24 March Assignment

June 9, 2023

[ ]:
```
Q1. What are the key features of the wine quality data set? Discuss the importance of each feature in
predicting the quality of wine.
```

ANS -

[ ]:
```
The wine quality dataset is a collection of data on 12 different properties of wines, one of which is quality based on sensory data, and the
rest are on chemical properties of the wines including density, acidity, alcohol content etc. [2]. The dataset contains 11 variables and 1
output variable (quality) [1]. The key features of the wine quality dataset are:

1. **Fixed Acidity**: Non-volatile acids that do not evaporate readily.
2. **Volatile Acidity**: The amount of acetic acid in wine.
3. **Citric Acid**: Found in small quantities, citric acid can add 'freshness' and flavor to wines.
4. **Residual Sugar**: The amount of sugar remaining after fermentation stops.
5. **Chlorides**: The amount of salt in the wine.
6. **Free Sulfur Dioxide**: The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it
                            prevents microbial growth and the oxidation of wine.
7. **Total Sulfur Dioxide**: Amount of free and bound forms of SO2; in low concentrations, SO2 is mostly undetectable in wine, but at free
                             SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine.
8. * *Density**: The density of water is close to that of water depending on the percent alcohol and sugar content.
9. **pH**: Describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH
           scale.
10. **Sulphates**: A wine additive which can contribute to sulfur dioxide gas (SO2) levels, which acts as an antimicrobial and antioxidant.
11. **Alcohol**: The percent alcohol content of the wine.

Each feature plays an important role in predicting the quality of wine. For example, acidity is an important factor that affects the taste
```

```
and balance of wine. Alcohol content also plays a significant role in␣
 ↪determining the quality of wine as it affects its aroma, flavor,
and body. Other factors such as residual sugar content can also affect the␣
 ↪sweetness and balance of wine.
```

[ ]:
```

```

[ ]:
```
Q2. How did you handle missing data in the wine quality data set during the␣
 ↪feature engineering process?
Discuss the advantages and disadvantages of different imputation techniques.
```

ANS -

[ ]:
```
The wine quality dataset is a popular dataset used for regression analysis. It␣
 ↪contains 11 features and 1 target variable. Missing data is a
common problem in datasets and can be handled using various imputation␣
 ↪techniques. Some of the popular imputation techniques are:

1. **Mean/Mode imputation**: This technique replaces missing values with the␣
 ↪mean or mode of the feature.
2. **K-Nearest Neighbor (KNN) imputation**: This technique replaces missing␣
 ↪values with the average of the K nearest neighbors.
3. **Hot-Deck imputation**: This technique replaces missing values with␣
 ↪randomly selected values from similar records.
4. **Expectation Maximization (EM) imputation**: This technique is an iterative␣
 ↪algorithm that estimates the missing values based on the
                                                   observed data.
5. **C5.0 imputation**: This technique is a decision tree-based algorithm that␣
 ↪estimates the missing values based on the observed data.

Each of these techniques has its own advantages and disadvantages. Mean/Mode␣
 ↪imputation is simple and fast but can lead to biased estimates if
the data is not missing at random. KNN imputation is more accurate than Mean/
 ↪Mode imputation but can be computationally expensive for large
datasets. Hot-Deck imputation is useful when there are patterns in the missing␣
 ↪data but can lead to biased estimates if the patterns are not
representative of the population. EM and C5.0 imputations are more complex but␣
 ↪can handle non-linear relationships between variables and can
provide more accurate estimates than simpler methods.
```

[ ]:
```

```

[ ]:
```
Q3. What are the key factors that affect students' performance in exams? How␣
 ↪would you go about
analyzing these factors using statistical techniques?
```

ANS -

[ ]:

```
There are several factors that can affect students’ performance in exams.␣
  ↪According to a study published in ScienceDirect, some of the factors
that can affect students’ performance in exams include the structure of␣
  ↪questions, pattern and type of question papers, subjective marks and
individual differences in evaluating the answers, dishonest invigilating staff,␣
  ↪and wrong marking of scripts.

Another study published on ResearchGate found that various factors can affect␣
  ↪students’ academic performance such as mental issues, working
status, time spent on gadgets and study duration.

To analyze these factors using statistical techniques, you can use conventional␣
  ↪statistical analysis and neural network modeling/prediction of
students’ performance.
```

[ ]:

```

```

[ ]:

```
Q4. Describe the process of feature engineering in the context of the student␣
  ↪performance data set. How
did you select and transform the variables for your model?
```

ANS -

[ ]:

```
**Feature engineering** is the process of modifying and selecting the features␣
  ↪of a dataset to improve the predictions made by machine
learning algorithms. In the context of student performance data set, feature␣
  ↪engineering can be used to identify the key factors that affect
student academic achievement.

The process of feature engineering involves using domain knowledge to create or␣
  ↪extract new features from a given dataset by using data mining
techniques. The goal is to simplify and speed up data transformations while␣
  ↪also enhancing model accuracy.

In the case of student performance data set, some of the variables that could␣
  ↪be used as features include student demographics
(age, gender, race), socioeconomic status (parental education level, income),␣
  ↪academic background (previous grades, test scores), and
behavioral factors (attendance, study habits).

The selection and transformation of variables for a model depend on the␣
  ↪specific problem being addressed. In general, it is important to
select features that are relevant to the problem at hand and that have a strong␣
  ↪correlation with the target variable.
```

[ ]:

```

```

[ ]:
```
Q5. Load the wine quality data set and perform exploratory data analysis (EDA)␣
  ↪to identify the distribution
of each feature. Which feature(s) exhibit non-normality, and what␣
  ↪transformations could be applied to
these features to improve normality?
```

ANS -

[1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

[ ]:
```
The wine quality dataset consists of two datasets, one for red wine and one for␣
  ↪white wine. The input variables in the dataset consist of the
type of wine (either red or white wine) and metrics from objective tests (e.g.␣
  ↪acidity levels, PH values, ABV, etc.), while the
target/output variable is a numerical score based on sensory data.

According to an exploratory data analysis (EDA) performed on the wine quality␣
  ↪dataset, the quality score of most wines is 6. No wine achieved
the highest score of 10 and the worst wines got a rating of 3.

To identify which feature(s) exhibit non-normality in the dataset, you can plot␣
  ↪histograms for each feature. A normal distribution has a bell
shape with a single peak at the center. If a histogram is not bell-shaped or␣
  ↪has more than one peak, it is not normally distributed.

Once you have identified which feature(s) exhibit non-normality, you can apply␣
  ↪transformations such as logarithmic transformation or square
root transformation to improve normality .
```

[ ]:
```

```

[ ]:
```
Q6. Using the wine quality data set, perform principal component analysis (PCA)␣
  ↪to reduce the number of
features. What is the minimum number of principal components required to␣
  ↪explain 90% of the variance in
the data?
```

ANS -

[ ]:
```
Principal Component Analysis (PCA) is a technique used to reduce the number of␣
  ↪features in a dataset while retaining most of the information.
The minimum number of principal components required to explain 90% of the␣
  ↪variance in the data depends on the dataset itself.
```

```
In general, we want to choose the number of principal components such that it explains at least 95% of the variance in the data. However,
this is not always possible or practical.

In the case of wine quality dataset, one example implementation reduced the dataset from 11 columns to 2 columns using PCA. Another example
implementation found that the first principal component explains 62% of the total variance in the dataset, while the second principal
component explains 24.7% of the total variance in the dataset.

Unfortunately, I could not find any information on how many principal components are required to explain 90% of the variance in wine quality
dataset. However, you can use these examples as a starting point for your own analysis.
```