VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

Review Report - 0

Programme: BTech

Course: CSE2004-DBMS

Slot: D2

Faculty: Prof M. Premalatha

Component: J

Title: STUDENT MARKS PREDICTION

Team Member(s):

DOLLY AGARWALA (20BCE1863)

NAYAN KHEMKA (20BCE1884)

HARSH CHAUHAN (20BCE1886)

Abstract
[What, Why, How]

This project attempts to predict grade of students for Maths and Portuguese subjects based on data collected from two Portuguese schools using classification and prediction models. The goal is to go beyond knowing what has happened to providing a best assessment of what will happen in the future. Predictive models use known results to develop (or train) a model that can be used to predict values for different or new data. Modelling provides results in the form of predictions that represent a probability of the target variable based on estimated significance from a set of input variables.

Introduction
[Brief explanation about the project]

1. **Data Set Description**:

   Source: - https://archive.ics.uci.edu/ml/datasets/student+performance

   Description of Dataset:-This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por).

   Attribute Information: There are 32 attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course).
   Number of rows =

2. **Methodology and Algorithm used:**

   This project presents main characteristics of data sets in concise form. It covers following two basic data mining techniques:

   - Regression: Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as **temperature, age, salary, price,** etc.

   - Prediction: The word prediction in machine learning refers to the output of a trained model, representing the most likely value that will be obtained for a given input. Prediction in machine learning has a variety of applications, from chatbot development to recommendation systems. The model is trained with historical data, and then predicts a selected property of the data for new inputs. Prediction is used in lots of different areas, since it allows us to make highly accurate guesses about many things, such as predicting what the stock markets will do on any given day, predict results in sports, or even help the medical industry predict diseases. The algorithms for prediction are classified as supervised learning algorithms since they need a training dataset with correct examples to learn from them.

- Classification: As the name suggests, Classification is the task of "classifying things" into sub categories.But, by a machine. If that doesn't sound like much, imagine your computer being able to differentiate between you and a stranger. Between a potato and a tomato. Between an A grade and a F.In Machine Learning and Statistics, Classification is the problem of identifying to which of a set of categories (sub populations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.

Algorithms used

i. DECISION TREES:

Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression).

ii. XG BOOST:

XGBoost is an open-source software library which provides the gradient boosting framework for C++, Java, Python, R, and Julia. It works on Linux, Windows, and macOS. From the project description, it aims to provide a "Scalable, Portable and Distributed Gradient Boosting (GBM, GBRT, GBDT) Library". Other than running on a single machine, it also supports the distributed processing frameworks Apache Hadoop, Apache Spark, and Apache Flink. It has gained much popularity and attention recently as it was the algorithm of choice for many winning teams of a number of machine learning competitions.

iii. SUPPORT VECTOR MACHINE:

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well.