VIT®

Vellore Institute of Technology

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

School of Information Technology and Engineering

Submitted towards lab component of the course

Data Mining Technique

ITE2006

Under the guidance of

Dr. Ranichandra C

Associate Professor, SITE

Submitted by

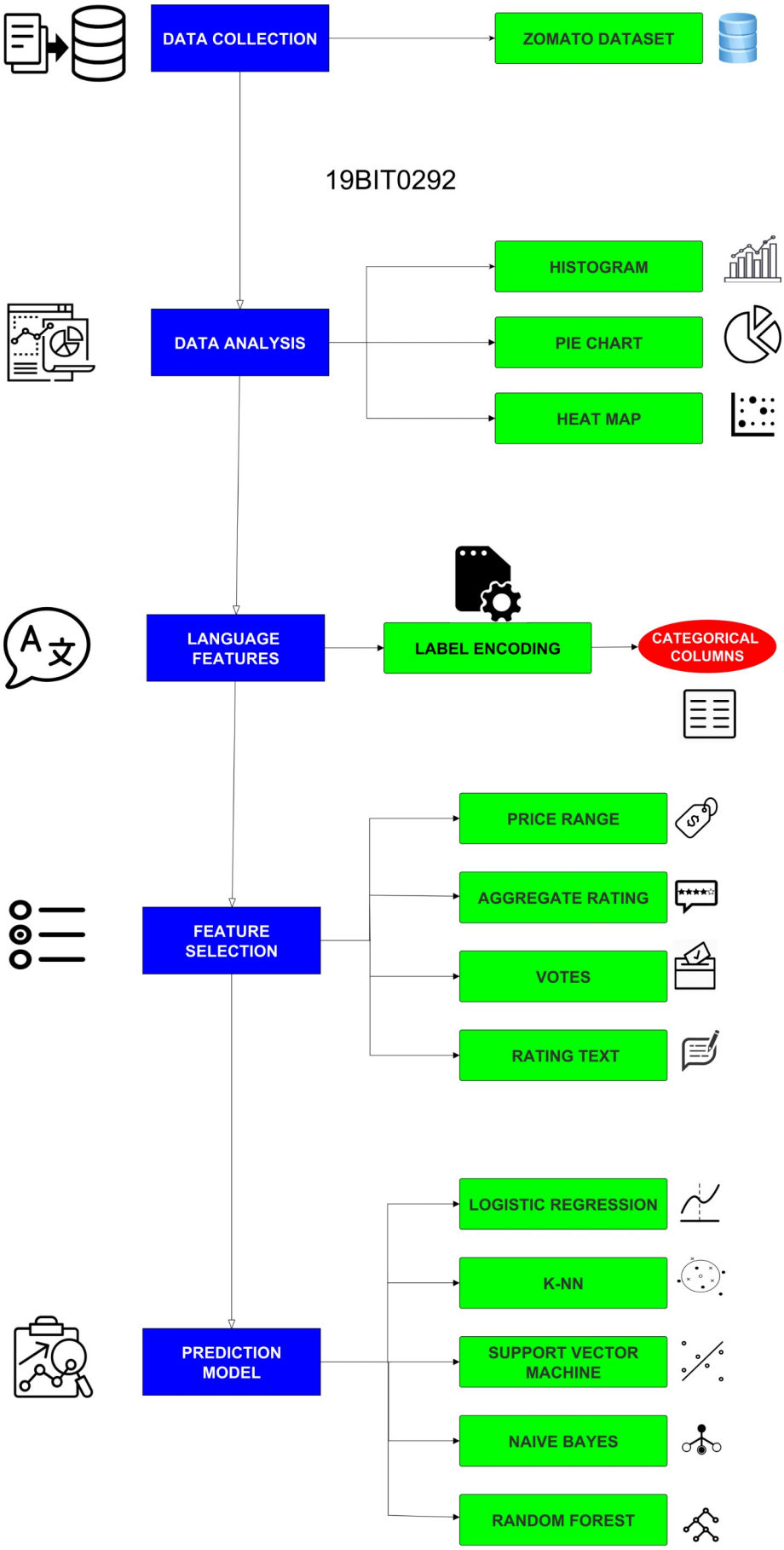
Bhaumik Tandan (19BIT0292)

Dhaval Mavani (19BIT0316)

Parinita Parihar (19BIT0287)

Review 2

Architecture diagram



Modules -Explanation

The proposed system for predictive analysis consists of language features and embedded classification models. The dataset (Zomato Dataset) is split into test and training data. The first step is to generate the ground truth, and is performed based on the training reviews and ratings. We perform Exploratory data Analysis and gain an insight into the contents of dataset and their relation to each other through various graphs and charts. The next step in the model was to perform Data Cleaning, Feature Selection and Feature Engineering. In this step we also deal with missing data. In the process of building models, we extract:

(a) Statistical features of users and reviews, and

(b) Language features extracted from the review corpus. The next step was to test and evaluate the various classifiers which were used. We also decided to test and check for over fitting and find out the trend of the training vs true error to pick the best model.

