LONDON METROPOLITAN UNIVERSITY

islington college
(इस्लिङ्टन कलेज)

Module Code & Module Title

CU6051NI - Artificial Intelligence

Assessment Weightage & Type 80%

Individual Coursework

Year and Semester 2020-21

Spring

Student Name: Shasank Shakya

London Met ID: 18029951

College ID: np01cp4a180303

Assignment Due Date: 7th February 2021

Assignment Submission Date: 7th February 2021

I confirm that I understand my coursework needs to be submitted online via Google Classroom under the relevant module page before the deadline in order for my assignment to be accepted and marked. I am fully aware that late submissions will be treated as non-submission and a marks of zero will be awarded.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

Artificial intelligence is coined as the capability of a computer or a machine to have human brain like intelligence. Some modern applications of A.I are to follow problem solving algorithms, object recognition, natural language processing, search algorithms, etc. These were once just possible in movies only but now it is very commonly seen. (IBM Cloud Education, 2020). According to Gartner a leading research company the business value of A.I in the market was $1.2 trillion in 2018 and is estimated to reach $3.9 trillion in 2022. (Gartner, Inc, 2018).

Machine learning which a part of A.I has been divided by the types of algorithms it follows. They are described as follows:

1.1.1 Supervised Learning

The process of supervised learning algorithm includes training the algorithm according to the data set with labels. Once done so the model is able to predict the label of a data with certain amount of approximation. Some examples of supervised learning algorithms are Naïve Bayes, Decision Tree, Linear Regression, etc. (Fumo, 2017)

1.1.2. Unsupervised Learning

For unsupervised learning there are no labelled data provided to the algorithm. Instead it is helps to identify different patterns made by the data and also helps to keep the data in groups. Example of unsupervised learning algorithms is k-means clustering. (Fumo, 2017)

1.1.3. Semi-supervised Learning

By the name itself it denotes that semi-supervised learning is an algorithm classified under the above two algorithms. The data set used for this algorithm is composed of less labelled data since it is expensive and more unlabelled data. As there is two types of data both above mentioned algorithms are used here. (Gupta, 2019)

1.1.4. Reinforcement Learning

This is an algorithm which is different form the above three. Reinforcement learning is an ongoing process in which the algorithm or machine keeps learning form the given environment. It keeps testing different things in the environment and keeps learning

from it. Example of reinforcement learning are Q-Learning, Deep adversarial Networks. (Fumo, 2017)

1.2. Algorithm used for this project

I have decided to go with k-nearest neighbour (KNN) algorithm which fall under supervised learning. It works based on finding labels with help of labels that already have been defined by measuring the distance between them. Euclidean distance is the most popular measure used for KNN. Also KNN has two sub categories within it. They are Classification and Regression. Classification is for distinct label in which label are classified in the group they belong to. Whereas, Regression is for continuous outcome which visualized on a graph. (scikit-learn developers, 2020)

KNN is called Non parametric and lazy learning algorithm. Non parametric because it can be used in all types of data distributions as it does not make any assumptions. And lazy learning algorithm because it does not learn in its training phase. The training data is rather used for testing. (Speck, 2017)

There are different distance metrics used in KNN for measuring distance between neighbours. Three of the most popular metrics are:

- Manhattan Distance

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Figure 1 Manhattan distance formula

- Euclidean Distance

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Figure 2 Euclidean distance formula

- Minkowski Distance

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^c\right)^{\frac{1}{c}}$$

Figure 3 Minkowski distance formula

For this project I am going to use KNN Classifier implementing Euclidean distance. (Fiori, 2020)

Figure 4 KNN definition

(Pradeepsingam, 2020)

1.3. Car Evaluation

Car is an expensive purchase for many people so obviously it is a tough decision to buy one. As it being expensive we all want the car to not give much trouble when driving. It is not something we just get rid of in case we get a bad quality. Hence there should be proper inspection done for selecting a well-conditioned car. There are many factors to be considered for a good car. Like mileage, ground clearance, luggage size, etc. Also car expense itself is an important factor to be considered.

This is why it is important for a car to be evaluated.

2. Background

Here are some popular algorithms used to evaluate a car.

