Capstone 3 Final Report:

Heart Attack Analysis & Prediction

**Problem Statement:**

According to the American Heart Association, Cardiovascular disease (CVD), listed as the underlying cause of death, accounted for 874,613 deaths in the United States in 2019. Using some indicators to predict the risk of Heart attack.

**Data Wrangling:**

There is no need to do data wrangling, because the dataset is well organized and clean. The only thing needs to do is to group the age into groups for EDA.

**Exploratory Data Analysis:**

* To count all the categorical data, to understand some of elements.

Chart, waterfall chart

Description automatically generated

For above, we understand in this dataset:

* There are more females
* Patient don’t have more exercise induced angina
* Most the patient with typical angina
* Most patient were having fasting blood sugar >120 mg/dl

Compare the categorical data with outputs:

* Gender Vs Output

Chart, bar chart

Description automatically generated

* Chest pain type vs output: 0= typical angina, 1=atypical angina, 2=non-anginal pain, 3=asymptomatic

Chart, bar chart

Description automatically generated

* Exercise Induced Angina vs Output: 1= yes, 0=no

Chart, bar chart

Description automatically generated

* Number of major vessels vs output:

Chart, bar chart

Description automatically generated

* Resting electrocardiographic results vs output: 0=normal, 1=Having ST-T wave abnormally, 2= showing probable or definite left ventricular hypertrophy by Estes' criteria

Chart, bar chart

Description automatically generated

* Age Group vs output

Chart, bar chart

Description automatically generated

Analysis on some continue data:

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Heart attack risk vs Chol, Resting Blood Pressure, Maximum HR Achieved:

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Check the relationship between all variables:

A picture containing table

Description automatically generated

Treemap chart

Description automatically generated

Check the Boxplot on some continued variables with output:

Graphical user interface, calendar

Description automatically generated

**Model Selection:**

I test three different Machine Learning models: Logistic Regression, Gradient Boosting, and SVC.

I use the default value to do the model and split the data to 30/70.

When using the Gradient Boosting, tune the best learning rate. Learning rate at 0.5 and 0.75 have the same Accuracy score. I use o.5 on the model.

**In Conclusion:**

Compare to three model below:

|  |  |  |
| --- | --- | --- |
|  | Accuracy Score | AUC |
| Logistic Regression | 0.824 | 0.899 |
| Gradient Boosting | 0.81 | 0.875 |
| SVC | 0.802 | 0.922 |

Even the SVC has the highest in AUC, I choose Logistic Regression as the best model to use, because it has the highest accuracy score.