National Research University Higher School of Economics

Graduate School of Business

Master’s Programme “Big Data Systems” (“Business Analytics and Big Data Systems” for 1-st year students)

**Project report**

Big Data Analytics for Educational Organizations: Student’s Performance Prediction

Performed by student

**Project supervisor:**

*Kawshiqe Ahamed*

*Associate Professor,*

*Morozova Yulia Alexandrovna*

*\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*(15.06.2022) (signature)*

**Moscow 2022**

**Assessment sheet**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(project title)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(project type)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(project duration)

|  |  |  |  |
| --- | --- | --- | --- |
| **Project supervisor**  Full name, position | |  | |
| **Student[[1]](#footnote-1)**: | |  | |
| Full name | |  | |
| Master’s Programme | | Big Data Systems | |
| Группа № | |  | |
| Components of the final grade[[2]](#footnote-2) | | Grade on a 10-point scale | | Notes (if necessary) | |
| **О пр** – Grade for the project result (product) | |  | |  | |
| **О сп** Grade for the methods and technologies used | |  | |  | |
| **О р** Grade for the implementation of the project’s work | |  | |  | |
| **О к** Grade for the developed competencies | |  | |  | |
| **О гр** Grade for the student's individual contribution to group work | |  | |  | |
| **О ком** Grade for team work | |  | |  | |
| **О з** Grade for the presentation, project defense | |  | |  | |
| **О вз** Grade given by other project participants (peers evaluation) | |  | |  | |
| **О с** Student self-assessment | |  | |  | |
| Grade calculation (formula with weight coefficient) | |  | | | |
| **Final grade for the project** | |  | | Project supervisor’s signature | |
| **Number of credits** | | 3 | |

Date\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Table of Contents**

Abstract……………………………………………………………………………04

1. Introduction……………………………………………………………………05
2. Background……………………………………………………………………06
3. Methodology…………………………………………………………………..09
   1. Data Visualization…………………………………………………………10
      1. Data Mining…………………………………………………………10
      2. Data Cleaning………………………………………………………..10
      3. Visualization………………………………………………………...10
   2. Machine Learning Models for Prediction………………………………….11
      1. What is Linear Regression ………………………………………….11
      2. Why Linear Regression ……………………………………………..12
      3. Key Assumption of Effective Linear Regression …………………..13
   3. Classification Models ……………………………………………………..13
      1. TensorFlow …………………………………………………………14
      2. GridSearchCV ………………………………………………………15
      3. Logistic Regression …………………………………………………16
      4. K-Nearest neighbors ………………………………………………..16
      5. Gaussian Naïve Bayes ………………………………………………17
      6. SVM ………………………………………………………………..18
      7. Decision Tree Classifier ……………………………………………18
      8. Random Forest Classifier …………………………………………..19
4. Experiments and Results ……………………………………………………...19
   1. Data Preprocessing for Visualization………………………………………20
   2. Visualization Results ……………………………………………………...22
   3. Data Preprocessing for Machine Learning Models………………………29
   4. Creation and Implementation of Regression Models……………………..31
   5. Predictions and Accuracies with Linear Regression……………………...33
   6. Data Preprocessing for Classification …………………………………….34
   7. Creation and Implementation of Classification Models………………….35
   8. Classification Report and Accuracy……………………………………….38
5. Conclusions ……………………………………………………………………41
6. References ……………………………………………………………………..42

**Abstract**

Student’s performance monitoring is an essential part a good education system which has been followed by all big educational institutes. To understand student’s performance and making curriculum and customized syllabus and care has been provided by student’s individual performance. Achieving well organized and best practices of technology has been applied to students according to their performance. Undoubtedly predicting performance[20] gives precautions to take initiatives both for students and organizations. Visualization tools and machine learning models[14] has been implemented to understand better predicting students’ performance. In this article we will demonstrate different models of machine learning with various python libraries such as pandas, numpy, pyplot, seaborn, sklearn to predict student’s performance, test model’s accuracy and cross model verification.

**1 Introduction**

Our main objective is to predict student’s performance where we need to focus more on scientific and applied methods and tools to get optimal results. To manipulate the results, we can us different tools of visualization and predictions. Most famous and user-friendly tools for visualization is Microsoft power BI, Tableau python and etc. For machine learning tools to predictions IBM Watson, Azure and obviously python are familiar and well renounced. Except python most of the tools are not customized and open source to use where python gives a lot of libraries for data mining, visualization and machine learning approaches[12]. So, we will use python to perform our scientific reports in a well-structured manner.

Traditional DM techniques are used in educational data mining (EDM) to address issues in the field of education. EDM refers to the application of DM techniques to educational data, such as student data, academic records, test results, participation in class, and the frequency of questions raised by students[5]. EDM has recently developed into a useful technique for seeing hidden patterns in educational data, forecasting academic success, and enhancing the learning and teaching environment.

Through the application of EDM, learning analytics has acquired a new dimension. Learning analytics includes the several facets of gathering student data together, comprehending the learning environment via examination and analysis, and identifying the optimal student/teacher performance. Learning analytics is the collection, evaluation, and reporting of data on students and their settings in order to comprehend and improve educational processes. It also covers the institutions' use of fresh tactics.

Predicting student academic achievement, identifying patterns of system access and navigational behaviors, and identifying students who may be at danger of failing are all aspects of learning analytics[29]. A student's potential behavior may be assessed by looking at the digital data left by learning management systems (LMS), student information systems (SIS), intelligent teaching systems (ITS), MOOCs, and other web-based education systems. These data may be used to analyze the behaviors of successful students and those who are at danger of failure using the EDM technique, to create remedial actions based on academic achievement, and ultimately to help instructors create pedagogical methods.

