Predictive Model for Students Admission Uncertainty Using Naïve Bayes Classifier and Kernel Density Estimation (KDE)

https://doi.org/10.3991/ijet.v17i08.29827

Nasim Matar(✉), Wasef Matar, Tirad Al Malahmeh
University of Petra, Amman, Jordan
nmatar@uop.edu.jo

Abstract—Uncertainty of getting admission into universities / institutions is one of the global challenges in academic environment. The students are having good marks with high credential but not sure about getting their admission into universities / institutions. In this research study the researcher built a predictive model using Naïve Bayes Classifiers –machine learning algorithm to extract and analyze hidden pattern in students' academic records and their credentials. The main objective of this research study is to reduce uncertainty for getting admission into universities / institutions on the basis of their previous credentials and some other essentials parameters. This research study presents a joint venture of Naïve Bayes Classification and Kernel Density Estimations (KDE) approach to predict student's admission into universities or any higher institutions. The predictive model is built on training dataset of students' examination score such as GPA, GRE, RANK and some other essential features that offered the admission with a predictor accuracy rate of 72% and has been experimentally verified. To improve the quality of accuracy of predictive model the researcher used the Shapiro-Walk Normality Test and Gaussian distribution on large datasets. The predictive model helps in reducing the admission uncertainty and enhances the universities decision making capabilities. The significance of this research study is to reduce human intervention for making decisions with respect to students' admission into universities or any higher academic institutions, and it demonstrates that many universities and higher-level institutions could use this predictive model to improve their admission process without human intervention.

Keywords—KAPPA, machine learning, KDE, WEKA

1 Introduction

Admission uncertainty is one of the major concerns in universities or higher institutions in today's academic environment. Universities / institutions are having parameters and organization constraints for the admission process such as through the entrance exam, GRE and other criteria which are essential to the admission process [2]. Different number of students are applying for getting admission into universities / institutions on the basis of their credentials and GRE scores. Uncertainty of getting admission is one

of the major problems amongst students and their parents, and it is one of the critical decision-making processes for the managerial committee [18]. In order to solve uncertainty problems related to getting admission into universities / institutions, machine learning algorithms are playing significant roles for confirming and giving assurance for getting student admission decisions [18].

Implementation of Machine learning techniques and tools is a fast-growing predictive analysis with vast application in universities and higher education institutions [22]. The importance of machine learning algorithms is to find interesting and hidden pattern in volumes of data. Academic authorities and students could benefit from the application of machine learning techniques on large datasets from universities [12]. One area of concern in universities is to provide students with necessary guidance in the learning process. Machine learning algorithms such as Decision Tree, Neural Networks, Regression, Logistic Regression, and Clustering are the useful techniques for discovering the knowledge in a large dataset that can be used for prediction regarding admission of students into universities.

Students uncertainty of admission into universities or higher educational institutions after giving their examination is one of the significant research issues, as research can be formulated on how things can be simulated in order to provide some convenient environment to students in regard of their chances of getting admission into universities [19]. Universities must have predictive model that is able of providing prediction related to chances of admissions on the basis of students examination score into universities or higher educational environment [9].

The researcher developed a very simple predictive model to predict the chances of admission into universities on the basis of examining scores and some essentials credentials. The proposed model breaks the uncertainty in the process of getting admission in universities or higher educational institutions [20]. This research study is focused on one of the machine learning algorithms, Naïve Bayes algorithm in order to predict student's status of either getting or failing admission in universities. Naïve Bayes algorithm is the predictor classification algorithms and used to study the historical data to generate comprehensive and precise analysis [19].

In this research study the researcher used the Kernel Density Estimation (KDE), which is a non-parametric way to estimate the probability of a random variables. The fundamental approach of kernel density estimation is handling data smoothing problems where inferences about samples are made, which is based on finite data samples (Lestari, 2020). The Kernel density estimation is one of the probability techniques used for estimation that must be enabled for data in order to have a better analysis of the studied probability distribution in comparison with the use of traditional histogram [11].

This research study is based on predictive model using Naïve Bayes Classifiers to predict student's admission in case of uncertainty into universities / institutions. Admission uncertainty is one of the significant research issues related to universities / institutions, for taking decisions with respect to student's admission. The researcher proposed and built a predictive model to reduce admission uncertainty for getting admission into universities / institutions. This research found that the use of Naïve Bayes Classifiers is more suitable and efficient machine learning algorithm for prediction with

accuracy level of 72 % which is more reliable and beneficial towards decision by universities / Institutions.

The researcher contributed significant efforts to formulate the research problem after reviewing several researcher articles found in the literature and current research issues and interests. The researcher investigated literature review and built predictive model using Naïve Bayes Classifiers to predict the admission into universities / Institutions.

This research study is divided into seven sections, the introduction section describes the basic and elementary information about data attributes and showing the research areas, the literature section is divided into two sub section, the subsection- A gives the theoretical background of research study whereas subsection- B gives the machine learning algorithms that are significant towards this research study. In section three the researcher proposed a framework of the research study with variables description, in section four the researcher formulated the research problem objectives. Section five is based on research design, methodology and provides details about different phases of Naïve Bayes Classifiers. Section six shows results and data analysis report, section seven is related to the summary, conclusion, limitation, and future research study.

2 Literature review

2.1 Theoretical background of predictive model

In a recent research study by [2], the researcher stated that admission uncertainty is one of the major concerns for getting student admission into universities / institutions. Moreover, his research study worked on having accurate prediction by employing the use of Naïve Bayes Classifier that is well known for generating minimum error towards the decision making process, the researcher developed a prediction model with accuracy of 97% using WEKA for data processing. Another study emphasized that the use of machine learning is multidisciplinary approach in the field of higher educational management system, and it have a broad concern about knowledge discovery and pattern identification. The researcher developed a proposed model using Naïve Bayes classifier to predict the quality of education to evaluate the rules which are studied for educational performance [8].

