STUDENT MARK PREDICTION USING MACHINE LEARNING

A PROJECT REPORT

submitted By

BASIL NARAKODAN

TVE21MCA2028

to

the APJ Abdul Kalam Technological University

in partial fullfilment of the requirements for the award of the degree

of

Master of Computer Applications

Department of Computer Applications

College of Engineering

Trivandrum-695016

NOVEMBER 2022

Declaration

I undersigned hereby declare that the project report titled **"Student mark prediction. using Machine Learning"** submitted for partial fulfillment of the requirements for the award of degree of Master of Computer Applications of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of Smt. Baby Syla L, Asst.Professor. This submission represents my ideas in my words and where ideas or words of others have been included. I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity as directed in the ethics policy of the college and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and/or University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title.

Place : Trivandrum **Basil narakodan**

Date : 14/11/2022

DEPARTMENT OF COMPUTER APPLICATIONS

COLLEGE OF ENGINEERING
TRIVANDRUM

CERTIFICATE

This is to certify that the report entitled **Student mark prediction. using Machine Learning** submitted by **Basil narakodan** to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree of Master of Computer Applications is a bonafide record of the project work carried out by him under my guidance and supervision. This report in any form has not been submitted to any University or Institute for any purpose.

Internal Supervisor External Supervisor

Head of the Dept

Acknowledgement

First and for most I thank **GOD** almighty and to my parents for the success of this project. I owe a sincere gratitude and heart full thanks to everyone who shared their precious time and knowledge for the successful completion of my project.

I am extremely thankful to **Dr Suresh Babu**, Principal, College of Engineering Trivandrum for providing me with the best facilities and atmosphere which was necessary for the successful completion of this project.

I am extremely grateful to **Dr. Deepa S S**, HOD, Dept of Computer Applications, for providing me with best facilities and atmosphere for the creative work guidance and encouragement.

I express our sincere thanks to **Smt. Baby Syla L**, Asst. Professor, Department of Computer Applications, College of Engineering Trivandrum for her valuable guidance, support and advice that aided in the successful completion of my project.

I profusely thank other Asst. Professors in the department and all other staffs of CET, for their guidance and inspirations throughout my course of study.

I owe my thanks to my friends and all others who have directly or indirectly helped me in the successful completion of this project. No words can express my humble gratitude to my beloved parents and relatives who have been guiding me in all walks of my journey.

Basil narakodan

Abstract

Getting to a prestigious institution is dream of all students. But this is not a easy task. Students may need crack highly competitive exams inorder to get into it. Tracking their preparation individualy and guiding them is not easy. Some times students may fail even after good preparation. It will be better if it is able analyse their performance based on their previous results and predict the chances of qualifying the main exam There may be a incredible need to expand the proper strategy to help students retention. Early grade prediction is one of the answers that have a propensity to screen students development within the courses on the university and will result in enhancing the students getting to know system based totally on predicted grades. Using machine learning can enhance the gaining knowledge of method of students. Distinct models may be advanced to predict students grades within the enrolled publications, which offer valuable facts to facilitate students' retention in those publications. This facts can be used to early perceive students at-chance based totally on which a machine can recommend the instructors to offer special interest to those college students. This statistics also can help in predicting the students grades in unique publications to reveal their overall performance in a higher way that could beautify the results.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Right now, Education is very important issue regarding development of any country. The main objective of educational institutions or college is to provide high quality education to their students. One way to solve this is by predicting student's academic performance early and thereby taking early steps to improve student's performance and teaching quality. So, with the mindset that learn by doing is the most effective technique, I set out to do an AI project using Different Regression as my machine learning model of choice.

Student marks Prediction is a way of predicting a student marks based on his/her previous marks. This also makes the teachers to know whether students are in a position to reach his/her expected marks or not. If this model shows that student needs to improve then that student can prepare more for that semester so that he/she can reach their expected marks or grade.

Main objective for this project is to help teachers to analyze students performance easily and if needed they can help her/him to improve their student's performance by taking some actions like increasing their reading hours, giving some assignments etc..

I like to focus on using real-world data, and in this project, we will be exploring student performance data collected from a Portuguese secondary (high) school. The data includes personal and academic attributes of the students along with final semester grades. Our objective will be to create a model that can predict grades based on the student's information. This dataset, along with many other useful things for testing models or trying out machine learning techniques, is available on the UCI Machine Learning Repository.

