Department of Computer Science
FAST National University of Computer and Emerging Sciences
Karachi Campus

Predictive Analytics on the Academic Record of NUCES

16th October 2020

Submitted by :

Obaid Ur Rehman
17k-3848

Areeka Ajaz
17k-3913

Tooba Shahid
17k-3731

Supervisor :

Dr Jawwad Ahmed Shamsi

ABSTRACT

Predictive Analytics is the process of using past data to make future predictions . The past data is used to capture important trends with the help of a mathematical model and the model is then used to make predictions on current data . We aim to perform predictive analytics on the academic record of NUCES . Our project will use the past academic record of NUCES to make insights and find out correlations between different attributes and then build a predictive model , based on the statistical inference , to make predictions on the current data. Finally we aim to integrate our analytics with a system , developing a fully functional web portal . The web portal will support the features such as importing data , query processing on data , visualizing data and predicting results .

INTRODUCTION

Each year a number of students take admission in FAST NUCES . The students taking admission in FAST NUCES are from different educational backgrounds and different regions . Their academic performance throughout their university life is a reflection of different factors , not only but including their educational background , their previous academic records , the region/district from where they belong etc . The project aims to answer numerous questions about how these factors are related to the performance of students at FAST throughout their educational period and will finally use this information to produce a model to predict the student's CGPAs and categorize them into different groups.

PROJECT BACKGROUND AND MOTIVATION

Predictive analytics has become an influencing factor in improving educational experiences for students. The result of predictive analytics on academic record plays a big role in a way to achieve the highest level of quality of education. This analytics can not only be used to better understand student performance but also to boost graduation rates. Moreover , the predictive model may also help to identify the students who are subject to low performance at an early stage and do the necessary intervention. Hence , early student performance prediction can help universities to provide timely actions, like planning for appropriate training to improve students' success rate.

LIST OF STAKEHOLDERS

The primary stakeholder of this predictive analytics is the academic institution which in our case is FAST NUCES .

LITERATURE REVIEW

To discover different patterns that can improve students' performance, many studies have been conducted. Especially during the last few years lots of research has been carried out to predict students' academic performance. The research begins with identifying the important factors that affect the students' academic performance. Different researchers have identified different factors that affect academic performance . Abeer Badr El Din Ahmed et. al. , in his study , used the course of the student, mid-term marks, Lab test grade, assignment, attendance, homework, student participation. Another research was carried by Fadhilah Ahmad and Azwa Abdul Aziz in which they used nine parameters like gender, race and hometown, GPA, family income, university entry mode, and grades in related courses. Raheela Asif and Mahmood K. Pathan in their study used data from four academic batches of Computer Science & Information Technology (CS & IT)

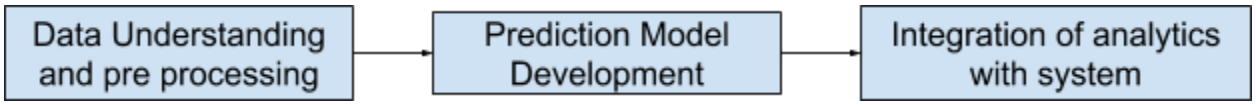
department at NED University, Pakistan. They used HSC marks , Maths marks in HSC , and marks in programming courses like Logic design, OOP, DBMS and Data Structures .

The next step involves building the predictive model. In this step different machine learning are taken into consideration like classification, clustering and regression analysis. Two types of data mining models are commonly used: predictive and descriptive . Predictive models apply supervised learning functions to provide estimation for expected values of dependent variables according to the features of relevant independent variables. Descriptive models are used to produce patterns that describe the fundamental structure, relations, and interconnectedness of the mined data by applying unsupervised learning functions on it. Typical examples of predictive models are classification and regression, while clustering and association , produce descriptive models. Classification is the most used method, followed by regression and clustering. The most commonly used classification techniques are Bayesian networks, neural networks, decision trees. Common regression techniques are linear regression and logistic regression analysis . Clustering uses techniques like neural networks, K-means algorithms, fuzzy clustering and discrimination analysis .

METHODOLOGY

Overall , the methodology of this project consists of the following main phases :

