Application of Naive Bayes to Students' Performance Classification

Abstract

Naive Bayes Classifier is a strong tool or model in classifying students' performance based on various factors. Thus, this research developed a classification model that can accurately classify students into different academic performance categories. The study utilized data, collected from 1,422 students at the University of Ibadan, Nigeria. Descriptive statistics and data visualization techniques were used to gain insights into the distribution and relationships among the variables. Subsequently, a Naive Bayes classifier model was built using 70% of the data for training and 30% for testing. In addition, a Support Vector Machine (SVM) model was built to compare with the performance of the Naive Bayes model. The results of the descriptive statistics show that the respondents comprise of 846 females and 576 males. From the female respondents, 144 of them had First Class grade, 432 had Second Class Upper, 252 had Second Class Lower, and the remaining 18 had Third Class. From the male respondents, 144 of them had First Class grade, 198 had Second Class Upper, 216 had Second Class Lower, and the remaining 18 had Third Class. The Naive bayes model achieved an overall accuracy of 87%, while the SVM model achieved an overall accuracy of 85%. The results highlighted that department, grade in the first year, and monthly allowance were the most crucial features for classifying performance outcomes, while gender, age group and whether or not the respondents' parents are educated, exerted the least significant influence on the models. Thus, on average, the Naive Bayes model outperformed the SVM in the classification of students' performance based on the data collected. Also, the early academic performance, and financial support are significant factors in determining students' overall performance in the Institution.

Keywords: Naive bayes, Support vector machine, Data mining, Classification, Model accuracy.

1 Introduction

For all educational institutions, raising educational standards and student performance is of vital significance. Over the last ten years, there has been an increase in interest in determining the key variables affecting students' academic success in higher education, particularly when employing data mining approaches. The application of such research in assisting in the early identification of low-performing students to overcome their learning challenges and improve their learning outcomes, which in turn serves the institutional goals of providing high-quality educational ecosystems, is credited with motivating this interest (Abu Saa, Al-Emran and Shaaran, 2019). Student academic performance refers to how far a student, teacher, or institution has come in achieving their short- or long-term educational objectives (Alturki et'al 2022). There is no consensus on how it should be evaluated or which components are most significant, despite the fact that it is frequently measured by examination or continuous evaluation. Academic success is crucial in establishing the value of graduates who will be in charge of the nation's social and economic development. Student performance or achievement is crucial in higher education settings in Nigeria. This is so because a great track record of academic success is one of the requirements for a high-caliber university. According to Usamah, Buniyamin, Arsad, and Kassim (2013), co-curriculum and learning evaluation measurements may be used to determine a student's performance. All educational institutions place the highest priority on improving student performance and education

quality. A deep study of the students' prior records can be quite important for providing high-quality instruction (Shahiri, Husain, and Rashid., 2015). The classification of student performance is a crucial topic for enhancing the educational process. Many variables, such as the father's profession, the student's gender, and their average test scores over the preceding years, may have an impact on the student's performance level, (Arkana 2023). Early student performance classification may contribute to bettering the educational process (Chang et'al 2022). Data mining techniques applied to educational data sets can be used to classify students' performance (Amra and Maghari, 2017, and Pavan 2023). The use of data mining tools has expanded in recent years. These methods are frequently employed in the educational sector to draw out buried data and identify trends in educational datasets. Knowledge Discovery in Databases (KDD), often known as data mining, is the process of extracting meaningful information from massive data sets and a variety of techniques and models are used to extract patterns from stored data. Statistical methods and data mining techniques are the two ways. Classification, clustering, classification, association rules, neural networks, decision trees, and the nearest neighbor method are some of the data mining approaches (Thant, et'al 2021).

In order to better understand students and the learning environment they are in, educational data mining, a relatively new field of study, uses data-driven approaches to evaluate educational datasets such as student, professor or instructor, course, and school data. Educational data mining entails analyzing and improving the techniques used to classify student performance. Typically, educational data is gathered through computer-assisted or web-based learning tools, or the management of the school or institution will supply the information. It is frequently noted that the data are intricate and intricately connected to one another. It is crucial to identify underachievers within the first few weeks of the semester so that teachers may intervene appropriately, for as by providing mentorship or going over material with the student. To be able to recognize these pupils, faculty members will require useful tools (Bibireddy, 2017). According to projection findings, if the demands of the students are met promptly, the overall outcome and performance will improve year after year, (Anwarudin et al 2022). Important characteristics and prior performance data of pupils are gathered for the goal of performance analysis and classification. To get deeper insights and make more accurate forecasts, a variety of studies pertinent to student learning has been successfully modelled using educational data mining techniques and classification models. Additionally, models improve their accuracy year after year and are verified to become more generalizable over time. Numerous significant pedagogical advancements have been made as a result of education research. The use of computer-based technology has changed how we study and live. Today, a second wave of revolution in all areas of learning and accomplishment is supported by the utilization of data gathered through these technologies (Ashraf, Sajid, and Gufran, 2018). The goal of EDM is to enhance the educational system, lower the failure rate, identify critical traits, and take into account student performance and achievement. Additionally, it makes it possible for us to create beneficial classification models for performance classification. It offers information and insights for the following year's preparation of the educational process in addition to assisting in taking urgent action for the welfare of at-risk pupils. Numerous data mining approaches and classification models, including Naive Bayes, Decision Trees, Neural Networks, Outlier Detection, and Advanced Statistical approaches, have been applied in recent years. These methods are used on student data to get information, support decision support systems, and uncover patterns, among other things (Ashraf, Sajid and Gufran, 2018 and Udomboso et'al 2019).

The Grade Point Average (GPA) is a widely used measure of academic achievement. Many institutions have minimum GPA requirements that students must adhere to. As a result, academic planners continue to utilize GPA as their primary yardstick for assessing students'

academic progress. Throughout their time in college, a student's ability to achieve and maintain a high GPA that accurately represents their overall academic achievement may be hampered by a variety of issues. By monitoring the development of their performance, faculty members might focus on these characteristics when devising ways to enhance student learning and boost their academic success. The crucial qualities for future classification may be found using the data mining technique's clustering model and decision tree. The technique of extracting previously undiscovered, reliable, strategically relevant, and concealed patterns from big data sets is known as data clustering. Educational databases are storing an ever-growing amount of data. The most popular strategy for classifying the future is clustering. Clustering's major objective is to divide pupils into homogenous groups based on their traits and skills. The quality of education may be improved with the aid of these applications for both students and teachers (Islam and Haque, 2012).

