

PROJECT REPORT ON DATA SCIENCE

SUBMITTED BY : SINO LEWIS NZAU

REG.NO : BCSC01/0071/2018

This report comprises of three main sections:

- a model demonstrating supervised machine learning,
- unsupervised learning algorithms ,
- text mining as applied in data science.

SUPERVISED LEARNING

1.1 Car Evaluation PREDICTION MACHINE LEARNING

INTRODUCTION.

Looking for the best project to show supervised learning. I decided to go with Car Evaluation being one of my favourite hobbies. I will show how algorithms have helped build model to classify whether a car is unacc (acceptable), acc (acceptable), good, vgood.

1.2 PROCEDURE USED

This section introduces as to step by step process that I followed in order to come up with this model. It involves:

1) Understanding the goal of the project.

First and foremost a data science, its required that before you solve a problem is to define exactly what you want. In this case my main aim is to clearly show how supervised learning algorithms have been adapted. In this case, I'll have to create a machine learning model using supervised learning. I picked on Car Evaluation prediction model to illustrate supervised learning algorithm.

2) COLLECTING DATA

The dataset to be used is from archive.ics.uci.edu. [Archive.ics.uci.edu](http://archive.ics.uci.edu) is a website for data science course. It has various datasets for various machine learning tasks. The data set that was used to train the model to predict Car Evaluation was gathered from an open source data shared by a data scientist at a repository. The data set contained information about cars. Dat hs been labeled **car.data**

Below is a snapshot of car dataset.

```
1729 lines (1729 sloc) | 50.7 KB
1 buying,maint,door,persons,lug_boot,safety,class
2 vhigh,vhigh,2,2,small,low,unacc
3 vhigh,vhigh,2,2,small,med,unacc
4 vhigh,vhigh,2,2,small,high,unacc
5 vhigh,vhigh,2,2,med,low,unacc
6 vhigh,vhigh,2,2,med,med,unacc
7 vhigh,vhigh,2,2,med,high,unacc
8 vhigh,vhigh,2,2,big,low,unacc
9 vhigh,vhigh,2,2,big,med,unacc
10 vhigh,vhigh,2,2,big,high,unacc
11 vhigh,vhigh,2,4,small,low,unacc
12 vhigh,vhigh,2,4,small,med,unacc
13 vhigh,vhigh,2,4,small,high,unacc
14 vhigh,vhigh,2,4,med,low,unacc
15 vhigh,vhigh,2,4,med,med,unacc
16 vhigh,vhigh,2,4,med,high,unacc
17 vhigh,vhigh,2,4,big,low,unacc
18 vhigh,vhigh,2,4,big,med,unacc
19 vhigh,vhigh,2,4,big,high,unacc
20 vhigh,vhigh,2,more,small,low,unacc
21 vhigh,vhigh,2,more,small,med,unacc
22 vhigh,vhigh,2,more,small,high,unacc
23 vhigh,vhigh,2,more,med,low,unacc
24 vhigh,vhigh,2,more,med,med,unacc
25 vhigh,vhigh,2,more,med,high,unacc
26 vhigh,vhigh,2,more,big,low,unacc
27 vhigh,vhigh,2,more,big,med,unacc
28 vhigh,vhigh,2,more,big,high,unacc
29 vhigh,vhigh,3,2,small,low,unacc
30 vhigh,vhigh,3,2,small,med,unacc
31 vhigh,vhigh,3,2,small,high,unacc
32 vhigh,vhigh,3,2,med,low,unacc
33 vhigh,vhigh,3,2,med,med,unacc
34 vhigh,vhigh,3,2,med,high,unacc
35 vhigh,vhigh,3,2,big,low,unacc
```

Figure 1 Dataset screenshot

3.) PROCESS DATA FOR ANALYSIS

This includes data preprocessing. Now that I already have the dataset, my next step is to keenly analyze and preprocess the data set. Below are some ways that I implemented to help in preprocessing:

3.3.1 MANUAL EXPLORATION

This step is very important in the development of machine learning algorithms because we analyze the dataset and label a car unacc (acceptable), acc (acceptable), good, vgood.