(1) The data analysis and exploration of the dataset consisted of Displaying the ratings of restaurants by country code and sorting them according to that criteria. We explored the data further by calculating the percentage of restaurants that were not rated and compared these restaurants in terms of percentage to whole of the dataset. From this analysis we explored and analysed that lower priced food was often not rated. Therefore, by similar case imputation, we replaced the 'not rated' with the average rating of lower priced food. We also noted that lower priced food is considered average for the most part, while more expensive food is rated more highly. Expensive food seems rarer than low cost food. We further explored the data with the

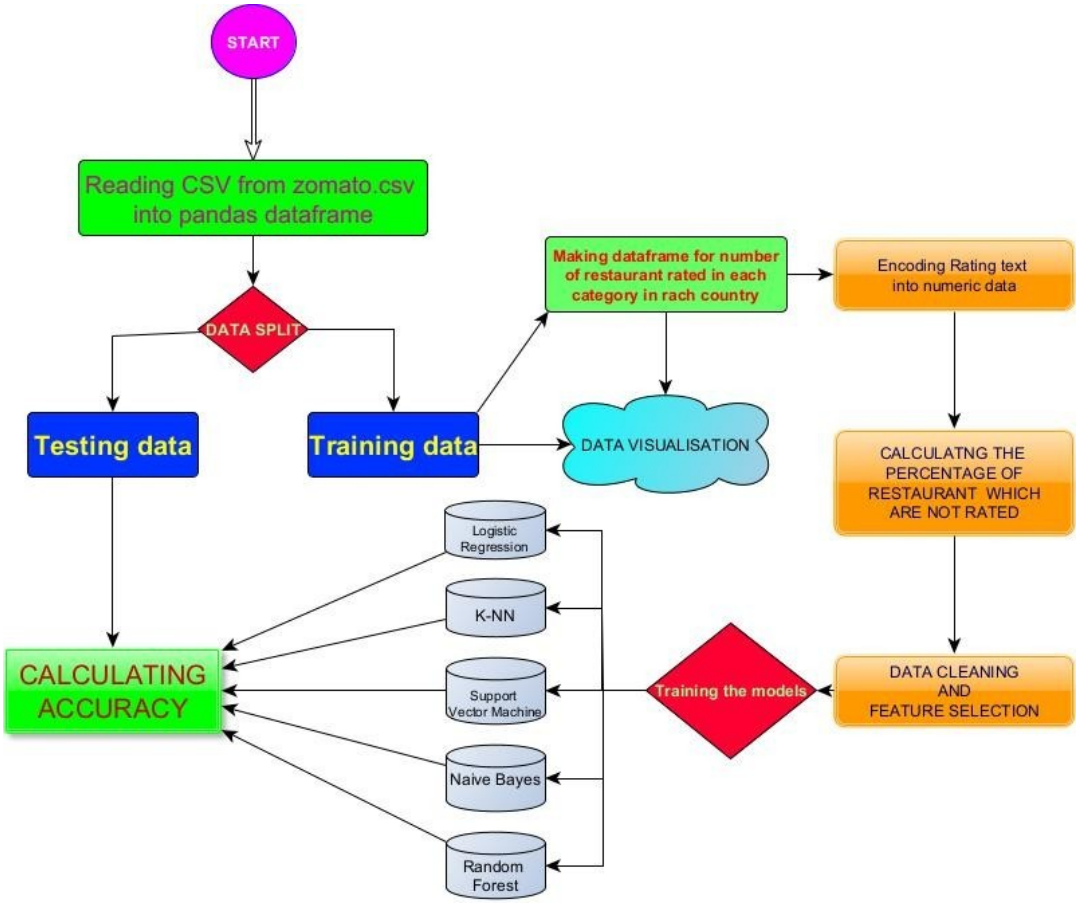
help of various charts and compared the relationships of the various attributes of the dataset.

(2) Data Cleaning step of the model consisted of dealing with the attributes not important for the prediction of the final output and hence these attributes were dropped.

(3) Feature Engineering and Feature selection: In this step we analysed all the restaurants that were not rated and were given a rating based on their price range. As an aggregate rating of 2.5-3.5 translates to a rating of "Average", the "Not Rated”, the text can be replaced with "Average" in the dependent variable. The next step in feature engineering was dealing with the 'Rating Text' as it Is given as a string. This needed to be encoded into numeric data so that the algorithm can perform calculations. This was done using the LabelEncoder tool from sklearn. We classified the rating text into the form where excellent = 1, very good = 4, good = 2, average = 0, and poor = 3.

