Student Dropout Prediction using machine learning

Midhilesh Dusanapudi
700732415
dept.Computer Science
University of Central Missouri
mxd24150@ucmo.edu

Sai Sree Chikkala
700726877
dept.Computer Science
University of Central Missouri
sxc68770@ucmo.edu

Preethi Bukka
700741930
dept.Computer Science
University of Central Missouri
PXB19300@ucmo.edu

Yamini Garikapati
700732498
dept.Computer Science
University of Central Missouri
yxg24980@ucmo.edu

[1].

***Abstract*—There are different factors that trigger student dropout rate in Universities or in schools.Now educational institutions are in search of finding the answers what factors trigger the most.As per the studies there are plethora of factors that can cause this situation.But manually analyzing the causes is a cumbersome task.A simple prediction tool for classification of dropout or consistent student helps the organization improve the dropout rates.Machine learning offers simple and user-friendly techniques to predict using complex dataset. Student dropout prediction is a natural problem in educational institutions and is current challenge to overcome. This dropout rate causes vary depending on the type of institution online or offline, student economical background, student participation in the curricular activities, instructor's role etc.In this paper we are implementing 3 types of features selection methods Correlation matrix, Univariate feature selection and wrapper methods. These methods are implemented using scikit-learn library. After applying the feature selection methods classification of student data into drop out or consistent is performed using machine learning algorithms Naive Bayes and KNN(K Nearest Neighbor). To conclude the project we conducted the comparative analysis of the 3 methods. The data used in the project is a private dataset and the features are mainly categorized into academic and non academic parameters.Academic parameters are:SEM(1-8)SGPA,SEM(1-8) KT,Hours on, assignment, Hours on studies, 2 hr lectures focus score and the non academic parameters are Social skills, Mode of transportation, Traveling time, Internet availability, Internet speed, Attendance, 5 hr practice and Teacher's feedback.**

***Index Terms*—Feature selection, KNN(K Nearest Neighbors), KT(keeping terms), Semester Grade Point Average(SGPA), Correlation matrix**

I. Introduction

One of the major challenges in the education system is to limit the dropout rates of the students.However the root cause of the problem varies depending on the data and situation. Sometimes we can not find out the reason by manual analysis. For any educational institution a robust student dropout prediction model helps the management and even the government.Predicting this kind of problems involves identifying the patterns in the data and selecting the most important features. Machine learning methods help to analyze the features visually and are helpful in drawing the conclusions in any situation without manual intervention.The objective of our project is to select the features that affect the classification of two classes "likely dropout" and "normal" student using correlation , filter and wrapper feature selection methods.In the end a comparative analysis is conducted to choose the best model. The features of the project is diversity of the data which includes different categories of the data. The datasets include the academic records to predict the dropout rates. As the research suggests, not only academic performance, other factors like traveling model , lack of interest and socioeconomic parameters also contribute to the dropout rate.

In this paper we are implementing various feature selection methods to select important features and classification algorithms.For experimental analysis we have used a private dataset.Data is collected through google form in the process of information gathering various categories of information is collected.This can be categorized into student scores, economical background, family details, student participation in non-academic events and so on.Since the data contains both numerical and categorical values feature selection process is implemented individually for each category.These feature selection methods are adopted from scikit-learn library.Selected features are feed to the classification algorithms.In this case multiple classification algorithms are designed.Expected output for this model is whether the student is a dropout or consistent one.

Feature selection methods in machine learning improves the model in following ways:

- Avoids overfitting of model
- Eliminates the complexity of the data
- Refines the accuracy of the model
- Reduces the training time of the model

There are 3 types of feature selection methods: filter, wrapper and hybrid. One of the feature selection methods is filter method.Filter method selects the features using correlation with dependent variables.A correlation matrix is constructed internally and selects a subset of most correlated features .On the other hand wrapper method generates subset of features while training the model.Hybrid method is a combination of

[1]https://github.com/Saisree16/Final-Project

both wrapper and filter methods.Each method has its pros and cons.In this paper we are proposing all the methods and conduct a comparative analysis .Student prediction dataset contains multiple columns which correlated with the dependent values.These can be boiled down to reduced set using these methods.Hybrid feature selection method ensures the best of both the algorithms i.e. filter and wrapper.Again this methods are subdivided into feature selection for categorical values and numerical values.For categorical values chi-squared test is implemented.Chi-square test is statistical procedure to measure how good the fit is.This is one of the best method to calculate the subset of corpus.Mutual Information calculates how independent feature depends on independent feature if there is no relation then the Mutual Information value is set to zero.For numerical values Mutual Information and Pearson correlation is used. Training time of the model depends on various parameters. One of the most important factors is complexity of the data and dimension of the input features.Reducing the dimensions of the input features improves the training performance.

