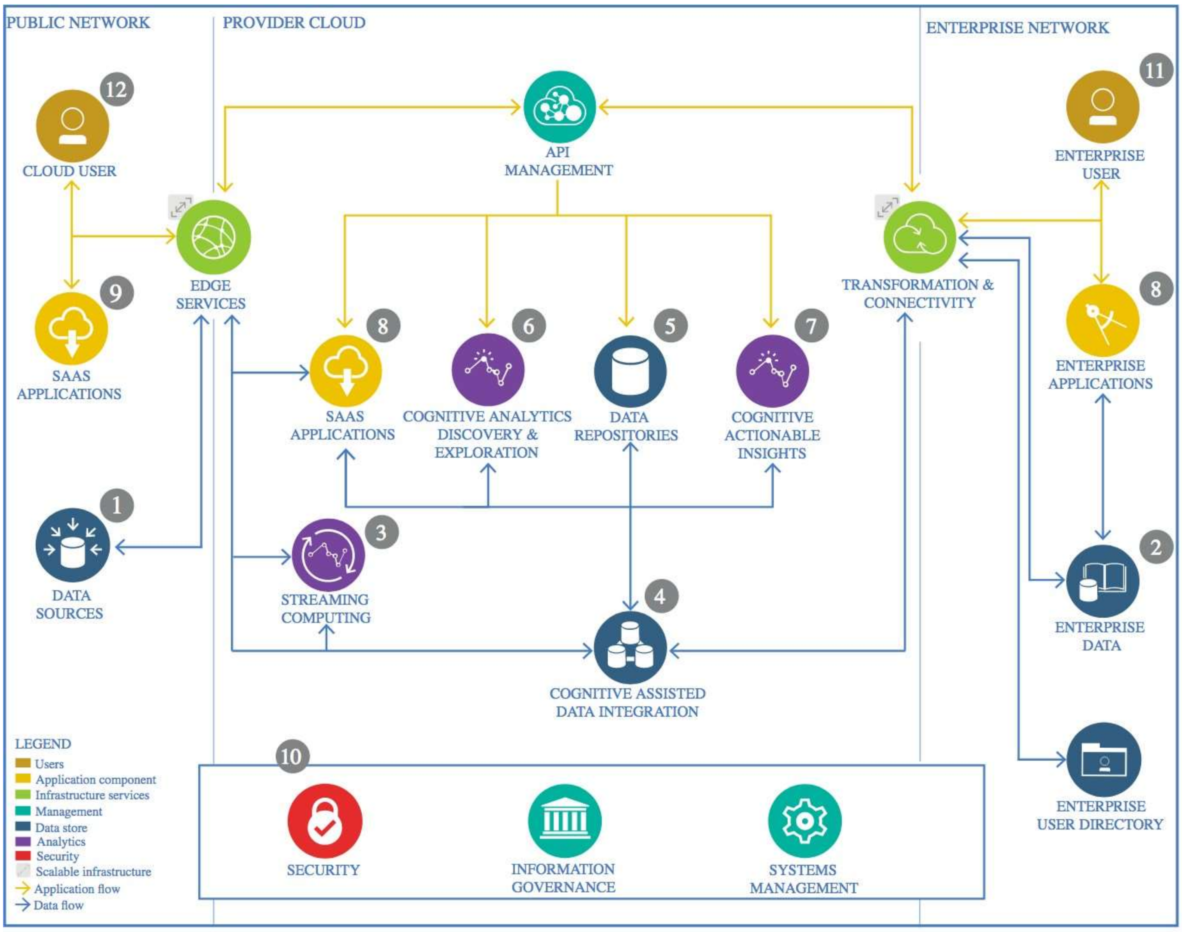
Breast Cancer Classification

1. Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1. Data Source

1.1.1. Technology Choice

The dataset selected for this project represents Breast Cancer from the Wisconsin Breast Cancer Database.

It contains data of 357 benign and 212 malignant cancers.

Link to database:
https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

1.1.2. Justification

CSV file with the results of science research. It is a typical format for the stable research data.

1.2. Enterprise Data

1.2.1. Technology Choice

This component is not needed in this project.

1.2.2. Justification

We are using a CSV file as a data source since our data does not require frequent updates, therefore, a cloud-based solution for enterprise data is not needed.

1.3. Data Integration

1.3.1. Technology Choice

- For data preprocessing, we handled the dataset as a Pandas data frame object, a 2-dimensional labeled data structure with the index for rows and columns.
- We used LabelEncoder for the target feature to one-hot-encode categorical features and RobustScaler, StandardScaler and MinMaxScaler to scale numeric features.
- Additionally, PCA Analysis and Feature Importance is performed to check dimensionality reduction

1.3.2. Justification

Panda's data frame structure is fast and has high productivity and performance, and it is suitable for our two-dimensional dataset. We used these preprocessing to perform the feature transformation and feature creation steps.

1.4. Data Repository

1.4.1. Technology Choice

The CSV file with the dataset is saved in the GitHub repository to be used in the analysis. It can be also downloaded on the above link website.

1.4.2. Justification

Based on the nature of the data and the objective of the analysis, the dataset will not require frequent updates for future models' training, therefore, the GitHub repository is a sufficient and efficient technology for our case.

1.5. Discovery and Exploration

1.5.1. Technology Choice

In order to visualize the data in a structured manner, the Pandas framework has been used. During the exploration of data, I have used Pandas, Seaborn, Plotly and Matplotlib Python packages to visualize the distribution of various features to understand their effectiveness for further processing.

1.5.2. Justification

These visualizations provide a brief idea about all the features along with the frequency of the corresponding values. During this phase, the following analysis has been performed.

- Feature Distribution
- Outliers
- Correlations

1.6. Actionable Insights

1.6.1. Technology Choice

We have used Scikit-Learn and Pyspark framework to develop different machine learning models. We have imported the following model classes along with the different accuracy measures.

- RandomForestClassifier
- accuracy_score
- classification_report
- confusion_matrix

1.6.2. Justification

- The machine learning classifier help to predict the diagnosis based on the selected features.
- The accuracy measures will be needed to perform the performance comparisons of these models.
- The classification report and confusion matrix will help to understand the model performance.

1.7. Applications / Data Products

1.7.1. Technology Choice

The data produced for this project is an ipynb file of the analysis generated from the “IBM_Capstone_Project.ipynb” Jupyter Notebook,

1.7.2. Justification

The main motivation behind this work is to build an intelligent system to predict breast cancer on the given data. Some major objectives include:

- Exploratory Data Analysis to get an insight of each feature.
- Correlation Analysis to check collinearity.
- Classification Analysis to predict the breast cancer.

1.8. Security, Information Governance and Systems Management

1.8.1. Technology Choice

This component is not needed.

1.8.2. Justification

The dataset used for this project is public, therefore no information governance and system management are needed.