CMPE 273

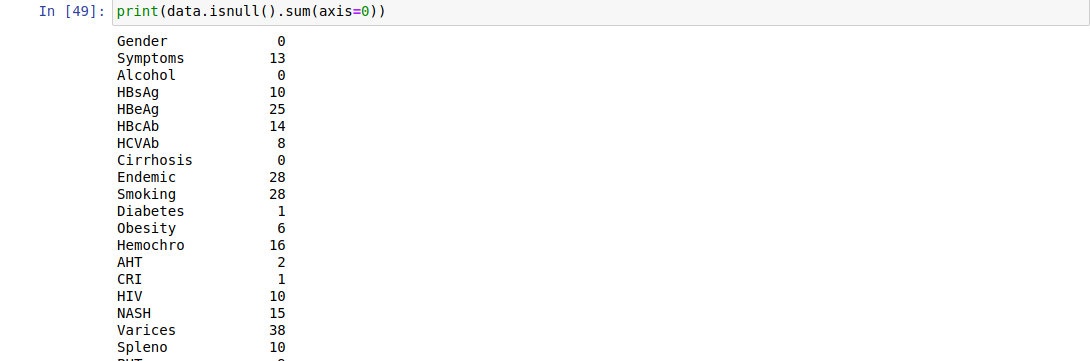
Lab 1

Pranav Karmalkar

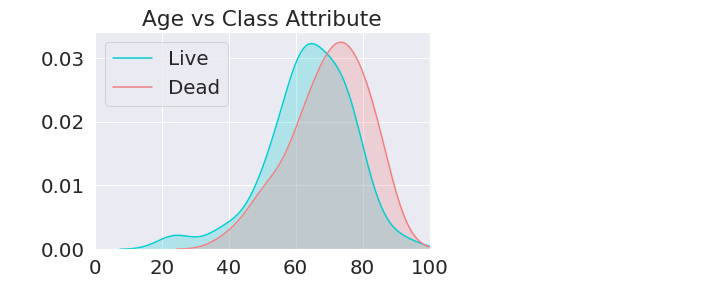
01469212

**Task 1:**

1. Data till 'Hallmark' column from start is nominal, 'Nodules' and 'Age' columns are an integer, 'Encephalopathy','PS', 'Ascites' are ordinal columns apart from above exceptions all other columns after 'Hallmark' are continuous.
2. The data set contains 110 rows and 50 feature columns. Also there are a lot of null values. Data in this set in numerical and stored in int or float.
3. By finding the number of null values we can infer that it is a sparse database



1. It can be inferred from the plots and dataset that:
   1. Max Age - 93
   2. Min Age - 23
   3. Avg Age - 66.081
   4. Median Age - 67.0
   5. With increase in age the probability of survival with HCC reduces.
   6. Survival rate is the percentage of people in a study or treatment group still alive for a given period of time after diagnosis. 100% of people have survived between the ages of 0-25(max survival rate).





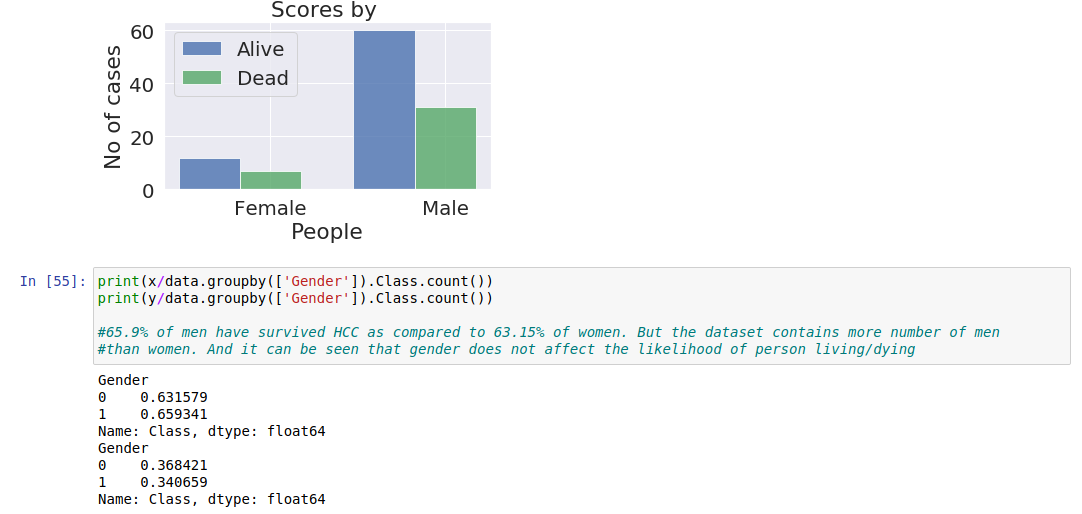
* 1. Age Groups:

65.9% of men have survived HCC as compared to 63.15% of women.