Commonly, a student's academic success is assessed using their prior CGPA, but there are several other significant factors that influence how well a student does overall. On student datasets, several empirical and statistical based-studies have recently been undertaken. Using pre-university data, Kabakchieva (2013) classified student performance using Bayes and decision classifiers. Other methods that have been suggested in the literature include neural networks, statistical techniques, and ID3 models. Overall, the ability to anticipate students' academic performance has the potential to improve educational outcomes by facilitating early interventions, individualized instruction, effective resource management, well-informed curriculum preparation, and data-driven decision-making. Classification modelling is a tool that administrators and teachers may use to improve educational procedures and help kids reach their full potential. The abundance of data in educational databases makes classifying students' success more difficult. Furthermore, to raise student accomplishment, a rigorous literature evaluation on forecasting student performance using data mining approaches is suggested (Shahiri et al., 2015).

Data mining's primary objective is to examine vast amounts of data in order to uncover new information and patterns. Regression, association, and classification are a few data mining techniques. Classification, cauterization, and regression are the building blocks of the classification model. Most often, categorization was utilized by researchers to forecast students' performance. Naive Bayes Classifier, a probabilistic classifier founded on the Bayes Theorem, is one of the classification models. The Nave Bayes classifier makes the assumption that the value of an attribute's impact on a particular class is unrelated to the value of other characteristics. Naive Bayes is a well-known method for classifying texts (Makthar, Nawang, and Shamsuddin, 2017). Furthermore, sometimes, there is a clear separation between instances of the different categories in vector space. Thus, when a new sample is received, it is mapped into the designated vector space, and depending on which side of the gap it falls, its label is categorised. By using the kernel approach, an SVM can successfully classify non-linear data. (Arcinas et al. (2021).

2. Methodology

Research Design

The research design for this study involves the utilization of a questionnaire-based survey to collect data from a sample of 1,422 students at the University of Ibadan. The target population for this study consists of undergraduate and postgraduate students enrolled at this Institution. The study employs a purposive sampling technique to select the participants. The design adhered to ethical guidelines and regulations to ensure participant confidentiality, anonymity, and informed consent. Confidentiality of the collected data was maintained by storing the data securely and using anonymized identifiers during data analysis and reporting.

2.1 The Naive Bayes Classifier

Naive Bayes classifiers operate under the assumption of feature independence. In contrast to numerous other classifiers that make the assumption of correlation between features within a given class, naive Bayes explicitly represents the features as conditionally independent given the class. Although it may appear to be an excessively simplistic limitation on the data, in practical applications, Naive Bayes demonstrates competitiveness with more advanced techniques and is supported by certain theoretical evidence of its effectiveness (Jean-François Boulicaut, 2004).

The Naive Assumption

One simplifying assumption made in this context is that the features exhibit conditional independence given the class (C_k). As illustrated herein, this approach effectively mitigates the issue of dimensionality by enabling the decomposition of the joint distribution $P(y_1,...,y_n, C_k)$ into n+1 factors, which consist of n features in addition to the class prior $P(C_k)$.

The Naive Bayes Model

Assuming a data point $Y = [y_1, y_2, \dots, y_n]$ of r features and the class $C_k, k = 1, 2, \dots K$, Naive Bayes classify the class C_k according to the probability:

According to Bayes' theorem, this can be expressed as:

$$P(C_k|Y) = \frac{P(Y|C_k)P(C_k)}{P(Y)} = \frac{P(y_1,\dots,y_n|C_k)P(C_k)}{P(y_1,\dots,y_n)} \quad (1)$$

Where, $P(C_k|Y) = P(C_k|y_1, y_2, \dots, y_n)\ for\ k = 1, 2, \dots K$

Applying the Chain Rule, $P(y_1, \dots, y_n|C_k)$ can be further decomposed as

$$P(y_1, \dots, y_n|C_k) = P(y_1|y_2, \dots, y_n, C_k)P(y_2|y_3, \dots, y_n, C_k) \dots P(y_{n-1}|y_n, \dots, y_n, C_k)P(y_n|\ C_k) \quad (2)$$

Here, the "naive" assumption of conditional independence is employed and can be formulated from the above decomposition as

$$P(y_i|y_{i+1}, \dots, y_n|C_k) = P(y_i|C_k) \rightarrow P(y_1, \dots, y_n|C_k) = \prod_{i=1}^{n} P(y_i|C_k) \quad (3)$$

Thus, for a data point $Y = [y_1, y_2, \dots, y_n]$ of r features and the class $C_k, k = 1, 2, \dots K$,

$$P(C_k|Y^r) = \frac{\prod_{i=1}^{n} P(Y_i^r|C_k)P(C_k)}{P(Y)}\ , \quad (4)$$

where r is the features,

$P(C_k|Y^r) = posterior\ probability,\ P(Y_i^r|C_k) = likelihood,\ P(C_k) = prior\ probability,$
$P(Y^r) = Evidence$

In this context of classifying the student students based on their grades,

$$P(Grade\_Class|Student\_Features) = \frac{\prod P(features\ |grade\ )P(Grade\ \_Class\ )}{P(Stud\ ent\ \_Features\ )} \quad (5)$$

Thus, the posterior probability of each class given the input vector and the class with the maximum posterior probability becomes the output class. That is, the output class is given as:

$$c = argmax_{k\in\{1,2,\dots,K\}} P(C_k|Y^r) \quad (6)$$

2.2 Support Vector Machine (SVM)

In SVM, the input data is represented as a set of feature vectors, each associated with a class label or a regression target value. The feature vectors capture the characteristics or attributes of the data points. A kernel function is chosen to transform the input feature space into a higher-dimensional space. This transformation helps in finding a decision boundary that can linearly separate the data or capture non-linear relationships. Common kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid. In cases where the data is not linearly separable in the original feature space, SVM uses the kernel trick. The kernel function implicitly maps the data into a higher-dimensional space, where it becomes linearly

separable. This allows SVM to find a decision boundary that is non-linear in the original feature space