Chapter 2

Problem Defnition and Motivation

The problem statement can be defined as follows "Given a dataset containing attribute of 396 students where using the features available from dataset and define classification algorithms to identify whether the student performs good in final grade exam, also to evaluate different machine learning models on the dataset."

The major motivation behind choosing the project was the drawbacks of the existing system. The system that we have now is a complete manual one. And which is followed from the very beginning.

The Main objective of this project is to study more about various parameters that depends the accuracy of the final outcome and to enhance the accuracy of prediction using machine learning methods.

Chapter 3

Literature Review

I have alluded to several research articles that are related to the thesis in order to specify the thesis as a well-structured thought. Conclusion information of few of the papers are as follows. This research study describes how the linear regression approach issued in predicting student's academic performance .

1. In this studies paper creator carried out the thesis the usage of SVM technique in java, selection tree, C4.5, Naive Bayes, Lib. SVM, Logistic Regression and Hybrid technique LMT and as compared the accuracy of overall performance prediction most of the hybrid approaches. I have alluded to several research articles that are related to the thesis in order to specify the thesis as a well-structured thought.

2. In this studies paper it's far discovered that the writer used a number of the maximum famous algorithms and regression algorithms. The experiment was conducted with administrative data from the University of Polo, which included 700 courses. The article concludes that decision Trees and SVM produce the best results. The main contribution of this work is to compare the levels of accuracy of several algorithms.

3. The studies is targeted on predicting student's overall performance the usage of personalised analytics. This paper presents two different approaches to work on the thesis. The author's initial technique is the Regression Algorithm, which is a data mining function. The root mean square method is also used to calculate the regression algorithm's error rate. In this paper the author worked on how to improve the prediction algorithms which are used to analyze and predict the student's performance. The decision trees algorithm is used in this paper's work.

4. This paper proposed the student Academic performance prediction using Support Vector

Machine. The author compared SVM to various machine learning approaches such as linear regression, Decision Trees, and KNN and determined that SVM outperformed them.

Chapter 4

Requirement Analysis

4.1 Purpose

In order to achieve a working project in the python environment some major requirements were done. Software Development require some important steps to be taken. In developing our web application was also done with similar steps

4.2 Overall Description

4.2.1 Software requirement

- Operating System : Linux/Windows
- Platform : Python3
- Librarie used : pandas, matplotlib, numpy, sklearn,
- Flask

4.2.2 Hardware Requirements

- Processor: Intel Pentium 4 or equivalent
- Minimum of 4 GB or higher

- Hard Disk: 10 GB or above
- Architecture: 32-bit or 64-bit
- M

4.3 Functional Requirements

The functional requirements includes all the activities or processes that should be achieved by the proposed system. It includes

- **Python:** Python is an interpreted, high-level, general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant white space. Its language constructs and object-oriented approach aims to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming
- **Pandas:** is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance productivity for users.
- **Numpy:** NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.
- **sklearn:** sk learn (formerly sci-kit learn and sometimes called sk learn) is a machine learning library can be used in python programming language. By using this library, we can implement various regression, classification and clustering algorithms such as random forest, support vector machine, k-means and DBSCAN. And the sk learn library is built in a way that it can work with various scientific and numeric libraries of python such as scipy and numpy.

- **matplotlib:** It's used for the visualisation of data in python programming language. It's implemented to work with the wider scipy stack and it's built on numpy arrays. It's a multi platform data visualization technique. It was developed in 2002 by John Hunter. Visualization is the most efficient way to understand the data. Using this library, we can represent our data in various plots such as line, bar, histogram, scatter etc.

4.4 Non Functional Requirements

4.4.1 Performance Requirements

- Accuracy : Accuracy in functioning and the nature of user-friendly should be maintained by the system.
- Speed : The system must be capable of offering speed.
- Low cost: This system is very cheap to implement and is also user-friendly.
- Less Time consuming: It uses very less time comparing to the existing sysytem .
- User Friendly: This proposed system is highly user friendly they enables to create a good environment.

4.4.2 Quality Requirements

- Scalability : The software will meet all of the functional requirements.
- Maintainability : The system should be maintainable. It should keep backups to atone for system failures, and should log its activities periodically.
- Reliability : The acceptable threshold for down-time should be large as possible. i.e. mean time between failures should be large as possible. And if the system is broken, time required to get the system backup again should be minimum.
- Availability: This system is easily available as the core equiments in building the sofware is easily obtained.
- High- Functionality: This system is highly functional in all environment since, They are highly adaptable.

Chapter 5

Methodology and System Design

The proposed system is used to analyse and predict mark by using a pre-trained model. The model is trained using various machine learning and deep learning algorithms upon the features extracted.