Based on [6] study, they focused on educational data mining in order to discover a new knowledge for academic counselling and improvement, through applying clustering algorithms on educational data. Their study employed Naïve Bayes Classification approach to predict students' performance. The predictor accuracy rate was 88% and deployed a proposed model for academic performance. According to [19], their study proposed manual classification for individuals with respect to different categories used in order to show the classification model of educational qualifications. The binary classification problems which are based on machine learning algorithms was used for commitments decisions on whether a student is going to be offered admission or not. The performance of the proposed algorithms was based on using the metrics of accuracy, precision, recall and operator curve. The use of machine learning algorithms made im-

provements on accuracy through providing classes with different category of qualification [12]. In another study, the researcher investigated machine learning techniques for deciding the choices path of students, by having dataset from universities. The results showed that the use of Naïve Bayes Classifier is one of the best techniques to predict academic performance with respect to student's admission [20].

The Naïve Bayes Classifier is a powerful algorithm to forecast students' performance in order to relate it to the prediction of university's management decision towards decision making process [1]. On the other hand, another research study focused on the educational data mining and the upcoming trends in the field of research. It was found that the educational data mining helps the educational establishments to make decisions related to student's performance. The researcher used the Naïve Bays Classification model to provide more accuracy over another method like regression, decision tree , comparison and prediction [15].

2.2 Machine learning methods for predictive analysis

Machine learning techniques are playing one of the significant roles for constructing classification models that could predict student's admission in any academic exposure. Different research studies proposed the use of Naïve Bayes Classification model for its accuracy of prediction [13]. The Naïve Bayes classifier provides accurate prediction results with a minimum error if compared with all other machine learning algorithms [3]. Another proposed study, used WEKA software for data processing, and it showed the accuracy of 97%, it also emphasized that the students feedback data can be used for decision making by the academic council. Moreover, they stressed out that employing machine learning can help in monitoring the effectiveness of academic programs. The researcher used K- Means clustering algorithms to outline the similarity and dissimilarity of students feedback [6]. In a study by [9], his research focused on using machine learning algorithms to predict academic performance in universities. The researcher developed a model for comparing the efficiency of other models. The researcher emphasized that C4.5 model is more suitable and efficient in predicting learner's academic performance. The researcher stated that using machine learning algorithms are of great value for educational planning. In another research study, [18], developed a predictive tool that classifies students' academic performance and enables faculty members to observe their grades and performances. The researcher studied the machine learning technique to predict employability skills that are based on students' performance in their academic exposure. The predicted model verifies the accuracy level at 67.67%. According to [10] research study, they developed machine learning techniques to predict students' grades in different courses, the researcher used matrix factorization, regression and classification model to evaluate the performance of students who got admission in graduates programs.

Another research study by [23], developed a predictive model for forecasting the future performance of students for academic excellence in their academic exposures using logistics regression. The researchers found that the forecasting model is having more accuracy and efficiency to predict the future performance of students, and those results were supported by another research study by [9].

The use of Naïve Bayes classifier as predictive model was used to measure student's performance through their study. The developed predictive model was more accurate and giving high level of efficiency in students' career exposure [19].

The previous research studies shows that Machine learning algorithms are having much efficiency in predicting students' performance in universities and higher educational environment. In another research study, the researcher used Naïve Bayes Algorithm on two different datasets with features engineered version. The study found that Naïve Bayes Algorithm is one of the best techniques to predict 98% accuracy with first dataset and 78% accuracy with second dataset [17]. Another study by [21], proposed the use of machine learning algorithms to predict the academic performance with respect to subject wise and semester wise. Their study employed Naïve Bayes Classifier and C4.5 decision tree techniques to explore the predictive model to measure students' performance in academic pursuit [21]. Another research study by [9], employed the use of Neural Network model for prediction of academic performance with respect to different class categories. their predictive used model generated high level of accuracy and efficiency. On the other hand, the use of Naïve Bayes Algorithms towards students achievement prediction was 84.3% accurate and this result is considered much better in comparison to the previously mentioned machine learning algorithms [9]. According to [22], his research study outlined the benefits of using data mining technique for predicting student's performance in universities. The results of his research study emphasized on the benefits to improve students' academic excellence through the use of classification algorithms such as decision trees. His research study analyzed the accuracy of predictive model and decision tree algorithms C4.5 that was based on internal assessment of students to predict their performance in final examination [14]. The accuracy of decision tree technique can be employed for predicting academic students' performance, which can be used with students that need special attention, in order to provides better chances for appropriate counselling by their teachers on timely manner [4]. Moreover, the use of decision tree algorithm can be employed to predict dropout features of students, which will enable academic counsellor into taking the appropriate interventions [16]. On the other hand, machine learning algorithms can be used to develop a predictive model based on Bayesian Network algorithms that can assist in classifying students' performance with respect to poor and bright students [18]. In a research study by [11], he used decision tree algorithm for academic data mining, were his study focused on evaluating past data achievements in order to analyze students' performance and provide the appropriate customized help for students. Another study by [5], emphasized that using decision tree as predictive model was successful in differentiating students' performances. Moreover, the use of decision trees was able of predicting academic performance on the basis of previous achievements, which proved beneficial for students in their career exposure [7].

3 Problem statement and research objectives

Nowadays, students are facing uncertain situations for getting their admission into universities / higher academic intuitions despite their high credential and good marks.

This research formulated the uncertainty for getting admission as one of the critical research issues in academic environment. Uncertainty is one of the major research concerns in academic environment to predict student's admission into universities or any higher-level academic institutions that are committed for the academic excellence. Moreover, it is becoming demanding for academic authorities to use new techniques and technologies to be able to predict future occurrence of different numbers and interests of students through the deployment of machine learning algorithms to study data patterns found in large datasets.

The significant research issues and objectives are:

1. To build a predictive model for getting admission into universities or higher educational institutions.
2. To evaluate the accuracy level of predicted model to provide more significance towards decision making process.

Apart from the previous objectives, this research stated the assumption between independent variables and dependent variable as it will be revealed shortly.