Our governments are looking for prevention of student dropouts in various educational institutions, schools, universities and online educational institutions. The scenario of the online educational system is completely a different story compared to physical institutions.Early prediction of potential dropout rate we can provide proper attention and focus on the specific area which the student is lacking . Research suggests that drop rate reasons are not only limited to the academic performance factors and include non academic reasons also.With the correlation matrix the selected features are attendance, internet availability, mode of transportation and social skills are the features that impact the target variable. After the feature selection method machine learning algorithms Logistic regression and Naive bayes classifiers are used to predict the results.

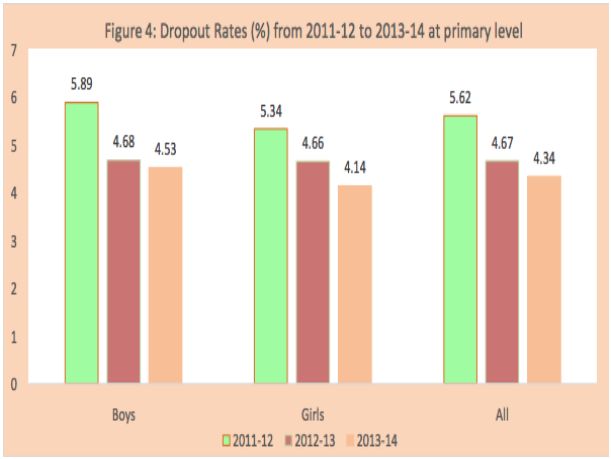

Figure 1. Student dropout rates

The above figure shows the dropout rates of the students it is in decreasing order but the rate of the dropouts is still a growing concern. When we see this in the level of urban and rural we can see the big picture of the problem. All the institutions are looking for the solution to this problem through an efficient method.

In general the data of this kind of problems comes with complexity. To break the complexity we need to find the patterns and remove the unnecessary things from the data. For this machine learning offers various feature selection methods to select the important features from the data. The choice of choosing the feature selection method depends on the input and output of the model. The input and output of the data are categorized into numerical and categorical.For example input can be numerical or categorical and at the same time output can be numerical or categorical. If the variables or input and output are categorical then we need to choose Chi-square method. In machine learning we don't need to build the feature selection methods from scratch sklearn library offers the feature selection methods those are ready use. Some of the feature selection methods that are offered by sklearn are : RFE(Recursive Feature Selection), SelectKBest and ANOVA feature selection methods. If we understand the working of feature selection methods they eliminate the unnecessary features and select the important features using different techniques. For example correlation is feature selection method where the features are selected based on the correlation score. First it will calculate the covariance of the features one by one if the score of the features is higher than the threshold value it will select that feature as highly correlated feature.

II. Motivation

Educational institutions also adapting the new technologies in decision making process. This is one of the industry where decision making is driven by data. It produces the data where it should be analysed before making the decisions. Machine learning algorithms with powerful data analysis and interpretation techniques makes the process easy. Data of student dropout prediction comes with data of different units and categories. To interpret the data we need powerful tools like analysis tools to understand the underlying structure of the data and classification tools to classify the given data into different categories. With the recent advent of AI and machine learning this process became easy. The data analysis tools like pandas and visualization tools seaborn and matplotlib gives the way to analyze the data deeper. Machine learning classification algorithms powerful with less computational power and we have the flexibility of choosing the various performance metrics to analyze the models.

III. Objectives

Main objectives of the project are:

- The objective of our project is to classify the dropout students from the given data
- Applying various feature selection methods to select the important features from the dataset.
- Training the machine learning binary classification models to predict the dropout student

- Testing the machine learning algorithms with performance metrics and conducting comparative analysis
- Deploying the best model into web application

IV. Related work

Huge amount of data is produced in different organizations. One of the organizations is educational institution. Huge amount of data is produced in the institutions that is data of the professors and students. In the students data category we have different kinds of data. Student dropout data is our main topic which can be used for the prediction of dropout of the institutions. Manually analyzing the data is a cumbersome task. So applying machine learning techniques on this data makes the process easy [2]. With the growing concern of this topic we can use Data mining techniques. With the help of this method we can draw conclusions from the data.One of the popular method of data mining is CRISP-DM crisp method in data mining. In this paper we have collected the Cobalto dataset and applied EDM(Educational Data Mining) on the data. And they achieved the accuracy of 77.24 percent with decision Tree algorithm. 85 percent accuracy on the Random Forest algorithm. The challenge of predicting the student dropout prediction is it depends on the multiple factors. The main task is to separate the most influential parameters corresponding persistent and dropout student.In the first step we have applied a feature selection method chi-square for selecting the features from the dataset. In this paper we have implemented the tree based algorithms those classification and regression tree algorithm and CHAID tree. After the experimental results we can conclude that CART(Classification and Regression Tree) outperformed the CHAID tree. In the implementation part we have implemented the 3 stages of CHAID and 4 levels of CART. In conclusion we can say that age, gender, ethnicity are the most influential parameters [11].