3.3.2 DATA PREPROCESSING

It's an important step in Machine learning as the quality of the data and the useful information that can be derived from it directly affects the ability of our model to learn.

3.3.3 FEATURES SELECTION

This is also called variable selection, it's the process of selecting a subset of relevant features for use in model construction. The classifier used to describe our data set are enumerated by the following attributes

Attributes(Columns):

buying: vhigh, high, med, low.

maint: vhigh, high, med, low.

doors: 2, 3, 4, 5more.

persons: 2, 4, more.

lug_boot: small, med, big.

safety: low, med, high.

Those are the best features selected that are suitable to train this model, the rest of these features we have deleted them.

3.3.4 DATA CLEANING

As we all know raw data is mostly not pure. There are several ways that I have used to clean my data.

i) **Find missing values.**

One of the first steps in data cleaning is to check missing values/ incomplete values and fill them out. Looking at our dataset, we will first of all get the **count** of the dataset provided and from that we will be able to identify features with missing values.

ii) **Removing rows or columns**

Just as explained under feature selection, I removed various columns that were not helpful to my model. Machine learning uses **drop ()** Function that helps in this

iii) **Label encoding**

It involves changing the data type from one form to another. i.e in this case features are label encoded from STRING type to DECIMAL base 4 form.

3.4 MODEL DEVELOPMENT

Is an iterative process in which many models are derived, tested and built upon a model fitting the desired criteria is built. After all the data analysis and preprocessing has been performed, my next step was to now build the model. It's in this stage we built and code from scratch. After data preprocessing, I divided data into 80% for training and 20% for testing. Several algorithms were also used to train and test to determine the most suitable for predicting well.

3.4.1 CLASSIFICATION MODELLING

I used 5 different machine learning algorithms to determine the one with the highest accuracy score. Below are discussions of the algorithms that I have adopted:

1. SUPPORT VECTOR MACHINE

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. In practice, SVM algorithm is implemented with kernel that transforms an input data space into the required form. SVM uses a technique called the kernel trick in which kernel takes a low dimensional input space and transforms it into a higher dimensional space.

2. DECISION TREES CLASSIFIERS

A decision tree has been a very successful classifier that has been applied in many domains. They are built using a recursive partition process in which data points are split at each node by using the selected split criteria. A path from the root node to a leaf is a rule which is used for the prediction

3. Logistic regression

It uses a logistic function to model the dependent variable. It clearly gives two possible classes. It's mostly used when the data has a binary output.

4. Random forest regressor

It's a classification algorithm consisting of many decision trees and uses debugging when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

5. Naïve Bayes

It's a probabilistic classifier. It is based on probability models that incorporate strong independence assumptions.

3.5 TESTING ACCURACY

All the five algorithms were implemented. Car dataset were trained for all the algorithms individually, after this all of them were tested. The most efficient algorithm was to be selected on their accuracy score.

To test for accuracy foreach algorithm, I first had to import the correct metrics for accuracy and used accuracy score to get the percentage for each algorithm. From this model, I found the results as follows:

Below is the figure showing accuracy of the algorithms in an ascending order:

```
In [78]: models = pd.DataFrame({
    'Model': ['Support Vector Machines', 'Logistic Regression',
              'Random Forest', 'Naive Bayes',
              'Decision Tree'],
    'Score': [acc_svc, acc_logreg,
              acc_randomforest, acc_gaussian, acc_decisiontree]})
models.sort_values(by='Score', ascending=False)
```

Out[78]:

	Model	Score
4	Decision Tree	97.424103
0	Support Vector Machines	96.596136
1	Logistic Regression	95.952162
2	Random Forest	90.038255
3	Naive Bayes	73.413063

I decided to finally train the model with Support Vector Machines.

1.3 CHALLENGES OF THIS MODEL

1. The main challenge I encountered in this model was having to build the model using logistic regression despite decision tree giving the highest accuracy score.
2. Another challenge was a lot of time was taken during preprocessing

UNSUPERVISED LEARNING

INTRODUCTION

Unsupervised machine learning involves learning patterns from untagged data. The model itself finds the hidden patterns and insights from the given data.