3 Results and Discussion

Table 1: Distribution of students' grade class by gender

	Female	Male
First Class	144 (50)	144 (50)
Second Class Lower	252 (53.8)	216 (46.2)
Second Class Upper	432 (68.6)	198 (31.4)
Third Class	18 (50)	18 (50)

The cross-tabulation in Table 1 presents the distribution of students' grade classes by gender. In the First-Class category, an equal distribution of 50% exists between female and male students. This indicates a balanced representation of high-performing students from both genders. In the Second-Class Lower category, 53.8% of the students are female, while the remaining 46.2% are male. This suggests a slightly higher proportion of female students achieving Second Class Lower grades compared to their male counterparts. The Second-Class Upper category displays a substantial gender disparity. Here, 68.6% of the students achieving Second Class Upper grades are female, while only 31.4% are male. This implies a significantly higher percentage of female students attaining this level of academic performance. Lastly, the Third-Class category demonstrates an equal distribution, with 50% of students being female and the remaining 50% being male. This indicates a balanced representation of students across genders in the Third-Class grade category. Figure 1 below represents the distribution of students' grade based on gender.

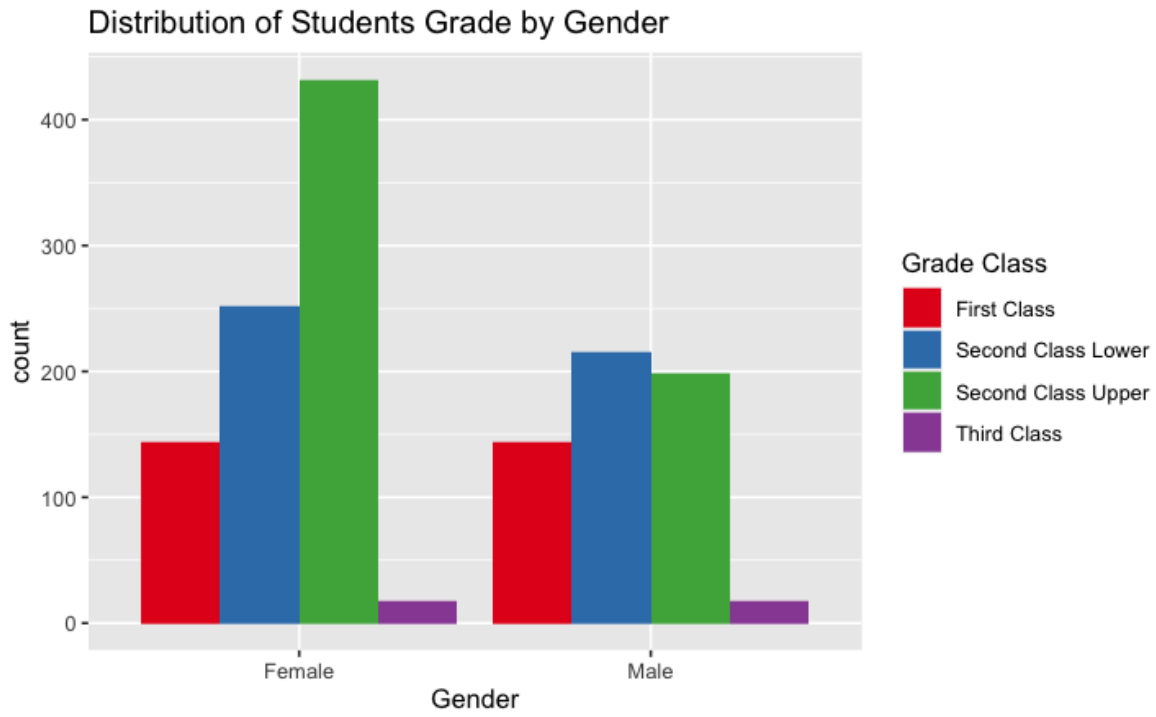

Figure 1: Distribution of Students' grades by gender

Figure 1 highlights notable gender differences in academic performance. While the distribution of First Class and Third-Class grades is fairly equal between female and male

students, there is an imbalance in the distribution of Second Class-Lower and Second-Class Upper grades, with a higher percentage of female students achieving Second Class Upper grades. These findings provide insights into the relationship between gender and academic achievement among the students in the study.

Table 2: Distribution of students' grade class by age group

	16-18	**19-21**	**22 and above**
First Class	18 (6.25)	72 (25)	198 (68.75)
Second Class Lower	18 (3.85)	72 (15.38)	378 (80.77)
Second Class Upper	36 (5.71)	180 (28.57)	414 (65.71)
Third Class	0 (0)	18 (50)	18(50)

Table 2 illustrates the distribution of students' grade classes across different age groups. Among students who attained a First-Class grade, the distribution across age groups is as follows: 6.25% of First-Class achievers are in the 16-18 age group, 25% in the 19-21 age group, and the majority, accounting for 68.75%, are in the 22 and above age group. The distribution of Second-Class Lower achievers across age groups reveals that 3.85% are in the 16-18 age group, 15.38% in the 19-21 age group, and the highest proportion, 80.77%, falls within the 22 and above age group. Among students who achieved a Second-Class Upper grade, the distribution across age groups is as follows: 5.71% in the 16-18 age group, 28.57% in the 19-21 age group, and 65.71% in the 22 and above age group. The data shows that no students in the 16-18 age group attained a Third-Class grade. Among students aged 19-21 and 22 and above, an equal percentage of 50% achieved a Third-Class grade. From the table, it is evident that the distribution of students' grade classes varies across age groups. The highest proportion of First-Class achievers is found among students in the 22 and above age group. In contrast, Second Class Lower achievers are predominantly observed in the 22 and above age group. The distribution of Second-Class Upper achievers is spread across age groups, with a notable increase in the proportion from the 16-18 age group to the 19-21 age group. Lastly, no Third-Class achievers are identified in the 16-18 age group. Figure 2 below represents the distribution of students' grade based on gender.