Predicting the dropout and prevention of the system is necessary for many reasons.In this paper we have proposed student dropout risk prediction system. In this we have used the data of 11,000 students recorded in the span of 5 years.With the data we have performed the classification task using c5.0 and created a system of SPA a dropout prevention system in Spanish. This model is in production currently [13].

Online coursed also experience the dropout rates. Major factors that effect the dropout is behavioural patterns. In this paper we have collected the behavioral patterns of the data and trained the Convolutional Neural Network to predict the risk of the dropout from the given data. FDD cup 2015 dataset is used to train and predict the results [20]. When we analyse the root cause of the student dropout problem we found out the some factors with the survey. People are enrolling the courses with some goal in mind but after sometime they are dropping from the course with various reasons. For any institution having the clear understanding of the reasons for the droput helps in working on the problem.In this paper we have implemented the SMOTE analysis on the dataset using machine learning and with the subsequent method PCA for efficient results. When compared the SMOTE-PCA model with individual methods the hybrid method gained the high accuracy [16].

In the recent years promoting the higher education is the key to improve the quality of education in the institutions. In this paper we have studied the important factors that effect the dropout in the institutions. Data is collected from King Mongkut's University of Technology Thonburi (KMUTT) on this data we have applied the machine learning algorithms RandomForest, Gradient Boosting and decision tree algorithms. At the end of the comparative analysis we have concluded that Gradient Boosting algorithm outperformed the other algorithms. In the feature selection step we have identified academic year, high school GPA, channels of university admission, gender as the top 5 important features [19].

In general when it comes to student dropout prediction the traditional models focus on the academic scores to classify the dropout students. But as per the recent studies we have observed that other factors like extra curricular activities also play an important role in the student dropout prediction. In this paper we are conducting a study on importance of the extra curricular activities [5].

To improve the performance of the models in machine learning in predicting the student dropout prediction system is using the auto encoders to denoise the encoder and it is used in CNN architecture to train the model. This system is used to alert the higher rates of the dropouts in a any organisation and also it helps to take the precautionary measure to reduce the dropout rates [8].

Data mining techniques are more popular methods to analyze the student data. With these techniques we can analyze the chances of higher education of the students. This helps the students future as well as the institution's reputation. This is not an intelligent machine learning model which can analyze the data it is a simple web application which can displays the chances of higher education using Naive Bayes classifier [3].

Student dropout rates are not only limited to the physical classrooms these are there in the online classes as well. To predict the dropout chances in virtual class rooms we have to use the robust networks for analyzing the data ANN(Artificial Neural Networks) are the deep neural networks with deep neural structures analyzes the data and gives the best results compared to the non neural networks [10].

The other reasons for student dropout rates are student teacher dropout rates. Students dropout from the institutions because of various parameters but in the recent surveys it is observed that dropping of the teachers also has an impact on the students dropout. Many countries dream is implementing the healthy learning environment. In this paper we are proposing the student-teachers risk of dropping out from the institutions. The model mainly focused on the least developed countries where the education system is at risk. This model helps the system to create a support system to avoid the potential dropouts from

the institutions and create a system to improve the learning in the institutions. For that we have used a multi stage Logistic regression to classify the results [18].

Before analysing the student dropout prediction system this prediction depends purely on the academic record so analysing the academic performance gives the picture of the dropout system. So in this paper we are analysing the student academic performance into two groups pass or fail. This system working not only limited to the simple classification of pass or fail of the data. It also groups the students into poor, good and average. In the second step we have used several feature selection and extraction techniques to create the model. In the third step we have calculated the correlation of the features through correlation matrix and heat map of the correlation matrix. The advantages of this system is the execution time is very less compared to the traditional models and with the accuracy of 95 percent [6].

Unlike physical classrooms in MOOC online platforms the dropout rates are categorized based on the different parameters. In this platform the dropout rates purely depend on the behavioral aspects. In this system we are using TSF(Time Series Forest) algorithm for classification of dropout or persistent student groups. In this we have used the dataset only to the 5 percent and achieved 85 percent accuracy and the experimental results show that the accuracy grows with the increase of dataset. This system helps the instructors to take the measuring steps towards the dropout [4].

Educational Data Mining (EDM) helps to analyze the different techniques related to educational institutional related problem solving techniques for example student performance analysis, student cognitive capabilities analysis, course delivery analysis etc., in this our topic of research is to classify the dropout rates. This technique offers different data analytical tools. Several traditional models offer the same techniques but they lack the interpret ability. The advantage of interpret ability is it makes the decision process easy. Having a clear interpret-ability in place gives the idea of cause and effect of the problem and the solution which is very important in considering the precautionary measures of the problem [1].