5.1 System Design

Design of a system can be defined as the process of applying various techniques and principles for defining a device, a process or a system in sufficient detail to permit its physical realization. Thus, system design is a solution, a "how to" approach to the creation of a new system. This important phase provides the understanding and procedural details necessary for implementing the system recommended in the feasibility study. The design step produces a data design, an architectural design and a procedural design. The data design transforms the information domain model created during analysis into the data structures that will be required to implement the software. The architectural design defines the relationships among major structural components into a procedural description of the software. Source code is generated, and testing is conducted to integrate and validate the software. From the project management point of view software design is conducted in two steps, preliminary design is concerned with the transformation of requirements into data and software architecture. Detailed data structure and algorithmic representation for software.

5.1.1 Methodology

The main process is the creation of the trained model. The major steps in the model creation Feature extraction, training, testing and model evaluation. The major steps in the model creation are mentioned below.

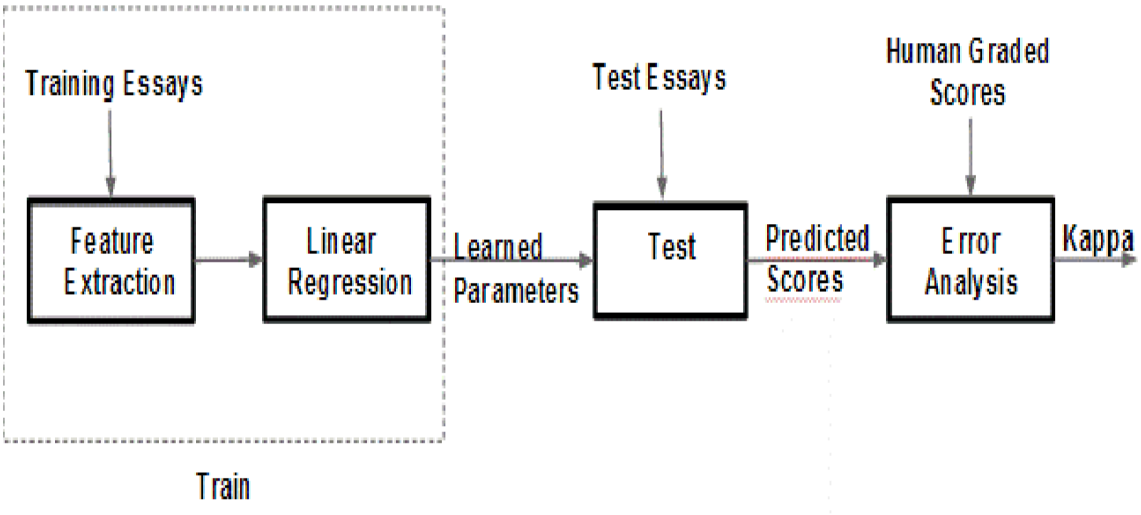

Figure 5.1: Architecture of model Creation

- **feature extraction:**Feature extraction is the most important part of any machine learning task and so is the case with us. Here we specify the mark and background features of students and they are extracted. We extract the features such as previous scores, parents education etc. as the features

- **Training:**Training is the most important step in machine learning. In training, you pass the prepared data to your machine learning model to find patterns and make predictions. It results in the model learning from the data so that it can accomplish the task set. Over time, with training, the model gets better at predicting.

- **Testing:** In testing phase we test the generated model with the remaining portion of the dataset. The data set is fed into the generated model and their results are recorded for the next stage which is the model evaluation and error analysis.

- **Model evaluation and error analysis:** The results of the testing data along with their original values are used for the error calculation. Various statistical measures can

be adopted for calculating the efficiency of the model. The various measures are accuracy, precision, recall, kappa score etc. If the results of these quantitative analyses are acceptable, then we move forward with the generated model. If the results are poor, the model is to be regenerated with more features so that the best result is obtained.

