0.579	0.514	0.674	0.61	0.555	0.632
		*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	#	$	*,#,!	#
	0.3	0.613	0.546	0.679	0.682	0.602	0.758	0.688	0.617	0.733	0.709	0.655	0.73
		*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	#
	0.4	0.701	0.626	0.793	0.762	0.681	0.79	0.77	0.694	0.8	0.783	0.729	0.802
		*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	#
	0.5	0.783	0.704	0.821	0.827	0.748	0.835	0.829	0.757	0.837	0.838	0.791	0.849
		*,#,!	*,#,!	#	*,#,!	*,#,!	!	*,#,!	*,#,!	!	*,#,!	*,#,!	#
$F - distribution$	0.1	0.497	0.38	0.148	0.573	0.549	0.394	0.612	0.632	0.476	0.673	0.712	0.651
		$	*,#,!	*	$	*.#	$	$	*,#,!	$	$	*,#,!	$
	0.2	0.681	0.563	0.375	0.755	0.705	0.624	0.776	0.735	0.683	0.803	0.781	0.788
		$	*,#,!	$	$	*,#,!	$	$	*,#,!	$	$	*,#,!	$
	0.3	0.772	0.67	0.539	0.832	0.775	0.738	0.842	0.793	0.782	0.853	0.83	0.845
		$	*,#,!	$	$	*,#,!	$	$	*,#,!	$	$	*,#,!	$
	0.4	0.801	0.732	0.663	0.87	0.819	0.805	0.878	0.829	0.834	0.886	0.866	0.88
		$	*,#,!	$	$	*,#,!	$	$	*,#,!	$	$	*,#,!	$
	0.5	0.838	0.784	0.762	0.901	0.851	0.843	0.907	0.863	0.874	0.913	0.89	0.907
		$	*.#	$	$	*,#,!	$	$	*,#,!	$	!	*,#,!	$

Source: Authors, (2025).

APPENDIX C

Table 7: The highest selectivity rates obtained in the study.

Selectivity rates													
Distribution	**Ratio**	$N = 100$			$N = 500$			$N = 1000$			$N = 10000$		
		NB	***C5.0***	***SVM***	***NB***	***NB***	***C5.0***	***SVM***	***NB***	***NB***	***C5.0***	***SVM***	***NB***
Normal	0.1	0.986	0.998	0.999	0.988	0.984	0.999	0.988	0.982	0.997	0.989	0.98	0.996
		*,#,!	$	*	*,#,!	$	*	*,#,!.df	$	*	*,#,!	df	*
	0.2	0.965	0.981	0.993	0.972	0.939	0.966	0.973	0.945	0.99	0.973	0.958	0.99
		*,#,!	$	*	*,#,!	df	!	*,#,!	df	*	*,#,!	df	*
	0.3	0.94	0.945	0.983	0.951	0.896	0.982	0.953	0.905	0.981	0.954	0.929	0.975
		*,#,!	$	*	*,#,!	df.$	*	*,#,!	df	*	*,#,!	df	*
	0.4	0.907	0.808	0.96	0.928	0.85	0.961	0.928	0.863	0.969	0.93	0.895	0.935
		*,#,!	*,#,!.df	*	*,#,!	*,#,!	*	*,#,!	df	*	*,#,!	df	*
	0.5	0.876	0.762	0.86	0.895	0.803	0.873	0.897	0.818	0.88	0.9	0.857	0.896
		*,#,!	*,#,!	#	*,#,!	*,#,!	!	*,#,!	*,#,!	!	*,#,!	*,#,!.df	!
χ^2	0.1	0.975	0.999	0.99	0.983	0.991	0.995	0.985	0.987	0.994	0.985	0.98	0.992
		*,#,!	$	!	*,#,!	$	$	*,#,!	$	*	*,#,!	df	*
	0.2	0.939	0.988	0.985	0.955	0.941	0.974	0.958	0.943	0.969	0.96	0.944	0.966
		*,#,!	$	$	*,#,!	df	*	*,#,!	df	*	*,#,!	df	*
	0.3	0.899	0.944	0.947	0.92	0.885	0.944	0.926	0.89	0.935	0.929	0.901	0.936
		*,#,!	$	df	*,#,!	df	*	*,#,!	df	*	*,#,!	df	*
	0.4	0.844	0.781	0.858	0.879	0.819	0.9	0.885	0.829	0.888	0.893	0.854	0.891
		*,#,!	df	df	*,#,!	df	*	*,#,!	$	*	*,#,!	$	*
	0.5	0.777	0.707	0.801	0.827	0.751	0.836	0.834	0.766	0.837	0.848	0.796	0.844
		*,#,!	*,#,!	#	*,#,!	*,#,!	!	*,#,!	*,#,!	!	*,#,!	*,#,!	!.$
$F - distribution$	0.1	0.982	0.995	0.998	0.991	0.993	0.998	0.992	0.994	0.998	0.996	0.996	0.998
		df	$	#.!	df	df	#	df	df	#	df	df	#
	0.2	0.962	0.975	0.991	0.978	0.977	0.993	0.981	0.982	0.993	0.991	0.992	0.995
		df	df	#	df	df	#	df	df	#	df	df	#
	0.3	0.924	0.934	0.976	0.95	0.948	0.979	0.957	0.958	0.981	0.98	0.979	0.981
		*.!	df	#	*.!	df	#	$	df	#	df	df	#
	0.4	0.905	0.842	0.945	0.934	0.897	0.925	0.938	0.911	0.926	0.946	0.937	0.943
		*	*,#,!	#	*.!	$	#	$	$	#	$	df	$
	0.5	0.875	0.789	0.782	0.908	0.867	0.857	0.917	0.887	0.888	0.925	0.911	0.921
		*.!	*,#,!	$	*.!	*,#,!	$	$	$	$	$	*,#,!	$