Although the MOOC's are the ever growing field in the category of educational institutions and the trend of the dropout rates also in the same manner. In a novel approach we are proposing a graph model to predict and classify the dropout rates.Graph models give the advantage of more interpret ability compared to the other traditional machine learning algorithms. In the coclusion part we have visualized the realtionship between the student behaviour and the dropout rates in the gaph based model [12].

With the growing industry of Information and Communications Technology(ICT) one of the industry that is growing rapidly is educational systems makes us to understand the data produced by the educational category. The approaches of data mining help Learning Management System of the institutions. In this paper we are focusing on the behavioural patterns of the students before and after the online course for example downloading the files from the resources, interactions with the instructor and practising the exercises after the course completion and knowledge gain from the course factors are recorded in the database and analysed to classify the dropout rates of the student [17]. In this paper we are proposing a System Engineering(SE) application to predict the dropout rates using the student data collected from 7 years of Columbia University. This data is preprocessed and applied various machine learning algorithms like Logistic Regression, Naive Bayes and Decision Trees. In this approach we have used Watson Analytic to analyse the data. The main objective of this project is to find the causes of the dropout rates for better decision making. Moreover this is used to analyse the data quality so that we can collect the data when we record the samples [15]. Dropout rate is a concerning thing in countries like Peru. So data is collected from one of the private university of Lima and this data consists of records of the 500 undergraduate students. Educational Data mining techniques are implemented on this data and the results show that Bayesian classifiers out perform the decision trees in the metrics like precision, recall and F1 score. In the comparative analysis we have found that Bayesian algorithms give 67 percent accuracy where as the decision trees gives the 62 percent accuracy [9]. Most of the existing models focus on the classification of the dropout rates. But in this approach we focus on the quantitative approach of finding the who and when the drop out will happen. It is also a better approach to know when the dropout will happen.In this paper we have proposed three stage model to classify the dropout prone students , non dropout students and the classification model to predict when the dropout will happen. In this study we have used the kappa value as performance metric to analyse the performance of the model [14].This is a challenge in distance education as well. In this paper we are implementing the algorithms to classify the early dropout rates in the distance education system [7].

V. Data Description

In our project we have used a private dataset that is collected from Github repository.As per the information provided in the source. The data is collected from various sources and situations. Feature description:

S.No.	Column name	Description
1.	SEM(1-8) SGPA	semester Grade Point Average of 8 semesters
2.	SEM(1-8) KT	8 semesters keeping terms
3.	Hours on assignment	Hours spent on assignment
4.	Hours on studies	Hours spent on studies
5.	2 hr lectures focus score	score from 2hrs continuous focus score
6.	5hr lecture focus score	5hr lecture score based on focus
7.	Coaching classes	Number of coaching classes
8.	Average pointer	Average grade pointer
9.	Social skills	score based on social skills
10.	Internet availability	Internet availability yes or no
11.	Internet speed	speed on the scale of
12.	Attendance	Attendance on the scale of
13.	5 hr practice	speed on the scale of
14.	Teacher's feedback	string consists of feedback

Table I
DATASET FEATURE DESCRIPTION

VI. Proposed Framework

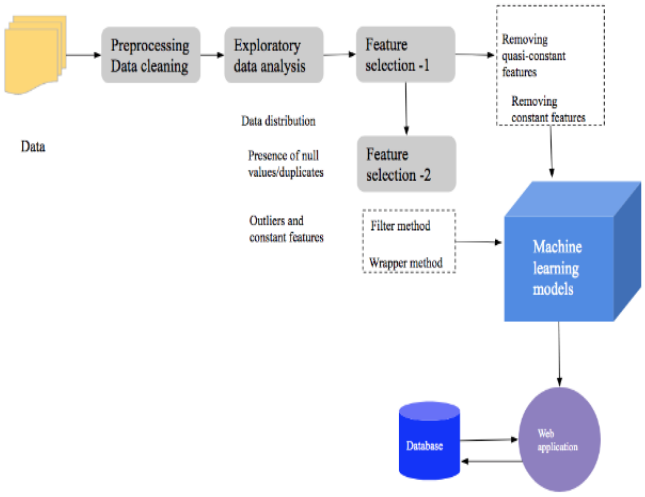

Figure 2. Workflow

Our project is divided into 4 stages: data cleaning, exploratory data analysis, feature selection and machine learning modeling. In the data cleaning part we have observed the statistical view and overall view of the dataset using Pandas library.

In the next step exploratory data analysis is conducted to observe the underlying structure of the dataset. In this step various graphical representations are used to check the relationship of the data.In the data exploration step we have observed presence of constant and quasi constant features in the dataset. So these can be removed using scikit-learn variance threshold method.