Learning management systems, easily implemented, have become an essential component of higher education, particularly during the epidemic time. The log data generated as a result of pupils using these technologies have become easier to access[21]. Universities should now increase their ability to use these data to forecast academic achievement and guarantee student advancement[6].

Consequently, by spotting hidden patterns in educational data, EDM gives educators fresh insights. To ensure the quality of education, several elements of the educational system may be assessed and modified using this approach.

**2 Background**

While collecting data of students and previous results is a big data challenging fact there predicting student’s performance is hard enough to be accomplished. When the necessity of modern technology used in education system previous analog system demolished automatically because of its inconvenience and deadlock strategy. Organizations and students are concerned about what are they doing and get feedback and assumption of what they are intentioned to be performed. That is why student’s skill measuring, learning patterns, sentiment analysis and predicting performance is a part and parcel of modern education system. Machine learning, Artificial Intelligence has been performed almost all of the sectors of modern science to get good output and service such as quantum physics, neuroscience, statistics and social network analysis. Here we are in the primary need of human kind the education system and this sector is also taking the advantages of machine learning and Artificial Intelligence[9]. From class schedule, mail, teacher’s performance, student’s performance, results and so on are dependent on machine learning. So, we are going to briefly discuss about machine learning models which is suitable for our work.

Asif [4] Focused on two facets of undergraduate students' performance utilizing DM techniques. Predicting students' academic success at the completion of a four-year program of study is the first component. The second is to assess pupils' growth and incorporate the findings into predictions. He split the class into groups with poor and high academic achievement. In order to provide early warnings, help struggling students, and provide guidance and chances to high-performing students, he has discovered that it is crucial for educators to concentrate on a limited number of courses that indicate exceptionally strong or bad performance. 16 demographic variables, including age, gender, class attendance, internet access, computer ownership, and the number of courses completed, were used to predict student academic achievement. Machine learning techniques including support vector machines, logistic regression, random forest, and k-nearest neighbors may accurately predict student performance to within 50 to 81 percent of the true value[11].

A machine learning-based algorithm was suggested by Ahmad and Shahzadi to determine if pupils were at danger of performing poorly in school. The forecast was generated with an accuracy of 85% using the students' learning abilities, study habits, and academic interaction traits[2]. The researchers came to the conclusion that identifying academically failed students could be done using the technique they suggested. A machine learning model based on learning techniques, perceptions of social support, motivation, socio-demographics, health status, and academic performance factors was suggested by Musso et al. in 2020[22]. He made predictions about academic performance and dropout rates using this approach. He came to the conclusion that learning methodologies had the biggest impact on predicting GPA, whereas background knowledge had the most impact on predicting dropouts.

Accurate prediction of student academic success necessitates a thorough comprehension of the variables and characteristics that affect student outcomes and student accomplishment[3]. Hellas evaluated 357 publications on student performance to determine the effects of 29 characteristics for this reason. These characteristics were mostly linked to psychomotor abilities, including course and prior performance, student involvement, demographics of the student body, such as gender, high school performance, and self-control[16]. However, factors such as student motivation, habits, social and financial problems, a lack of advancement, and professional changes have a significant impact on dropout rates.

According to the literature evaluation, it is essential to raise educational standards by foreseeing kids' academic performance and helping those who are at danger. In the literature, the prediction of academic performance was made using a wide range of variables, including student demographic information (gender, age, economic status, number of courses taken, internet access, etc.) and various digital footprints left by them online (browsing, lesson time, percentage of participation[13, 24]. Social support perception, motivation, socio-demography, health status, academic performance characteristics[10], homework, projects, quizzes, etc. are just a few of the factors that can affect how well students perform in school. Prediction accuracy in virtually all of the models created in these experiments ranges from 70 to 95 percent. However, gathering and analyzing such a wide range of data demands both a lot of time and specialized skills. Similar to this, Hoffait and Schyns (2017) argued that gathering so many data points is challenging and that socioeconomic statistics are superfluous[17]. Additionally, these socioeconomic or demographic statistics might not necessarily provide the best insight into how to avoid failure[6].

The study looks at forecasting pupils' academic performance only based on grades, without considering socioeconomic or demographic factors. The goal of this project was to create a novel model based on machine learning algorithms to forecast undergraduate students' final exam scores based on their midterm test marks, Faculty, and Department.

Using machine learning classification algorithms, it was possible to identify the classification algorithms that performed the best at predicting students' academic success. The decision to study Turkish Language I was made for the simple fact that it is a required course for all university enrollees. The final exam marks of the pupils were forecast using this methodology. These models will make it possible to create new instructional strategies and regulations to raise student achievement levels. Following the evaluations conducted after each midterm, the number of possibly failing students might be decreased in this way.

**3 Methodology**

We will discuss the steps and methods of the whole project in a nutshell for further appliances we make it brief to introduce the tools and methods to be followed. Firstly, we will mine data and preprocess it. Then we will be able to visualize data to understand which way we should predict. Then we will use machine learning tools to train the machine to get prediction from input data. So, let’s begin with the discussion with the steps we will follow to get prediction for student’s performance.

**3.1 Data Visualization**

Data visualization is the first attempt to work with any dataset to understand which data we are going to use for further application. Visualization shows the data in graphic representations which helps researchers and users to understand the whole perspective in a simple way where with row data it’s impossible to get an idea about the data. There are some steps to visualize data. Actually, data mining and cleaning is the 80% of work before visualizing the data and 20% of work is to visualize the data[28].

