**UCI HEART DISEASE DATASET ANALYSIS:**

Columns available in dataset:

* **id** = ID
* **age** = Age in years
* **sex** = Gender
* **dataset** = Location of dataset collection.
* **cp** = Chest pain type.
* **trestbps** = resting blood pressure.
* **chol** = cholesterol measure.
* **fbs** = fasting blood sugar (>120 mg/dL)
* **restecg** = Ekg observation at resting condition.
* **thalch** = Mximum heart rate achived.
* **exang** = Exercise induced angina (chest pain)
* **oldpeak** = ST depression induced by exercise replative to rest.
* **slope** = The slope of the peak exercise ST segment.
* **ca** = Number of major vessels (0-3) colored by flourosopy.
* **thal** = Thalassemia | Disorder that causes your body to have less hemoglobin than normal.
* **num** = Target 0=no heart disease; 1,2,3,4 = stages of heart disease.

Learnings and Findings:

it gives a real-world setting for machine learning specialists to create models that can foresee heart disease based on different variables. By preparing models on this dataset, people can learn how to analyze and interpret restorative information, a significant expertise within the field of healthcare and information science that would be helpful for future generations, and developing models for heart disease can help healthcare experts in early conclusion and mediation, possibly sparing lives. These models maybe helpful in creating new tools for cardiologist's and recognize people at higher danger and suggesting suitable preventive measures or treatment plans, this learnings would have potential affect beyond the education, enhance patient care, encourage early disease location, and contribute to the broader field of restorative investigate and innovation.

Usage: The model created would be utilize in any devices and software’s, for continuous tracking the important variable of a client/ human being and perform stages of data analysis and machine learning, goes thru algorithm and provide vital insights from the heart rate readings, the fun part is it can do it faster cause in today’s life a client wants to track his/her heart vitals and get insights it has to go through blood sample/device reading > healthcare laboratory > a parson has to do sample testing > it goes to doctor for validation and lastly to the client. With machine learning model firstly it will starting working as soon as its gets the reading and it would be already train on allot of data, and with good accuracy as well, it could generate the lab report in minutes, for the implementation we need to work on data preprocessing, feature engineering, model selection, hyperparameter optimizing and cross validation after all this, cross-validation of the working should be validate but specialist like senior cardiologist, than we can understand our model stands.

Learning: Though am into python for last 4-5 years, mostly spend my time in core-python and we development, even worked around machine learning model building but this project helped to think more about its not just model building, its about selecting best variable, parameters for algorithm, cross-validation the results, check for autoML as well. The understanding of which niche we are working on is important, here we are working on heart-rate data, we have to do analysis on which variable would be vital in our model building.

Insights:

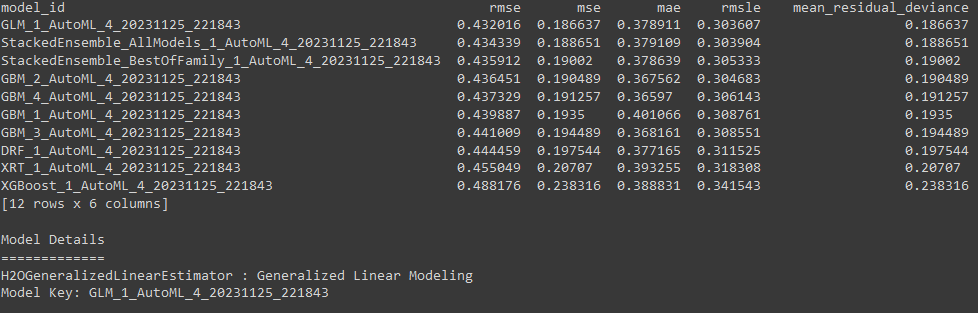
* The correlation matric shows chest pain, max heart rate and slope are important factors on target variable.
* Heart Attack has happened more to Male than Females.
* Age 40-70 has high chances of getting a Heart Attack.
* Cholesterol and Max HR lead to a heart disease.

Model:

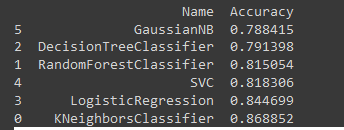
KNN(KNeighborsClassifier) is well-suited for scenarios where feature interactions are important. The variables ['age', 'rest\_blood\_pressure', 'cholesterol', 'max\_heart\_rate', 'oldpeak'] collectively represent different aspects of an individual's health, and their interactions could be crucial in understanding heart disease risk. KNN makes predictions based on the similarity of data points, and in this case, it may identify individuals with similar health profiles who are more likely to share similar heart disease outcomes, and with applying Hyperparameter (RandomizedSearchCV) the performance of the algorithm increases.

Best Parameter for KNeighborsClassifier:{'weights': 'distance', 'n\_neighbors': 8, 'metric': 'manhattan'} , with Hyperparameter Optimization the Accuracy is 0.8688524590163934 and without is 0.8448387096774195

Performed model building on KNeighborsClassifier, RandomForestClassifier, DecisionTreeClassifier, LogisticRegression, SVC, GaussianNB, the best accuracy got it from KNeighborsClassifier, with that installed and setup H20 AutoML library to get the leaderboard of the top-performing models generated by the H20 AutoML, it runs thru bundle of machine learning algorithm and check with each and every of them and create a leaderboard with accuracy, rmse values and H2O’s AutoML can also be a helpful tool for the advanced user, by providing a simple wrapper function that performs a large number of modeling-related tasks that would typically require many lines of code, and by freeing up their time to focus on other aspects of the data science pipeline tasks such as data-preprocessing, feature engineering and model deployment.



With other algorithm, the result is shown below:



Reference:

<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>

<http://archive.ics.uci.edu/dataset/45/heart+disease>

<https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html>