2.1. Using Decision Tree Classifier

Decision Tree falls under the category of supervised learning as well. A decision tree includes ‘leaves’ which are the output values and ‘decision node’ which are the node that split data and helps in deciding its class. (randerson112358, 2019) The general idea of this algorithm is to split the whole data into smaller chunks based on their ‘Gain’. This process keeps on repeating until the leaves/ outputs are free of entropy. Which means until the group of output fall under the same category. (Chakure, 2019)

The beginning process are similar as all the packages are imported. Then the data set is loaded for the car evaluation and kept in data frames. Data is split for training and a decision tree model made. The model is trained with the help of data set and it is ready to make the predictions. (randerson112358, 2019)

2.2. Using Logistic Regression

Logistic regression is an algorithm derived from statistics that can classify data having two outcomes (Binary Logistic Regression). (Brownlee, 2016) However there are other approaches for this algorithm to work on different data types. Multiclass logistic regression is used for data having multiple outcomes. (Revision 6c02c0fd, 2017)

The mechanism behind this algorithm is the output is calculated using logistic function [1 / (1 + e^-x)] where x are the inputs and e is the Euler’s constant number. All of the output is generated between 0 and 1 with 0.5 being the output of midpoint. By this way we can make a prediction of the number is above or below 0.5 it belongs to class of 1 or 0 respectively. As car evaluation being multiclass data this process is carried out multiple times. (Brownlee, 2016)

2.3. Similar System Case Study

2.3.1. Car Evaluation using KNN and Decision Tree (Research Paper 1)

This is the research work done on car evaluation using KNN and Decision Tree algorithm. The accuracy score of both algorithms are then compared. Firstly the unnecessary attributes in dataset are removed if possible as it will make the processing faster with less attributes. Then any outliers or missing value is handled which in this case there are not any. Dataset contains non integer values which need to be converted into integer to be processed by the algorithm. Independent value determine the dependent value which is the output. Each independent value is assumed whether it gives a positive result or not.

Then the data is randomized and split into training, testing with 80:20 ratio. Then it is fitted into KNN model with 5 and 7 nearest neighbour with 1 and 2 power value first. This gives highest accuracy of 93.64161849710982% with 7 nearest neighbour and 2 power value.

KNN			
Splitting Percentage (Training % and Testing %)	50-50	60-40	80-20
n_neighbors	5	5	7
Power variable (p)	2	1	2
Testing accuracy	91.20370370370371%	92.61939218523878%	93.64161849710982%
Classification error rate	8.79629629629629%	7.380607814761222%	6.358381502890175%

Figure 5 KNN scores 1

Confusion Matrix	[[577 6 3 0] [46 167 3 3] [4 3 21 1] [1 2 4 23]]	[[459 4 2 0] [28 148 3 0] [3 5 15 0] [1 5 0 18]]	[[231 2 1 0] [14 73 0 0] [1 1 10 0] [0 2 1 10]]
precision	0.8465658156996925	0.8996017827059918	0.9270637898686679
recall	0.809500828387994	0.8040215824237817	0.8572060123784262
f1 score:	0.8247259934493818	0.8457758780971122	0.8875617588932807

Figure 6 KNN scores 2

Second, the data is fitted into decision tree model with max depth of 6. Which give highest accuracy of 93.64161849710982% as well.

Decision trees max_depth=6			
Splitting Percentage (Training % and Testing %)	**50-50**	**60-40**	**80-20**
Testing accuracy	92.24537037037037%	93.19826338639653%	93.64161849710982%
Classification error rate	7.754629629629628%	6.80173661360347%	6.358381502890175%

Figure 7 Decision Tree scores 1

Confusion Matrix	[[568 16 2 0] [16 179 21 3] [0 0 24 5] [0 4 0 26]]	[[449 14 2 0] [2 156 18 3] [0 0 19 4] [0 4 0 20]]	[[224 9 1 0] [1 78 7 1] [0 0 11 1] [0 2 0 11]]
precision	0.7868611018471237	0.7800093405644288	0.8242653161281193
recall	0.8702219370468116	0.8741300168982008	0.9041592985558503
f1 score	0.8178696121130331	0.815354746873236	0.8545574400650303

Figure 8 Decision Tree scores 2

In terms of accuracy score both perform the same but there was a difference in f1 score. F1 score is the score calculated by combining precision and recall. When comparing with f1 score KNN performs better than decision tree.

This research paper is available at: https://www.ukessays.com/essays/computerhttps://www.ukessays.com/essays/computer-science/car-evaluation-using-machine-learning.phpscience/car-evaluation-using-machine-learning.php

2.3.2. Car Evaluation using KNN and others (Research Paper 2)

This is a research paper about car evaluation using KNN, Logistic Regression and Decision Tree. Here too the metric used for comparison of algorithm is accuracy score. As above the dependencies, data are imported .Then the data is preprocessed in a similar way which is the missing values and outliers are taken care of. Then here data transformation is done which are normalization and aggregation. Normalization is for making the data scalable to features. Again the data is split into training and testing.