**3.1.1 Data Mining**

To visualize the row data such as pdf, csv, json or other format of file we need to get the data using pandas data frame and store it as a variable and mining the data. This work can be completed in a number of ways.. One of the most common function used to get data is read\_csv function to get data from csv file. We are going to use pandas library for mining the data.

**3.1.2 Data Cleaning**

However, from the row data there are lots of unstructured values and fields which interrupt to get the accurate results for visualization or predictions. Because of this problem data cleaning is so important[18]. Regarding this problem there are several ways. We can use .loc and pandas dataframe and python code to shape our data in an usable condition so that we can get better and accurate output.

**3.1.3 Visualization**

Data visualization depends on the clear understanding of taking decisions as what and how to explain data in a chart, bar, pi etc. Visualization can be build based on three categories comparison, time-based and categorical. We are going to use seaborn and pyplt libraries for visualization. These two libraries combinations gives better outcomes.

**3.2 Machine Learning Models for Prediction**

To predict the results from a dataset machine learning methods are used to train the machine to predict the future variables. To do so we have to select a model which generally split the data between test and train. Train data is used to give the input for machine to analyze the data to train itself and test data used to check the results of train data. This method is used for supervised machine learning where output variable is already given to machine to learn an accurate the result of prediction. We are going to use supervised machine learning method to predict one variable where other variables will be given to the model[27]. We will use Linear Regression model to train and test our machine to get prediction.

**3.2.1 What is Linear Regression**

A variable's value can be predicted using linear regression analysis based on the value of another variable. The dependent variable is the one you want to be able to forecast. The independent variable is the variable you are using to forecast the value of the other variable[30]. With the help of one or more independent variables that can most accurately forecast the value of the dependent variable, this model calculates the coefficients of the linear equation. The difference between the output values anticipated and those obtained is minimized by linear regression by fitting a straight line or surface. In order to get the best-fit line for a set of paired data, there are straightforward linear regression calculators that employ the "least squares" approach. So, using Y (independent variable), we may extrapolate the value of X (the dependent variable).

Linear regression can be performed by several ways some are here:

* R linear regression
* MATLAB linear regression
* Sklearn linear regression
* Linear regression Python
* Excel linear regression

**3.2.2 Why Linear Regression**

The mathematical technique used in linear-regression models is straightforward and may be used to make predictions. Numerous corporate and academic disciplines can benefit from the use of linear regression[19].

From the biological, behavioral, environmental, and social sciences to business, linear regression is employed widely. Future predictions may now be made scientifically and with high reliability using linear-regression models. The features of linear-regression models are well understood and can be trained extremely rapidly since linear regression is a statistical technique that has been around for a very long time. The use of linear regression techniques can help business and organizational leaders make better decisions. Organizations gather vast amounts of data, and linear regression enables them to use that data, rather than depending on experience and intuition, to better manage reality. It is possible to turn enormous volumes of raw data into useful knowledge.

By revealing patterns and links that your coworkers may have previously seen and assumed they already knew, you may also utilize linear regression to offer superior insights. For instance, analyzing sales and purchase data might reveal specific buying trends on certain days or at particular times. Business executives may predict periods when their company's products will be in great demand using insights from regression analysis[19].

**3.2.3** **Key assumptions of effective linear regression**

Several presumptions must be taken into account for linear regression analysis to be successful. Lets consider 95-percent-confidence intervals for each regression coefficient, variance-covariance matrix[25], variance inflation factor[15], tolerance, Durbin-Watson test[31], distance measures (Mahalanobis, Cook and leverage values), DfBeta, DfFit, prediction intervals and case-wise diagnostic information[8].

**Think about scatterplots, partial plots, histograms, and normal probability plots when describing plots.**

**Data:** Dependent and independent variables should be quantitative. It is necessary to recode categorical data into binary (dummy) variables or other contrast variables, such as religion, primary field of study, or location of residence.

**Other assumptions:** For each value of the independent variable, the distribution of the dependent variable must be normal. For all possible independent variable values, the variance of the distribution of the dependent variable should remain constant.

**3.3 Classification Models**

We have different level of education and grades. We have set combined score for classification. The level of classification is determined as 0 to 49 as 0, 50 to 59 as 1, 60 to 69 as 2, 70 to 79 as 3, 80 to 89 as 4 and 90 to 100 as 5 so that we can understand each grade and classify the encoded dataset. To get the encoded dataset we change our dataset with “numpy where” method in Pandas data frame. Then we splited data for training and testing to fit our models. We used seven models and compare the results. We used several models those are briefed below.

**3.3.1 TensorFlow**

TensorFlow is a language for describing computations as stateful dataflow graphs[1]. Machine learning models that are described as dataflow graphs perform better during training[23]. Backpropagation makes it simple to extract gradients of dataflow graphs, allowing for quick parameter adjustments. Second, independent nodes of the computational graph may be distributed across independent machines, including GPUs and TPUs, and run in parallel TensorFlow became a powerful tool for machine learning and deep learning as a result of these computational advantages[26]. TensorFlow uses tensors for the directed edges and operations (ops) for the nodes to create this dataflow graph. A rank n tensor is essentially an n-dimensional array for our purposes. Tensors are also associated with a data type in TensorFlow, such as integer or string. Tensors are a convenient method of thinking about data; in machine learning, iteration over the members of a dataset is frequently reserved for the first index. Additional indices can be used to signal the use of several filters, such as in convolutional neural networks with multiple feature maps. An op is a function that maps input tensors to output tensors in general. Ops can act on zero or more input tensors, but it must always produce at least one output tensor. The addition op, for example, takes two tensors and outputs one tensor that represents the elementwise sum of the inputs, whereas the constant op takes no tensors and acts as a root node in the dataflow graph. TensorFlow's backend has the structure of a directed acyclic graph thanks to the mix of ops and tensors[1].

