The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

# Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

## Data Source

### Technology Choice

The database chosen was a CSV file from the Dutch government website (<https://data.overheid.nl/>). The dataset is available at the following link: <https://www.duo.nl/open_onderwijsdata/databestanden/vo/leerlingen/leerlingen-vo-7.jsp>. The data already comes in a very neat and organized format, therefore, no ETL operations are required except for loading the dataset into a pandas dataframe.

### Justification

Since the main question in this project is to predict performance of a student on the final exam based on his level of education as well as school performance, this dataset provides all the necessary data. It includes different levels of education for students across the Netherlands together with their grades for school exams and their grades throughout studies. It also shows their grades for the final exams.

## Enterprise Data

### Technology Choice

IBM Watson, Jupyter notebook, IBM cloud storage, GitHub

### Justification

This platform is supported by coursera learning community. Additionally, it provides enough storage space to keep all the data on the cloud. Furthermore, it provides an environment for analyzing and manipulating data.

## Streaming analytics

### Technology Choice

Streaming data was not chosen.

### Justification

School grades is data that changes only once a year: when new exams are taken. Therefore, streaming data, in this case, is not available. But it is also not useful, because school exams and final school grades are only captured once per year. Therefore, the model based on streaming analytics is not necessary.

## Data Integration

### Technology Choice

Some columns were removed. Data types were checked. Repetitive values were checked. The ranges of different features were checked. Fields where only specific values were allowed were checked.

### Justification

As a result, columns with a lot of missing values were removed. Additionally, columns that did not correlate with the label values were removed. As for the rest, the data types were correct everywhere, repetitive values were found not to be an issue. Fields where only specific values were allowed did not contain inconsistent values.

Standardization was not required (because grades were already at the same scale), but it was still carried out for all the numeric fields used in the model to ensure that variables are on the same scale and the model performs better. Moreover, all the categorical variables were OneHotEncoded to ensure they get fed properly in the model.

## Data Repository

### Technology Choice

GitHub

### Justification

As GitHub provides free storage for data accessible for everyone if access is shared, it was chosen as the persistent storage depository for data.

## Discovery and Exploration

### Technology Choice

Histograms, statistical measures, graphs, python coding, pandas dataframe

### Justification

Histograms were used to visualize the distributions of different numerical fields such as grade distribution. It is the best way to see any anomalies within variables. Additionally, statistical measures were used to check for validity of the items, whether there were any unusual fields or spikes in the dataset. Furthermore, python was used to check for unique fields and missing values. The data at hand was discovered to be data on the current school grades, school province as well as the level of education. This were the main fields used for training the model.

## Actionable Insights

### Technology Choice

Logistic Regression, R squared

### Justification

Logistic Regression was chosen because it enables prediction of continuous variables in a supervised learning environment. At the same time, the evaluation method that will be used is R2. For regression, this evaluation method is the best.

As a result of this project, a Linear Regression model was created that can predict the performance of students on the future exam with 76% accuracy was created. It was determined that normalizing the values has a negative effect on the model performance. Additionally, standartization has a positive effect on the model performance. Furthermore, dropping the geography OneHotEncoding decreases model performance, however it is a spurious correlation, therefore, it is needed.

The Linear Regression Model was made available through GitHub in the public repository.

A Deep Learning model was also created. However, this models' performance was very low. It only improved slightly when different activation function was used and different loss was applied. Loss changed as a result of this. However, the model still stayed low quality because its accuracy is just 1%.

## Applications / Data Products

### Technology Choice

Jupyter notebook, video presentation

### Justification

The Jupyter notebook with a nicely mapped out code and steps is easy for a business user to follow and understand. Additionally, a video presentation that shows all the steps in the analysis and the results was created for the stakeholders.

## Security, Information Governance and Systems Management

### Technology Choice

The machine learning algorithm can be accessed by anyone without restrictions.

### Justification

The dataset used for feeding the algorithm is publicly available. All the data is open to public, therefore, there is nothing restricted. Thus, this machine learning algorithm can be used by anyone.