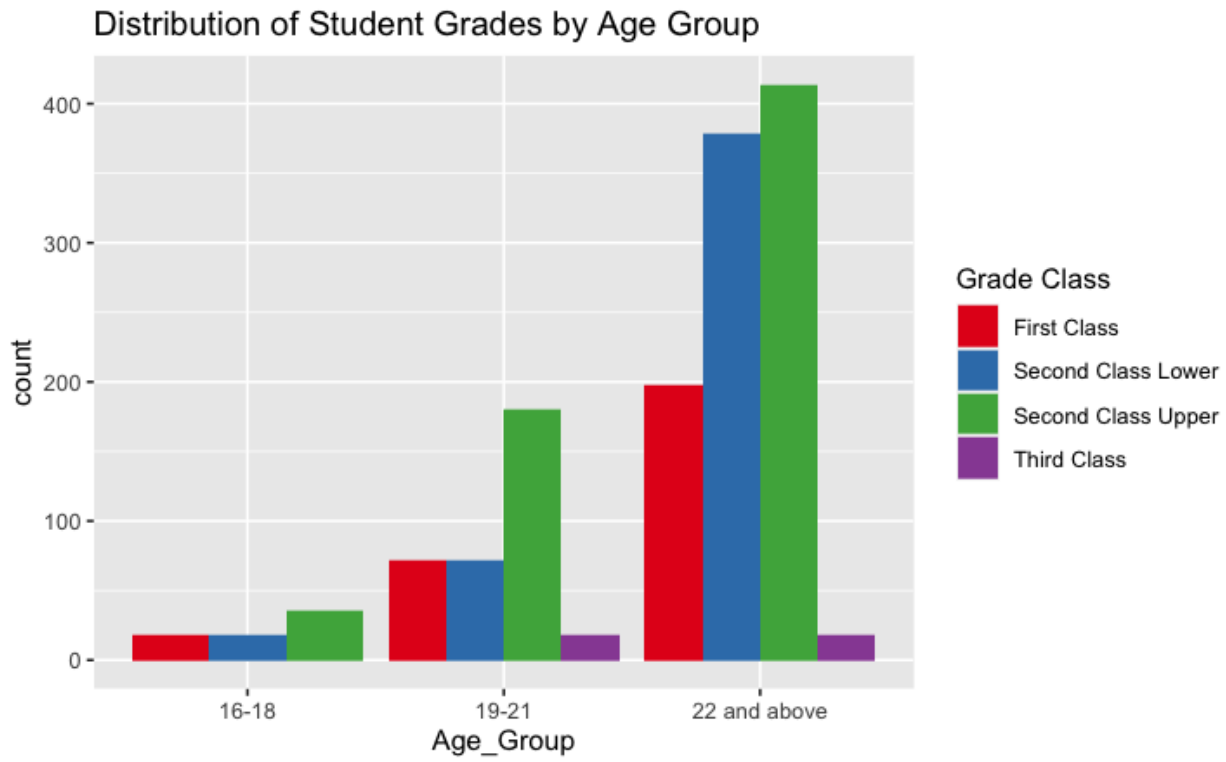

Figure 2: Distribution of Students' grades by age group

Table 3: Distribution of students' grade class by Level

	100	200	300	400	500	Post Grad
First Class	54 (18.75)	18 (6.25)	108 (37.50)	90 (31.25)	0 (0)	18 (6.25)
Second Class Lower	18 (3.85)	18 (3.85)	90 (19.23)	324 (69.23)	0 (0)	18 (3.85)
Second Class Upper	36 (5.71)	108 (17.14)	198 (31.43)	234 (37.14)	18 (2.86)	36 (5.71)
Third Class	0 (0)	0 (0)	18 (50)	18 (50)	0 (0)	0 (0)

Table 3 above presents the distribution of students' grade classes across different levels of study, including 100 level, 200 level, 300 level, 400 level, 500 level, and Post Grad. Among students achieving a First-Class grade, the distribution across levels of study is as follows: 18.75% of First-Class achievers are at the 100 level, 6.25% at the 200 level, 37.50% at the 300 level, 31.25% at the 400 level, and 6.25% among Post Grad students. There are no First-Class achievers at the 500 level. The distribution of Second-Class Lower achievers across levels of study reveals the following percentages: 3.85% at the 100 level, 3.85% at the 200 level, 19.23% at the 300 level, 69.23% at the 400 level, and 3.85% among Post Grad students. There are no Second-Class Lower achievers at the 500 level. Among students achieving a Second-Class Upper grade, the distribution across levels of study is as follows: 5.71% at the 100 level, 17.14% at the 200 level, 31.43% at the 300 level, 37.14% at the 400 level, and 5.71% at the 500 level.