5.1.2 Linear Regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

5.1.3 Lasso Regression

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of muticollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

5.1.4 Decision Tree Regression

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

5.1.5 Dataset

Dataset is the most important part in any Machine Learning projects, This project uses the same dataset that are use in reffered research paper. The data were obtained in a survey of students math courses in secondary school. It contains a lot of interesting social, gender and study information about students. You can use it for some EDA or try to predict students final grade.

school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

sex - student's sex (binary: 'F' - female or 'M' - male)

age - student's age (numeric: from 15 to 22)

address - student's home address type (binary: 'U' - urban or 'R' - rural)

famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at$_h ome' or' other')

$Fjob - father's job(nominal :' teacher',' health' carerelated, civil' services'(e.g.administrativeorpolice),'$

$reason - reasontochoosethisschool(nominal : closeto' home', school' reputation',' course' preferenceor' oth$

$student' sguardian(nominal :' mother',' father' or' other')$

$traveltime - hometoschooltraveltime(numeric : 1 - < 15min., 2 - 15to30min., 3 - 30min.to1hour, or4 - > 1hour)$

$studytime - weeklystudytime(numeric : 1 - < 2hours, 2 - 2to5hours, 3 - 5to10hours, or4 - > 10hours)$

$failures - numberofpastclassfailures(numeric : nif1 <= n < 3, else4)$

$schoolsup - extraeducationalsupport(binary : yesorno)$

$famsup - familyeducationalsupport(binary : yesorno)$

$paid - extrapaidclasseswithinthecoursesubject(MathorPortuguese)(binary : yesorno)$

$activities - extra - curricularactivities(binary : yesorno)$

$nursery - attendednurseryschool(binary : yesorno)$

$higher - wantstotakehighereducation(binary : yesorno)$

$internet - Internetaccessathome(binary : yesorno)$

$romantic - witharomanticrelationship(binary : yesorno)$

$famrel - qualityoffamilyrelationships(numeric : from1 - verybadto5 - excellent)$

