**Question 3**

**Answer：**

*Colab Link:*

*https://colab.research.google.com/drive/1WKCtE2ZBHUtkNfpi8mxUvQPpys\_1n9o\_?usp=sharing*

White wine is known to improve heart health and may prevent heart diseases.

The dataset provided for this ECA is a dataset for white wines which contains 11 chemical aspects of the sample, and 1 score (0-10) for measuring wine quality.

The aim of this question is to take this white wine data, perform some data analysis, exploring what factors influence the quality of white wine in (a) and (b), and then in (d), (c) and (e) build a neural network model to predict the quality of white wine by 11 chemical aspects.

**Step 1 for importing data and overview – (a) & (b)**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

sns.set(style="darkgrid")

pd.set\_option('precision',3)

df = pd.read\_csv("/content/winequality-white.csv",sep = ';')

# number of rows & columns in the dataset

df.shape

Output: (4898, 12)

# first 5 rows of the dataset

df.head()

Output:

图形用户界面

描述已自动生成

df.info()

Output:

一些文字和图片的手机截图

中度可信度描述已自动生成

This dataset has a total of 4898 samples of white wines, included 11 chemical characteristics and 1 wine quality score.

The data are complete, with no missing values.

# View the distribution of quality

df["quality"].value\_counts()

Output:

表格

描述已自动生成

can be found that the quality of white wines mostly concentrated in the 5-7 point range, with almost no 9 point.

# View the distribution of each variable

# bar chart

df.describe()

colnm = df.columns.tolist()

plt.figure(figsize = (10, 8))

for i in range(12):

    plt.subplot(4,3,i+1)

    df[colnm[i]].hist(bins = 100,)

    plt.xlabel(colnm[i],fontsize = 12)

    plt.ylabel('Frequency')

plt.tight\_layout()

print('\nFigure 1: Bar chart')

The following by calculating the mean, variance, minimum maximum, etc., with the diagram to show how data distributed.

df.describe()

output:

电脑萤幕的截图

描述已自动生成

# Boxplots

colnm = df.columns.tolist()

fig = plt.figure(figsize = (10, 6))

for i in range(12):

    plt.subplot(3,4,i+1)

    sns.boxplot(df[colnm[i]], orient="v", width = 0.5, )

    plt.ylabel(colnm[i],fontsize = 10)

print('\nFigure 2: Boxplots')

Plotting the box line shows that all of the variables except alcohol have outliers and generally have a larger number of larger outliers.

(1) Of these, the box for chloride was the flattest, indicating that 50% of the data fluctuated around the mean, but there were a large number of larger outliers that deviated from the mean distribution.

(2) In contrast, the outliers for residual sugars, total sulphur dioxide, density, and quality scoring were relatively few in comparison to the other characteristics and were generally much larger than the mean.

(3) fixed citric acid and pH ad a large number of large outliers, but also had some smaller outliers at the same time.

(4) volatile acid, chloride, free sulphur dioxide and sulphates generally had large outliers, with only chloride having a few small outliers.

(5) There were no outliers for alcohol

below for the chemical aspects most associated with quality

corr\_matrix = df.corr()

corr\_matrix["quality"].sort\_values(ascending= False)

Output:

手机屏幕截图

描述已自动生成

# cochemical Properties and Wine Quality by Boxplot

plt.figure(figsize=(10,8))

colnm = df.columns.tolist()[:11] + ['total acid']

for i in range(11):

    ax = plt.subplot(4,3,i+1)

    sns.boxplot(data= df ,x ='quality', y= colnm[i], width = 0.6)

    plt.ylabel(colnm[i],fontsize = 12)

plt.tight\_layout()

print("\nFigure 3: cochemical Properties and Wine Quality by Boxplot")

Output:

日历

描述已自动生成

See the relationship between the quality of red wines and their chemical characteristics.

In summary -

Positive correlations: citric acid, sulphates, alcohol, total acid;

The highest correlation is between alcohol and quality.

Negative correlations: volatile acidity, density, pH;

A good quality white wine has lower volatile acids, density, and pH

Residual sugars, chloride and sulphur dioxide do not seem to have a significant impact on the quality of white wine.

# Correlation Matrix Heatmap

f, ax = plt.subplots(figsize=(10, 6))

corr = df.corr()

hm = sns.heatmap(round(corr,2), annot=True, ax=ax, cmap="coolwarm",fmt='.2f',

                 linewidths=.05)

f.subplots\_adjust(top=0.93)

t= f.suptitle('WhiteWine Attributes Correlation Heatmap', fontsize=12)

Output:

图表

描述已自动生成

The gradients in the heat map vary according to the strength of the correlation, and spot potential attributes that are strongly correlated with each other, lighter colours indicate a weaker linear correlation between the variables.

In summary -

The quality scores of white wines had a strong linear relationship with alcohol and density. There was a positive correlation between quality score and alcohol with a strength of 0.44, the higher the alcohol content, the higher the quality score; And a negative linear correlation with density of -0.31, the higher the content, the lower the quality score of the white wine.