When the data was fitted into models the accuracy score for logistic regression was 66.4% which later became 71% after cross validation score. Secondly there was KNN which scored 90%. But this time the decision tree had better accuracy than KNN which was 96.5%.

This research paper is available at: https://www.ijraset.com/fileserve.php?FID=22012

2.3.3. Car Evaluation using KNN (Article 1)

This article too show the process and result of car evaluation but using KNN only. Rather, it is compared with taking different values of nearest neighbours. And again the metric used for comparison is accuracy score as it works best for classifying algorithms. The information of the data is checked for pre-processing. Then the independent attributes are identified and converted into integer for the algorithm to process. After that independent variable and target variable is selected and sent to training, testing split.

KNN model is trained and it scores accuracy of 94%. Also the misclassified data is calculated here by [count_misclassified = (y_test != y_pred).sum()] formula which calculates 32.

```
('Accuracy is ', 83.62235067437379, '% for K-Value:', 1)
('Accuracy is ', 80.15414258188824, '% for K-Value:', 2)
('Accuracy is ', 89.21001926782274, '% for K-Value:', 3)
('Accuracy is ', 88.82466281310212, '% for K-Value:', 4)
('Accuracy is ', 93.83429672447014, '% for K-Value:', 5)
('Accuracy is ', 92.8709055876686, '% for K-Value:', 6)
('Accuracy is ', 92.8709055876686, '% for K-Value:', 7)
('Accuracy is ', 89.78805394990366, '% for K-Value:', 8)
('Accuracy is ', 90.94412331406551, '% for K-Value:', 9)
('Accuracy is ', 88.82466281310212, '% for K-Value:', 10)
('Accuracy is ', 89.40269749518305, '% for K-Value:', 11)
('Accuracy is ', 88.6319845857418, '% for K-Value:', 12)
('Accuracy is ', 88.82466281310212, '% for K-Value:', 13)
('Accuracy is ', 89.01734104046243, '% for K-Value:', 14)
('Accuracy is ', 89.78805394990366, '% for K-Value:', 15)
('Accuracy is ', 88.6319845857418, '% for K-Value:', 16)
('Accuracy is ', 88.82466281310212, '% for K-Value:', 17)
('Accuracy is ', 88.4393063583815, '% for K-Value:', 18)
('Accuracy is ', 88.6319845857418, '% for K-Value:', 19)
('Accuracy is ', 88.6319845857418, '% for K-Value:', 20)
('Accuracy is ', 88.2466281310212, '% for K-Value:', 21)
('Accuracy is ', 89.01734104046243, '% for K-Value:', 22)
('Accuracy is ', 89.21001926782274, '% for K-Value:', 23)
('Accuracy is ', 89.01734104046243, '% for K-Value:', 24)
('Accuracy is ', 89.59537572254335, '% for K-Value:', 25)
```

Figure 9 KNN accuracy of 25 values of k

Hence among 25 values of k (nearest neighbours) 5 scores the highest.

This article is available at: https://bigishere.wordpress.com/2018/09/22/k-nearesthttps://bigishere.wordpress.com/2018/09/22/k-nearest-neighbour-knn/neighbour-knn/

3. Solution

From above we got the definition of KNN and here is how it actually works. When an unknown data is given this algorithm tries to find closest known data around it. The number of closest data i.e. nearest neighbor (K) can be customized accordingly. If the value of K is 1 it is too less for the algorithm to classify the data. Whereas, if the value of K is closer to the amount of data in dataset then the closer it gets to the mean of whole dataset. Either of them are not useful. One popular approach of finding a good value of K is the square root of total number of data in a dataset. (Band, 2020) After that there is a distance metric used to measure the distance between the nearest neighbours around its radius. I have used Euclidean Distance as mentioned above by which the algorithm identifies the input's nearest neighbours and classifies according to most labels that fall under its radius.

Here is an example of a simple classification process of KNN. Here the blue star is the unlabelled value. The graph consists two category of data i.e. the red ovals and the green rectangles. Hence to find out which category does the blue star fall under KNN is used.