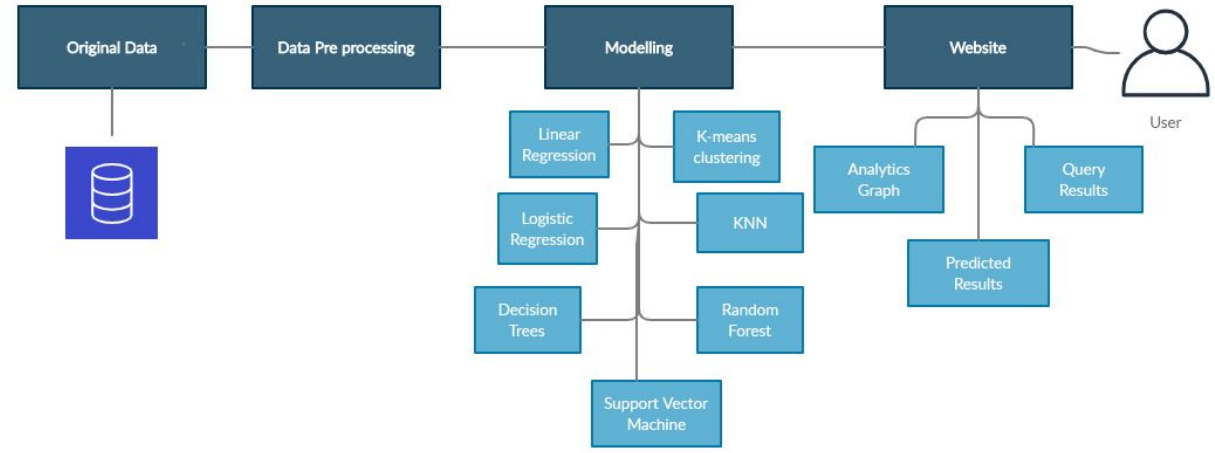
1) Data understanding and pre processing
2) Prediction Model Development
3) Integration of analytics with system (Website development)



After the collection of academic data from FAST NUCES the data will go through pre-processing , which includes data cleaning , data transformation and data reduction. Preliminary statistical analysis, especially through visualization, will be performed to better understand the data before moving to more sophisticated machine learning tasks and algorithms. With the help of statistical analysis the importance of attributes for the prediction process will be analyzed. In this phase we will be finding relationships between variables with the help of regression and correlation analysis. Scatter plots and regression lines will be used to visualize the data graphically.

After the phase of data understanding and pre processing , our project will enter the second phase i-e prediction model development by employing training and testing datasets with machine learning algorithms. We will use our analytics to predict student CGPA and we will also classify students in different categories. This project aims to explore different prediction , classification and clustering techniques to find out the model which gives the best accuracy . For our project we will be using linear and logistic regression , decision trees , random forest , support vector machine , kNN and k-means clustering .

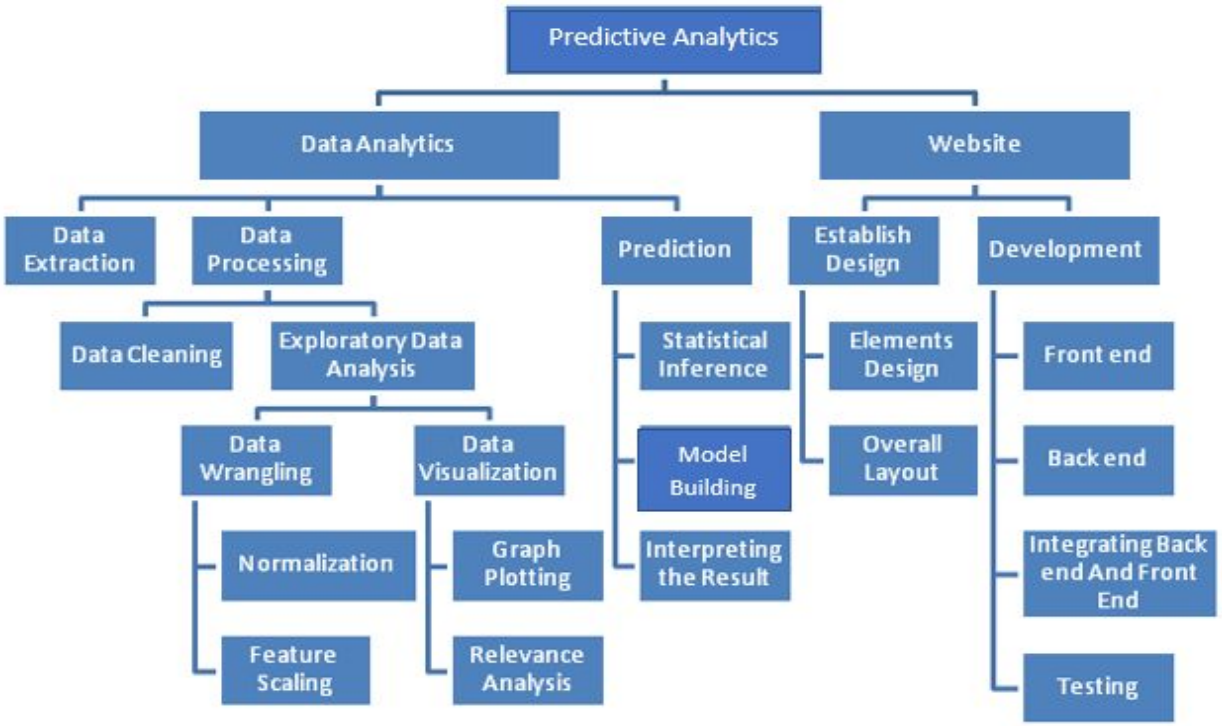
The models will then be integrated in an interface. This is where our project enters the last phase that is web development. The interface will give three sets of outputs which is predicted result , analytics graph and result of the query on the dataset .



