PUNE INSTITUTE OF COMPUTER TECHNOLOGY
DHANKAWADI, PUNE

DATA ANALYTICS MINI-PROJECT REPORT
ON

"TRIP HISTORY ANALYSIS "

SUBMITTED BY

Nachiket Erlekar   41434
Ashay Koradia   41429

Under the guidance of
Prof. Deepali Kadam

DEPARTMENT OF COMPUTER ENGINEERING
Academic Year 2020-21

Contents

1   Problem Statement

Use trip history dataset that is from a bike sharing service in the United States. The data is provided quarter-wise from 2010 (Q4) onwards. Each file has 7 columns. Predict the class of user.

Make use of at least two classification algorithms and provide comparative analysis.

2 Abstract

Data Analytics is the process of analyzing data and drawing conclusion from it. One such dataset is the trip history of capital bikeshares, which logs the travel history of its riders. The dataset is available for each quarter after 2010. We select the first quarter of 2017 for our analysis. The main goal is predict the class of the user as Member or Casual. We inspect various algorithms to achieve this goal and compare their performance.

3   Hardware and Software Requirements

3.1   Hardware Requirements

1. 500 GB HDD

2. 4GB RAM

3. Monitor

4. Keyboard

3.2   Software Requirements

1. 64 bit Open Source Operating System like Ubuntu 18.04

2. Python 3

3. Google Colab

4. Different Libraries

5. Libararies like sklearn, pandas, matplotlib

4   INTRODUCTION

The data includes:
1. Duration – Duration of trip

2. Start Date – Includes start date and time

3. End Date – Includes end date and time

4. Start Station – Includes starting station name and number

5. End Station – Includes ending station name and number

6. Bike Number – Includes ID number of bike used for the trip

7. Member Type – Indicates whether user was a "registered" member (Annual Member, 30-Day Member or Day Key Member) or a "casual" rider (Single Trip, 24-Hour Pass, 3-Day Pass or 5-Day Pass)

This data has been processed to remove trips that are taken by staff as they service and inspect the system, trips that are taken to/from any of "test" stations at warehouses and any trips lasting less than 60 seconds (potentially false starts or users trying to re-dock a bike to ensure it's secure).

We first preprocess the data. The Data fields are used to extract year, month, day, hour, week, minute and second. We add this as features to our dataset. Out of this we keep only month and hour field as rest fields have a uniform distribution across entire dataset.

We perform one hot encoding of the remaing data fields and use Random Forest Classifier with default parameters as our classificaton algorithm. We use train test split of 60-40 and report a classification accuracy of 90 %.

5   OBJECTIVE

- To analyse trip history dataset

- To predict the class of the user from given dataset..