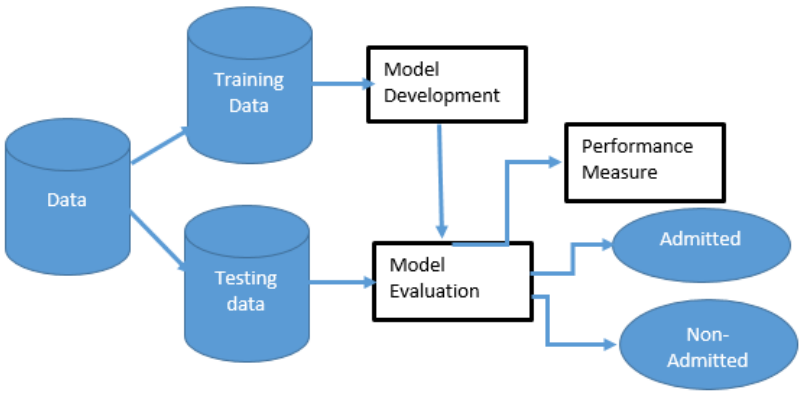

Fig. 1. Proposed predictive model using Naïve Bayes classifier

4 Proposed predictive model using Naïve Bayes classifier

The above proposed research study is based on students' datasets and their credential such as GRE, GPA, RANK and ADMIT. The specification of variables by nature that are (GRE, GPA and RANK), are the independents variables and ADMIT is the dependents variable. The researcher divided the entire datasets into two categories: training data and testing data. The ratio between training and testing data are 80:20 which is one of the standard mapping for predictive model using machine learning algorithms. The researcher proposed Naïve Bayes Classification model to predict student's admission in high level academic institutions or universities. The training datasets will be used by Naïve Bayes Classifier to evaluate accuracy level of this predictive model on the basis

of a large datasets. The reason for selecting Naïve Bayes Classifier is related to its accuracy and minimum errors.

Independents Variables: GRE, GPA, and RANK.

Dependent Variable: ADMIT (Categorical).

5 Methodology

This research is designed with the use of secondary data and machine learning algorithm- Naïve Bayes Classifiers. the secondary data provided a strong background of research study and formulated the research problem in case of admission uncertainty into universities / institutions. The researcher used the Naïve Bayes Classifiers for prediction as it was found accurate and produces minimum errors. Moreover, it was found easy, fast in providing prediction on large classes of data sets, performs better compare to other machine learning algorithms like in comparison with logistics regression [15]. The importance of this classifier is more relevance to accept categorical and numerical variables. The researcher categorized the entire research method into number of following segments.

5.1 Conditional probability model for classification

Conditional probability modeling attributed to classification can be used to solve predictive modeling problems and classify the categories of datasets as per problem specification. The Naïve Bayes classification predictive model accepts sample data as input mode such as training and testing datasets. The researcher specified the variables categories into independents and dependents variables which are referred to as X and Y. X represent Independent variables and Y represents dependent variables.

The general model of classification is such as Y=f(X), where Y is dependent variable and X is independent variables. In terms of generalizations the classification problem might have k class label and n input variables.

Y: y1, y2, y3, y4…..Yk. (Dependent Variables).

X: x1, x2, x3, x4……Xn. (Independent Variables).

The conditional probability can be calculated using joint probability, though that would be intractable. The principal of Naïve Bayes algorithm is specified in a very simple way using conditionality probability that follows as:

$$\mathrm{P}(A|B) = \frac{P(B|A)*P(A)}{P(B)} \tag{1}$$

Where researcher specify the P (A| B) probability of P (A) given by B, A and B are both independent events which could occur at any point of time. The marginal probability of event P (A) is called prior. The conditional classification and Bayes theorem are framed as:

$$P(Yi|X1, X2, X3, X4 \dots \dots Xn) \ \frac{P(X1, X2, X3 \dots Xn|Yi)}{P(X1,X2,X3 \dots Xn)} \tag{2}$$

operations are performed on each of the class labels with large probability for the given instances. The Naïve Bayes Algorithms simplification is widely used for classification and predictive modeling.

5.2 Calculations of prior and conditional probability

1. The prior probability P (Yi) is one of the straightforward estimates used by the frequency observations in the training datasets with class labels by the total number of rows in training datasets.

$$P(Yi) = \frac{\text{Occurrence of Yi}}{\text{Total Occurrence}} \tag{3}$$

2. The features value of conditional probability can also be estimated from the datasets with K-class, and K*n different probability combinations will be created with different number of datasets.

5.3 Naïve Bayesian classification algorithms

Naïve Bayesian Algorithms is a machine learning technique that is used to predict the classification model that is based on independents variables and correlated with dependent variables. The dependent variables would be categorical such admitted or not admitted students and represented by 0, 1. The use of machine learning algorithm provides one of the powerful tools and techniques to classify the dependent variables and predict results on the basis of appropriate datasets as input parameters.

The efficiency of Naïve Bayes Classification model is measures by Kernel density estimation (KDE) that is based on conditional probability and maximum likelihood of occurrences. This predictive algorithm is completed by three different stages.

1. Define the training datasets with N number of data items called the row or tuples and represented as 'K', dimensional attributes vector M, where
   M= [M1, M2, M3, M4, …….Mk],
2. Define the number of classes as C1, C2, C3, C4, ……..Cp, As per Naïve Bayes Classifier a Tuple T belong to Cx only it has higher conditional probability than any other class Cy, where X ≠ Y,
3. P(Cx) | T > P(Cy) | T and,

$$\mathrm{P(Cx|T)} = \frac{\mathrm{P}\left(\mathrm{T}\middle|\mathrm{Cx} * \mathrm{P(Cx)}\right)}{\mathrm{P}} \tag{4}$$

The conditional dependency assumes that:

$$\mathrm{P}(M|Cx) = 1 - \mathrm{Pk}\; i\; P(Mx|Cx) = \mathrm{P}(M1) * \mathrm{P}(M2) * \mathrm{P}(M3) \ldots * \mathrm{P}(M1) \tag{5}$$

Class Cx is predicted as the output class when:

$$\mathrm{P} = (M|Cx) * P(Cx) > P(M\backslash Cy) * P(Cy), where\ 1 \leq X, Y \geq P\ and\ X \neq Y \tag{6}$$

5.4 Splitting datasets into training and testing categories

The researcher divided entire datasets into two categories that is called training and testing which comprises about.