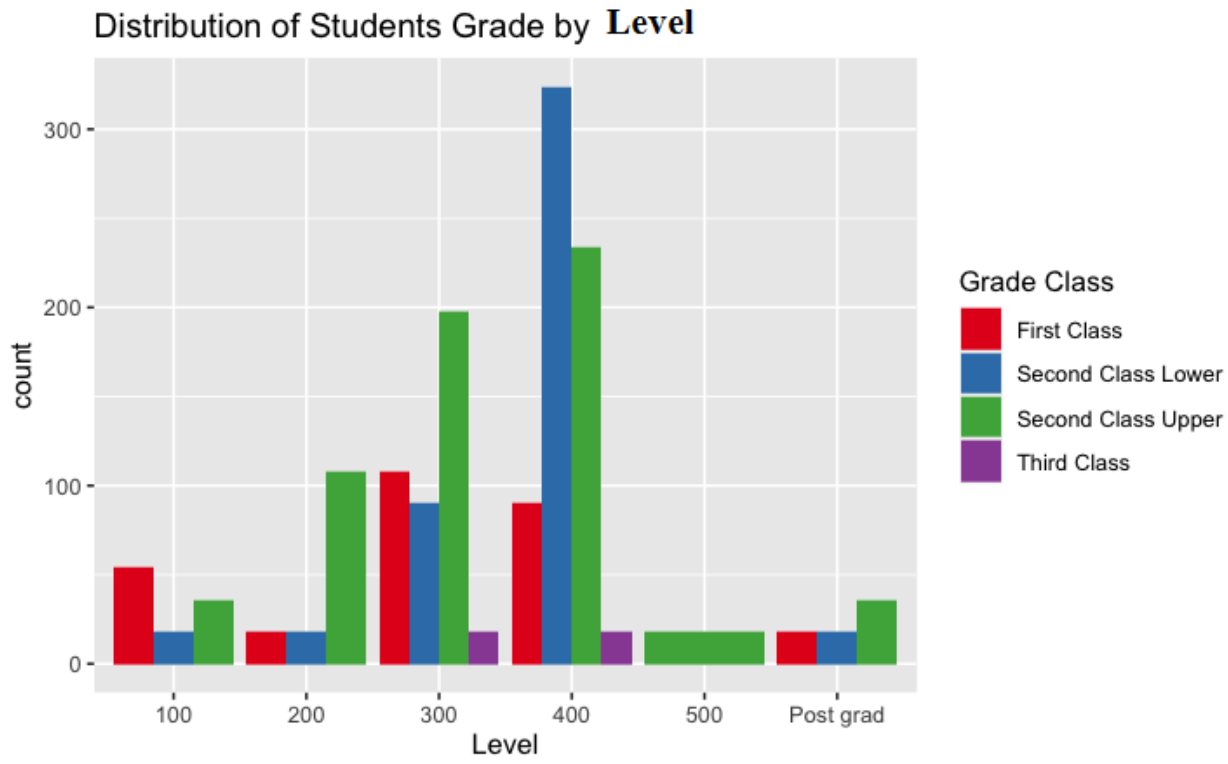

Figure 3: Distribution of Students' grades by Level

Additionally, 5.71% of Second-Class Upper achievers are among Post Grad students. No students at the 100 level or 200 level attained a Third-Class grade. Among students at the 300 level and 400 level, 50% of students achieved a Third-Class grade. There are no Third-Class achievers at the 500 level or among Post Grad students. However, it is important to note that these interpretations are based on the provided data and may not be generalized beyond the specific context of the study.

Modelling

In this section, the modelling phase is constructed. The primary objective is to develop a classification model using the Naive Bayes classifier to forecast/classify students' performance based on the available categorical variables. Additionally, it compares the performance of the Naive Bayes classifier with one other popular classification model [Support Vector Machine (SVM)] in this section. The classification modelling phase is a crucial step in this study as it allows the harnessing of the power of machine learning and Bayes Theorem to make classifications and gain a deeper understanding of the factors influencing students' performance.

The Independence Assumption

The Naive Bayes model assumes that the features (variables) used for classification are conditionally independent given the class label (Ray, 2019). Table 4 is the results of the Chi-Square test of independence for several pairs of categorical variables, with most having a p-values are higher than a significance level of 0.05. This suggests that there is not enough evidence to reject the null hypothesis of independence for these variable pairs.

Table 4: Test of independence for input variables

	Age	Gender	Level	Mon_All	Dept	Accom	Edu_Par	Sch_Pol	Curr_Act	Stu_tm	Grp_Disc	75%_Att	LS	LM
Age	0.00	0.08	0.00	0.10	0.05	0.36	0.69	0.80	0.09	0.32	0.58	0.43	0.02	0.63
Gender	0.08	0.00	0.55	0.03	0.60	0.08	0.89	0.62	0.14	0.17	0.65	0.01	0.70	0.64
Level	0.00	0.55	0.00	0.02	0.00	0.00	0.01	0.58	0.62	0.84	0.64	0.42	0.48	0.46
Mon_All	0.10	0.03	0.02	0.00	0.32	0.01	0.57	0.52	0.71	0.75	0.92	0.83	0.43	0.25
Dept	0.05	0.60	0.00	0.32	0.00	0.60	0.26	0.93	0.13	0.09	0.11	0.33	0.53	0.35
Accom	0.36	0.08	0.00	0.01	0.60	0.00	0.01	0.78	0.97	0.05	0.30	0.03	0.77	0.89
Edu_Par	0.69	0.89	0.01	0.57	0.26	0.01	0.00	0.95	0.48	0.77	0.90	0.66	0.82	0.31
Sch_Pol	0.80	0.62	0.58	0.52	0.93	0.78	0.95	0.00	0.20	0.97	0.00	0.78	0.24	0.71
Curr_Act	0.09	0.14	0.62	0.71	0.13	0.97	0.48	0.20	0.00	0.63	0.58	0.75	0.93	0.29
Stu_tm	0.32	0.17	0.84	0.75	0.09	0.05	0.77	0.97	0.63	0.00	0.76	0.06	0.17	0.95
Grp_Disc	0.58	0.65	0.64	0.92	0.11	0.30	0.90	0.00	0.58	0.76	0.00	0.62	0.32	0.49
75%_Att	0.43	0.01	0.42	0.83	0.33	0.03	0.66	0.78	0.75	0.06	0.62	0.00	0.80	0.59
LS	0.02	0.70	0.48	0.43	0.53	0.77	0.82	0.24	0.93	0.17	0.32	0.80	0.00	0.35
LM	0.63	0.64	0.46	0.25	0.35	0.89	0.31	0.71	0.29	0.95	0.49	0.59	0.35	0.00

Building and Evaluating the Model

The Naive Bayes model was built using R Programming, and the data was split into a training set comprising 70% of the data and a test set consisting of the remaining 30%. This approach allows for evaluating the performance and generalization ability of the model on unseen data (Nguyen et al., 2021).

Table 5: Confusion Matrix for the Naive Bayes model classifications

	First Class	**Second Class Lower**	**Second Class Upper**	**Third Class**
First Class	67	0	7	0
Second Class Lower	8	134	23	0
Second Class Upper	11	6	159	0
Third Class	0	0	0	11

Table 5 shows the classification made by the trained model using the 30% testing model, from which the performance of the model is computed.

