A Project Report On

Data Mining and Wear Housing

(Student Performance Analysis)

SUBMITTED BY

Nikhil Jain	**Roll No:41425**
Atharva Jagtap	**Roll No:41423**
Vaibhav Marathe	**Roll No:41433**

CLASS: BE-4

GUIDED BY

Dr. S.D.Kale

DEPARTMENT OF COMPUTER ENGINEERING

PUNE INSTITUTE OF COMPUTER TECHNOLOGY

DHANKAWADI, PUNE-43

SAVITRIBAI PHULE PUNE UNIVERSITY

2020-21

COMPUTER ENGINEERING DEPARTMENT

Academic Year: 2020-21

1. Abstract:

Data mining is the analysis step for discovering knowledge and patterns in large databases and large datasets. Data mining is the process of applying machine learning methods with the intention of uncovering hidden patterns in large data sets. Data mining techniques basically involve many different ways to classify the data. Such classified data is used to fast access to data and for providing fast services to the customers. Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.

2. Introduction

One of the most active and exhilarating research areas is Data Mining where it is extracting or mining knowledge from large amounts of data. Data mining is a natural result of the advancement of information technology. Data mining helps in discovering treasured information and patterns which can be applied to decision making in business, business management, marketing analysis, production control, science exploration and engineering design. Its wide applications bring data mining, an integral part of the business.

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. In other words, a cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters.

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks. Classification has many applications in customer segmentation, business modelling, marketing, credit analysis, and biomedical and drug response.

3. Problem Statement

Classification and analysis on student Performance predictive analysis techniques like logistic regression,random forest,SVM.

4. Objective

To classify students according to various features affecting conclusively to their performance in exams such as gender,parent's education,etc.

5. Scope

To predict the outcome of whether a student will be able to pass or fail and calculate the accuracies of predictions using different classification techniques.

6. Software and Hardware requirement

- Jupyter Notebook
- 4GB RAM
- 500 GB HDD
- OS(64 Bit) Windows/Ubuntu
- Programming Language- Python

7. Theory and Design

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.

In Real world **data** are generally incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate **data**.,Noisy: containing errors or outliers,Inconsistent: containing discrepancies in codes or names.

7.1. Steps in Data Preprocessing

Step 1 : Import the libraries

Step 2 : Import the data-set

Step 3 : Check out the missing values

Step 4 : See the Categorical Values

Step 5 : Splitting the data-set into Training and Test Set

Step 6 : Feature Scaling

7.2. Algorithm

1.Logistic Regression

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for a given set of features(or inputs), X.

For example, y is a categorical target variable which can take only two possible types:“0” or “1”.

2.Support Vector Machine

SVM is a supervised machine learning algorithm capable of performing classification, regression.

All the data points that fall on one side of the line will be labeled as one class and all the points that fall on the other side will be labeled as the second.

3.Random Forest Classifier

Random forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction.

