```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import sklearn as skl
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
from sklearn.model_selection import cross_val_score
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score
```

The dataset 'heart', is used to predict the presence or absence of heart disease based on several input features. The following is description of each feature.

• Sex: male or female (0 is female and 1 is male)
• Age: Age of the patient;
• cp: Chest Pain type: Between 0 to 3
• trestbps: Resting Blood Pressure (in mm Hg) : The blood pressure of the patient when they are at rest.
• Chol: Serum Cholesterol (in mg/dl):Blood cholesterol levels
• fbs: Fasting Blood Sugar. Whether the person has high fasting blood sugar levels >120mg/dl (1=T or 0=F)
• restecg: Resting Electrocardiographic Results (0=Normal, 1=ST Wave Abnormality, 2=left ventricular hypertrophy)
• thalach: Maximum Heart Rate Achieved: The maximum heart rate achieved during exercise stress testing.
• exang: Exercise Induced Angina. Whether or not the person experienced angina (chest pain) during exercise (1 = Yes, 0 = No).
• oldpeak: Depression Induced by Exercise Relative to Rest
• slope: Slope of the Peak Exercise ST Segment (0=Upsloping, 1=Flat, 2=Downsloping)
• ca: Number of Major Vessels Colored by Fluoroscopy (The number of major blood vessels 0 to 4 that are colored by fluoroscopy)
• thal: Thalassemia A blood disorder that affects the shape of red blood cells. The values typically are (3=Normal, 6=Fixed defect, 7=Reversible defect. If any other value don't worry about it)
• Target: (0= No heart disease (healthy) 1= Heart disease present (diseased))

```
data = pd.read_csv('')
data
```

1. Identify what are the data types of each column (Nominal vs Continous). (14 points)

Write your answer here
. • Sex: Male or Female: ?
• Age: ?
• cp: ?
• trestbps: ?
• Chol: ?
• fbs: ?
• restecg: ?
• thalach: ?
• exang: ?
• oldpeak: ?
• slope: ?
• ca: ? (number of blood vessels)
• thal: ?
• Target: ?

2. Create a correlation heatmap of the data. Set annot to true to check the values. (4 points)

```
#Enter your code here.
```

3. According to the heatmap, do you think some of the feature should not be count in the logistic regression? Which ones? How do you know? (Hint: Find the corrrelated pairs first and choose items with strong (around 0.4) correlation). (10 points)

```
#Enter your code here
```

Enter your answer here

4. How many null values are there in each feature?(4 points)

```
#Enter your code here.
```

5. Show the histogram of each factors and comment on the distribution of some of them. (5 points)

```
#Enter your code here
```

6. Split the data set into X (features) and y. The features to use are age, sex, cp, thalach, slope and restecg. (6 points)

```
#Enter your code here
#Hint: features=[?]
```

KNN

Finding the best value for K

7. Plot K Neighbors Classifier Scores for different K values:
   Perform 5-fold cross-validation to evaluate the model's performance for upto 50 neighbors. Then plot the results. Which K value gives the highest cross-validation score? (10 points)

```
#Modify this code
#Hint: Store neighbors in a list and iterate through each. Then plot on the same graph
```

8. Find the Average cross validation score for 11 neighbours. Hint use .mean() for average. (6 points)

```
#Enter your code here
#Hint: Use KNeighborsClassifier
```

9. Split the dataset into X_train, X_test, y_train and y_test. Split using 70% training and 30% test set. Use random state = 5. Then train the classifier. Then predict y using this. (6 points)

```
#Enter your code here
```

10. Find the accuracy, precision and recall upto 4 decimal places. Explain what each of these mean. (15 points).

```
#Enter your code here
# Accuracy


# Precision


# Recall
```

Explain here.

Real-Life Example 1

(20 points, 10 points each)

11. Given a 50 year old male who has chest pain type 3, maximum heart rate 222, and UpSloping ST Segment (slope=0) and resting ecg having LVT (restecg=2), predict if the person has a chance of developing heart disease.

```
#Enter your code here
#Hint: create a tuple to store these values, convert to an array and reshape so you have one row and n columns. Then predict on that.

# Hint if predicted value is 1 then what does it mean?
```

Real-Life Example 2

12. Given a 63 year old female who has chest pain type 1, maximum heart rate 100, and Downsloping ST Segment (slope=2) and a normal resting ecg (restecg=0), predict if the person has a chance of developing heart disease.

```
#Enter your code here
#Hint: create a tuple to store these values, convert to an array and reshape so you have one row and n columns. Then predict on that.

# Hint if predicted value is 1 then what does it mean?
```