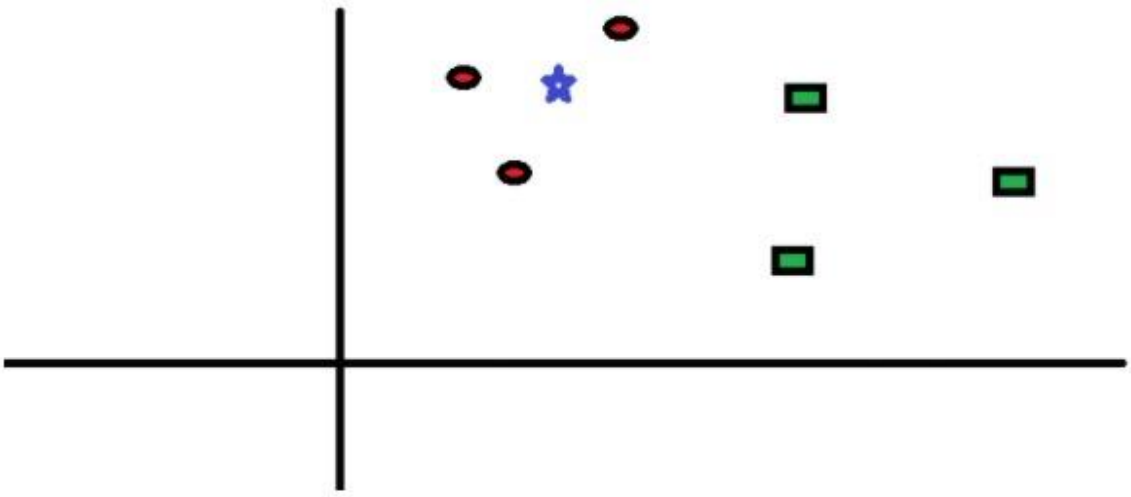

Figure 10 KNN example 1

Taking the value of K= 3 which means to find 3 nearest neighbour of the blue star. The nearest neighbour is found using the distance metric.

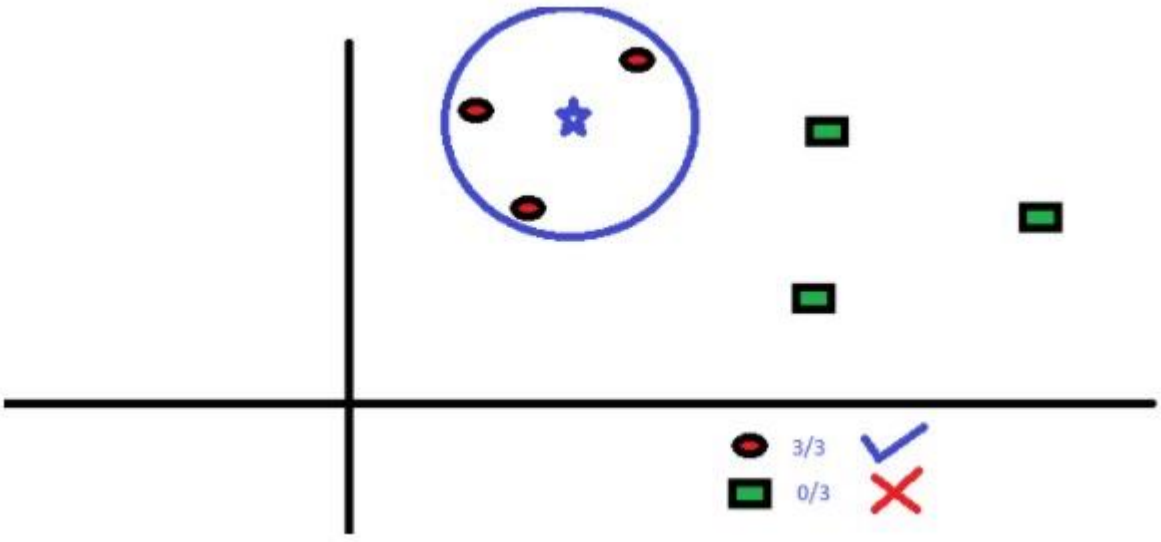

Figure 11 KNN example 2

Here it is calculated that the 3 nearest neighbours of blue star are 3 red ovals and 0 green rectangles. As the number of red ovals are higher than green rectangles in the radius of the blue star, it is classified as the red oval as well. (Srivastava, 2018)

3.1. Algorithm of the solution

Step1: The dataset is imported and inspected for handling null values and outliers.

Step2: The independent values and dependent values are uniquely identified.

Step3: The unique values are given an integer value so that it can be processed in the algorithm.

Step4: Data is split into Train and Test.

Step5: The split data is fitted into a KNN model.

Step6: Data is analyzed and visualized using graphs.

Step7: Predictions are ready to be made.