$freetime - freetimeafterschool(numeric : from1 - verylowto5 - veryhigh)$

$goout - goingoutwithfriends(numeric : from1 - verylowto5 - veryhigh)$

$Dalc - workdayalcoholconsumption(numeric : from1 - verylowto5 - veryhigh)$

$Walc - weekendalcoholconsumption(numeric : from1 - verylowto5 - veryhigh)$

$health - currenthealthstatus(numeric : from1 - verybadto5 - verygood)$

$absences - numberofschoolabsences(numeric : from0to93)$

$Thesegradesarerelatedwiththecoursesubject, Math$

$G1 - firstperiodgrade(numeric : from0to20)$

$G2 - secondperiodgrade(numeric : from0to20)$

$G3 - finalgrade(numeric : from0to20, outputtarget)$

5.2 Screenshots of user interface

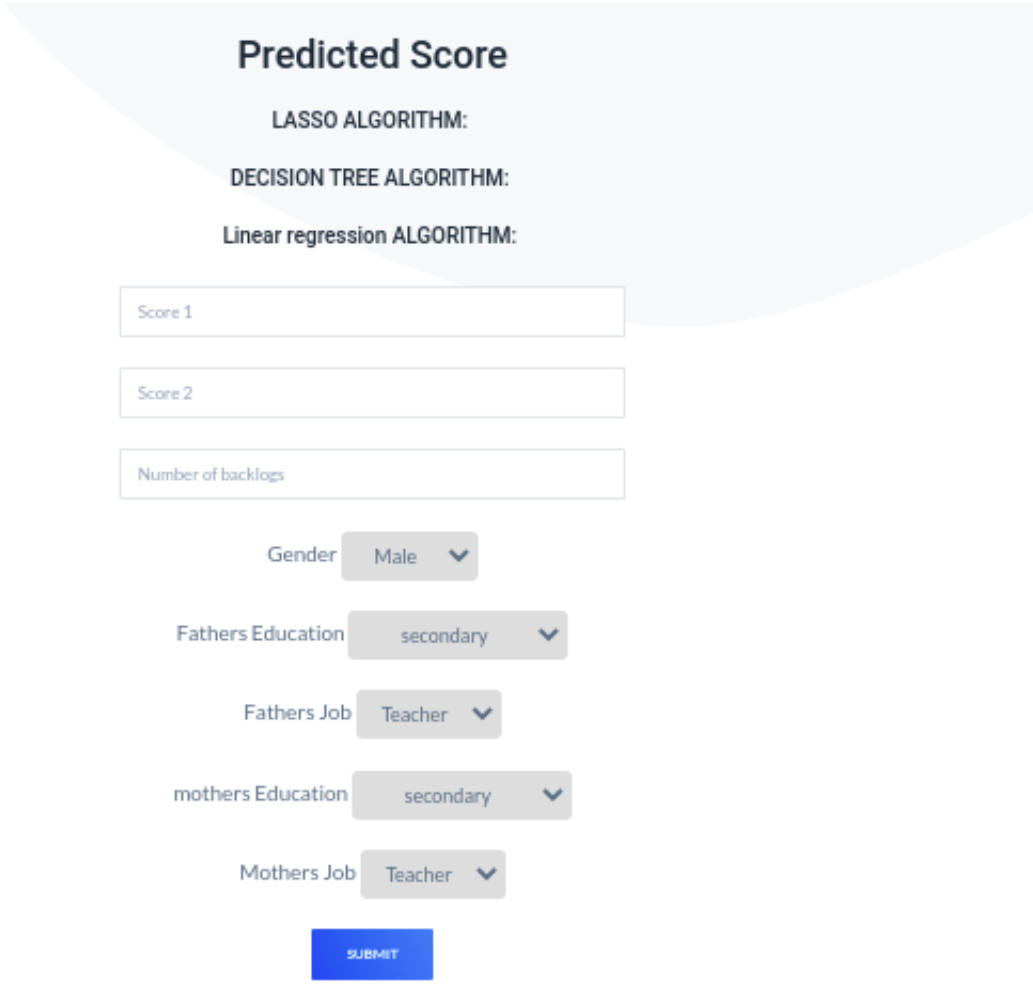

Figure 5.2: User interface

Chapter 6

Coding

Algorithm 1 Algorithm for Creating the model:

1: Split the data set into training data set and testing dataset. 80% of the dataset is used for training and the remaining 20% is used for testing to obtain better result.

2: The training dataset is used for the preprocessing stage and the preprocessed data is further used to extract the features.

3: The extracted features are then analysed to find out which features has the higher influence on the results. Then the features with high dependency is used to create the model.

4: The model is created using different algorithm with the selected features. The model evaluates how much the dependent values depend upon the independent values.

5: The testing dataset is feeded into the created model and their results are noted down.

6: The result of testing dataset evaluated using the created model is then compared with the actual values of the testing dataset to evaluate the efficiency of the model. Various statistical measures such as Accuracy, Precision, Recall, Kappa score etc. can be used to evaluate the model.

7: Further tuning is performed upon the created model to improve the efficiency of the model.

6.1 Source code

```
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
import os,csv
from sklearn.metrics import confusion_matrix
%matplotlib inline
from keras.utils import np_utils
from keras import regularizers
from keras.models import Sequential,load_model
from keras.utils.np_utils import to_categorical
from sklearn import preprocessing
from keras.layers import Dense, Dropout
```

Figure 6.1: impoting libraries

```
data=pd.read_csv('/content/drive/MyDrive/STUDENT_MARK_PREDICTION2/student-mat.csv',delimiter=',')
```

Figure 6.2: loading and reading the dataset

```
[ ] X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.4)

[ ] from sklearn.linear_model import LinearRegression, Ridge, Lasso
    model=Lasso()
    model.fit(X_train,y_train)
    print(model.score(X_test,y_test)*100)

    79.77315117646006
```

Figure 6.3: model building splitting our dataset using lasso

```
[ ]  model1=LinearRegression()
     model1.fit(X_train,y_train)
     print(model.score(X_test,y_test)*100)

     86.05049407970576
```

Figure 6.4: loading and reading the dataset using linear regression

```
[ ] #After multiple attempts, I was able to achieve best model with below configuration.
    model = Sequential()
    model.add(Dense(64, activation='relu',  kernel_regularizer=regularizers.l2(0.001),input_shape = (17,)))
    model.add(Dense(64, kernel_regularizer=regularizers.l2(0.001), activation='relu'))
    model.add(Dense(32, kernel_regularizer=regularizers.l2(0.001), activation='relu'))
    model.add(Dense(5,  kernel_regularizer=regularizers.l2(0.001),activation='softmax'))
    model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
    history  = model.fit(x_train,y_train, epochs = 200, batch_size = 5, validation_data = (x_test,y_test))
    model.save('UCI_model_A.h5')
```

Figure 6.5: loading and reading the dataset using multiclass learning

Chapter 7

Testing and Implementation

7.1 Testing and various types of testing used.

Once a software is developed, the major activity is to test whether the actual results match with the experimental results. This process is called testing. It's used to make sure that the developed system is defect free. The main aim of testing is to find the errors and missing operations by executing the program. It also ensure that all of the objectivs of the project are met by the developer. The objective of testing is not only to evaluate the bugs in the created software but also finding the ways to improve the efficiency, usability and accuracy of it. It aims to measure the functionality, specification and performance of a software program. Tests are performed on the created software and their results are compared with the expected documentation. When there are too much errors occurred, debugging is performed. And the result after debugging is tested again to make sure that the software is error free. The major testing processes applied to this project are unit testing, integration testing and system testing. In unit testing, our aim is to test all individual units of the software. It makes sure that all of the units of the software works as it intended. In integration testing, the combined individual units are tested to check whether it met the intended function or not. It helps us to find out the faults that may arise when the units are combined. In system testing the entire software is tested to make sure that it satisfies all of the requirements. The tables shown below describes the testing process occurred during the development of this project "Student mark prediction. using Machine Learning". This defines the various steps took to create the project error free.