Tools and Technologies :

1. All our work of EDA will be performed in R and the model building will be done in Python . R and Python both are the most famous and commonly used languages in data science .
2. Web development will be done using MERN (MongoDB, Express.js, React.js, Node.js) stack . It is a free and open source javascript software stack for building dynamic websites and web applications.

WORK BREAKDOWN STRUCTURE



DELIVERABLES

FYP I :

1. Relationship of different attributes with the academic performance through scatter plots and regression lines .

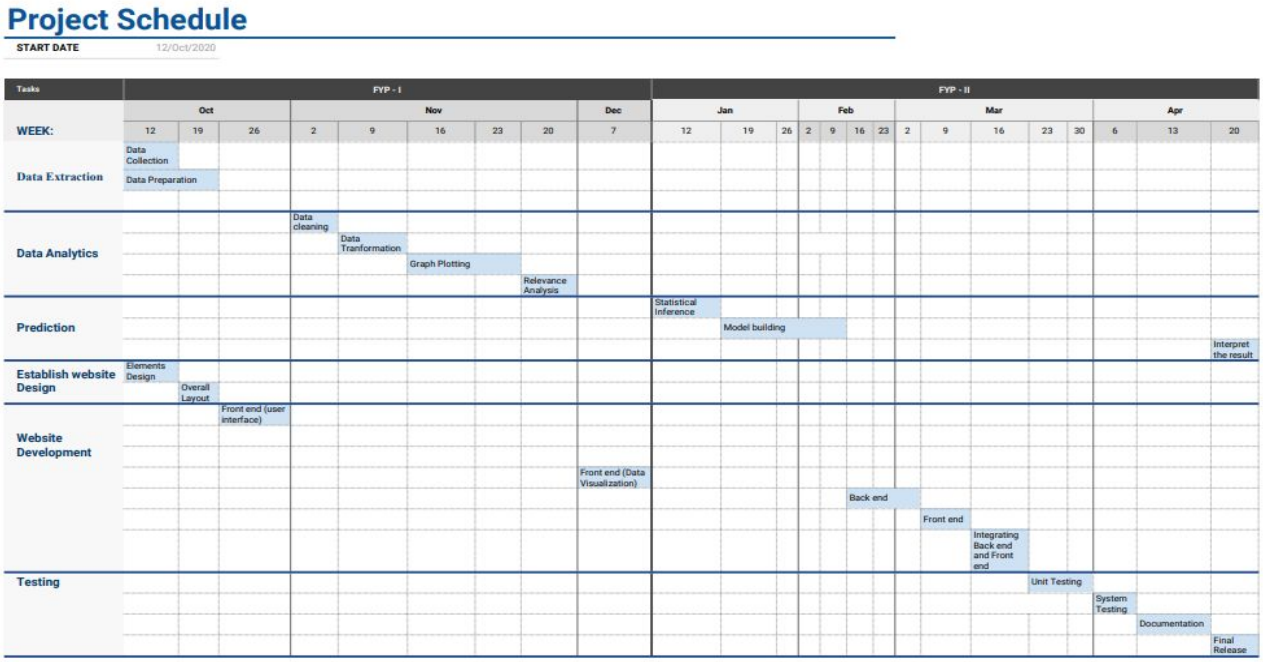
   Few exemplary questions that will be answered :

   - Does there exist any correlation between the city/district (a person belongs to) and their academic performance?
   - Does the NU test score affect the academic performance at FAST? What is the correlation between these two?
   - What is the correlation between intermediate / equivalence grade and the performance at FAST?
2. The above mentioned deliverables will be integrated in the website with features such as importing data and visualizing analytics graphs .

FYP II :

1. Model Building : Linear Regression and Logistic Regression , Decision Trees , Random Forest , Support Vector Machine , kNN and k-means clustering .
2. Which of the models implemented : linear regression , logistic regression, decision tree, random forest , or support vector machines , KNN and k-means clustering provide the best result .
3. Fully functional website showing all the analysis and prediction along with the feature of query processing on the data set .

PROPOSED TIMELINE



REFERENCES

1. Hajra Waheed , Saeed-Ul Hassan , Julie Hardman , Salem Alelyani , Raheel Nawaz , ‘“Predicting academic performance of students from VLE big data using deep learning models” , Volume 104, March 2020
2. Eyman Alyahyan , Dilek , “Predicting academic success in higher education: literature review and best practices” , International Journal of Educational Technology in Higher Education , 10 February 2020
3. Muhammad Yunus Iqbal Basheer, Sofianita Mutalib, Nurzeatul Hamimah Abdul Hamid, Shuzlina Abdul-Rahman, Ariff Md Ab Malik , “Predictive analytics of university student intake using supervised methods” , IAES International Journal of Artificial Intelligence , Vol. 8, No. 4, December 2019, pp. 367~374
4. Sonali Rawat , “Predictive Analytics for Placement of Student- A Comparative Study” , International Research Journal of Engineering and Technology (IRJET) , Volume: 06 Issue: 06 , June 2019
5. Ashish Dutt , Maizatul Akmar Ismail , Tutut Herawan , “A Systematic Review on Educational Data Mining” , IEEE Access , Volume: 5 , January 2017