Source: Authors, (2025).

APPENDIX D

Table 8. The highest positive predictive values obtained in the study.

Positive Predictive Values												
Distribution	**Ratio**	$N = 100$			$N = 500$			$N = 1000$			$N = 10000$		
		NB	$C5.0$	SVM	NB	NB	$C5.0$	SVM	NB	NB	$C5.0$	SVM	NB
Normal	0.1	0.828	0.471	0.837	0.858	0.604	0.826	0.86	0.652	0.83	0.865	0.742	0.851
		*,#,!	*,#,!	#	*,#,!	df	!	*,#,!	df	!	*,#,!	df	!
	0.2	0.853	0.603	0.825	0.872	0.677	0.842	0.874	0.716	0.849	0.873	0.79	0.866
		*,#,!	*,#,!	#	*,#,!	df	!	*,#,!	df	!	*,#,!	df	!
	0.3	0.863	0.678	0.841	0.876	0.731	0.853	0.88	0.75	0.86	0.881	0.814	0.877
		*,#,!	*,#,!	!	*,#,!	*,#,!	!	*,#,!	df	!	*,#,!	df	!
	0.4	0.868	0.724	0.844	0.888	0.772	0.862	0.887	0.785	0.871	0.889	0.833	0.885
		*,#,!	*,#,!	!	*,#,!	*,#,!	!	*,#,!	*,#,!	!	*,#,!	df	!
	0.5	0.883	0.773	0.865	0.895	0.805	0.874	0.89	0.817	0.88	0.898	0.854	0.895
		*,#,!	*,#,!	#	*,#,!	*,#,!	!	*,#,!	*,#,!	!	*,#,!	*,#,!	!
χ^2	0.1	0.559	0.362	0.718	0.721	0.47	0.744	0.749	0.524	0.753	0.77	0.626	0.759
		*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	df	!	*,#,!	df	!
	0.2	0.694	0.516	0.772	0.765	0.604	0.775	0.78	0.634	0.778	0.793	0.688	0.789
		*,#,!	*,#,!	!	*,#,!	$	!	*,#,!	df	!	*,#,!	df	!
	0.3	0.743	0.602	0.782	0.789	0.662	0.798	0.801	0.682	0.803	0.812	0.731	0.81
		*,#,!	*,#,!	!	*,#,!	df	!	*,#,!	df	!	*,#,!	$	!
	0.4	0.767	0.669	0.802	0.811	0.713	0.817	0.819	0.724	0.82	0.829	0.765	0.827
		*,#,!	*,#,!	!	*,#,!	#	!	*,#,!	*,#,!	!	*,#,!	$	!
	0.5	0.789	0.718	0.817	0.829	0.753	0.837	0.834	0.766	0.839	0.846	0.793	0.844
		*,#,!	*,#,!	#	*,#,!	#	!	#.!	*,#,!	!	*,#,!	*,#,!	!
$F - distribution$	0.1	0.538	0.576	0.894	0.756	0.786	0.94	0.795	0.833	0.938	0.871	0.874	0.935
		$	*,#,!	!	$	df	#	$	!	#	*.!	df	#
	0.2	0.698	0.708	0.879	0.846	0.794	0.879	0.857	0.833	0.923	0.893	0.88	0.934
		$	$	#	$	df	!	$	*,#,!	#	#.!	!	#
	0.3	0.786	0.754	0.859	0.879	0.84	0.884	0.894	0.84	0.891	0.904	0.883	0.902
		*	*,#,!	#	$	$	#	$	*,#,!	#	$	*.#	$
	0.4	0.829	0.777	0.824	0.898	0.84	0.851	0.906	0.862	0.875	0.916	0.895	0.911
		*	*,#,!	#	$	$	$	$	*,#,!	$	$	*,#,!	$
	0.5	0.861	0.802	0.79	0.908	0.868	0.856	0.917	0.886	0.887	0.924	0.91	0.92
		*	*,#,!	$	$	*,#,!	$	$	*,#,!	$	$	*,#,!	$