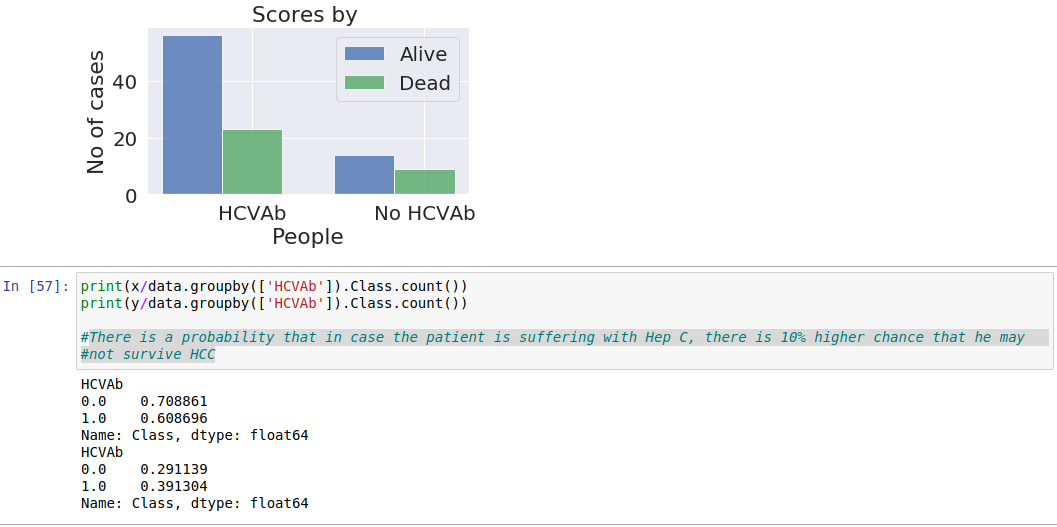
But the dataset contains more numbers of men than women. And it can be seen that gender does not affect the likelihood of person living/dying

Hepatitis C & B:

Worldwide, the most common risk factor for liver cancer is chronic (long-term) infection with hepatitis B virus (HBV) or hepatitis C virus (HCV). These infections lead to cirrhosis of the liver and are responsible for making liver cancer the most common cancer in many parts of the world.

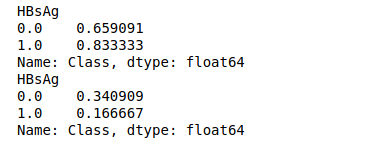


There is a probability that in case the patient is suffering with Hep C, there is 10% higher chance that he may not survive HCC



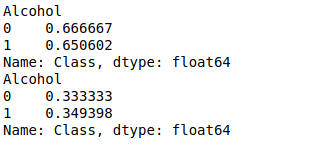
There is a probability that in case the patient is suffering with Hep C, there

is 10% higher chance that he may not survive HCC



Alcohol:

There is a probability that in case the person does not consume alcohol,there is 1% higher chance that he may survive HCC



* 1. Plot Corelation Matrix

(unable to attach in Doc)

From the below matrix we get to know the correlation of different columns with each other and also the result class. Based on this various columns can be removed which are not co-related to the outputclass.