3.2. Pseudocode

```
data = import (Dataset/car.data) data.view()

for i = 0 to data.LastColumnIndex:

        for each data.column[i].UniqueValues:

                dataInteger[i][each] =  assign unique integer


X = dataInteger[0 to (LastColumnIndex-1)]

Y = dataInteger[LastColumnIndex]


Split (X, Y, test size= 0.2)

        Xtest = 0.2 * X

        Ytest = 0.2 * Y

        Xtrain   =   (1-0.2)   *   X

Ytrain = (1-0.2) * Y


classifier = KNeighborsClassifier()


classifier.fit(Xtrain, Ytrain)

classifier.showAccuracy() prediction

= classifier.predict(Xtest)

prediction.view()
```

3.3. Flow Chart

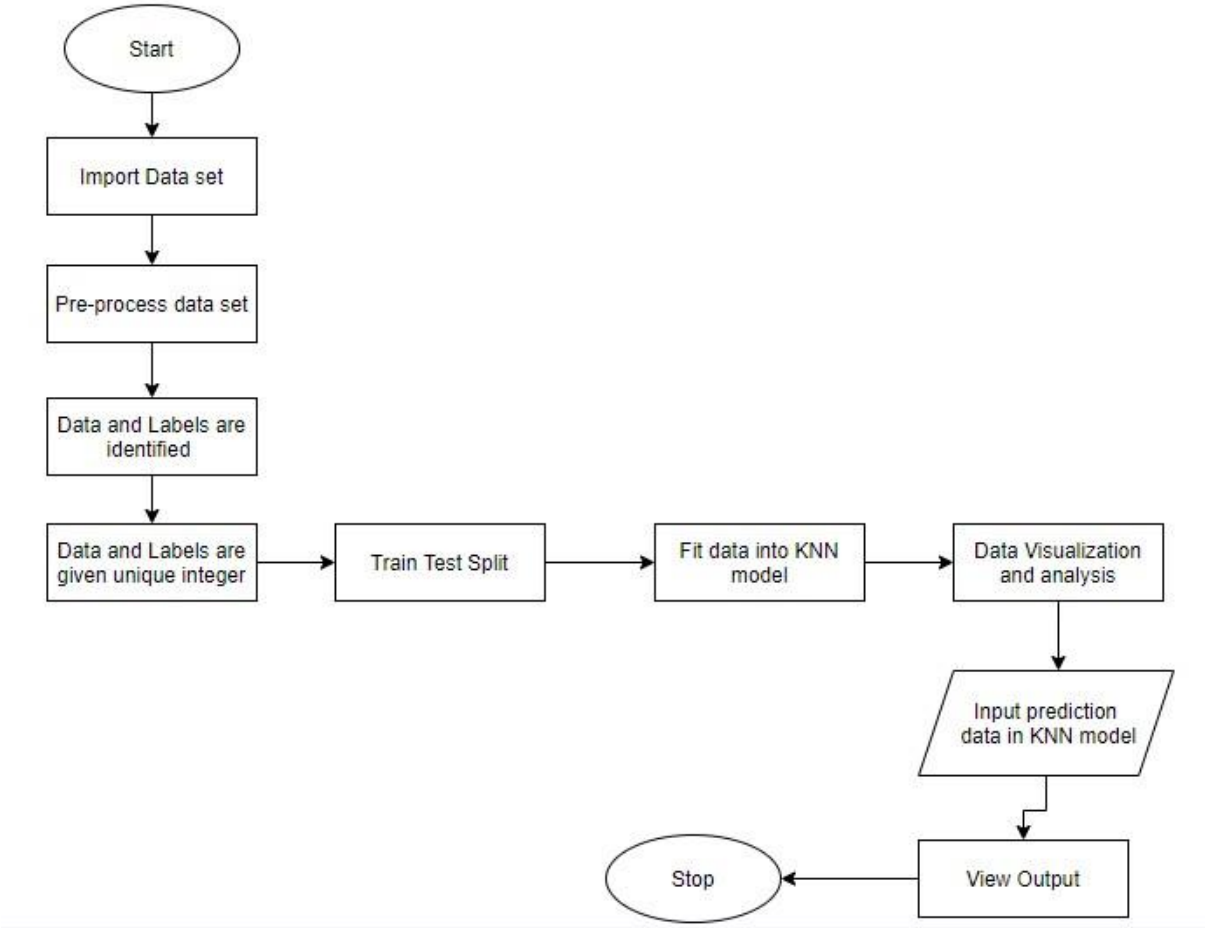

Figure 12 Flow Chart

3.4. Development Process

Here it is explained about the development process for the prototype of the chosen A.I. topic.