Constant features are the values same throughout the dataset i.e. constant.These constant features do not add any value to the machine learning models. Removing these values is the first step in feature selection.sklearn variance threshold removes the constant values which doesn't meet the given threshold value

Removing quasi constants is the next step in feature selection. Quasi constant features the values present majority of the time in the dataset.For this we are using the same method like removing constants i.e. variance threshold with threshold value 98 percent. We should be very careful while removing the quasi constant features it may remove the outliers of the data. Removing constant and quasi constant features is the basic steps.After this we have to apply filter, wrapper methods to select the best features.

In the filter method there are various types of feature selection methods.In our project we are using Correlation matrix, ANOVA feature selection, Uni variate selection methods.In the wrapper methods we have used Recursive Feature Elimination(RFE) and SFS(Sequential Forward Selection and backward selection) methods.

To check the accuracy of the selected features only Naive Bayes classifier is implemented. In future these results are going to be published on a web application.Web application is designed using python programming language for back end and HTML and CSS for the front end part.

A. Exploratory data analysis

In the Exploratory Data Analysis we have explored the following things:

- Dataset contains balanced classes of 0 and 1
- Here 1 represents persistent student
- 0 represents the dropout student Internet availability count plot shows the internet availability is high for persistent student compared to the dropout student and the graphs of SEM-KT for each shows that till the 4th semester we have KT increasing but from the 5th semester it is constant. These graphs show the importance of the features to retain or eliminate while training the model. All the visualizations are drawn using seaborn visualization tool.

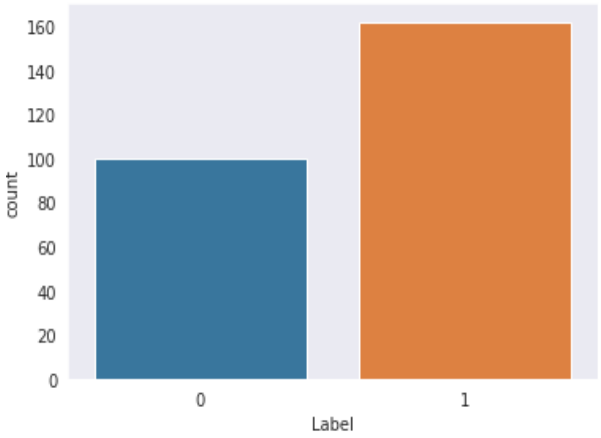

Figure 3. Target values distribution

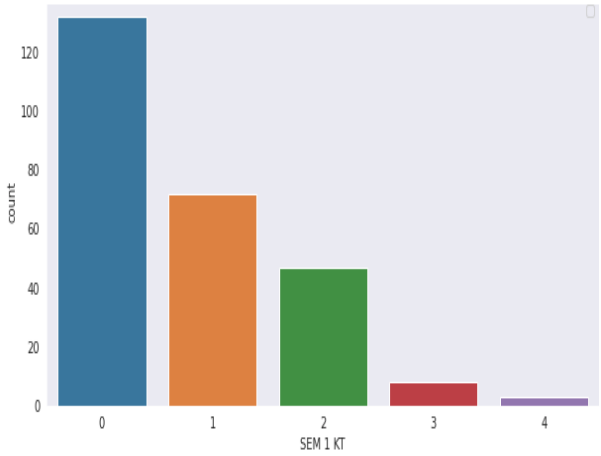

Figure 4. KT for SEM1

From the above visualizations we can observe that the target values are not equally distributed to avoid this we have used cross validation techniques to test the model. From each

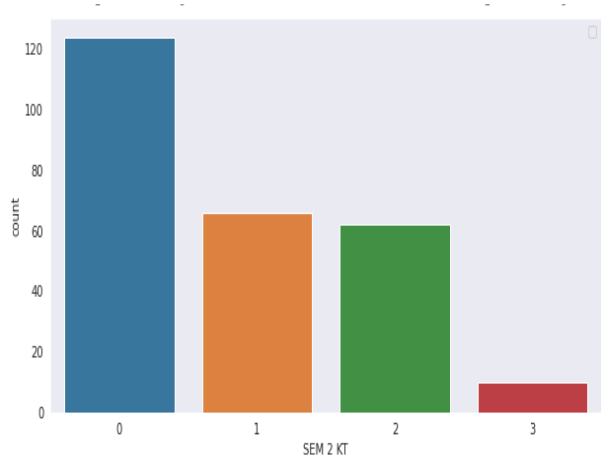

Figure 5. KT for SEM2

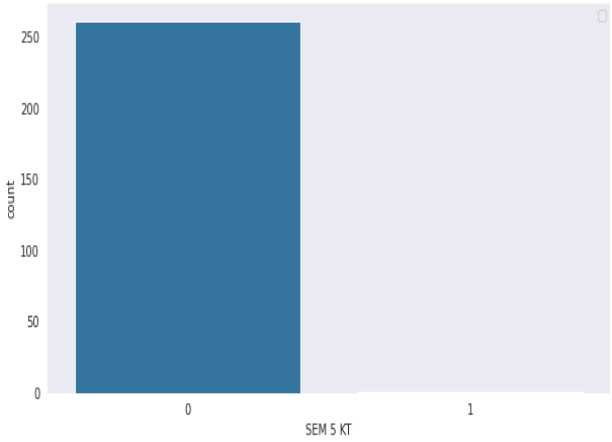

Figure 6. KT for SEM 5

cross validation we have checked the score and performed the optimization of the algorithm if required.

