CMPE 257
Lab 1

Predict Cardio Vascular Disease

Due date: October 7th

Competition Link

Task 1 (Kaggle dataset)

Submission - Kaggle & canvas

Dataset – cardio-train.csv and cardio-test.csv

Data Visualization: In this step, you will analyze the datasets and try to find a relationship between the attributes. This is the most important step in any machine learning problem.

About data:

The data will be found in cardio-train.csv. There are 12 variables:

1. Age(days)
2. Height(cm)
3. Weight(kg)
4. Gender
5. Systolic blood pressure
6. Diastolic blood pressure
7. Cholesterol
8. Glucose
9. Smoking
10. Alcohol intake
11. Physical activity
12. Presence(1) or absence(0) of cardiovascular disease (Target Variable)

Each following row contains the information of one patient. There are 500 samples in total.

1. Identify the dataset columns into nominal, categorical, continues etc. categories
2. Use dataframe.info and dataframe.describe to get the insights about the data.
3. Find the number of null values for each columns

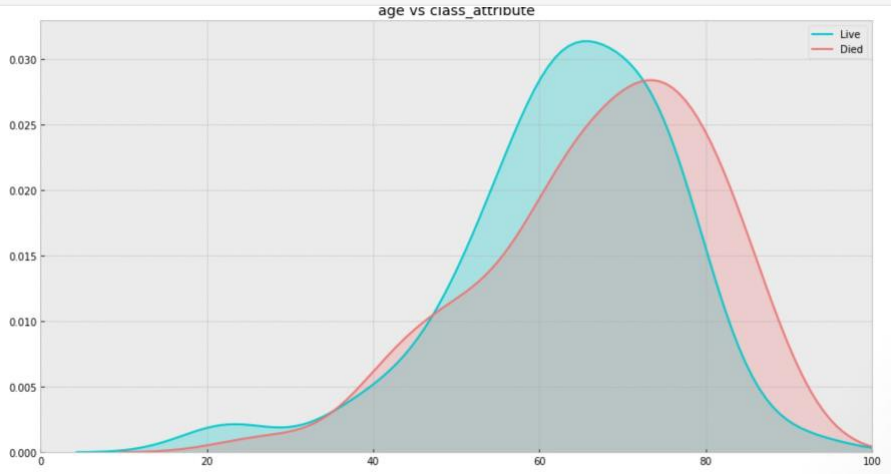
4. Know about the patients (Example of analysis for ages)
    a. Find the oldestperson
    b. Find the youngestperson
    c. Find the average age group
    d. Find median age
    e. Find the relationship between the cardio and ages(the cardio column is your prediction variable)

```
plt.figure(figsize=(15,8))

sns.kdeplot(
    data.age[data.class_attribute == 1],
    color="darkturquoise",
    shade=True
)

sns.kdeplot(
    data.age[data.class_attribute == 0],
    color="lightcoral",
    shade=True
)

plt.legend(['Live', 'Died'])
plt.title('age vs class_attribute')
plt.xlim(0,100)
plt.show()
```



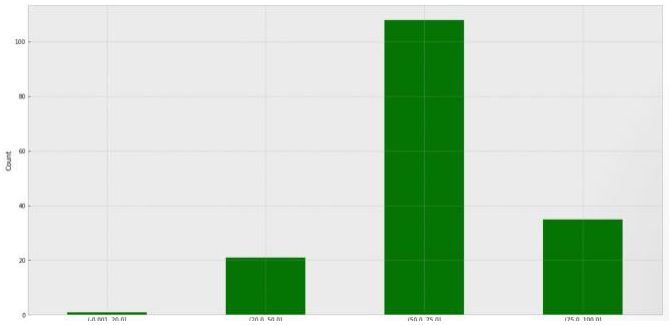
    f. Find the age groups whose survival rate is the largest

```
bins = [0, 20, 50, 75, 100]

out = pd.cut(
    data.age,
    bins=bins,
    include_lowest=True
)

ax = out.value_counts(sort=False).plot.bar(
    rot=0,
    color="g",
    figsize=(20,10)
)

plt.xlabel('Age bins')
plt.ylabel('Count')
plt.show()
```



g. Find similar relationships for at least **3-4 columns** that you think can play a role in prediction (For example, systolic BP, cholesterol etc.)
h. Get more visuals on data distributions
    i. Use plotCorrelationMatrix
    ii. plotScatterMatrix
    iii. plotPerColumnDistribution

    Use information from the plots to getan intuition for selecting feature variables
i. Find missing values
    i. Get the count of missing values
    ii. Plot a heat map for missing values
j. Applying a different technique to handle missing values (For each technique verify your prediction results)
    i. Use dropna
    ii. Use replace NA with zero or max value
    iii. Use replace NA with mean
    iv. Search for additional techniques to handle null values, excluding the above threeandtest.(Include allthetechniquesthatyou used inyour report.)
k. Applying the feature scaling technique if you think it is required. (Optional)
l. Applying the regression models that you think is most suited for this problem.
m. At least one of the models used to compute should be your own implementation using NumPy.
n. Upload your test data **predictions** to Kaggle competition in the correct submission format.
o. Use the cardio-validation.csv and cardio-train.csv as well to make your final prediction.

Task 2 (Use cardio-complete.csv.)

Submission - canvas

1. Split the dataset in train and test samples
2. Applying the regression model that you think is most suited for this problem.
3. Compare your prediction result with the first technique.

Comparison technique:

We will use confusion matrix to evaluate the performance
Compute Precision, Recall and F1 score for both Task 1 and Task 2

Task 3

Submission - canvas

a. Apply feature transform on the features used in task 1
    a. Does varying the polynomial degree change your accuracy?
    b. Can you identifyifyourmodel isunderfitting oroverfitting? (Hintuse cross-validation error and in-sample error plot to identify high bias and high variance.) Plot therelationships.

Sample code for polynomial regression.

```
# pass the order of your polynomial here degree is 2
poly = PolynomialFeatures(2)

# convert to be used further to linear regression

X_transform = poly.fit_transform(X_train)
```

Submission details:

- Jupyter Note files (You can have one file to show task 1-3)
- Prediction file submitted to Kaggle
- A detailed report of your analysis and finds. Add plots and describe your findings on data analysis and model prediction. Compare the results for Tasks 1,2 and 3.

GRADING

Task 1 (50):

- Data visualization (cardio-train.csv):20
- Own implementation of ML algorithm: 10
- Kaggle leaderboard standing:20

Task 2 (20):

- Repeat steps to compare: 10
- Confusion matrix: 10

Task 3 (10)

Report (10)

QA (10)

NOTE: If your code and submission result on Kaggle does not match, no points will be awarded for LAB 1.