3.4.1. Technologies and Platform

3.4.1.1. Python Programming Language

Python is a high level language for general purpose programming released in 1991. It has a very good library support for Web development, Data visualization, Machine learning etc. And also one of the most popular language in current scenario of programming.

3.4.1.2. Anaconda Distribution

Anaconda is an environment for data science and Artificial Intelligence. It is a distribution of Python and R language with all the necessary libraries for machine learning. (Waseem, 2019)

3.4.1.3. Jupyter Notebook

Jupyter Notebook is web based environment which runs Python. It is known for data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, etc.

3.4.1.4. Python Libraries

Pandas – It is a library for importing and manipulating data which is helpful for data preprocessing phase.

Numpy – This library is used to handle arrays in an easy way and also helps with different mathematical functions.

Matplotlib – This library focuses of visualizing data using graphs.

Sklearn – It is a very popular library for machine learning purpose as it consists many algorithm models.

seaborn – This library focuses on visualizing data as well but rather in a different graphs than matpolotlib.

warnings – This library is used to ignore the warnings given by python code.

3.4.1.5. Dataset

The data set being used here is the UCI Car Evaluation Database created by Marko Bohanec and Blaz Zupan. It has six attributes divided into three different concepts. The six attributes are buying price, maintenance price which fall under **PRICE** concepts. Under **TECH** concept falls safety and another sub concept **COMFORT**. And under COMFORT falls doors, person and luggage boot. Here are the information about the attributes:

Table 1 Attribute Information

Attribute Names	Types			
Class Value(Label)	unacc	acc	good	vgood
Buying	vhigh	high	med	low
Maint Price	vhigh	high	med	low
Doors	2	3	4	5 or more
Persons	2	4	more	
Luggage Boot	small	med	big	
Safety	low	med	high	

```
[6]: import pandas as pd

[7]: data = pd.read_csv("car.data")

[8]: data.head()
```

[8]:

	vhigh	vhigh.1	2	2.1	small	low	unacc
0	vhigh	vhigh	2	2	small	med	unacc
1	vhigh	vhigh	2	2	small	high	unacc
2	vhigh	vhigh	2	2	med	low	unacc
3	vhigh	vhigh	2	2	med	med	unacc
4	vhigh	vhigh	2	2	med	high	unacc

Figure 13 Dataset

Dataset link: https://archive.ics.uci.edu/ml/datasets/car+evaluation

3.4.2. Data Cleaning and Pre-processing

Once the data set has been imported it was missing columns. So each of the columns were given a name using pandas column function.

```
[91]: data = pd.read_csv("car.data")
      data.columns = ['Buying', 'Maintainance_Cost', 'No_of_Doors', 'Person_Capacity','Luggage_Size','Safety','Evaluation']

[92]: data.head()
```

[92]:

	Buying	Maintainance_Cost	No_of_Doors	Person_Capacity	Luggage_Size	Safety	Evaluation
0	vhigh	vhigh	2	2	small	med	unacc
1	vhigh	vhigh	2	2	small	high	unacc
2	vhigh	vhigh	2	2	med	low	unacc
3	vhigh	vhigh	2	2	med	med	unacc
4	vhigh	vhigh	2	2	med	high	unacc

Figure 14 Adding columns

All of the data were in string which is not computable by the algorithm hence, each of the unique data of attributes are given an integer value. This is process is called “Integer Encoding” which is simply done by replace keyword. Now all of the independent columns are separated out. They are Buying, Maintainance_Cost, No_of_Doors, Person_Capacity, Luggage_Size and Safety. Now that the dependent column is Evaluation it is kept separately as it is used as a label for this data set. This separating of independent and dependent columns has been done by pandas loc function.

3.4.3. Train Test Split and KNN Model Training

x_train, x_test, y_train, y_test variables are used to train test split and store the data. x_train, x_test are the independent data and y_train, y_test are the labels. The amount separated for train size is 90% and rest 10% is for test size. This is possible through train_test_split form sklearn.model_selection.

Using a for loop the number of nearest neighbour is set form 1 to 10 in KNN model and for each time its accuracy is checked. KNeighborsClassifier form sklearn has been used as the KNN model. The number having the highest accuracy score is selected and used further.

Predictions are made with the data separated for test. Also the algorithm can be tested using customized data.

3.4.4. Achieved results

A graph with number of nearest neighbour and accuracy score is plotted in a graph using matplotlib.