2.1 Car Evaluation USING K-NN

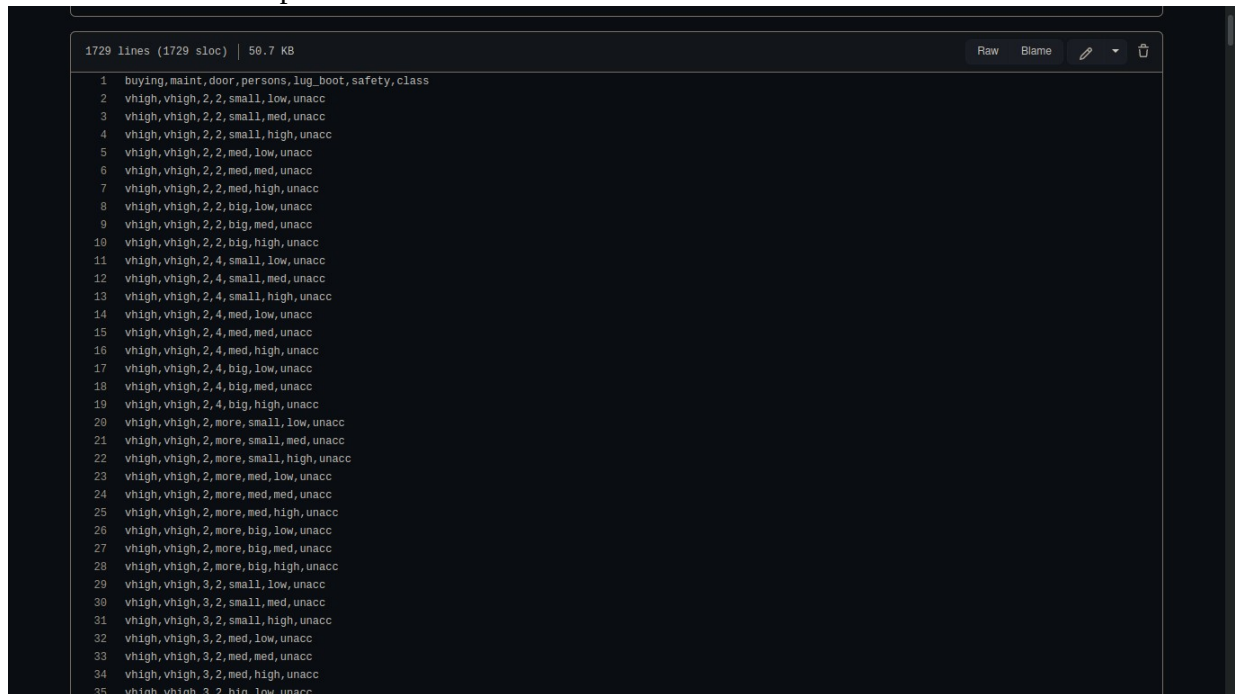
K-means being an unsupervised learning algorithm its used mostly in clustering and its also known as segmentation. It works in a way for putting the data points into a predefined number of clusters labelled as k. the k in k-means algorithm is the input since you'll realize in the algorithm the number of clusters you want to identify in the data. Each item of the data used gets assigned to the nearest cluster center called the centroids, the procedure of clustering may be repeated as many times as possible until the clusters are well defined.

2.2 PROCESS FOLLOWED

- 1.) Understanding the goal of the project- same as discussed above, the main goal of this second part was to demonstrate unsupervised machine learning algorithm. I decided to show this by using K-nearest neighbors.