The model overall accuracy score is presented in Table 6 below

Table 6: Model Performance for the Naive Bayes

Overall Accuracy	**95% CI**	**P-Value**
87%	(0.84, 0.90)	.000

As observed in Table 6 above, the Naive Bayes model achieved an overall accuracy of 87% in classifying the students' grade classes. This means that 87% of the classifications made by the model matched the actual grade classes of the students. Looking at the confusion matrix in Table 5, we can observe the distribution of classified grade classes against the actual grade classes.

The model was further evaluated based on its performance on each class of the classified variable. The performance is presented in Table 7 below

Table 7: Other Performance metrics for the Naive Bayes

	Sensitivity	Specificity	Precision	F1-Score
First Class	78%	98%	91%	81%
Second Class Lower	96%	89%	81%	84%
Second Class Upper	84%	93%	90%	85%
Third Class	100%	100%	100%	100%

The performance of the Naive Bayes model can be further assessed using additional evaluation metrics such as sensitivity, specificity, precision, and F1-Score. These metrics provide more insights into the model's ability to correctly identify positive and negative instances, as well as the balance between precision and recall (Chang et al., 2022).

Looking at the sensitivity (also known as recall), it can be seen that the model achieved 78% sensitivity for classifying "First Class." This means that 78% of the actual "First Class" instances were correctly identified by the model. Similarly, for "Second Class Lower," the model achieved a sensitivity of 96%, indicating that 96% of the actual "Second Class Lower" instances were correctly classified. The sensitivity for "Second Class Upper" was 84%, and for "Third Class," it was 100%, meaning that all instances of "Third Class" were correctly identified by the model. Specificity measures the model's ability to correctly identify negative instances. The model achieved high specificity for all grade classes, with values of 98% for "First Class," 89% for "Second Class Lower," 93% for "Second Class Upper," and 100% for "Third Class." These high specificity values indicate that the model was able to accurately classify instances that did not belong to the respective grade classes.

Precision represents the proportion of correctly classified instances out of all instances classified as a specific grade class (Chang et al., 2022). The model achieved precision values of 91% for "First Class," 81% for "Second Class Lower," 90% for "Second Class Upper," and 100% for "Third Class." These values indicate the model's ability to minimize false positives and provide accurate classifications within each grade class. The F1-Score combines precision and recall into a single metric and provides a balanced measure of the model's accuracy (Chang et al., 2022). The F1-Score values for the Naive Bayes model were 81% for "First Class," 84% for "Second Class Lower," 85% for "Second Class Upper," and 100% for "Third Class." These values indicate a reasonable balance between precision and recall for each grade class, with the highest score achieved for "Third Class."

Comparison Between Naive Bayes and Support Vector Machine (SVM)

In the model comparison between Naive Bayes and Support Vector Machine (SVM), the performance results indicate varying levels of accuracy. Naive Bayes achieved an overall accuracy of 87%, suggesting that the model correctly classified the class labels for 87% of the instances in the dataset. This indicates a reasonably good performance, although it is worth noting that the model may have some misclassifications. In comparison, Support Vector Machine (SVM) achieved an overall accuracy of 85.9%. While this accuracy is lower than that of Naive Bayes models, it still indicates a reasonably good performance. In general, Naive Bayes stands out with between the two models, suggesting stronger classification capabilities. Figure 4 below indicates the performance of each model in making their classifications.

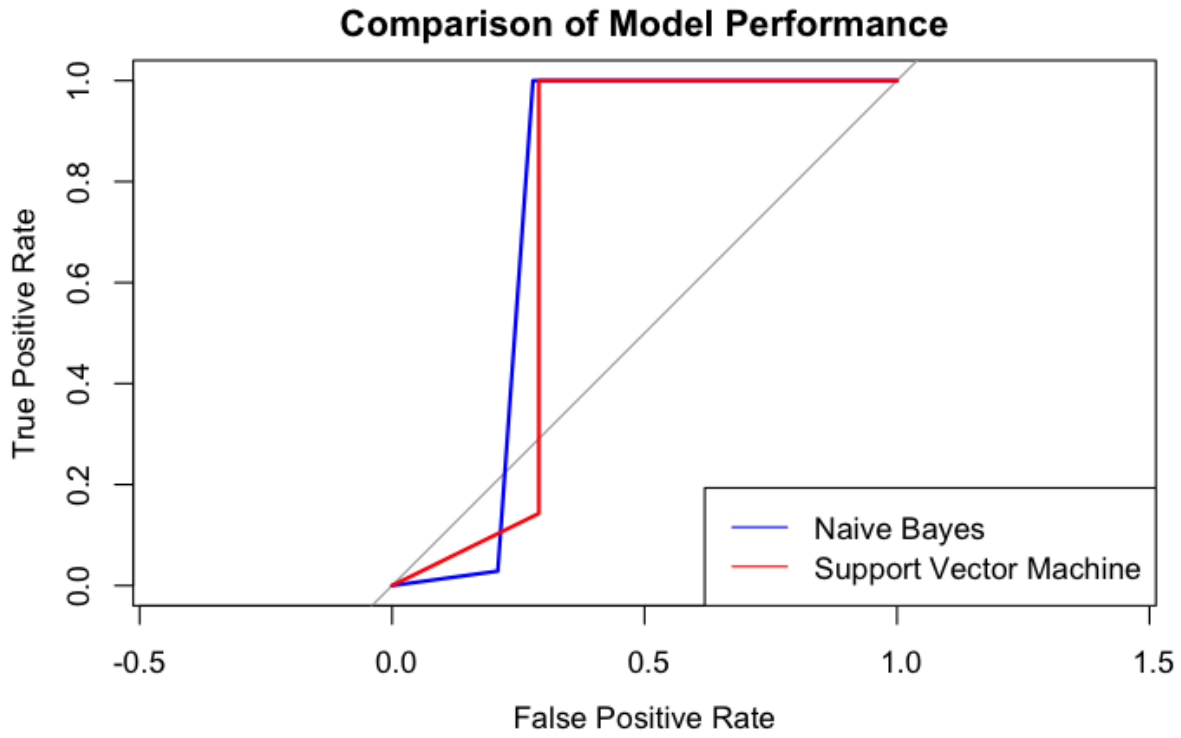

Figure 4: ROC Curve for Comparison of model performance

Table 8: Confusion Matrix for the SVM model classifications

	First Class	Second Class Lower	Second Class Upper	Third Class
First Class	61	0	0	0
Second Class Lower	0	120	9	6
Second Class Upper	25	20	180	0
Third Class	0	0	0	5