The general form of Naïve Bayes Theorem:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)*\mathrm{P}(A)}{\mathrm{P}(B)} \tag{7}$$

The posterior probability for admitted is given by

$$\mathrm{P}(admit|data) = \frac{\mathrm{P}(admit)*\mathrm{P}(GPA|admit)*\mathrm{P}(GRE|admit)}{\mathrm{P}(data)} \tag{8}$$

And for non-admitted:

$$\mathrm{P}(!\,admit|data) = \frac{\mathrm{P}(!admit)*\mathrm{P}(GPA|!\,admit)*\mathrm{P}(GRE|!\,admit)}{\mathrm{P}(data)} \tag{9}$$

Whereas

- P(admit) / P(!admit) represents the prior probability
- P(GRE) | admit) * P(GRE) | admit) represents the likelihood, the researcher assumed Zero correlation between GPA, GRE and rank of universities

The researcher stated the second assumption is that all explanatory variables follow a normal distribution into the associated the normal density function, parameterized by corresponding estimates of mean.

The researcher used the data set with 400 entries where 325 for training dataset and 75 entries for testing dataset for the efficiency of classifier to classify the data in terms of 70:30 ratio. The entire datasets are represented as follows:

1. The complete datasets are represented by D={D1 ,D2, D3, D4…….D400},
2. The training datasets are presented as Train={D1 ,D2, D3, D4…….D325},
3. The test data are represented as Test={D326 ,D327, D328, D328…….D400},

The splitting of dataset is based on a random manner and the system automatically divided the two datasets in terms of ration 70:30 manner, which is one of the standard mapping to train and test the machine learning model.

5.5 Algorithm

The researcher used Naïve Bayes Algorithm to predict the n admission and uncertainty in academic institutions and universities.

- Starting Point
- Nc → define no. of classes
- Na → define no. of attributes
- N → define total sample
- For each class C: do calculate prior probability.

- $P(Ci) = \frac{\sum Ci}{\sum N, I \in \{1, Nc\}}$
- For each class Ci do for each attributes Aj do Calculate the conditional probability A (Ai | Ci) = ∑Ci with Ai / ∑Ci, i € {1,2,3,4,…..nc} and j ∈ {1,2,3,4,5…….na}
- For each class Ci do calculates the conditional probability of the tuple K i.e P(K | Ci)=P(A1| Ci) * P(A2 | Ci) *……..P(Ana | Ci)
- For each class Ci do calculate the posterior probability of the tuple K i.e P (Ci) * P (K | Ci). Prediction if ((P(Cp) * P(K | Cp) > (*P*(*Cq*) * *P*(*K*|*Cq*))*)* Prediction ->Cp Else Prediction -> Cq, where p, q ∈ {1, 2, 3, nc} and p≠q
- End.

This research used the kernel density estimation (KDE) to improve the efficiency of predictive model and accuracy rate of probability.

$$f\hat{}(x) = 1/n \sum_{n=1}^{n} K\left(\frac{x - x(i)}{h}\right) \tag{10}$$

The value x represents the features of the data item, estimate the f(x) integrate to 1where kernel function K is usually chosen to be a smooth uni-model function with a peak at 0, K denotes as a kernel function and its bandwidth by h, the estimate density at any point is x, n denotes the total number of observations of data points having local neighbor data point or sample. The contribution of data points X (i) to estimates some points X*.

5.6 Shapiro-Walk normality test

This research used the Shapiro-Walk Normality Test to test the hypothesis and its significant effect on dependent variable at 0.05 significant level. This test gives the probability of statistics P-Value, it compares it with the significant level 0.05, if it is less than the assumption between independents variables and dependent variable, then it is rejected and the result is significant. If it is greater than 0.05 significant level, in that case the research declares to fails the null assumption and results are not significant. In this research paper the researcher used the probability of statistics P-Value to find out whether the independent variables GRE, GPA, Rank are having significant impact on dependent variable admission or not.

Gaussian distribution:

$$fx = \frac{1}{\sigma\sqrt{2\pi}\, e^{-1/2(x-\mu)_2}} \tag{11}$$

f(x) = Probability Density Function

σ Standard Deviation

μ= Mean

The researcher used Gaussian distribution that is also called a normal distribution or normal deviate, to map the real world random variables whose distribution are completely not known. The Gaussian distribution that approximates the exact binomial distribution of events which are often represented in bell shaped curve.

6 Results and discussion

Source of datasets: Kaggle (Admission dataset in Universities / Institutions)

This research used students' data sets from Kaggle with attributes GRE, GPA ADMIT and RANK with 400 different samples. The nature of data attributes such as ADMIT, GRE, and RANK belong to integer category whereas GPA is in category of decimal numbers. The data were compiled by R-programming using Naïve Bayes Classifiers and Kernel Density Estimation (KDE) tools. The description of data and its categories are given below. The use of Naïve Bayes Classification approach was to predict students' performance [6]. The predictor accuracy rate was 88% and it deployed the proposed model for the academic performance. In other context, the use of Naïve Bayes Classifier was chosen as it produces minimum error for making policies in academic environment towards the decision-making process. Different research studies proposed a prediction model with accuracy of 97% using WEKA for in data processing [2], The researcher studied the different machine learning technique to predict employability skills which is based on student's performance in their academic exposure. The predicted model verified the accuracy level 67.67% by [10]. The researcher found that Naïve Bayes Algorithm is one of the best techniques to predict 98% accuracy with first dataset and 78% accuracy with second dataset [17]. This predictive model generated high level of accuracy and efficiency to handle all possible categories available in the datasets. The Naïve Bayes Algorithm achieved 84.3% accuracy in prediction and performed much better in comparison with other machine learning algorithms [9].

The structure of data are defined as numerical and categorical which are the essential features of Naïve Bayes Algorithm in handling such data. This research categorized Admit, Rank as categorical variables whereas GPA belong to numerical variables as per the Naïve Bayes Algorithm (Table 1).

Table 1. Description of variables

Data	Frame':	400 obs. of 4 variables
$ admit	:int	0 1 1 1 0 1 1 0 1 0 ...
$ gre	: int	380 660 800 640 520 760 560 400 540 700 ...
$ gpa	: num	3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...
$ rank	: int	3 3 1 4 4 2 1 2 3 2 ...