Tensorial data format is not to be confused with Tensor Networks[7], which are a mathematical tool used to efficiently express many-body quantum states and actions in condensed matter physics and quantum information research. which are a mathematical tool used to efficiently express many-body quantum states and actions in condensed matter physics and quantum information research. Recently, libraries for creating such Tensor Networks in TensorFlow have become available and associate reader to the corresponding blog post for understanding the difference between TensorFlow tensors and the tensor objects in Tensor Networks[32]. The recently announced TensorFlow 2[33, с. 2]. One new feature is the Python function decorator @tf.function , which automatically converts the decorated function into a graph computation. Also relevant is the native support for Keras[34], which provides the Layer and Model constructs. These abstractions enable for the succinct development of dataflow graph-based machine learning models that ingest and analyze data. TensorFlow 2.0 is a suitable environment for data-driven machine learning research because of its high-level abstractions and efficient dataflow graph backend.

**3.3.2 GridSearchCV**

In recent times, the terms "machine learning," "artificial intelligence," and "deep learning" have appeared in the greatest number of locations on the Internet. Everyone wants to experiment with various Machine Learning and Deep Learning models in order to get the best outcomes. For several of the models, there are computational restrictions. Hyperparameter tuning is a technique used in machine learning to obtain the optimal model. In order to determine the ideal settings, GridSearchCV essentially takes into account all possible combinations of candidates. When there are more parameters and values to tweak, this would also take a very long period. There is a method we can use to speed up this procedure. This is the primary task that takes up the majority of machine learning time. Let's quickly review the fundamentals of GridSearchCV and parallel computing principles before moving on to the method section.

**3.3.3 Logistic Regression**

A statistical analysis technique called logistic regression uses previous observations from a data set to predict a binary result, such as yes or no. By examining the correlation between one or more already present independent variables, a logistic regression model forecasts a dependent data variable. A logistic regression might be used, for instance, to forecast whether a candidate for office will win or lose, or if a high school student would be accepted into a certain institution or not. These simple choices between two options allow for binary results. Multiple input criteria may be taken into account using a logistic regression model. The logistic function can take into account the student's grade point average, SAT score, and number of extracurricular activities in the event of college admittance. It then rates new instances according to their likelihood of falling into one of two result groups based on historical information about past outcomes using the same input criteria.

In the field of machine learning, logistic regression has grown in significance. It enables machine learning algorithms to categorize incoming input based on previous data. The algorithms get more accurate at predicting classes within data sets when new pertinent data is added. By enabling data sets to be staged for analysis throughout the extract, transform, and load (ETL) process by placing them into certain preset buckets, logistic regression may also be used in data preparation operations.

**3.3.4 K-Nearest Neighbors**

The k-nearest neighbors algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single data point. Although it may be applied to classification or regression issues, it is commonly employed as a classification method since it relies on the idea that comparable points can be discovered close to one another. A class label is chosen for classification issues based on a majority vote, meaning that the label that is most commonly expressed around a particular data point is adopted. The term "majority vote" is more frequently used in literature even though this is formally referred to as "plurality voting." The difference between both terms is that "majority voting" informally calls for a majority of more than 50%, which often only applies when there are only two options. You don't absolutely need 50% of the vote to draw a conclusion about a class when there are numerous classes, such four categories; you might give a class label with a vote of more than 25%. This is succinctly stated by the University of Wisconsin-Madison using an illustration.

**3.3.5 GaussianNB**

Using my preferred machine learning library scikit-learn, we will create the Naive Bayes classifier in Python in our experiments. The trained Naive Bayes (supervised classification) model will then be used to forecast the Census Income. Conditional probability is the foundation of Bayes' theorem. We may calculate the likelihood that something will occur given that another event has already occurred using the conditional probability. The naive Bayes classifier utilizes the Bayes theorem to function. The naive Bayes classifier makes the assumption that each characteristic is distinct from the others. even if the features rely on one another or on the other features' existence. A unique kind of NB algorithm is a Gaussian Naive Bayes algorithm. When the features contain continuous values, it is specially employed. Additionally, it is presumed that each feature has a gaussian distribution.

**3.3.6 SVM**

Support vector machines (SVMs) are robust yet adaptable supervised machine learning techniques used for classification, regression, and the detection of outliers. SVMs are frequently employed in classification issues and are extremely effective in large dimensional areas. Because they only employ a portion of the training points in the decision function, SVMs are well-liked and memory-efficient.

To determine a maximum marginal hyperplane (MMH), SVMs split datasets into several classes, which may be accomplished in the following two steps: First, Support Vector Machines will initially iteratively create hyperplanes that best divide the classes. And then the hyperplane that properly separates the classes will then be selected. SVC, NuSVC, and LinearSVC are the three classes offered by Scikit-Learn that can do multiclass-class classification. In our case we will only use SVC which uses libsvm as the foundation for its C-support vector classification implementation. Scikit-learn makes use of the sklearn.svm.SVC module. According to a one-vs-one system, this class manages multiclass support.

**3.3.7 Decision Tree Classifier**

Decision tree classifiers have a wide range of effective applications. Their capacity to extract descriptive decisionmaking information from the provided data is their key strength. Training sets can be used to create a decision tree. After it has been constructed, the decision tree can handle choice circumstances that are not included in the training set. The production rules can thus be stated as generalized statements for this reason. As has been demonstrated, using decision trees is straightforward and just as efficient as doing analysis using a strict mathematical model. The connectionist systems approach is another another instrument that may be very useful in the process of knowledge retrieval for AMS information flow and assessment. Any decision tree classifier's intelligence is based on the splitting criteria principle. Similar to a flow chart, decision trees have a tree structure and classify cases based on the values of their features. In a decision tree, a node represents an instance, a branch represents test results, and a leaf node symbolizes the class label.