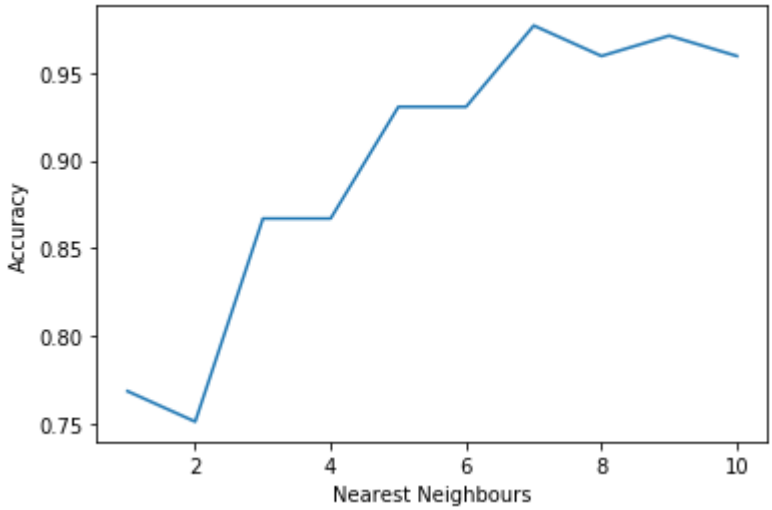

Figure 15 Accuracy score graph

After all the predictions have been made from the test data a seaborn countplot graph is plotted. It represents the number of acceptable, unacceptable, very good and good cars in the testing data.

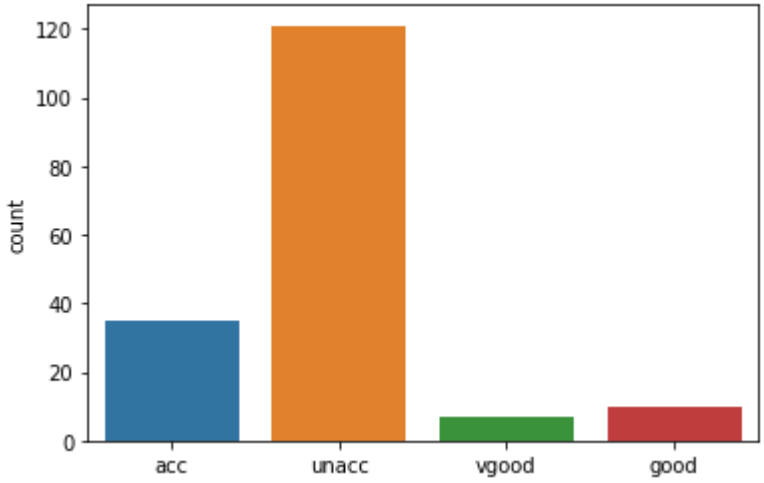

Figure 16 Evaluation count graph

Also, the model can be used to predict external input data to see under what category of evaluation it falls. All of the data have to be inputted using integer value looking at the table of maximum and minimum value.

```
[107]: x_test.describe()
```

[107]:

	Buying	Maintainance_Cost	No_of_Doors	Person_Capacity	Luggage_Size	Safety
count	173.000000	173.000000	173.000000	173.000000	173.000000	173.000000
mean	2.653179	2.554913	2.537572	1.942197	2.028902	1.942197
std	1.164371	1.143049	1.138659	0.826248	0.810023	0.819181
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	2.000000	2.000000	2.000000	1.000000	1.000000	1.000000
50%	3.000000	3.000000	2.000000	2.000000	2.000000	2.000000
75%	4.000000	4.000000	4.000000	3.000000	3.000000	3.000000
max	4.000000	4.000000	4.000000	3.000000	3.000000	3.000000

```
[153]: #Predicting own custom value
prediction1 = model.predict([['1','1','3','3','3','1']])
j = (prediction1[0])
print (labels[j-1])
```

```
unacc
```

Figure 17 Custom data prediction

4. Conclusion

To conclude, this project is based on car evaluation through KNN classifier. And the metric system being used here is Euclidian distance. Some common algorithms used for car evaluation are decision tree and logistic regression. There is a possibility that decision tree might outperform KNN. Talking about the most suitable number of nearest neighbour it seems 5 to 8 happens to be giving higher accuracies.

4.1. In Real World Scenario

Talking about KNN's use in real world scenario first of all it is highly accurate algorithm hence it is somewhat reliable to make big decisions like car evaluation. It would specially be applicable in second-hand market as those types of cars are the once that people need to be more concern about the condition. However, even buying a car for the first time the attributes can be changed with the new and essential car features like mileage, engine CC, torque, etc.