B. Methods

After the visualization we can choose the algorithms to perform the classification of the data and feature selection methods. In this paper we are proposing the 3 kinds of feature selection methods those are filter, wrapper and hybrid feature selection methods. In this project we have two stages of feature selection methods the first stage is basic feature selection from the data that is removing the constant and quasi constant features from the data. In the second stage we are removing the unnecessary features from the data and we select the important feature based on the score of the features. These selection methods are implemented using scikit-lean. The purpose of each feature selection method is different . In the end each feature selection method is compared to check the performance of the each model. After feature selection various machine learning algorithms are trained on the reduced features.

VII. Results Analysis

A. Filter methods

- 1. Filter methods generally a part of preprocessing of data(removing constants and removing quasi constants)
- 2. These methods generally do not use machine learning algorithms to select the important features instead they use statistical tests
- 3. filter methods usually include: Correlation matrix. ANOVA(Analysis of variance). Univariate selection methods.
- 4.The characterstics of filter methods are they are computationally inexpensive Well suited for quick removal of irrelavant features and low prediction power

Correlation matrix calculates how close two variables are to have linear relationship When two features have high correlation they will have same effect on dependent variable hence we can drop one of them. From the correlation matrix we have finalised the features SEM 1 SGPA, SEM 1 KT, SEM 2 KT, SEM3 KT, SEM 4 KT, SEM 6KT,Hours On Assignment,Hours On Studies,Travel Time,Attendance, Internet Availability ,Mode Of Transportation',2 hrs lect, Submissions, 5 hrs lect, 5 hrs pracs, Coaching classes, Social Skills. After selecting the features using correlation we have traine two models Naive Bayes and KNN algorithm. Test results show that Naive Bayes achieved 98 percent accuracy and KNN algorithm achieved 100 percent accuracy on the test data.

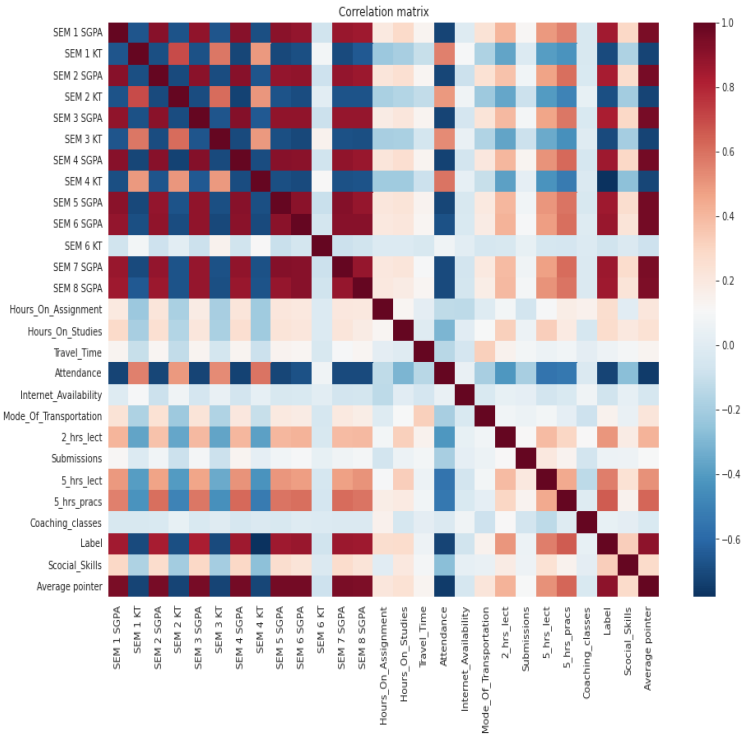

Figure 7. Correlation matrix

B. Uni variate Feature selection

Since correlation does not capture the non linear relationship between the variables. We are selecting the uni variate feature selection. In the uni variate feature selection we have implemented SelectKBest algorithm with f calssif function using scikit-learn library. Train data is used to train the models with the selected features. Following figure shows the scores of the features.A highly correlated feature is given high value and less correlated features is given low weight.