(4) Having pre-processed the data, We now provided the data to different classifiers and see which one performs better in creating a model of classification for this data. We compared performance using the cross-validation. Python has the cross_val_score class from the sklearn.model_selection library to perform cross validation. We utilized supervised learning techniques for classifying high-dimensional data in order to predict.

Module Diagrams



Data pre-processing

1. **Data Cleaning:** The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc. We will drop the unnecessary columns in our training data. Like 'Restaurant ID', 'Restaurant Name', 'Country Code', 'City','Address', 'Locality', 'Locality Verbose', 'Longitude', 'Latitude', 'Cuisines', 'Average Cost for two', 'Currency', 'Has Table booking', 'Has Online delivery', 'Is delivering now', 'Switch to order menu', 'Rating colour', etc. We will replace the restaurants which are not rated with the mean rating in the price range.

2. **Data Integration:** This step involves integrating data from different resources, we will try to include countrycode.csv also in this project which includes the country name to the specific country.

3. **ENCODING:** This is done so as to get numerical data out of object data because the machine learning models fit only numerical data. We will encode the rating text with a value.

Data Mining -algorithms

1. **Logistic regression** is a classification algorithm, used when the value of the target variable is categorical in nature.
2. **K nearest neighbours** is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has 18 been used in statistical estimation and pattern recognition as a non-parametric technique.
3. **Support vector machines** (SVM) represents examples as points in space for nonlinear classification. Different kernels lead to differences in the performance based on the dataset.
4. **Naïve Bayes** (NB) is a probabilistic model with maximum likelihood to classify categories. We use NB to compare the results of different distributions.
5. **Random forest** (RF) is an ensemble technique that fits a number of decision tree classifiers on various sub-samples of the dataset. We have experimented with different number of decision trees as estimators.

Pattern Evaluation

We compared performance using the cross-validation. Python has the cross_val_score class from the sklearn.model_selection library to perform cross validation. We utilised supervised learning techniques for classifying high-dimensional data in order to predict.

By default, cross_val_score uses KFold cross-validation. This works by splitting the data set into K equal folds. Say we have 3 folds (fold1, fold2, fold3), then the algorithm works as follows:

i. Use fold1 and fold2 as your training set in svm and test performance on fold3.
ii. Use fold1 and fold3 as our training set in svm and test performance on fold2.
iii. Use fold2 and fold3 as our training set in svm and test performance on fold1.

So each fold is used for both training and testing.

Now to second part of your question. If you increase the number of rows of data in a fold, you do reduce the number of training samples for each of the runs (above, that would be run 1, 2, and 3) but the total number of training samples is unchanged.

Generally, selecting the right number of folds is both art and science. For some heuristics on how to choose your number of folds, I would suggest this answer. The bottom line is that accuracy can be slightly affected by your choice of the number of folds. For large data sets, you are relatively safe with a large number of folds; for smaller data sets, you should run the exercise multiple times with new random splits.

Knowledge representation

We have plotted some of the below listed graphs:

(1) Bar graph

A bar graph is a chart that plots data using rectangular bars or columns (called bins) that represent the total amount of observations in the data for that category. Bar charts can be displayed with vertical columns, horizontal bars, comparative bars (multiple bars to show a comparison between values), or stacked bars (bars containing multiple types of information).

- Bar graphs can be created to show data in multiple, highly visual ways.
- Bar graphs have an x- and y-axis and can be used to showcase one, two, or many categories of data.
- Data is presented via vertical or horizontal columns.
- The columns can contain multiple labeled variables (or just one), or they can be grouped together (or not) for comparative purposes.

(2) Histogram

A histogram is a graphical representation that organises a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

- A histogram is a bar graph-like representation of data that buckets a range of outcomes into columns along the x-axis.
- The y-axis represents the number count or percentage of occurrences in the data for each column and can be used to visualize data distributions.

- In trading, the MACD histogram is used by technical analysts to indicate changes in momentum.

(3) Heat map

A heat map is a two-dimensional visual representation of data using colours, where the colours all represent different values.

A heat map can be used with all sorts of data, from the real estate market representing the number of foreclosures to the spreads of credit default swaps (CDS) to webpage analysis reflecting the number of hits a website receives.

- A heat map is a graphical representation of data in two-dimension, using colours to demonstrate different factors.
- Heat maps are a helpful visual aid for a viewer, enabling the quick dissemination of statistical or data-driven information.
- On the downside, heat maps provide only selective information, therefore clouding the big picture on an issue; heat maps are also often prepared when only preliminary information is available.
- While heat maps are used in a variety of industries and circumstances, they are commonly employed to show user participation on a website.