Table 9: Model Performance for the SVM

Overall Accuracy	95% CI	P-Value
85.9%	(0.82, 0.89)	.000

Table 10: Other Performance metrics for the SVM

	Sensitivity	Specificity	Precision	F1-Score
First Class	71%	100%	100%	83%
Second Class Lower	86%	95%	88%	87%
Second Class Upper	95%	81%	80%	87%
Third Class	45%	100%	100%	62%

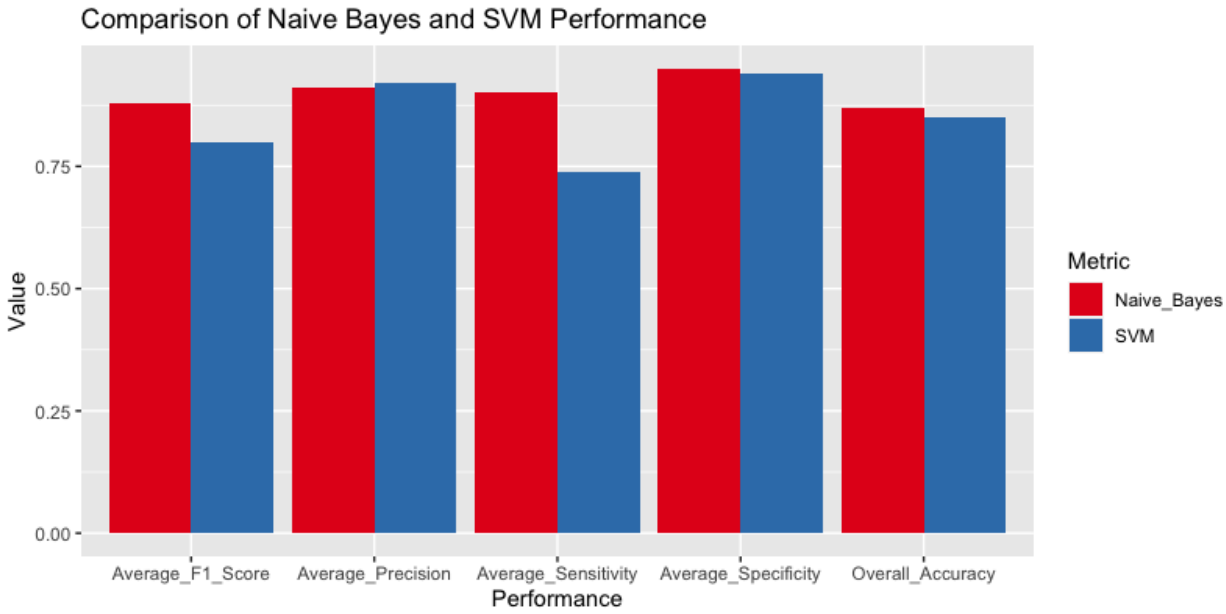

Figure 5: Metrics for comparison of Naive Bayes and SVM

The comparison between the Naive Bayes model and the Support Vector Machine (SVM) model in Figure 5 above provides valuable insights into their respective performances in classifying the target variable. The comparison of performance metrics beyond accuracy further validates the strong performance of the Naive Bayes model. While accuracy provides an overall measure of correct classifications, exploring additional metrics allows for a more comprehensive evaluation of the model's effectiveness. The strong performance of the Naive Bayes model across multiple evaluation metrics reaffirms its reliability and robustness in classifying the target variable. Its ability to handle categorical features, account for feature independence given the class variable, and generate accurate classifications contributes to its favourable performance.

Feature Importance

Table 11 reveals that the "Department" as a variable emerges as the most important feature for classifying the students' performance. Following "Department," the variable "Grade_100L" is identified as the second most important feature. Furthermore, the "Monthly allowance" is ranked as the third important variable. This suggests that the amount of monthly allowance received by students may affect their performance. These findings highlight the multifaceted nature of factors that can influence academic outcomes, encompassing aspects such as academic progression, study habits, and satisfaction with the learning environment

Table 11: Feature importance percentages

Features	**Percentage of Importance**
Department	**24.3%**
Grade_100L	**17.5%**
Monthly_Allowance	**10.6%**
Level	**6.8%**
Weekly_Study_Time	**6.1%**

Satisfied_wth_LS	**5.3%**
Engage_Grp_Disc	**4.6%**
Engage_Sch_Pol	**3.8%**
Attain_75%_Att	**3.6%**
Engage_Curr_Act	**3.5%**
Learning_Method	**3.3%**
Accomodation_Type	**2.9%**
Gender	**2.7%**
Age Group	**2.7%**
Educated Parent	**2.3%**