Plot Scatter Matrix

(unable to attach)

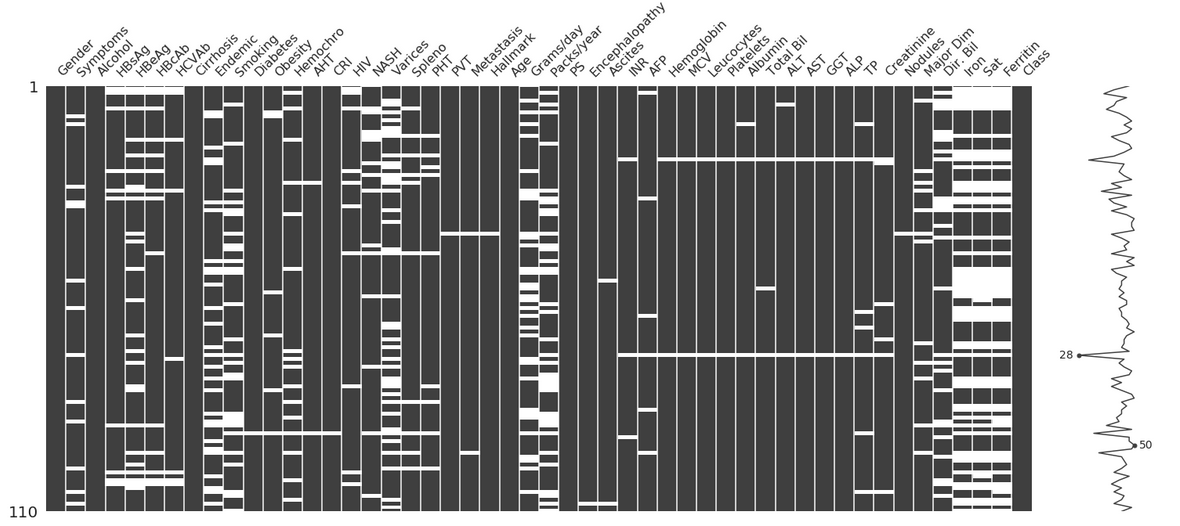
This gives us the understanding how the data is scattered in between features and class variable. Letting us understand the scatter.

PerColumnDistribution

This gives us the understanding of how data is distributed against each axis. It also gives us the understanding of the median and mean in a particular column.

* 1. HeatMap and missing values

From this graph and the missing value data it can be inferred that Sat, Ferritin and Iron are three columns with most missing values i.e. more than 50 in a 110 dataset and might not yield good results.



* 1. Insertions of various values:

After considering the below accuracy values, KNN Imputer was chosen in this scenario, with neighbour size equal to 9

Zero:

SKLearn Accuracy is 0.8636363636363636

Max:

SKLearn Accuracy is 0.6363636363636364

Mean:

SKLearn Accuracy is 0.7727272727272727

KNN Imputer:

SKLearn Accuracy is 0.8181818181818182

Simple/Iterative Imputer:

SKLearn Accuracy is 0.7272727272727273

* 1. Min-Max scaler was used for feature scaling was used because the nominal columns are mostly under 0 and 1
  2. Logistic regression model was chosen instead of SVM as it produced a higher accuracy.

SVM - SKLearn Accuracy is 0.5

Logistic Regression - SKLearn Accuracy is 0.8636363636363636

* 1. Own implementation of Logistic regression algorithm produced a lower accuracy score than SK Learn

Accuracy is 0.5454545454545454

* 1. Random Forest was used in case of predicting accuracy on Kaggle Dataset, with 150 trees and random state of 3. Rest of the features cause a greater RMS value.

**Task 2:**

Logistic Regression was used in this case as it was yielding the highest accuracy. But accuracy with hcc-complete-balanced.csv dataset was lower than the hcc-train.csv.

**Balance Dataset**

TN-29 FP-2 FN-8 TP-23

Recall= 0.7419354838709677

Precision= 0.92

F1= 0.8214285714285714

**Train Dataset**

TN- 5 FP- 1 FN- 2 TP- 14

Recall= 0.875

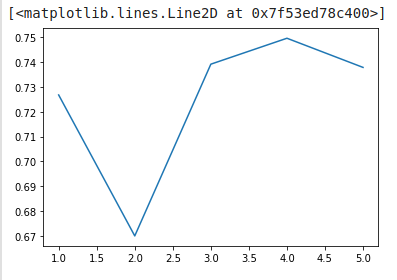
Precision= 0.9333333333333333

F1= 0.9032258064516129

**Task 3:**

The accuracy is getting reduced after every iteration of polynomial degree. The model is **overfitting** as with every iteration the accuracy becomes hundred person in case of in-sample accuracy. And the out-of-sample accuracy reduces, which shows the model is overfitting.

Cross-Validation Score vs Degree



In-Sample Accuracy vs Our-of-Sample Accuracy