- 2.) Data collection

The dataset used in this case was from archive.ics.uci.edu and its called car in a data file. Below is a snapshot of it:



```
1729 lines (1729 sloc) | 50.7 KB
Raw Blame
1 buying,maint,door,persons,lug_boot,safety,class
2 vhigh,vhigh,2,2,small,low,unacc
3 vhigh,vhigh,2,2,small,med,unacc
4 vhigh,vhigh,2,2,small,high,unacc
5 vhigh,vhigh,2,2,med,low,unacc
6 vhigh,vhigh,2,2,med,med,unacc
7 vhigh,vhigh,2,2,med,high,unacc
8 vhigh,vhigh,2,2,big,low,unacc
9 vhigh,vhigh,2,2,big,med,unacc
10 vhigh,vhigh,2,2,big,high,unacc
11 vhigh,vhigh,2,4,small,low,unacc
12 vhigh,vhigh,2,4,small,med,unacc
13 vhigh,vhigh,2,4,small,high,unacc
14 vhigh,vhigh,2,4,med,low,unacc
15 vhigh,vhigh,2,4,med,med,unacc
16 vhigh,vhigh,2,4,med,high,unacc
17 vhigh,vhigh,2,4,big,low,unacc
18 vhigh,vhigh,2,4,big,med,unacc
19 vhigh,vhigh,2,4,big,high,unacc
20 vhigh,vhigh,2,more,small,low,unacc
21 vhigh,vhigh,2,more,small,med,unacc
22 vhigh,vhigh,2,more,small,high,unacc
23 vhigh,vhigh,2,more,med,low,unacc
24 vhigh,vhigh,2,more,med,med,unacc
25 vhigh,vhigh,2,more,med,high,unacc
26 vhigh,vhigh,2,more,big,low,unacc
27 vhigh,vhigh,2,more,big,med,unacc
28 vhigh,vhigh,2,more,big,high,unacc
29 vhigh,vhigh,3,2,small,low,unacc
30 vhigh,vhigh,3,2,small,med,unacc
31 vhigh,vhigh,3,2,small,high,unacc
32 vhigh,vhigh,3,2,med,low,unacc
33 vhigh,vhigh,3,2,med,med,unacc
34 vhigh,vhigh,3,2,med,high,unacc
35 vhigh,vhigh,3,2,big,low,unacc
```

- 3.) Data analysis.

Under this section, deep analysis on the dataset was done. Several steps were done to the dataset that included:

- ❖ Getting more information about the dataset.
- ❖ Checking on the frequency distribution of the dataset etc.

4.) IMPLEMENTING K-MEANS

K-Means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. In K-Means, each cluster is associated with a centroid.

How to Implementing K-Means Clustering ?

- ✓ Choose the number of clusters
- ✓ Select k random points from the data as centroids
- ✓ Assign all the points to the closest cluster centroid
- ✓ Recompute the centroids of newly formed clusters
- ✓ Repeat steps 3 and 4

5.) TESTING ACCURACY

Accuracy in this case is tested using accuracy score and we are shown that the accuracy score is 0.946

2.3 CHALLENGES FACED IN THIS MODEL

- ✓ Coming up with the number of k in the clusters was hard because it determines the accuracy of the model.
- ✓ determining the value of K was complex some time
- ✓ Some outliers affected the overall accuracy significantly.

TEXT MINING

BACKGROUND

Text mining is a process of extracting useful information and nontrivial patterns from a large volume of text databases. There exist various strategies and devices to mine the text and find important data for the prediction and decision-making process. The selection of the right and accurate text mining procedure helps to enhance the speed and the time complexity also.

Pre-processing and data cleansing tasks are performed to distinguish and eliminate inconsistency from the data. The data cleansing process makes sure to capture the genuine text, and it is performed to eliminate stop words stemming (the process of identifying the root of a certain word and indexing the data).

Gathering unstructured information from various sources accessible in various document organizations, for example, plain text, web pages, PDF records, etc.

Analysis Methods for text:

Text categorization: to assign a category to the text amount categories predefined by users

Sentiment analysis (or opinion mining) is a natural language processing (NLP) technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs. Sentiment analysis is the process of detecting positive or negative sentiment in text. It's often used by businesses to detect sentiment in social data, gauge brand reputation, and understand customers

Text clustering: to segment texts into several clusters depending on the substantial relevances.

Text summarization: to extract its partial content reflection its whole content automatically.

EXAMPLE

The example used for this project consist of a Text summarization AI .