**3.3.7 Random Forest Classifier**

AdaBoost and random forests are both equally accurate, but random forests are more resistant to mistakes and outliers. As long as there are many trees in the forest, the generalization error converges. Overfitting is thus not a concern. The effectiveness of a random forest is influenced by the potency of each classifier individually as well as a measure of their interdependence. The aim is to preserve the power of each classifier while reducing their correlation. We now introduce random forests, another ensemble technique. Consider the ensemble of classifiers as a "forest" where each decision tree classifier serves as a classifier. To decide the split, a random selection of characteristics is used at each node to construct each distinct decision tree. In more technical terms, every tree in the forest is reliant on the values of a random vector that was sampled individually, with the same distribution across all trees. Each tree casts a vote during categorization, and the top class is shown.

**4 Experiments and Results**

Experimenting our proposed solution, we are going to use data which records student’s performance so that we can execute our methodology to get desired output. We collected our data from kaggle which records student’s performance in different subjects. See data in reference[35].

**4.1 Data Preprocessing for visualization**

The dataset we have been chosen is consisting of the marks scored by the students in different subjects. This data source is in csv format. So, we will use Pandas libraries to read the data with data frame. Fig1 is the libraries we have used and fig 2 contains the data we have read from csv file, rows of the data frame and shape. Fig 3 is explaining the basic information of data such as max, min, % and standard deviation. From those figures we already understand the whole data now we will delete the data which are violating the rules and fixed the pass mark as 40 and visualize the data with different parameter. And then we inserted one column as combined score which define as the total marks.

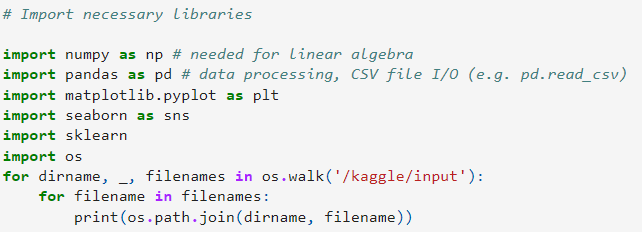
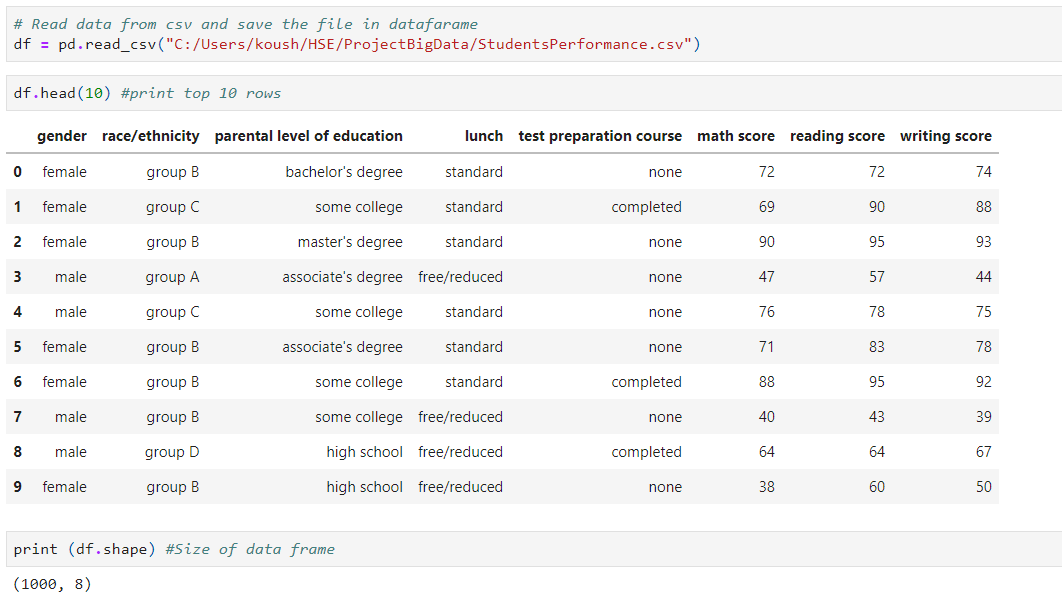


Fig1

Fig 2

Graphical user interface

Description automatically generated with medium confidence

Fig 3

**4.2 Visualization Results**

We are going to visualize data frame with the help of matlotlib through pyplt. In fig 4 we can see the visualization of education level as “parent\_ed” on X axis and combined score as Y axis. This bar chart represents the values of total scores according to different level of education.

Chart, bar chart

Description automatically generated

Fig 4

For math scores we plotted students acquire marks in the fig 5 accordingly where X axis represents the mark 0-100 and Y axis represents the number of students. And in fig 6 we assume the pass mark as 40 and demonstrate the graph of number of students (Y axis) passed and failed in examination by different level of education (X axis). Where number of passed students are 960 and failures are 40 as P and F in fig 6.

Chart, bar chart, histogram

Description automatically generated

Fig 5

Chart

Description automatically generated with low confidence

Fig 6

For reading scores we plotted students acquire marks in the fig 7 accordingly where X axis represents the mark 0-100 and Y axis represents the number of students. And in fig 8 we assume the pass mark as 40 and demonstrate the graph of number of students (Y axis) passed and failed in examination by different level of education (X axis). Where number of passed students are 974 and failures are 26.