4.2. Further Work

It is also possible modify to the KNN for improving accuracy. Also KNN can be used alongside another Machine Learning algorithm to add more functioning. Like it can be combined with user based collaborative filtering so it can recommend the evaluated car according to user's information.

And a UI can be made for filtering input data more effectively. Rather than making it tedious with coding it can be made attractive and easy to use as well.

Also for better visualization of data a scattered plot with different area of evaluation and data falling under each category can be made.

5. References

Band, A. (2020) *How to find the optimal value of K in KNN? | by Amey Band | Towards Data Science* [Online]. Available from: https://towardsdatascience.com/how-to-find-the-optimal-value-of-k-in-knn-35d936e554eb#:~:text=The%20optimal%20K%20value%20usually,be%20aware%20of%20the%20outliers. [Accessed 4 February 2021].

Brownlee, J. (2016) *Logistic Regression Tutorial for Machine Learning* [Online]. Available from: https://machinelearningmastery.com/logistic-regression-tutorial-for-machine-learning/ [Accessed 8 January 2021].

Chakure, A. (2019) *Decision Tree Classification. A Decision Tree is a simple… | by Afroz Chakure | The Startup | Medium* [Online]. Available from: https://medium.com/swlh/decision-tree-classification-de64fc4d5aac [Accessed 8 January 2021].

Fiori, L. (2020) *Distance metrics and K-Nearest Neighbor (KNN) | by Luigi Fiori | Medium* [Online]. Available from: https://medium.com/@luigi.fiori.lf0303/distance-metrics-and-k-nearest-neighbor-knn-1b840969c0f4 [Accessed 4 Feburary 2021].

Fumo, D. (2017) *Types of Machine Learning Algorithms You Should Know | by David Fumo | Towards Data Science* [Online]. Available from: https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861 [Accessed 6 January 2021].

Gartner, Inc. (2018) *Gartner Says Global Artificial Intelligence Business Value to Reach $1.2 Trillion in 2018* [Online]. Available from: https://www.gartner.com/en/newsroom/press-releases/2018-04-25-gartner-says-global-artificial-intelligence-business-value-to-reach-1-point-2-trillion-in-2018 [Accessed 4 January 2021].

Graff, C. & Dua, D. (2019) *UCI Machine Learning Repository: Car Evaluation Data Set* [Online]. Available from: https://archive.ics.uci.edu/ml/datasets/car+evaluation [Accessed 8th January 2021].

Gupta, A. (2019) *ML | Semi-Supervised Learning - GeeksforGeeks* [Online]. Available from: https://www.geeksforgeeks.org/ml-semi-supervised-learning/ [Accessed 7 January 2021].

IBM Cloud Education. (2020) *What is Artificial Intelligence (AI)? | IBM* [Online]. Available from: https://www.ibm.com/cloud/learn/what-is-artificial-intelligence [Accessed 4 January 2021].

Pradeepsingam. (2020) *KNN(K_Nearest Neighbors). KNN is a supervised machine learning… | by Pradeepsingam | Analytics Vidhya | Medium* [Online]. Available from: https://medium.com/analytics-vidhya/knn-k-nearest-neighbors-1add1b5d6eb2 [Accessed 7 January 2021].

randerson112358. (2019) *Car Classification Using Python & Machine Learning | by randerson112358 | Medium* [Online]. Available from: https://randerson112358.medium.com/car-classification-89ad60204acf [Accessed 8 January 2021].

Revision 6c02c0fd. (2017) *Logistic Regression — ML Glossary documentation* [Online]. Available from: https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html#multiclass-logistic-regression [Accessed 8 January 2021].

scikit-learn developers. (2020) [Online]. Available from: https://scikit-learn.org/stable/modules/neighbors.html# [Accessed 7 January 2021].

Speck, M. (2017) *What is K-Nearest Neighbors? In data science, predictions can take… | by Matthew Speck | Medium* [Online]. Available from: https://medium.com/@mjspeck/what-is-k-nearest-neighbors-c9b4cdf9f35c#:~:text=KNN%20is%20a%20non%2Dparametric,that%20are%20close%20to%20it.&text=Non%2Dparametric%20means%20that%20KNN,the%20data%20it%20is%20modeling. [Accessed 2 Feburary 2021].

Srivastava, T. (2018) *K Nearest Neighbor | KNN Algorithm | KNN in Python & R* [Online]. Available from: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/ [Accessed 10 January 2021].

Waseem, M. (2019) *Python Anaconda Tutorial | Getting Started With Anaconda | Edureka* [Online]. Available from: https://www.edureka.co/blog/python-anaconda-tutorial/ [Accessed 2 February 2021].