6.1 Summary of data

This research outlined the summary of data with respect to mean first quarterly, median, 3rd quarterly and max nature of attribute's values. The summary of data statistics showed that the average mean of GRE is 587, mean of GPA is 3.390 and rank mean is 2.485. The statistical results also outlined that the max and min value of GRE is 220, 800, GPA min and max value is 2.260 and 4.00, whereas rank mean and max value is 1.000, 4.000. The description of median value of given data frame of GRE is 580.0, GPA median is 3.395, and rank median is 2.00. The previous data statistics is calculated by using R programming language (Table 2).

Table 2. Summary of data

Admit-1/Non Admit-0	GRE	GPA	Rank
Min. :0.0000	Min. :220.0	Min. :2.260	Min. :1.000
1st Qu.:0.0000	1st Qu.:520.0	1st Qu.:3.130	1st Qu.:2.000
Median :0.0000	Median :580.0	Median :3.395	Median :2.000
Mean :0.3175	Mean :587.7	Mean :3.390	Mean :2.485
3rd Qu.:1.0000	3rd Qu.:660.0	3rd Qu.:3.670	3rd Qu.:3.000
Max. :1.0000	Max. :800.0	Max. :4.000	Max. :4.000

6.2 Correlation matrix

As this research study used the Naïve Bayes Classification model, it was important to investigate the correlation of independent variables and see how strongly they are correlated or not. The statistics shows that the independents variables GRE, GPA and Rank are not strongly correlated to admission data with references as dependent variables (Table 3).

The researcher observed that Admit and Rank are numerical variables, and they were used as factors for categorical variables. The researcher used the factor to convert integer variables to categorical with respect to factor values. The boxplot shows that significant amount of overlap in observation between the distributions of Admitted and not Admitted students. The results shows that GRE score are higher for students who have been admitted in comparison to those who are not admitted. The researcher used the density plots to show significant of GRE and GPA score for admitted students in university or higher-level institutions (Figure 2).

Table 3. Correlation matrix

Admit-1/Non Admit-0	GRE	GPA	Rank
admit 1.0000000	0.1844343	0.17821225	-0.24251318
gre 0.1844343	1.0000000	0.38426588	-0.12344707
gpa 0.1782123	0.3842659	1.00000000	-0.05746077
rank -0.2425132	-0.1234471	-0.05746077	1.00000000

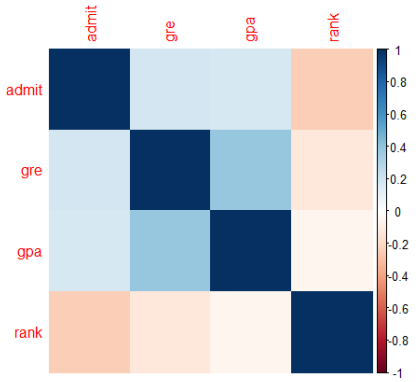

Fig. 2. Correlation matrix

6.3 Data distribution

The boxplot shows the comparison of distribution of admission, on the basis of GRE and GPA score. The data shows students that are admitted in universities or any higher institutions are having higher score of GPA and GRE in comparison to students that are not admitted. With the help of boxplot, the researcher identified the outliers that are too far from centroids, outliers with high values and unexpected values (Figure 5).

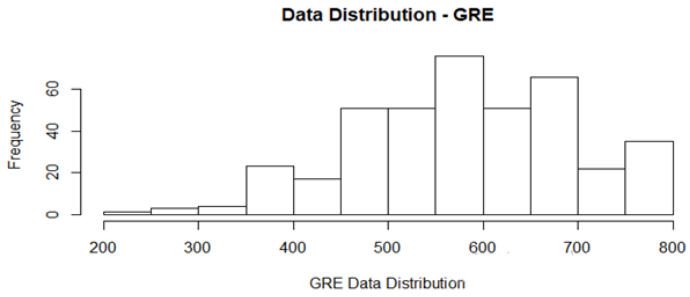

Fig. 3. GRE data distribution

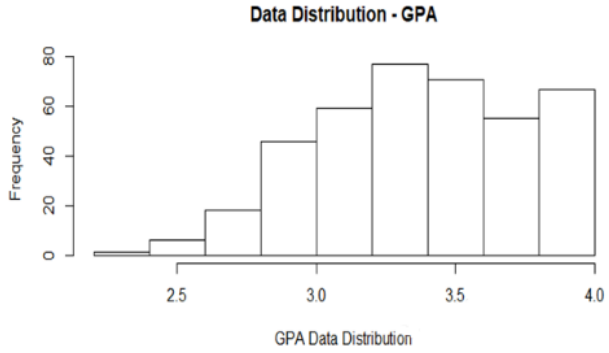

Fig. 4. GPA data distribution

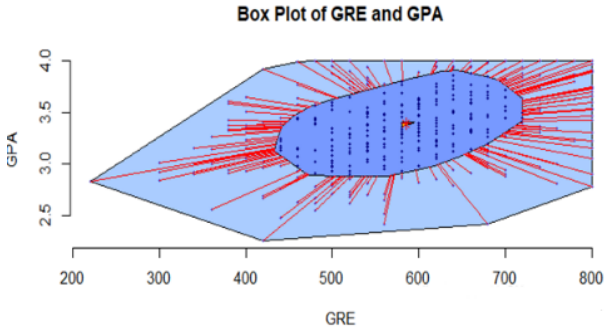

Fig. 5. Boxplot of GRE and GPA

6.4 Density plot of GRE and GPA

The density plot shown in blue are skewed to the right as it had higher GPA/GRE score for admitted students in comparison with students that were not admitted. Moreover, there can be some potentials to develop a classification problem, but in terms of accuracy it may be comprised as there is overlap between the two score. Thus, in order to move in forward direction, this research study built a predictive model on the basis of training and testing data (Figure 6).

The entire dataset was divided into 325 observations of training data and 75 observations are considered as testing data sets. The Naïve Bayes Classification model is based on assumption that features GPA and GRE are independent variables with respect to Admit as a dependent variable. The researcher first considered the training dataset to train the predictive model-1 using Naïve Bayes Algorithm (Figure 7).