6   Scope

We select only the 2010 file of records for our analysis. This is because the size of whole dataset makes it difficult to run the model on limited memory, currently we are making use of 115,000 records on the datasets to make our models work.
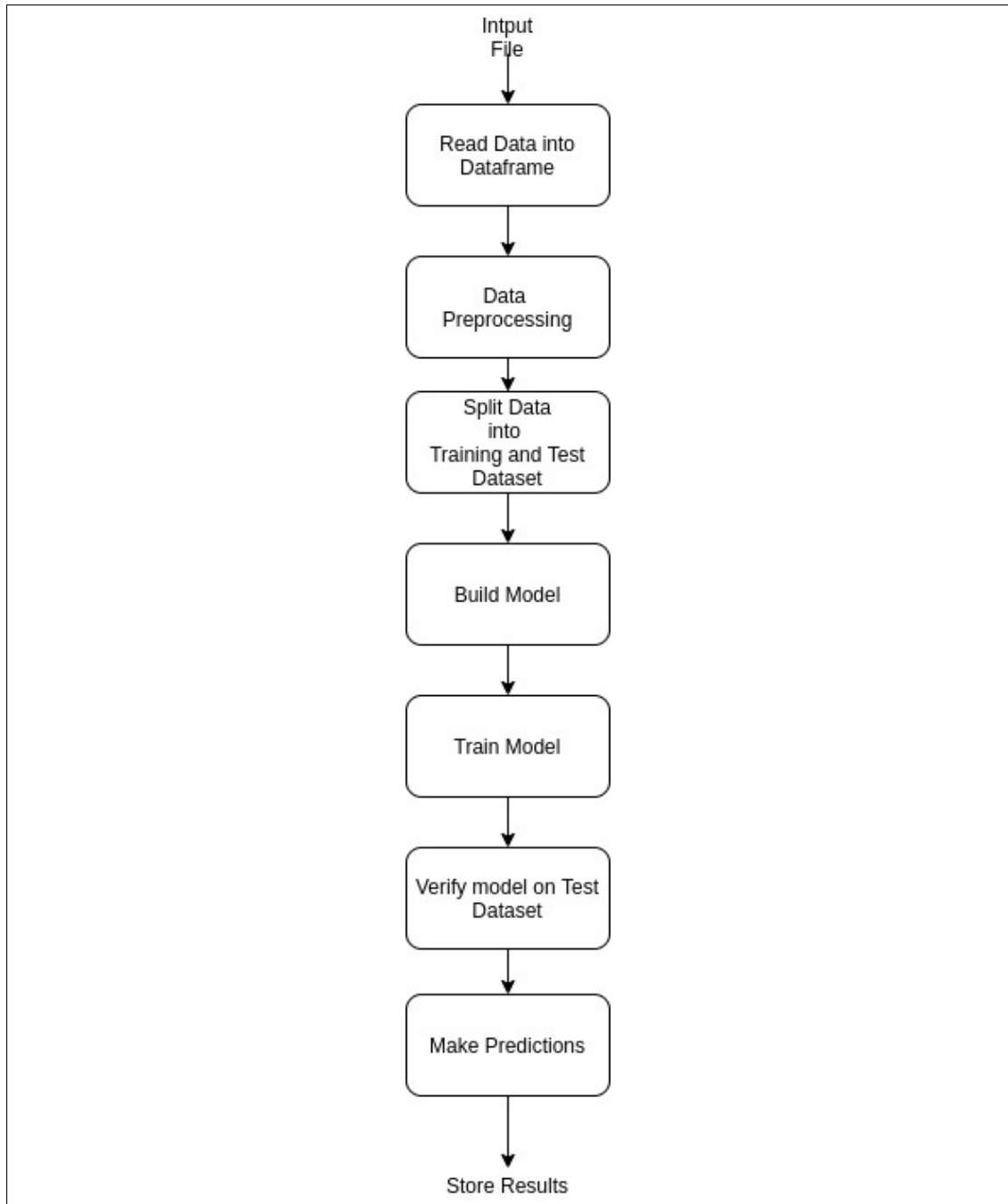
7 System Architecture

Figure 1: System Architecture

8 Result

The Cross Validation Scores for Various models are:

Model	Cross-Validation Score
DecisionTree	0.7924939083293684
RandomForest	0.8701789293077843
GradientBoostingClassifier	0.8647865392101735
XGBoost	0.8642098129965253
KNeighborsClassifier	0.8464033911501363
GaussianNB	0.8398503395475583
SVC	0.8501232752281673
LinearSVC	0.8334054241100394

Table 1: Cross Validation Scores for vaious Models

We see that Random Forest Classifier gives the best score. We then use this model to perform training and testing of the model. After training, the model gives an accuracy of 87 %.

9   Conclusion

We presented classification of trip hsitory dataset to predict the clas of user using Random Forest Classifier. We report an classification accuracy of 87%.

References

[1] https://www.capitalbikeshare.com/system-data

[2] https://scikit-learn.org/stable/modules/generated/
   sklearn.ensemble.RandomForestClassifier.html