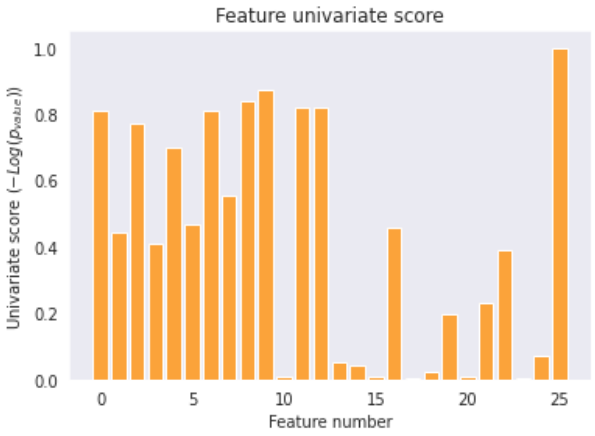

Figure 8. Features vs feature scores univariate feature selection

C. Wrapper method

Unlike filter methods Wrapper methods use series of models to generate subset of features. Based on the previous training model observations to will add or remove the new features. Recursive Feature Elimination with cross validation:RFE eleminates the features 0 to N iteratively based on accuracy score using cross validation. Recursive Feature Selection method is implemented using scikit-learn library. This is trained using stratified cross validation with 5 folds. For each fold the accuracy is calculated. After the cross validation the optimal number of features are selected that is 5. 2. In the wrapper method in addition RFE we have implemented the forward and backward feature selection methods. Features selected by forward sequential selection: 'SEM 1 KT' 'SEM 3 KT' 'SEM 4 KT' 'SEM 8 SGPA' '2 hrs lecture' '5 hrs practice','Average pointer'. Features selected by backward sequential selection: 'SEM 1 SGPA' 'SEM 1 KT' 'SEM 3 KT' 'SEM 4 KT' 'SEM 8 SGPA' '2 hrs lecture','5 hrs practice'. These two are implemented using sequential forward selection method in sklearn in the SFS forward or backward or mentioned using direction parameter.If the direction is forward then it is called as forward SFS and if the direction is backward then it is called. In the recursive feature elimination method it creates the subsets of the features and recursively eliminates the features. In the Sequential feature selection features are eliminated or selected based on the correlation eliminated in backward or forward fashion.

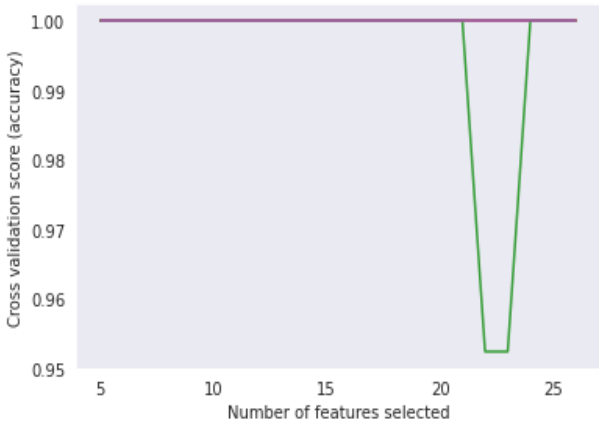

Figure 9. Cross validation score on the features

	precision	recall	f1-score	support
0	0.94	1.00	0.97	16
1	1.00	0.97	0.99	37
accuracy			0.98	53
macro avg	0.97	0.99	0.98	53
weighted avg	0.98	0.98	0.98	53

Figure 10. Classification report Naive bayes

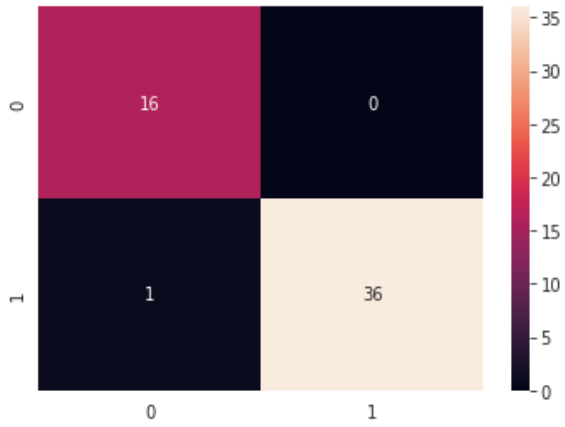

Figure 11. Confusion matrix naive Bayes

VIII. Results summary

With the experimental analysis on feature selection method all the methods performed filter, univariate and wrapper methods.Amongst all the algorithms Naive Bayes performed well without overfitting.

REFERENCES

[1] Manjari Chitti, Padmini Chitti, and Manoj Jayabalan. Need for interpretable student performance prediction. In *2020 13th International Conference on Developments in eSystems Engineering (DeSE)*, pages 269–272. IEEE, 2020.

[2] Alexandre G Costa, Emanuel Queiroga, Tiago T Primo, Júlio CB Mattos, and Cristian Cechinel. Prediction analysis of student dropout in a computer science course using educational data mining. In *2020 XV Conferencia Latinoamericana de Tecnologias de Aprendizaje (LACLO)*, pages 1–6. IEEE, 2020.

