Students Marks Prediction Using Linear Regression

Abstract:

Analyzing and prediction of academic performance is important for any education institutions. Predicting student performance can help teachers to take steps in developing strategy for improving performance at early stages. With the advancement of machine learning supervised and unsupervised techniques developing these kinds of applications are helping teachers to analyze students in better way compare to existing methods. In this student marks prediction using Linear regression project students' academic performance is prediction considering input as previous students marks and predict next subject marks and accuracy of the model is calculated.

Problem statement:

Analyzing and prediction of marks for students was done based on guess and students' personal marks details are not considered for academic evaluation.

Objective:

Machine learning based data mining techniques are used to automate process of student performance prediction using linear regression technique.

Existing system:

- Researches has done work on Grading systems which final examination marks are used for giving grades for students and evaluation of each student is done.
- Association rule mining and apriori algorithms are used for classifying students based on their marks

Disadvantages:

- Most of these methods work on data mining techniques which are based on after completing data.
- Early stage evaluation is not possible in these methods.

Proposed system:

- Students marks of other subjects are taken as input for evaluation students' performance. Data set is pre-processed and features and labels are extracted from dataset then dataset is split in to test and train sets then linear regression is applied to dataset for prediction.

DATA ANALYTICS AND PREDICTIVE MODELING:

Data analytics is the art of exploring raw data for the purpose of deriving valuable insights to end up with conclusions about that information. Almost many of the companies and organization have already started implementing data analytics to make improved decisions in their field of interests. Data mining and data analytics are different from each other by the scope, function and focus of the analysis. Data analytics basically aims on interpretation, the process by which conclusion is derived from what is previously known by the researcher.

REGRESSION ANALYSIS:

Regression analysis permits scientists to form numerical models that can be used to forecast the value of one variable from the information of another variable. There are a number of specific regression techniques that can be used by the scientists to model and drive real-world insights [34][35][36]. Linear and Logistic regression analysis are commonly the major algorithms scientists study in predictive modeling. Hence many analysts rationalize that the above algorithms are the only form of regressions because of the frequent use of these algorithms in the analysis. But actually there are numerous forms of regression models with slight variations and specific conditions that can be used for the effective analysis. In this research paper, we apply the following regression models to predict the student’s grade: 1) Boosted trees regression, 2) Decision trees regression, 3) Linear regression 4) Random forest regression. In the field of information technology, predictive model is most widely used to predict the future insights.

Linear regression:

In predictive modeling, one of the most commonly used modeling technique is the linear regression and also researchers use this regression analysis when learning predictive analysis for the first time. Here in this regression analysis, the dependent variable is continuous, independent variables can be isolated, and the nature of the regression line is only straight. Linear regression analysis uses the Least Square method for fitting a regression line and calculates the best-fit line for the experimental data by reducing the sum of the squares of the vertical deviations from each data point to the line. Hence, the deviations are first shaped and when added there is no cancelling out between positive and negative values.

OVERALL PROCESS:

The student marks from 4 semesters are collected as an anonymous dataset, which is used to create a regression to predict the student final exam marks. This dataset includes 6 different internal exam marks as features, which will be used to predict the final exam marks.

Analyze the Dataset:

The sample dataset is analyzed to get fine details about the various features and their relationships. The individual grades and their relationship with the target is given in the below box plot. The box plot gives the finite details for the target based on the mean, second and third quartiles. shows the linear relationship between the mid exam and the final exam marks based on their final grades.

Advantages:

- Before final marks of all subjects are evaluated prediction can be performed.
- Using machine learning process automation of marks prediction can be done.

SOFTWARE REQUIREMENTS:

Operating system	: Windows XP/7/10
Coding Language	: python
Development environment	: anaconda, Jupiter
Dataset	: students marks dataset
IDE	: Jupiter notebook

Source code:

```
#importing required libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

#Reading the csv file and printing 1st five rows

data = pd.read_csv('student_info.csv')

data.head()

#printing the no of rows and column in dataset

data.shape

data.isnull().sum()

#plotting with use of matplot library

plt.scatter(x = data.study_hours , y=data.student_marks)

plt.title("Student Data")

plt.xlabel("Student Study Hours")

plt.ylabel("Student Marks")

plt.show()

#printing the mean

data.mean()

#makiing the data nill

data = data.fillna(data.mean())
```

```
data.isnull().sum()

X = data.drop(columns = 'student_marks')

y = data.drop(columns = 'study_hours')

X.shape , y.shape

#importing sklearn's train and test

from sklearn.model_selection import train_test_split

X_train , X_test , y_train , y_test = train_test_split(X, y ,
random_state=51 , test_size=0.2)

X_train.shape , y_train.shape , X_test.shape , y_test.shape

#importing linear regression from sklearn

from sklearn.linear_model import LinearRegression

#implementing linear regression

lr = LinearRegression()

lr.fit(X_train , y_train)

lr.score(X_test , y_test)

lr.intercept_

pred = lr.predict(X_test)

pred

y_test

#printing the predicted marks

pd.DataFrame(np.c_[X_test , y_test , pred] ,columns =[ 'Study hours' ,
'Original Marks' , 'Predicted Marks'])
```

```
# Fine Tune Model

plt.scatter(X_train, y_train)

plt.plot(X_train ,lr.predict(X_train) , color='r')

import joblib

joblib.dump(lr , 'Student_Marks_Prediction_Model.pkl')

#predicting the required members marks

model.predict([[ 1 ]])[0][0]

data.describe()
```

REFERENCES: Google, Statistical machine learning, dataset from kraggle etc.

CONCLUSION:

Free and open source analytics software plays a key role in this current era to predict useful information from the raw datasets. The free and open source software tools like python, R are driving force in analytical modernizations and help researchers to predict better insights from the huge datasets. Data Analytics and predictive modelling is gaining its acceptance in almost all applications of real world. One of the data analytics techniques i.e., regression model is an interesting topic to the researchers as it is precisely and competently extract the data for future insights and interpretations. This research work extracted the dataset related to the students’ marks for a particular course. This will really help students and teachers to improve the performance of the students.