8. Testing/Result and Analysis

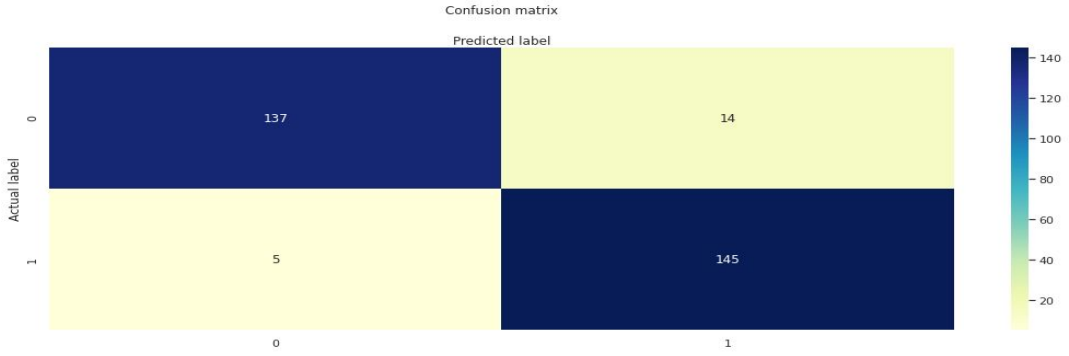
1. Logistic Regression

1.1 Confusion Matrix

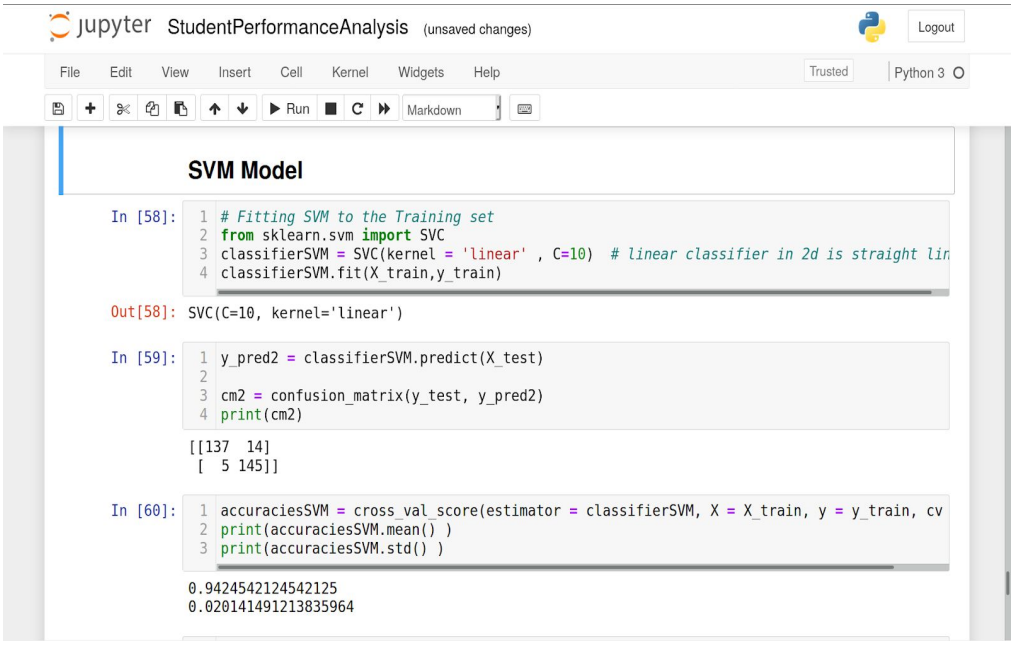
[[137 14]
[5 145]]

Accuracies(Mean) **0.9424297924297924**

Accuracies(std) **0.023105863332232503**



2. Support Vector Machine Analysis



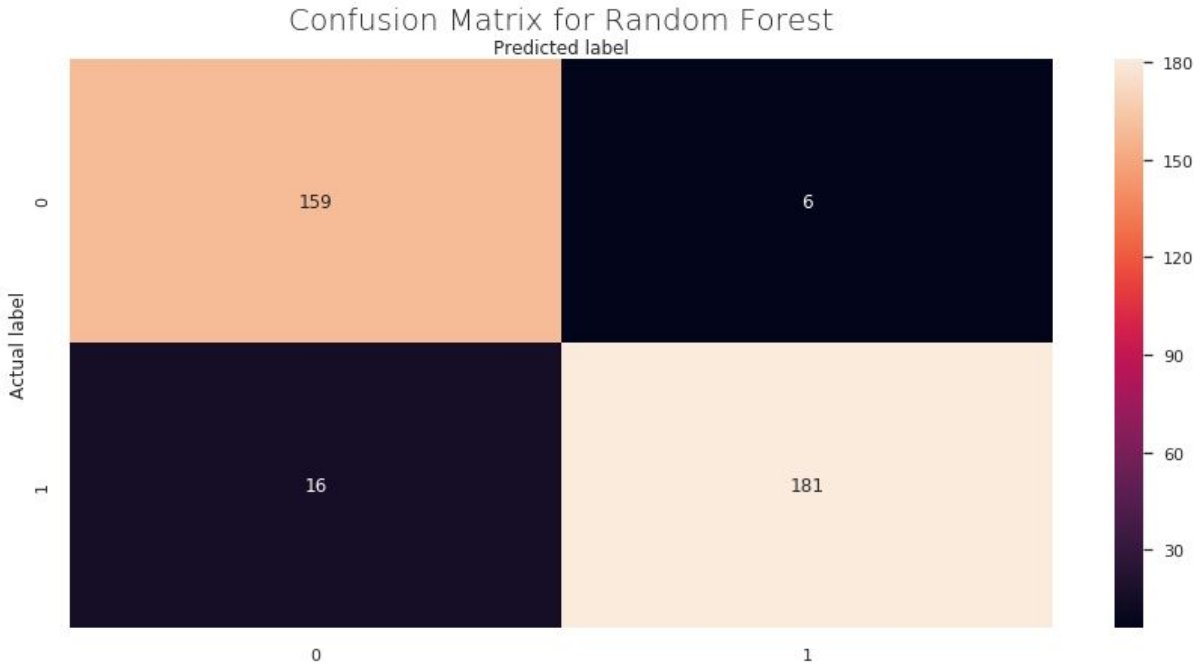
3. Random Forest Analysis

3.1 Confusion Matrix

[[159 6]
[16 181]]

Training Accuracy : 99.5249406175772

Testing Accuracy : 93.92265193370166



9. Conclusion and Future Enhancements

Random Forest,SVM,Logistic regression algorithms can be used to classify the given data points into several classes and predict the desired result. Classification using various algorithms is analysed and their prediction accuracies are compared.

10. References

1.https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f
2.https://www.statisticssolutions.com/what-is-logistic-regression/
3.https://towardsdatascience.com/support-vector-machine-python-example-d67d9b63f1c8
4.https://www.kaggle.com/roshansharma/student-performance-analysis/data