**Step 2 for Neural network Model building and testing – (c) & (d) & (e)**

from pandas import Series,DataFrame

from sklearn import svm

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import accuracy\_score

import numpy as np

import matplotlib.pyplot as plt

import keras

from keras.datasets import fashion\_mnist

from keras.models import Sequential

from keras.layers import Dense, Dropout

from tensorflow.keras.optimizers import RMSprop

from tensorflow.keras.optimizers import Adam

# Variable (11 chemical aspects of white wine)

x = DataFrame(df.drop("quality",axis=1))

# Target (quality scoring 1~10)

y = DataFrame(df["quality"])

# Divided into training data and test data respectively.

x\_train,x\_test,y\_train,y\_test = train\_test\_split(x,y,test\_size=0.2)

# Shape the data

x\_train = x\_train.astype(np.float)

x\_test = x\_test.astype(np.float)

y\_train = keras.utils.np\_utils.to\_categorical(y\_train,10)

y\_test = keras.utils.np\_utils.to\_categorical(y\_test,10)

*Overview:*

|  |  |
| --- | --- |
| Hidden layer | 3 |
| Number of units in Hidden layer | 50 |
| Output size | 10 |
| Function | relu |
| Error function | mean\_squared\_error |
| Optimizers | RMSprop() |
| Number of study | 1000 |

model = Sequential()

model.add(Dense(50, activation='relu', input\_shape=(11,)))

model.add(Dropout(0.2))

model.add(Dense(50, activation='relu', input\_shape=(11,)))

model.add(Dropout(0.2))

model.add(Dense(50, activation='relu', input\_shape=(11,)))

model.add(Dropout(0.2))

model.add(Dense(10, activation='softmax'))

model.summary()

# Neural network Model

model.compile(loss='mean\_squared\_error',optimizer=RMSprop(),metrics=['accuracy'])

# Neural network studying

history = model.fit(x\_train, y\_train,batch\_size=200,epochs=1000,verbose=1,validation\_data=(x\_test, y\_test))

# model evaluation

score = model.evaluate(x\_test,y\_test,verbose=1)

print("\n")

print("Test loss:",score[0])

print("Test accuracy:",score[1])

def plot\_history(history):

    print(history.history.keys())

Output:

图片包含 文本

描述已自动生成

# summarize history for accuracy

plt.plot(history.history['accuracy'])

plt.plot(history.history['val\_accuracy'])

plt.title('Model Accuracy')

plt.xlabel('Epoch')

plt.ylabel('Accuracy')

plt.legend(['acc', 'val\_acc'], loc='lower right')

plt.show()

# summarize history for loss

plt.plot(history.history['loss'])

plt.plot(history.history['val\_loss'])

plt.title('Model Loss')

plt.ylabel('Loss')

plt.xlabel('Epoch')

plt.legend(['acc', 'val\_acc'], loc='upper right')

plt.tight\_layout()

plt.show()

# Study history diagram

plot\_history(history)

Output:

图表

描述已自动生成

**Testing**

sample = [7.9, 0.35, 0.46, 5, 0.078, 15, 37, 0.9973, 3.35, 0.86, 12.8]

print(sample)

sample = np.array(sample)

predictions = (model.predict(sample.reshape(1,-1),batch\_size=1,verbose=0) > 0.5).astype("int32")

print(np.argmax(model.predict(sample.reshape(1,-1)), axis=1))

Output:

图片包含 日历

描述已自动生成

The predict white wine quality scoring for given simple is 5

The model is not very accurate, accuracy rate only 55.8%, as the two curves do not yet overlap according to the learning progress graph.

Further actions I would like to take to improve accuracy of model are:

1. Increase the amount of data learned, increase Epoch to 2000, to see whether 2 curves can overlap.
2. Cross Validation: try to leave a sample on which you do not train the model and test the model on this sample before finalizing the model. This method helps us to achieve more generalized relationships.
3. Algorithm Tuning: for example, turning n\_estimators. The default value of n\_estimators is 100. So can start n\_estimators at 50 and add 10 each time up to 300 and run it to see what happens to the average score.
4. Feature selection: finding which variable best explains the relationship between the target variable and each of the independent variables. For example, specialist knowledge of white wine quality identification combined with the selection of which variable has a greater influence on the target variable based on the previous Boxplot plot.
5. Multiple algorithms: when there is more data, if the model complexity is not enough, no matter adding amount of data and training to the model, it would not work. So at that point increase the complexity of the model.

图示

描述已自动生成

**And**

sample = [7, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]

print(sample)

sample = np.array(sample)

predictions = (model.predict(sample.reshape(1,-1),batch\_size=1,verbose=0) > 0.5).astype("int32")

print(np.argmax(model.predict(sample.reshape(1,-1)), axis=1))

Output：

手机屏幕截图

中度可信度描述已自动生成

If data for other beverages (e.g. water PH7 density 1) is input, the machine does not make a judgement on it. The result here still shows quality 4

As an improvement it is hoped to include various liquids including water and juice in the learning data in the future. Also, the machine is not a substitute for a human being to do quality checks and the accuracy of the data and result need to be judged by a human being.