PUNE INSTITUTE OF COMPUTER TECHNOLOGY
DHANKAWADI, PUNE

DATA MINING AND WAREHOUSING MINI-PROJECT REPORT
ON

"PREDICTING FAKE NEWS USING VARIOUSMODELS"

SUBMITTED BY

Tejas Dahad	41171
Varun Karwa	41174
Prachi Wagh	41176

Under the guidance of
Prof. M.S.Chavan

DEPARTMENT OF COMPUTER ENGINEERING
Academic Year 2021-22

Problem Statement

Consider a labeled dataset belonging to an application domain. Apply suitable datapreprocessing steps such as handling of null values, data reduction, discretization. For prediction of class labels of given data instances, build classifier models using different techniques (minimum 3), analyze the confusion matrix and compare these models. Alsoapply cross validation while preparing the training and testing datasets.

Abstract

Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered)class labels. For example, we can build a classification model to categorise bank loan applications as either safe or risky. Such analysis can help provide us with a better understanding of the data at large. In this project we use multiple classification models to analyse if the news is fake or not. Use apply suitable data preprocessing steps.We then compare performance of classification models tofind which one is the best

Hardware and Software Requirements

1. Hardware Requirements

1. 500 GB HDD
2. 4GB RAM
3. Monitor
4. Keyboard

2. Software Requirements

1. 64 bit Open Source Operating System
2. Python 3
3. Google Colab
4. Different Libraries
5. Libararies like sklearn, pandas, matplotlib

INTRODUCTION

We have been provided with the data regarding various aspects of the news. The Data fields are

1. Id − Unique id given to each news.
2. Title - Title of the news
3. Text - Content of the news
4. Subject - Type of the content
5. date − Date on which the news was published.

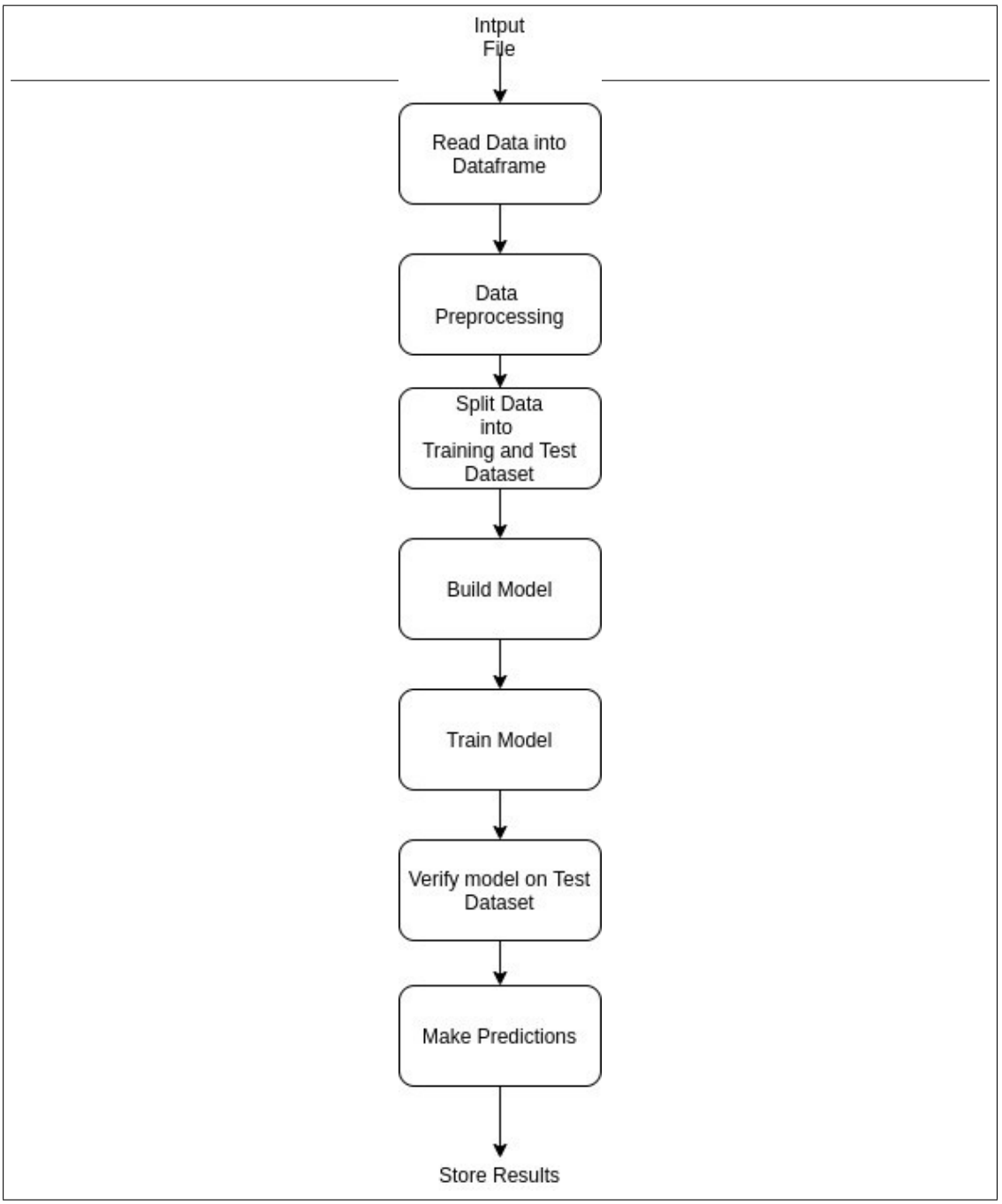
The train set contains 30000 records while the test set contains 20000 records.We added the target column in our analysis.

5 OBJECTIVE

- To understand data preprocessing
- To perform classification on dataset and predict labels for test dataset.

6 Scope

We select dataset of news from various sources. We try to apply various models and compare which one is the best model amongst them.



7 System Architecture

Figure 1: System Architecture

9 Result

The Accuracy for Various models are:

Model	Accuracy
Naive Bayes Classifier	94.08
LogisticRegression	98.73
SVC	99.39

Table 2: Accuracy of vaious Models

We see that Support Vector 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 99.39 %.

Figure 6: Comparison of various models

10 Conclusion

We have analysed the news dataset and performed data pre-processing steps.We have experimented multiple classification models and found out the best performer amongst them. We presented classification of news to predict the true/fake using Support Vector Classifier. We report a classification accuracy of 99.39