[3] Tismy Devasia, TP Vinushree, and Vinayak Hegde. Prediction of students performance using educational data mining. In *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, pages 91–95. IEEE, 2016.

[4] Liu Haiyang, Zhihai Wang, Phillip Benachour, and Philip Tubman. A time series classification method for behaviour-based dropout prediction. In *2018 IEEE 18th international conference on advanced learning technologies (ICALT)*, pages 191–195. IEEE, 2018.

[5] Tomas Hasbun, Alexandra Araya, and Jorge Villalon. Extracurricular activities as dropout prediction factors in higher education using decision trees. In *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*, pages 242–244. IEEE, 2016.

[6] Abhinav Jain and Shano Solanki. An efficient approach for multiclass student performance prediction based upon machine learning. In *2019 International Conference on Communication and Electronics Systems (ICCES)*, pages 1457–1462. IEEE, 2019.

[7] Georgios Kostopoulos, Sotiris Kotsiantis, Omiros Ragos, and Theodoula N Grapsa. Early dropout prediction in distance higher education using active learning. In *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pages 1–6. IEEE, 2017.

[8] Jong Yih Kuo, Chia Wei Pan, and Baiying Lei. Using stacked denoising autoencoder for the student dropout prediction. In *2017 IEEE International Symposium on Multimedia (ISM)*, pages 483–488. IEEE, 2017.

[9] Erik Cevallos Medina, Claudio Barahona Chunga, Jimmy Armas-Aguirre, and Elizabeth E Grandón. Predictive model to reduce the dropout rate of university students in perú: Bayesian networks vs. decision trees. In *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–7. IEEE, 2020.

[10] Hermel Santiago Aguirre Montaño and Ma Carmen Cabrera-Loayza. Early prediction of dropout in online courses using artificial neural networks. In *2020 XV Conferencia Latinoamericana de Tecnologias de Aprendizaje (LACLO)*, pages 1–6. IEEE, 2020.

[11] Mohammad Nurul Mustafa, Linkon Chowdhury, and Md Sarwar Kamal. Students dropout prediction for intelligent system from tertiary level in developing country. In *2012 International Conference on Informatics, Electronics & Vision (ICIEV)*, pages 113–118. IEEE, 2012.

[12] Izumi Nitta, Ryo Ishizaki, Masafumi Shingu, Satoshi Nakashima, Koji Maruhashi, Arseny Tolmachev, and Masaru Todoriki. Graph-based massive open online course (mooc) dropout prediction using clickstream data in virtual learning environment. In *2021 16th International Conference on Computer Science & Education (ICCSE)*, pages 48–52. IEEE, 2021.

[13] Alvaro Ortigosa, Rosa M Carro, Javier Bravo-Agapito, David Lizcano, Juan Jesús Alcolea, and Oscar Blanco. From lab to production: Lessons learnt and real-life challenges of an early student-dropout prevention system. *IEEE transactions on learning technologies*, 12(2):264–277, 2019.

[14] Daniel A Gutierrez Pachas, Germain Garcia-Zanabria, Alex J Cuadros-Vargas, Guillermo Camara-Chavez, Jorge Poco, and Erick Gomez-Nieto. A comparative study of who and when prediction approaches for early identification of university students at dropout risk. In *2021 XLVII Latin American Computing Conference (CLEI)*, pages 1–10. IEEE, 2021.

[15] Boris Perez, Camilo Castellanos, and Dario Correal. Applying data mining techniques to predict student dropout: a case study. In *2018 IEEE 1st colombian conference on applications in computational intelligence (colcaci)*, pages 1–6. IEEE, 2018.

[16] M Revathy, S Kamalakkannan, and P Kavitha. Machine learning based prediction of dropout students from the education university using smote. In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 1750–1758. IEEE, 2022.

[17] Edisson Sigua, Bryan Aguilar, Paola Pesántez-Cabrera, and Jorge Maldonado-Mahauad. Proposal for the design and evaluation of a dashboard for the analysis of learner behavior and dropout prediction in moodle. In *2020 XV Conferencia Latinoamericana de Tecnologias de Aprendizaje (LACLO)*, pages 1–6. IEEE, 2020.

[18] Harman Preet Singh and Hilal Nafil Alhulail. Predicting student-teachers dropout risk and early identification: A four-step logistic regression approach. *IEEE Access*, 10:6470–6482, 2022.

[19] Warit Tenpipat and Khajonpong Akkarajitsakul. Student dropout prediction: A kmutt case study. In *2020 1st International Conference on Big Data Analytics and Practices (IBDAP)*, pages 1–5. IEEE, 2020.

[20] Yafeng Zheng, Zhanghao Gao, Yihang Wang, and Qian Fu. Mooc dropout prediction using fwts-cnn model based on fused feature weighting and time series. *IEEE Access*, 8:225324–225335, 2020.