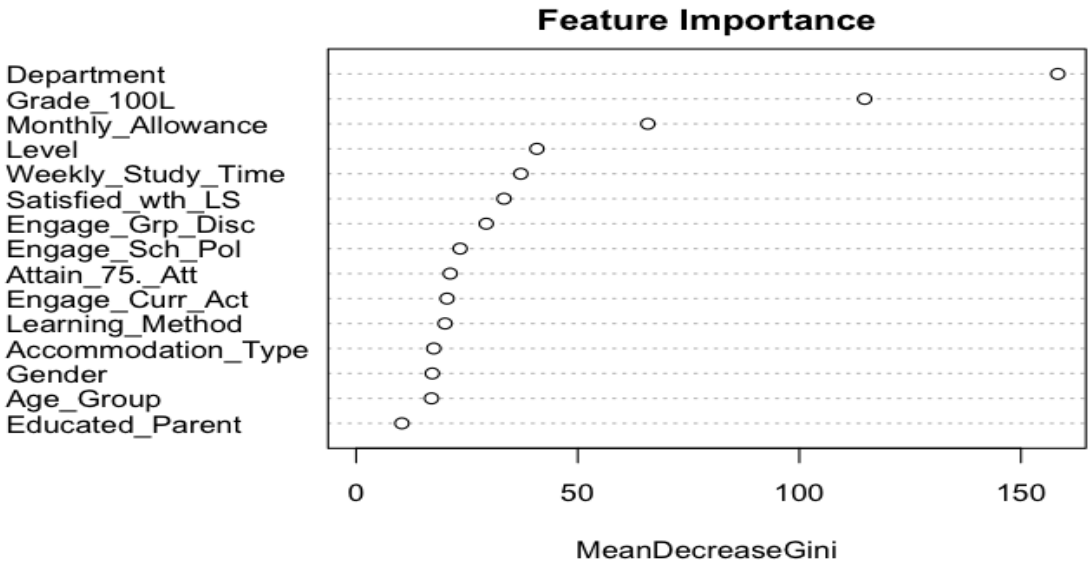

Figure 6: Important features in model classification

4 Conclusion

This study has examined the application of classification modelling technique, namely Naive Bayes, in classifying students' grade classes. There was also model comparison between Naive Bayes and SVM, as well as feature importance analysis that was conducted to identify the key factors influencing student performance. The results of this study demonstrated the effectiveness of Naive Bayes in classifying grade classes, achieving an overall accuracy of 87%. Furthermore, Support Vector Machine (SVM) achieved an overall accuracy of 85.9% in the study. SVM constructs a hyperplane to separate different classes and is particularly adept at handling high-dimensional data. While SVM's accuracy was lower compared to Naive Bayes, it still exhibited reasonably good performance. In conclusion, these findings can guide educational institutions and policymakers in identifying key areas for intervention and support to enhance student outcomes by focusing on crucial factors like department affiliation, early academic performance, and financial well-being; and targeted interventions can be designed to improve student success and overall educational quality.

Consent

As per international standard or university standard, Participants' written consent has been collected and preserved by the author(s).

5 References:

Abu, A., & Maghari, A. Y. A. (2017). Students performance classification using KNN and Naive bayesian. *IEEE Xplore*, 909–913. https://doi.org/10.1109/ICITECH.2017.8079967

Alturki, S., Cohausz, L., & Stuckenschmidt, H. (2022). Classifying Master's students' academic performance: an empirical study in Germany. *Smart Learning Environments*, *9*. https://doi.org/10.1186/s40561-022-00220-y

Amra, I. A. A., & Maghari, A. Y. A. (2017). *Students performance classification using KNN and Naive Bayesian*. IEEE Xplore. https://doi.org/10.1109/ICITECH.2017.8079967

Anwarudin Anwarudin, Widyastuti Andriyani, Bambang Purnomosidi DP, Dommy Kristomo (2022). The Prediction on the Students' Graduation Timeliness Using Naive Bayes Classification and K-Nearest Neighbor. Journal of Inteligent Softwares Systems. Vol. 1, No. 1, 75 – 88.

Arcinas, M., Sajja, G., Asif, S., Gour, S., & Okoronkwo, E., and Naved, M. (2021). Role Of Data Mining In Education For Improving Students Performance For Social Change. T¸rk Fizyoterapi ve Rehabilitasyon Dergisi/Turkish Journal of Physiotherapy and Rehabilitation. 32. 6519.

Arkana Yudhistira (2023). Student Performance Prediction Using Naive Bayes and Tree-Based Methods. https://rpubs.com/arkanayudhistira/students-performance-prediction

Ashraf, A., Sajid, A., and Gufran. M. (2018). A Comparative Study of Classifying Student's Performance by use of Data Mining Techniques. American Scientific Research Journal for Engineering, Technology, and Sciences. 44. 122-136.

Chang, V., Ganatra, M. A., Hall, K., Golightly, L., & Xu, Q. A. (2022). An assessment of machine learning models and models for early classification and diagnosis of diabetes using health indicators. *Healthcare Analytics*, 100118. https://doi.org/10.1016/j.health.2022.100118

Islam, H., & Haque, M. (2012). An approach of improving student's academic performance by using k-means clustering model and decision tree. *International Journal of Advanced Computer Science and Applications*, *3*. https://doi.org/10.14569/ijacsa.2012.030824

Jean-François Boulicaut . (2004). *Machine learning : ECML 2004 : 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004 : proceedings*. Springer-Verlag.

Kabakchieva, D., (2013). Classifying student performance by using data mining methods for classification. Cybernetics and information technologies, 13(1): p. 61-72.

Pavan Vadapall (2023). Naive Bayes Explained: Function, Advantages & Disadvantages, Applications. https://www.upgrad.com/blog/naive-bayes-explained/

Ray, S. (2019, September 3). *6 Easy Steps to Learn Naive Bayes Model (with code in Python)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Shahiri, Amirah Mohamed, Husain, W., & Rashid, N. A. (2015). A review on classifying student's performance using data mining techniques. Procedia Computer Science, 72, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Thant, Khin Shin, Ei, T., Khaing, Myat Mon, Myint, K. L., & Tin,. (2021). Evaluation of student academic performance using Naive bayes classifier. Advances in Computer and Communication, 1, 46–52. https://doi.org/10.26855/acc.2020.12.005

Usamah, M., Buniyamin, N., Arsad, P., and Kassim, R. (2013). An overview of using academic analytics to classify and improve studentsí achievement:A proposed proactive intelligent intervention, in: Engineering Education (ICEED), 2013 IEEE 5th Conference on, IEEE, pp. 126ñ130.

Udomboso C. G. Akanbi O. B. & Afolabi S. A. (2019). Application of Regression Type Estimator in Double Sampling Skills to Students' Enrollment in Oyo State. Global Scientific Journals. Vol. 7 no. 3\; 853 – 863.

Wati, M., Wahyu Indrawan, Joan Angelina Widians, & Novianti Puspitasari. (2017). Data mining for classifying students' learning results' https://doi.org/10.1109/caipt.2017.8320666