Chart, bar chart

Description automatically generated

Fig 7

Chart, bar chart

Description automatically generated

Fig 8

For writing scores we plotted students acquire marks in the fig 9 accordingly where X axis represents the mark 0-100 and Y axis represents the number of students. And in fig 10 we assume the pass mark as 40 and demonstrate the graph of number of students (Y axis) passed and failed in examination by different level of education (X axis). Where number of passed students are 968 and failures are 32.

Chart, bar chart

Description automatically generated

Fig 9

Chart, bar chart

Description automatically generated

Fig 10

For average scores we plotted students acquire marks in the fig 12 accordingly where X axis represents the average/percentage mark 0-100 and Y axis represents the number of students. And number of overall passed students are 949 and failures are 51 codes are in fig 11.

A picture containing chart

Description automatically generated

Fig 11

Chart

Description automatically generated

Fig 12

Now Let us assign the grades: above 80 = A Grade 70 to 80 = B Grade 60 to 70 = C Grade 50 to 60 = D Grade 40 to 50 = E Grade below 40 = F Grade (means Fail) to classify the CGPA. In fig 13 we classified the results of individuals according to the grade system [A, B, C, D, F] in GetGrade function which shows the number of students gets the CGPA. And in fig 14 visualize the results according to grades(X axis) by number of students(Y axis). Additionally fig 14 plotted grades(Y axis) by level of education(X axis).

Graphical user interface, text, application

Description automatically generated

Fig 13

Chart, bar chart

Description automatically generated

Fig 14

Chart, bar chart

Description automatically generated

Fig 15

**4.3 Data Preprocessing for Machine Learning Models**

We are going to use ordinal encoder library to transform our data frame into a matrix so that our model can execute the dataset. We transform varchar values of column to a decimal value to train the model. Fig 16 is an example of transferring male and female values of gender column to 0.0 and 1.0. Following the same way we transferred parent\_ed, lunch, prep with decimal values to fit our data model. Heat map for preprocessed data is in fig 17. Now we will create different columns for individual race column so that it concatenates and extend the matrix to give better results. Fig 18 shows the encoding, concatenation and effects.

Table

Description automatically generated with low confidence

Fig 16

A picture containing chart

Description automatically generated

Fig17

Text

Description automatically generated with medium confidence

Fig 18

Now for all the columns with marks except combined score we are going to set 0 to 49 as 0, 50 to 59 as 1, 60 to 69 as 2, 70 to 79 as 3, 80 to 89 as 4 and 90 to 100 as 5. We take this initiative to classify into 6 class to understand and balance each class.

A picture containing calendar

Description automatically generated

Now we will encode those columns and spread columns according to each value.

Application

Description automatically generated with medium confidence

We applied same procedures to math\_score and reading\_score as well to encode the whole dataset.

**4.4 Creation and implementation of Regression models**

To fit the data and splitting we need two variables like one will be X and another will be Y then we will split X and Y for train and test the model(fig 19) .

We used StandardScaler libraries to transform the data. Fig 19 is an example of the getting variable X and Y from one data frame where Y is the values for combined scores. Then we can see the values for X and Y. And latter splitting them as X\_train, X\_test, y\_train and y\_test in fig 20 and fit for linear regression model with test set 10% and train set 90% of whole dataset. We predicted values for each student’s and average predicted values of combined score and the the mean of predicted values. This methods will helps us to understand each students individual performance and overall status of the students.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Fig 19

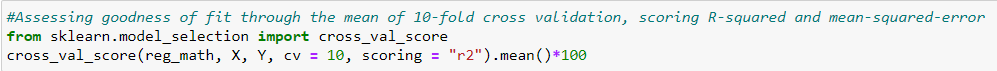
Graphical user interface, text

Description automatically generated

Fig 20

**4.5 Accuracy and Validation for Linear Regression**

Fig 21 shows the cross validation and mean square error of the predicted model for combined score. We got k-fold(10-fold is used) cross validation mean almost 97% and negative mean square error validation as -7. So, our result shows the accuracy of the model is satisfied with good performance.



Text

Description automatically generated

Graphical user interface, text

Description automatically generated

Fig 21

**4.6 Data Preprocessing for Classification**

To accrue classification, we need to be levelling the combined score. For the reason we set six levels with the values from 0 to 49, 50 -59, 60 to 69, 70 to 79, 80 to 89 and 90 to 100 respectively 0, 1, 2, 3, 4 and 5 where 0 to 50 assumed to be fail and 1 to 5 as grades (fig 22). After setting those parameters will set X and Y. After that we convert the column into int64 data type so that it can be fit to our model. Finally, we split the data as X\_train, X\_test, y\_train and y\_test for classification models where we set 80% data for training and 20% data for testing and set random state to shuffle and get result neutrally from the models (fig 23).