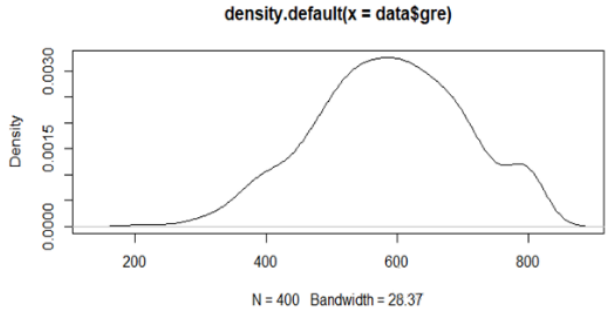

Fig. 6. Density plot of GRE

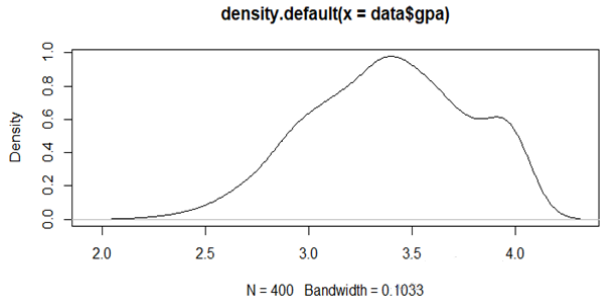

Fig. 7. Density plot of GPA

6.5 Predictive model-1

Table 4. A priori probabilities

0-(Non- Admitted)	1-(Admitted)
0.6861538	0.3138462

Table 5. GRE(Gaussian)

GRE	0-(Non- Admitted)	1-(Admitted)
Mean	578.6547	622.9412
Std. dev.	116.3250	110.9240

Table 6. GPA (Gaussian)

GRE	0-(Non- Admitted)	1-(Admitted)
Mean	3.3552466	3.5336275
Std. dev.	0.3714542	0.3457057

Table 7. Rank (Categorical)

Rank	0-(Non- Admitted)	1-(Admitted)
1	0.10313901	0.24509804
2	0.36771300	0.42156863
3	0.33183857	0.24509804
4	0.19730942	0.08823529

Table 8. Prediction- model-1

S. No	0	1	Admit	GRE	GPA	Rank
1	0.8449088	0.1550912	0	380	3.61	3
2	0.6214983	0.3785017	1	660	3.67	3
3	0.2082304	0.7917696	1	800	4.00	1
4	0.8501030	0.1498970	1	640	3.19	4
5	0.6917580	0.3082420	1	760	3.00	2
6	0.6720365	0.3279635	1	560	2.98	1

6.6 Confusion matrix- predictive-model-1

The below confusion matrix represented the accuracy of predictive model-1 and miss classifiers. The statistics shows that the predictive model truly predicted 196 in not admitted and 33 are admitted, whereas the predictive model -1 fails to truly predict about 27 and 69 which are identifed as miss classifiers. The percentage rate of misclassification is 0.29538461538 or 29.53% (Table 9).

Table 9. Confusion matrix- predictive-model-1

Probability	Not Admitted(0)	Admitted(1)
Not Admitted	196	69
Admitted	27	33

$$\text{Accuracy} = \frac{(196+33)}{(196+69+27+33)} \tag{12}$$

$$\text{Accuracy} = 0.70461538461 \text{ or } 70.46\%$$

$$\text{Misclassifies} = 29.53\%$$

This research used the Shapiro-Wilk normality test to the significance level of attributes to predict admission uncertainty. The probability of statistics shows P-value of GPA is 6.68e-06 which is less than 0.05 significant level and thus the result is significant. other attribute P-value of GRE is 0.0006268 which is also less than 0.05 significant level and thus the result is significant. At this level the researcher concluded that null assumptions are rejected about independent variable of GPA and GRE with respect to dependent parameter admission. The Attributes GRE and GPA are playing a significant role for getting students into University and Higher Academic Institutions (Table 10).

As the statistics showing that accuracy of predictive model -1 is 70.46% and miss classification is presented by the system is 29.53% which is not very much reliable model for predicting uncertainty of students admission. For more accuracy and to improve our model the researcher decided to move to kernel Density Estimate (KDE) which is stated as Predictive Model-2 (Table 10).

Table 10. Shapiro-Walk normality test

Data	**Data $ GPA**
W = 0.97736,	p-value = 6.68e-06
Data	**Data $ GRE**
W = 0.9859,	p-value = 0.0006268

6.7 Predictive model-2

Table 11. A priori probabilities

0-(Non- Admitted)	**1-(Admitted)**
0.6861538	0.3138462

Table 12. GRE::0 (KDE)

Density. default(x = x, na.rm = TRUE) Data: x (223 obs.); Bandwidth 'bw' = 35.5	
X	**Y**
Min. :193.5	Min. :6.010e-07
1st Qu.:371.7	1st Qu.:2.924e-04
Median :550.0	Median :1.291e-03
Mean :550.0	Mean :1.401e-03
3rd Qu.:728.3	3rd Qu.:2.405e-03
Max. :906.5	Max. :3.199e-03

Table 13. GRE::1 (KDE)

Density. default(x = x, na.rm = TRUE) Data: x (102 obs.); Bandwidth 'bw' = 39.59	
X	**Y**
Min. : 181.2	Min. : 1.145e-06
1st Qu.: 365.6	1st Qu.: 2.007e-04
Median : 550.0	Median : 1.129e-03
Mean :550.0	Mean : 1.354e-03
3rd Qu.: 734.4	3rd Qu.: 2.375e-03
Max. : 918.8	Max. : 3.465e-03

Table 14. GPA::0 (KDE)

Density. default(x = x, na.rm = TRUE) Data: x (223 obs.); Bandwidth 'bw' = 0.1134	
X	**y**
Min. : 2.080	Min. : 0.0002229
1st Qu.: 2.645	1st Qu.: 0.0924939
Median : 3.210	Median : 0.4521795
Mean : 3.210	Mean : 0.4419689
3rd Qu.: 3.775	3rd Qu.: 0.6603271
Max. : 4.340	Max. : 1.1433285

Table 15. GPA::1 (KDE)

Density. default(x = x, na.rm = TRUE) Data: x (102 obs.); Bandwidth 'bw' = 0.1234	
X	**y**
Min. : 2.25	Min. : 0.0005231
1st Qu.: 2.78	1st Qu.: 0.0800747
Median : 3.31	Median : 0.4801891
Mean : 3.31	Mean : 0.4710851
3rd Qu.: 3.84	3rd Qu.: 0.8626207
Max. : 4.37	Max. : 1.0595464

Table 16. Rank (Categorical)

Rank	**0-(Non- Admitted)**	**1-(Admitted)**
1	0.10313901	0.24509804
2	0.36771300	0.42156863
3	0.33183857	0.24509804
4	0.19730942	0.08823529

6.8 Confusion matrix- predictive-model-2

The below confusion matrix is based on predictive model -2 where 203 and 33 data attributes are truly predicted as non-admitted students, whereas 20 and 69 data items are showing not truly predicted by the predictive model-2, the researcher is representing it as a miss classifiers. The accuracy level of this predictive model is 72.61% that is better in result compared with predictive model-1(Table 17).