7.1.1 Unit Testing

Unit testing focuses verification error on the smallest unit of software design the module. Using the procedural design description as a guide, important control paths are tested to uncover errors with the boundary of module. The relative complexity of test and uncovered errors is limited by the constrained scope established for unit testing. The unit test is normally white box oriented and the step can be conducted in parallel for multiple modules. The module interface is tested to ensure that information properly flows into and out of the program unit under test. The local data structure is examined to ensure that data stored temporarily maintains its integrity during all steps in an algorithm's execution. Boundary conditions are tested to ensure that the module operates properly at boundaries established to limit or restrict processing. All independent paths through the control structure are exercised to ensure that all statements in a module have been executed at least once. And finally, all handling paths are tested.

Text Cases and Result

Sl No	Procedures	Expected result	Actual result	Pass or Fail
1	Feeding user data	Model should access the inputted values	Same as expected	Pass
2	Predicting the destination	Model should predict the output	same as expected	Pass
3	Should keep accuracy	Model should be accurate	Same as expected	Pass

Table 7.1: Unit test cases and results

7.1.2 Integration Testing

Also known as integration and testing, is a step in software testing in which individual modules are combined and tested as a single block.Integration testing is conducted to evaluate its possibilities with specified functional requirements.

Text Cases and Result

Sl No	Procedures	Expected result	Actual result	Pass or Fail
1	The training model integrated to project	Predictions should me made	Same as expected	Pass

Table 7.2: Integration cases and result

7.1.3 System Testing

System testing is the process in which the system undergoes experimental testing so as to check that the system does not fail i.e., to check whether the required system is running according to specification and user expectation.

Text Cases and Result

Sl No	Procedures	Expected result	Actual result	Pass or Fail
1	Correct output is given for input	Correct output is obtained	Same as expected	Pass

Table 7.3: System test cases and results

Chapter 8

Results and Discussion

Main objective of this project is to study more about various parameters that depends the accuracy of the final outcome and to enhance the accuracy of prediction using machine learning methods.

8.1 Accuracy

Machine learning model accuracy is the measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input, or training, data. In my project the accuracy score obtained is around 85 percent.

Sl No	Model	Accuracy
1	Linear Regression	86.05
2	Lasso Regression	79.77
3	Decition tree Regresser	70.10
4	Multiclass Classification	70.00
5	Multiclass Classification by dropping G2	56.67
6	Multiclass Classification by dropping G1 and G2	31.60

Table 8.1: Accuracy

8.2 Advantages and Limitations

The proposed system is a machine learning model to evaluate the input and predict the score. We also found out that all the algorithms that I have applied in my research returned almost similar accuracy level with linear Regression. The decition tree regression Model and multiclass classification model was not able to perform well. Like every other system, this model also have it's own disadvantages. But they are negligible while comparing with the advantages and they can be overcame in future.

Chapter 9

Conclusion and Future Scope

To sum up, we can say that surprisingly the variables Dalc (Workday alcohol consumption) and Walc (Weekend alcohol consumption) were not significant, so I can say that the consumption of alcohol don't impact the grades of the students much and the variables like failures, health, free time were much more impactful and they were also much far significant than Walc and Dalc. We also found out that all the algorithms that I have applied in my research returned almost similar accuracy level with linear Regression performing best with 92 percentage Accuracy and the decition tree regression Model and multiclass classification model was not able to perform well.

In this project we have applied many techniques and got almost similar results but in our research we found out that the linear Regression performed best with 92 percentage Accuracy. In the future researchers can apply clustering techniques onto the dataset to gain some new insights. After the use of clustering, we will also come to know if the students with similar interests and similar knowledge got into same cluster or not. In the future, researchers can apply clustering techniques like K means and Hierarchical Clustering.

Bibliography

[1] Scikit learn: https://scikit-learn.org/

[2] Python: https://www.python.org

[3] Flask : https://www.flaskproject.com/

[4] kaggle : https://www.kaggle.com/