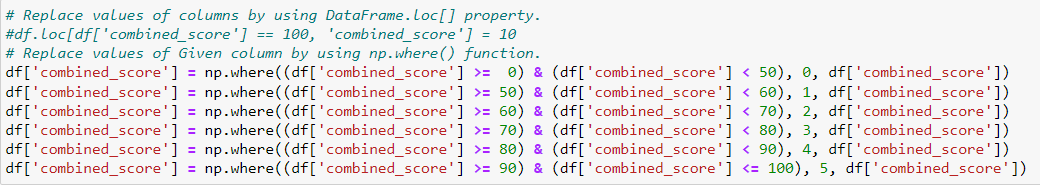


Fig 22

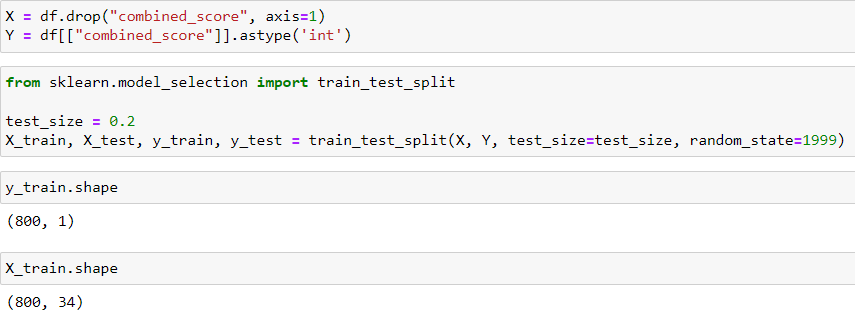


Fig 23

**4.7 Creation and Implementation of Classification Models**

Here we are going to show the comparison of the classification models we have mentioned in methodology section of classification. Regarding that first we installed and imported necessary libraries for the models (fig 24). Then we created function to reuse and reduce our manual work and get all the reports together. Then we set the parameters to those models and setting the parameters globally for individual models to be abstracted respectively in fig 25, 26 and 27 respeectively. Then we create\_models() function from where we will call the models we have created before as a function and fit the models with parameter (fig 28).

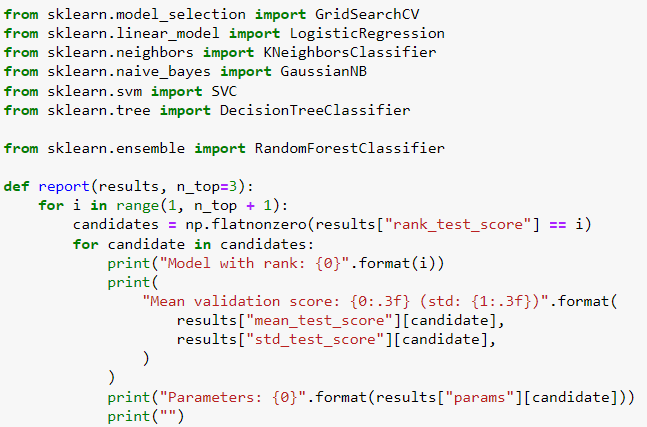


Fig 24



Fig 25

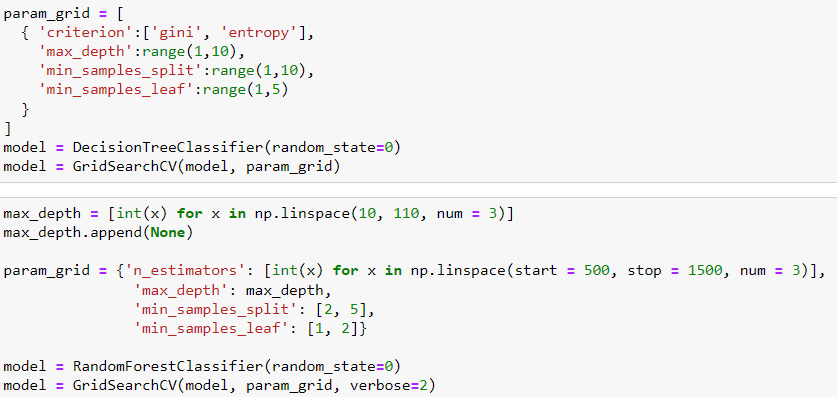


Fig 26



Fig 27

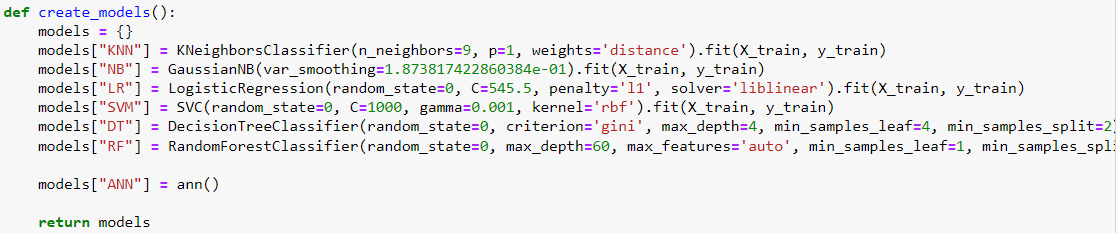


Fig 28

**4.8 Accuracy and Classification Report**

We use accuracy score from matrics library then we create a function to evaluate the models and from that function we get accuracy score for each model by calling those functions (fig 29)

Text, letter

Description automatically generated

Text

Description automatically generated

Fig 29

Now we will compare all of the models we used for classification. Where each model will have 0 to 5 rows which is the classification we have made before and four columns respectively precision, recall, f1-score and support. We have additionally macro average and weighted average accuracy. We will now get the classification report for all models in fig .

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

**5 Conclusions**

Predicting student’s performance help students and organization to take initiative if necessary. This time we used linear regression model for predicting values and the accuracy and mean square values are quite well still we can improve the model to get more accuracy for predicting results. Furthermore, we have applied TensorFlow, Random Forest, Decision Trees, Logistic Regression, Naïve bayes, K-Neighbors algorithm to get accurate and comparable accuracy for prediction and classification. We get the results almost 97% accuracy for predicting by Linear Regression which is satisfied our approach. For Classification weighted accuracy achieved 83% for SVM and Decision Tree which are great output indeed. Using Prediction and classification models we can segment we can easily understand which strategy we have to follow to get student’s future results. Students and Universities will apply machine learning technology in all sparce of educational system to prevent unexpected carry out and having nice and open environment for technology and accept it’s fruits.