The miss classifiers of this predictive model is found sd 27.38% which is also less than the previous predictive model-1.

Table 17. Confusion matrix- predictive-model-2

Probability	Not Admitted	Admitted
Not Admitted	203	69
Admitted	20	33

$$\text{Accuracy} = \frac{(203+33)}{(203+20+69+33)} \tag{13}$$

Accuracy = 0.72615384615 or 72.61%

Miss-classifiers: 0.2738462 or 27.38%

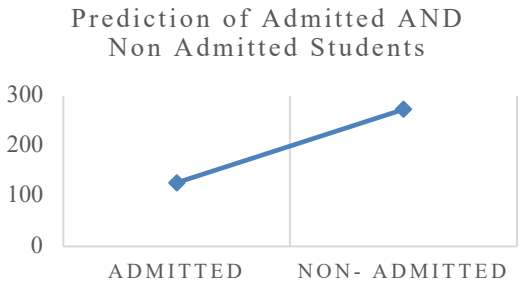

Fig. 8. Prediction of admit and non- admitted students

6.9 Findings

The previous data classification shows that 2/3rd set of data belongs to non-admitted, whereas 1/3 belongs to admitted category. The researcher observed that there is a huge difference between admitted and non-admitted students. During the data interpretation the researcher found that there are many factors that directly affects the admission process into universities or any higher academic institutions. The results shows that GRE, GPA and in some cases university ranks are the significant factors for the admission process as shown in (Figure 8).

In the case of imbalanced data for developing the predicting model the results are dominated by the data in the class. In terms of accuracy, the accuracy of the model will

be better when predicting for class 0 (Non- Admitted category) and 1 for admitted category (Figure 8).

The researcher had a positive response from the predictive model and it showed more accurate results at accuracy level of 72.61%. The researcher used a very limited data attributes such as GRE, GPA, and Rank that are globally defined and mainly used parameter for getting admission into universities / institutions.

According to a previous research study, it was found that the predicted model verified the accuracy level at 67.67% by [10], The researcher found that Naïve Bayes Algorithm is one of the best technique to predict 98% accuracy with first dataset and 78% accuracy with second dataset [17]. The predictive model produced high level of accuracy and was found highly efficient to handle all possible categorical and numerical data sets. The use of Naïve Bayes Algorithm achieved 84.3% accuracy and predicted that it is much better than the other machine learning algorithms [9]. Finally, this research study concluded that using different machine learning algorithms such as decision tree along with Naïve Bayes Classification for differentiating continuous students' performance, would results in better prediction performance for each students into getting admission probabilty into universities / higher academic institutions [7].

7 Summary and conclusion

In this research study the researcher built a predictive model to reduce uncertainty for getting admission into universities / institutions. This research used the Naïve Bayes classifiers and Kernel Density Estimations (KDE) machine learning algorithms to develop a predictive model that showed more reliability and efficiency to classify the admission category (Admitted / Not Admitted) into universities / institutions. During this research study the researcher found that accuracy performance of the predictive model-1 and predictive model-2 are very close to each other, which are stated as predictive model -1 has 70.46% truly predicted where as 29.53% are the misclassifies that is wrongly predicted. The accuracy performance of predictive model-2 has 72.61% truly predicted, whereas 27.38% are wrongly predicted that intended to misclassification. To improve the accuracy of predictive model the researcher used the kernel density estimations (KDE) tool, and Shapiro-Walk normality test to test the assumption between dependent variable ADMIT and independent variables GRE, GPA, Rank. With the help of Shapiro-Walk normality test, the researcher found that P-value is less than 0.05 significant level and thus the results are significant, moreover, the researcher found that independents variables have significant impact on dependent variable.

Finally, this research study concluded that there is huge differences between admitted and non – admitted students, the probability of statistics shows that 2/3 students are not admitted whereas 1/3 students are in admitted category. The differences are generated by the predictive model because of imbalance data sets. This research study strongly recommends for further research study using random forest algorithms in case of imbalanced datasets.

7.1 Limitation of study

The limitation of this research study regarding the developed predictive model is being not applicable for getting admission into all courses. Moreover, this research study used limited data attributes such as GRE, GPA and RANK that does not define the admission into all possible courses. It is stated that:

1. It is applicable only for student's admission into universities / institutions.
2. The researcher used only three attributes of admission parameters that are: GRE, GPA and RANK in order to have more focus on investigating and enhancing the used approach rather than the data complexity associated with wider selection of attributes.

7.2 Future research work

The use of machine learning techniques and algorithms should be considered by universities/ institutions in order to develop predictive models that would be more accurate to reduce admission uncertainty. Due to time and data availability constraints, this research study suggests for future research to increase the datasets, data attributes and the cross-validation process during the training process of predictive model. Being able to overcome the current constraints found in this research will increase the probability of having more accuracy and efficiency towards the decision-making process for getting admission into universities / institutions. The researcher stated the main significant research issues on admission uncertainty can be improved for future studies by:

1. Including more factors and attributes related to admission uncertainty and its significant role for getting admission into universities / institutions.
2. Investigating the accuracy provided by using different predictive models and to outline their significance towards decision making for getting admission into universities / institutions.

8 References

[1] A Daveedu Raju, Ch. Gaayathre, G Leela Deepthi, 2019. Prediction of Students Performance for a Multi Class Problem Using Naïve Bayes Classifiers, International Journal of Innovative Technology and Engineering Research(IJITEE), Vol.8, Issue 7, ISSN: 2278-3075.