Source: Authors, (2025).

APPENDIX E

Table 9. The highest negatives prediction values were obtained in the study.

Negative Predictive Values													
Distribution	**Ratio**	$N = 100$			$N = 500$			$N = 1000$			$N = 10000$		
		NB	***C5.0***	***SVM***	***NB***	***C5.0***	***SVM***	***NB***	***C5.0***	***SVM***	***NB***	***C5.0***	***SVM***
Normal	0.1	0.946	0.93	0.932	0.958	0.939	0.954	0.959	0.943	0.958	0.961	0.952	0.959
		*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	!
	0.2	0.924	0.887	0.913	0.935	0.902	0.934	0.936	0.907	0.932	0.937	0.92	0.935
		*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	!
	0.3	0.907	0.844	0.897	0.916	0.868	0.906	0.917	0.875	0.91	0.918	0.892	0.916
		*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	!	*,#,!	*,#,!	!
	0.4	0.891	0.805	0.889	0.9	0.834	0.886	0.898	0.843	0.892	0.9	0.866	0.898
		*,#,!	*,#,!	#	*,#,!	*,#,!	!	*,#,!	*,#,!	!	*,#,!	*,#,!	!
	0.5	0.877	0.763	0.844	0.883	0.799	0.869	0.882	0.808	0.872	0.884	0.838	0.882
		*,#,!	*,#,!	!	*,#,!	*,#,!	!	*,#,!	*,#,!	!	*,#,!	*,#,!	!
χ^2	0.1	0.924	0.921	0.916	0.934	0.929	0.938	0.936	0.931	0.943	0.941	0.936	0.948
		$	*,#,!	#.!	*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!.$	*,#,!	#
	0.2	0.88	0.865	0.882	0.897	0.879	0.917	0.901	0.883	0.92	0.908	0.894	0.912
		*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	#
	0.3	0.849	0.815	0.874	0.872	0.836	0.897	0.874	0.842	0.889	0.882	0.858	0.888
		*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	#
	0.4	0.815	0.767	0.868	0.848	0.795	0.864	0.853	0.802	0.869	0.86	0.825	0.869
		*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	#	*,#,!	*,#,!	#
	0.5	0.792	0.716	0.827	0.829	0.752	0.838	0.831	0.761	0.838	0.84	0.792	0.847
		*,#,!	*,#,!	#	*,#,!	*,#,!	!	*,#,!	*,#,!	!	*,#,!	*,#,!	#
$F - distribution$	0.1	0.945	0.934	0.914	0.954	0.951	0.936	0.958	0.96	0.944	0.964	0.969	0.962
		$	*,#,!	*.!	$	*,#,!	$	$	*,#,!	$	*,#,!.$	*,#,!	$
	0.2	0.923	0.897	0.864	0.941	0.928	0.912	0.946	0.936	0.924	0.952	0.947	0.948
		$	*,#,!	$	$	*,#,!	$	$	*,#,!	$	$	*,#,!	$
	0.3	0.906	0.869	0.829	0.93	0.907	0.894	0.934	0.914	0.91	0.939	0.929	0.935
		$	*,#,!	$	$	*,#,!	$	$	*,#,!	$	$	*,#,!	$
	0.4	0.875	0.833	0.802	0.916	0.883	0.876	0.921	0.89	0.893	0.926	0.913	0.922
		$	*,#,!	$	$	*,#,!	$	$	*,#,!	$	$	!	$
	0.5	0.85	0.797	0.78	0.903	0.856	0.846	0.908	0.869	0.877	0.913	0.893	0.908
		$	*,#,!	$	$	*,#,!	$	$	*.#	$	*,#,!	*,#,!	$

Source: Authors, (2025)