Use Logistic Regression to predict Heart Disease

Architectural Decisions Document

# Architectural Components Overview

The project was completely developed in Watson Studio available in the IBM Cloud. Specifically I’ve created my model in Jupyter Notebooks and used Python 3.5 with Spark as a development environment.

## Data Source

The data source used in this project is the Heart Disease UCI data set from Kaggle (<https://www.kaggle.com/ronitf/heart-disease-uci>).

### Technology Choice

In order to analyze the data I used Watson Studio. Then first step was to copy the data into the Cloud ObjectStorage. Then I used a SparkSession in the Jupiter notebook to read the data and save them into a dataframe for further analysis.

### Justification

I decided to use Watson Studio because it is an integrated environment for Data Scientist that offer lots of possibilities. It offers many options to connect to the data, it offers many development environments that can be selected based on specific needs, it offers many alternatives for creating ML and DP models.

## Enterprise Data

No Enterprise Data available in this case.

## Streaming analytics

No need to use streaming analytics because I used not streaming data.

## Data Integration

No need to integrate data because they come from a single data source.

## Data Repository

### Technology Choice

Data are stored in the Cloud Object Storage.

### Justification

The reason is that Cloud Object Storage is fully integrated with Watson Studio. Training a machine learning or a deep learning model using IBM Watson relies on using IBM Cloud Object Storage for reading input (such as training data) as well as for storing results.

## Discovery and Exploration

### Technology Choice

The Discovery and Exploration phase was done in a Jupyter Notebook. To have a better understanding of the data I did:

- print my data frame and verify number od columns and rows

- print Schema to check datatype in each column

- execute statistical measures to get an idea on the value distribution of the data

I discovered that the datatype of each column is String. Then I had to cast columns in int or double. All columns contains number and that is exactly what I need to buil my model.

### Justification

Please justify your technology choices here.

## Actionable Insights

### Technology Choice and Justification

Interesting to compare the ML and DL model. The pyspark Logistic Regression reaches an accuracy of 60% against the 82% of the regression in Keras model.

This was a subset of the full data set. Once we move to the complete one I would suggest to:

* Test both model, even if regression in Keras seems to be preferred.
* Save the data into a DB and create the connection to it in the Notebook. In case of more data the Cloud Object Storage could not be a good choice.

## Applications / Data Products

### Technology Choice and Justification

Both ML and DL model were deployed in the form of a Jupyter Notebook. As alternative solution could be deployed to Watson Machine Learning.

## Security, Information Governance and Systems Management

### Technology Choice

Not required

### Justification

Not Required