[2] Allsela Meiriza, Endang Lestari, 2019. Advances in Intelligent Systems Research, International Conference on IT and its Applications, Vol.172, Siconian.

[3] Allsela Meiriza1 and Endang Lestari, 2020. Predicting Graduate Students Use Naïve Bayes Classifier, International Conference on IT and its Application, Advances in Intelligent System Research, Vol.172, and Atlantis Press SARL. https://doi.org/10.2991/aisr.k.200424.056

[4] Baradwaj, B. K., & Pal, S., 2012. Mining Educational Data to Analyze Students Performance, ArXiv Pre-print ar Xiv: 120.3415.

[5] Bresfelean, V. P., 2007. Analysis and Prediction on Students Behaviors Using decision Tree in WEKA Environment. In Proceeding of the ITI, PP.25-28. https://doi.org/10.1109/ITI.2007.4283743
[6] Delali Kwasi Dake1, Esther Gyimah, 2017.Students Grades Predictor using Naïve Bayes Classifiers- A Case Study of University of Education, International Journal of Innovative Research in Science, Engineering and Technology, Vol.6, Issue 10, ISSN:2319-753.
[7] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From Data Mining to Knowledge Discovery in Databases, AI Magazine, Vol.17, Issue 3, PP.37-45.
[8] Hussain, S., Muhsion, Z. F., Salal, Y. K., Theodorou, P., Kurtoglu, F., & Hazarika, G. C. (2019). Prediction Model on Student Performance based on Internal Assessment using Deep Learning. iJET, 14(8), 4-22. https://doi.org/10.3991/ijet.v14i08.10001
[9] Imran, M., Latif, S., Mehmood, D., & Shah, M. S. (2019). Student Academic Performance Prediction using Supervised Learning Techniques. International Journal of Emerging Technologies in Learning, 14(14). https://doi.org/10.3991/ijet.v14i14.10310
[10] Iqbal, Z, Qadir, J., and Kamiran, F., 2017. Machine Learning Based Students Grade Prediction: A Case Study, Available on : https://arxiv.org/pdf/1708.08744.pdf
[11] Kabra, R. R., Bichkar, R. S., 2011. Performance Prediction of Engineering Students, Using Decision Trees, International Journal of Computer Applications, Vol. 36, Issue 11, PP.-12.
[12] Kanadpriya Basu and Treena Basu, 2019. Predictive Model of Student College Commitment Decisions Using Machine Learning, Vol.2, Issue 4, MDPI. https://doi.org/10.3390/data4020065
[13] Khin Lay, Aung Cho, 2019. Using Naïve Bayesian Classifier for Predicting performance of Students, International journal of Trend in Scientific Research and Development(IJTSRD), Vol.3, Issue , e-ISSN:2456-6470.
[14] Kumar, S. A., Vijayalakshmi, M. N., 2011. Efficiency of Decision Tree in Predictive Students Academic Performance, First International Conference on Computer Science, Engineering and Applications, Vol.2, PP.335-343.
[15] Ms. Ashna Sethi1, Mr. Charanjit Singh, 2017. Data Mining for Prediction and Classifications of Engineering Students Achievement Using Improved Naïve Bayes Algorithms, International Journal of Advanced Research in Computer Engineering and Technology (IJARCTE), VOl.6, Issue 7, ISSN: 227-1323.
[16] Nghe, N. T., Janecek, P., Haddawy, P., 2007. A Comparative Analysis of Techniques for Predicting Academic Performance, Conference Global Engineering: Knowledge without Boarders, Opportunity without Passports, IEEE-2007, T2G-7.
[17] Nuankaew, P., Nuankaew, W., Teeraputon, D., Phanniphong, K., & Bussaman, S. (2020). Prediction Model of Student Achievement in Business Computer Disciplines: Learning Strategies for Lifelong Learning. International Journal of Emerging Technologies in Learning, 15(20). https://doi.org/10.3991/ijet.v15i20.15273
[18] Quadri, M. M., Kalyankar, N. V., 2010. Drop Out Features of Students Data for Academic Performance Using Decision Tree Techniques, Global Journal of Computer Science and Technology, Vol.10, Issue 2.
[19] S. Karthika and N. Sairam, 2016. A Naïve Bayes Classifiers for Educational, Indian Journal of Science and Technology, Vol.8, Issue 16, ISSN: 0974-5645. https://doi.org/10.17485/ijst/2015/v8i16/62055
[20] Shashi Sharma, Sunil Kumar Pandey, and Kumkum Garg, 2020. Machine Learning for Prediction in Academics, International Journal of Recent Technology and Engineering (IJRTE), Vol.8, Issue 5, and ISSN: 2277-378. https://doi.org/10.35940/ijrte.E6965.018520

[21] Singh, M., and Singh, J., 2013. Machine Learning Techniques for Prediction of Subject Scores: A Comparative Study, International Journal of Computer Science and Network, Vol. 2, Issue 4, PP. 77-80.
[22] Undavia, J. N., Dolia, P. M., Shah, N. P., 2013. Prediction of Graduate Students for Master Degree Based on Their Past Performance using Decision Tree in WEKA Environment, International Journal of Computer Applications, Vol.74, Issue 21. https://doi.org/10.5120/12930-9877
[23] Vaidu, G., and Sornalakshmi, K., 2017. Applying Machine Learning Algorithms for Students Employability Prediction Using R programming Language, International Journal of Pharmaceutical Science Review and Research, PP.38-41.

9 Abbreviations

GRE : Graduate Record Examinations
GPA : Grade Point Average
KDE : Kernel Density Estimations
KAPPA : Coefficient is a statistical measure
WEKA : Waikato Environment for Knowledge Analysis

10 Authors

Nasim Matar is with University of Petra, Amman, Jordan, 961343.
Wasef Matar is with University of Petra, Amman, Jordan, 961343.
Tirad Al Malahmeh is with University of Petra, Amman, Jordan, 961343.

Article submitted 2022-01-30. Resubmitted 2022-02-25. Final acceptance 2022-02-26. Final version published as submitted by the authors.