**6 References**

1. Abadi M. [и др.]. TensorFlow : Large-Scale Machine Learning on Heterogeneous Distributed Systems / M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, [и др.]., 2015.

2. Ahmad Z., Shahzadi E. Prediction of Students’ Academic Performance Using Artificial Neural Network // Bulletin of Education and Research. 2018. № 3 (40). C. 157–164.

3. Alshanqiti A., Namoun A. Predicting Student Performance and Its Influential Factors Using Hybrid Regression and Multi-Label Classification // IEEE Access. 2020. (8). C. 203827–203844.

4. Asif R. [и др.]. Analyzing undergraduate students’ performance using educational data mining // Computers & Education. 2017. (113). C. 177–194.

5. Baker R. S. J. d, Yacef K. The State of Educational Data Mining in 2009: A Review and Future Visions // Journal of Educational Data Mining. 2009. № 1 (1). C. 3–17.

6. Bernacki M. L., Chavez M. M., Uesbeck P. M. Predicting achievement and providing support before STEM majors begin to fail // Computers & Education. 2020. (158). C. 103999.

7. Biamonte J., Bergholm V. Tensor Networks in a Nutshell 2017.

8. Borenstein M. [и др.]. Prediction Intervals 2021. C. 119–125.

9. Chen R., Chen C. Machine Learning Modeling 2022.C. 119–124.

10. Costa-Mendes R. [и др.]. A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach // Education and Information Technologies. 2021. № 2 (26). C. 1527–1547.

11. Cruz-Jesus F. [и др.]. Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country // Heliyon. 2020. № 6 (6).

12. Emad El-Den F. data mining // 2022.

13. Fernandes E. [и др.]. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil // Journal of Business Research. 2019. (94). C. 335–343.

14. Ghosh S., Dasgupta R. Introduction to the Machine Learning Models 2022.C. 45–50.

15. Gross J. Variance inflation factors // R News. 2003. (3). C. 13–15.

16. Hellas A. [и др.]. Predicting academic performance: a systematic literature review / A. Hellas, S. Liao, P. Ihantola, A. Petersen, V. Ajanovski, [и др.]., 2018. 175 c.

17. Hoffait A.-S., Schyns M. Early detection of university students with potential difficulties // Decision Support Systems. 2017. (101). C. 1–11.

18. Kapil S. Clean Python: Elegant Coding in Python / S. Kapil, 2019.

19. Kramer C. FUNDAMENTALS OF MULTIVARIATE ANALYSIS. LINEAR REGRESSION. // Journal of Quality Technology. 1972. (4). C. 177–180.

20. Li S., Liu T. Performance Prediction for Higher Education Students Using Deep Learning // Complexity. 2021. (2021). C. 9958203.

21. Macfadyen L. P., Dawson S. Mining LMS data to develop an “early warning system” for educators: A proof of concept // Computers & Education. 2010. № 2 (54). C. 588–599.

22. Musso M. F., Hernández C. F. R., Cascallar E. C. Predicting key educational outcomes in academic trajectories: a machine-learning approach // Higher Education. 2020. № 5 (80). C. 875–894.

23. Niu Y. TensorFlow Quantum: A Software Framework for Quantum Machine Learning 2020.

24. Rubin B. [и др.]. The effect of learning management systems on student and faculty outcomes // The Internet and Higher Education. 2010. № 1 (13). C. 82–83.

25. Schwertman N. C., Allen D. M. Smoothing an indefinite variance-covariance matrix // Journal of Statistical Computation and Simulation. 1979. (9). C. 183–194.

26. Shinde P. P., Shah S. A Review of Machine Learning and Deep Learning Applications 2018.C. 1–6.

27. Singh P. Linear Regression 2022.C. 53–74.

28. Tucker C., Pursel B., Divinsky A. Mining Student-Generated Textual Data In MOOCS And Quantifying Their Effects on Student Performance and Learning Outcomes Mining Student-Generated Textual Data in MOOCS and Quantifying Their Effects on Student Performance and Learning Outcomes // Computers in Education Journal. 2014. (5). C. 84–95.

29. Waheed H. [и др.]. Predicting academic performance of students from VLE big data using deep learning models // Computers in Human Behavior. 2020. (104). C. 106189.

30. About Linear Regression | IBM [Электронный ресурс]. URL: https://www.ibm.com/topics/linear-regression (дата обращения: 13.06.2022).

31. A Longitudinal Assessment of Tax Reforms and National Income in Nigeria: 1971-2014 // ResearchGate [Электронный ресурс]. URL: https://www.researchgate.net/project/A-Longitudinal-Assessment-of-Tax-Reforms-and-National-Income-in-Nigeria-1971-2014 (дата обращения: 14.06.2022).

32. Introducing TensorNetwork, an Open Source Library for Efficient Tensor Calculations // Google AI Blog [Электронный ресурс]. URL: http://ai.googleblog.com/2019/06/introducing-tensornetwork-open-source.html (дата обращения: 15.06.2022).

33. Effective Tensorflow 2 | TensorFlow Core // TensorFlow [Электронный ресурс]. URL: https://www.tensorflow.org/guide/effective\_tf2 (дата обращения: 15.06.2022).

34. Keras: the Python deep learning API [Электронный ресурс]. URL: https://keras.io/ (дата обращения: 15.06.2022).

35. Students Performance In Exams: EDA / predictions [Электронный ресурс]. URL: https://kaggle.com/doesnotcompile/students-performance-in-exams-eda-predictions (дата обращения: 14.06.2022).

1. For group projects, an assessment sheet is filled out for each group member [↑](#footnote-ref-1)
2. Only necessary components are used, if some component is not used in the assessment, then a dash is put in the corresponding line [↑](#footnote-ref-2)