Machine Learning Engineer Nanodegree

Using supervised learning we need to predict the student performance of there secondary education.

Akhila.Chitluri
June 27th ,2018

Proposal:

Domain Background:

I today's world every one are curious to know about future. Especially about our career and performance in further studies. For that purpose we need to evaluate the our past performance
for that supervised learning is very useful for prediction by regression and classification. This techinique3s are used in many streams that will to predict our future Eg : whether prediction,wholesale customer or retail customer,medical fields.

In our project we need to find the student performance based on the there primary education and then we will predict there performance in there secondary education.

Dataset Link : https://archive.ics.uci.edu/ml/datasets/student+performance

Problem Statement

Given the details of 649 students related to there primary education and some features related to there family and there economical status based on that now we need to predict the how a student secondary education would be.

Features and description :
these are the features related to dataset:

→ School
→ sex
→ age
→ family size
→ Father's Education
→ mother's Education
→ mother's job
→ Father's job
→ Internet
→ travel time to school to home
→ family educational support
→ wants to go to higher educational
→ quality of family relationships
→failures
→ absences
→ 1st time grade

→ 2nd term grade
By considering the above factors we will classify the data and then we will predict their future performance in there secondary education.

Dataset and Inputs :

The data set is based on the primary education and some additional information of about 649 students You can see the dataset by using this link: https://archive.ics.uci.edu/ml/datasets/student+performance. It contains of 33 attributes there are :

→ School
→ sex
→ age
→ address
→ family size
→ parents cohabitation status
→ Father's Education
→ mother's Education
→ mother's job
→ Father's job
→reason to choose this school
→ guardian
→ Internet
→ travel time to school to home
→ study time
→ extra classes
→ nursery
→ health status
→ family educational support
→ wants to go to higher educational
→ quality of family relationships
→failures
→ absences
→ 1st time grade
→ 2nd term grade

I took this data set and made some changes to it as some of the attributes are not necessary like address of the student and there parents cohabitational status ,guardian .

Solution statement:

The solution is to predict that the if any student data is given we will predict our there performance in there secondary education based on the our classification.

Benchmark model :

Initial we will use logistic regression . There F beta score and accuracy will be take as reference and the we will use other models to get better results for accuracy and F beta score .Generally ensemble methods will give best results in the case of more features to predict.

Evaluation Metrics:

In this we will use the Accuracy ,F beta score and others like recall,precision .

We define them as :

accuracy =(TP+TN)/(TP+TN+FP+FN)
precision = TP/(TP+FP)
recall = TP(TP+FN)
f beta score = (1+beta^2)*(precision*recall)/(beta^2*precision)+recall
beta =0.5

Project Design:

Data Analysis:
→ Initially we will read the csvfile into a data frame
→ we will find if there are any missing values

Data cleaning:
→ we will remove all the missing values and inconsistent data.
→ we will detect the outliers and we remove them.

Data preprocessing:
we will normalize the values using min-max scalar functions and log transform functions

Data modeling:
→ Then we split the data into training data and testing data.
→ we will find the some important features that predict the data will with the feature extraction functions in some ensemble methods and find the important features.
→ we will choose a benchmark algorithm we will train the data with that classifier and then we will fit the model on that specific features.
→ based on that we will find the F beta score and accuracy scores.
→ by doing histograms and we find the different algorithms for this which gives the better score and then finally we use that algorithm to model our data.

After all this we will use our testing data to predict and confirm our accuracy. As a result we will get good accuracy.