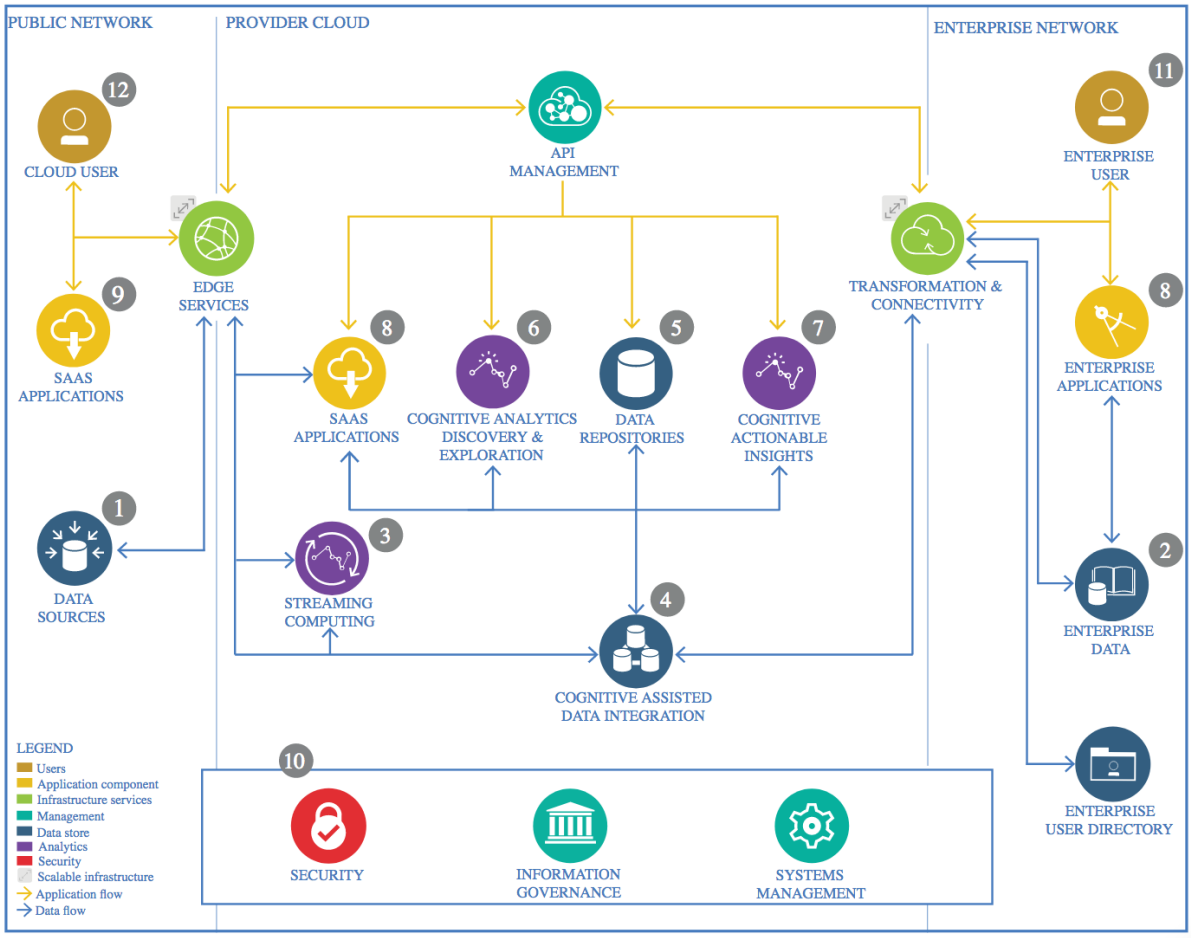
Predicting Heart Failure Survival with Machine Learning

Lightweight IBM Cloud Garage Method for Data Science

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 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.

1.1.2 Justification

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

1.2 Enterprise Data

1.2.1 Technology Choice

- Not needed

1.2.2 Justification

- Not needed

1.3 Streaming analytics

1.3.1 Technology Choice

- Not needed

1.3.2 Justification

- Not needed

1.4 Data Integration

1.4.1 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.

1.4.2 Justification

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

1.5 Data Repository

1.5.1 Technology Choice

- The dataset is saved in Github Repository.

1.5.2 Justification

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

1.6 Discovery and Exploration

1.6.1 Technology Choice

- Pandas, Matplotlib, Seaborn

1.6.2 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.

1.7 Actionable Insights

1.7.1 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

1.7.2 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.

1.8 Applications / Data Products

1.8.1 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”.

1.8.2 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.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

- Not needed

1.9.2 Justification

- The dataset used is open source.