Predicting Heart Failure Survival with Machine Learning

Lightweight IBM Cloud Garage Method for Data Science

# Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

## Data Source

### Technology Choice

* The dataset consists of medical records of 299 heart failure patients collected at Faisalabad Institute of Cardiology and Allied Hospital in Faisalabad, Punjab, Pakistan, during April–December 2015.
* The data is downloaded from Kaggle.
* Pandas is primarily used to load the data.

### Justification

* The dataset is in .csv format and pandas package provides convenient way to load the data and perform basic statistics of the dataset.

## Enterprise Data

### Technology Choice

* Not needed

### Justification

* Not needed

## Streaming analytics

### Technology Choice

* Not needed

### Justification

* Not needed

## Data Integration

### Technology Choice

* For loading and extracting the statistics of the dataset, pandas dataframe object was used.
* For scaling the dataset, sklearn package, StandardScaler() was used.

### Justification

* Both packages provide an easy and efficient way of performing computations with minimal arguments.

## Data Repository

### Technology Choice

* The dataset is saved in Github Repository.

### Justification

* Since there are no additional dataset or live dataset, Github provides easy way of storing data.

## Discovery and Exploration

### Technology Choice

* Pandas, Matplotlib, Seaborn

### Justification

* Pandas is used to explore the statistics of the dataset which provides metrics like mean, std, 25,50,75 % quantiles, min and max values.
* Matplotlib and Seaborn is used to visualize the data, specifically to plot distribution plots, correlation plots, histograms etc.

## Actionable Insights

### Technology Choice

* Scikit-Learn framework was used to split the dataset into train and test, scaling, training the ML models and printing the classification reports. Following methods and libraries are imported from sklearn,
  + train\_test\_split, GridSearchCV, StandardScaler
  + RandomForestClassifier, AdaBoostClassifier, GradientBoostingClassifier
  + classification\_report, confusion\_matrix, accuracy\_score

### Justification

* Scikit-Learn framework provides good collection of most of the baselines ML models to preprocess, train the data, perform hyperparameter tuning and also print performance reports.
* For hyperparameter tuning, the sklearn library provides easy way to get the best parameters for the given dataset.
* For assessing the model performance, the classification report provides a comprehensive report of evaluation metrics like Precision, Recall and F1 scores. Along with this, sklearn library also provides an easy function to print Confusion Matrix.

## Applications / Data Products

### Technology Choice

* Jupyter notebook is used for maintaining the code and performing experiments. Along with the code, explanations are also provided as a report within the .ipynb file.
* This project is carried out as an assignment from the “IBM Advanced Capstone Project in Data Science”.

### Justification

* Jupyter notebook provides an easy framework to load the packages, libraries, perform exploratory data analysis, ML model training and also writing the findings/analysis in the form of a report.

## Security, Information Governance and Systems Management

### Technology Choice

* Not needed

### Justification

* The dataset used is open source.