**Predicting Diabetes Mellitus with IBM Watson Machine Learning model.**

**1.Introduction:**

**1.1 Overview**

Diabetes mellitus is a chronic disease characterized by hyperglycemia. It may cause many complications. According to the growing morbidity in recent years, in 2040, the world’s diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes. There is no doubt that this alarming figure needs great attention. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health. It is a chronic disease or group of metabolic disease where a person suffers from an extended level of blood glucose in the body, which is either the insulin production is inadequate, or because the body’s cells do not respond properly to insulin. The constant hyperglycemia of diabetes is related to long-haul harm, brokenness, and failure of various organs, particularly the eyes, kidneys, nerves, heart, and veins. The objective of this research is to make use of significant features. Based on the few available health parameters I can predict weather the woman will get diabetic or Non Diabetic after her delivery. For this predication I can consider the few parameters of the like Glucose, Blood Presser, Skin thickness, Insulin, BMI, Diabetes Pedigree function, Age, Number of Pregnancies of the person.

**1.2 Purpose:**

To review the long-term effects of the diabetic pregnancy on the offspring among the Pima Indians of Arizona.

**2. Experimental Investigation:** Predict diabetes diagnosis for Pima Female Indians

**2.1 Data Understanding**: Data Set Properties: 9 attributes representing 769 Pima Female Indians.

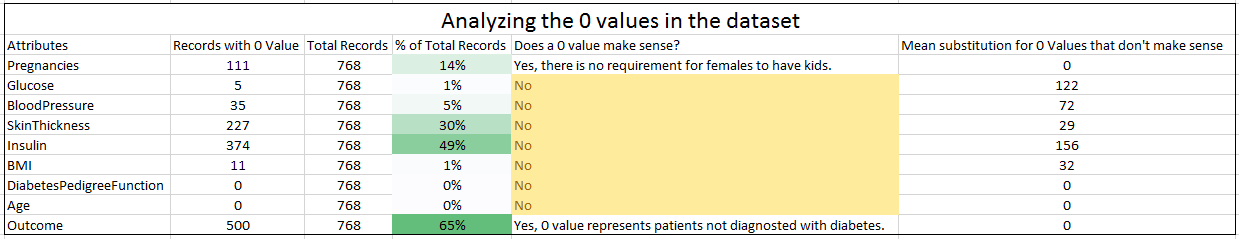
**2.2 Data Set Info**: Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

**2.3 Dataset Attributes:**  
 1 Pregnancies: Number of times pregnant  
 2 Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test  
 3 Blood Pressure: Diastolic blood pressure (mm Hg)  
 4 Skin Thickness: Triceps skin fold thickness (mm)  
 5 Insulin: 2-Hour serum insulin (mu U/ml)  
 6 BMI: Body mass index (weight in kg/(height in m)^2)

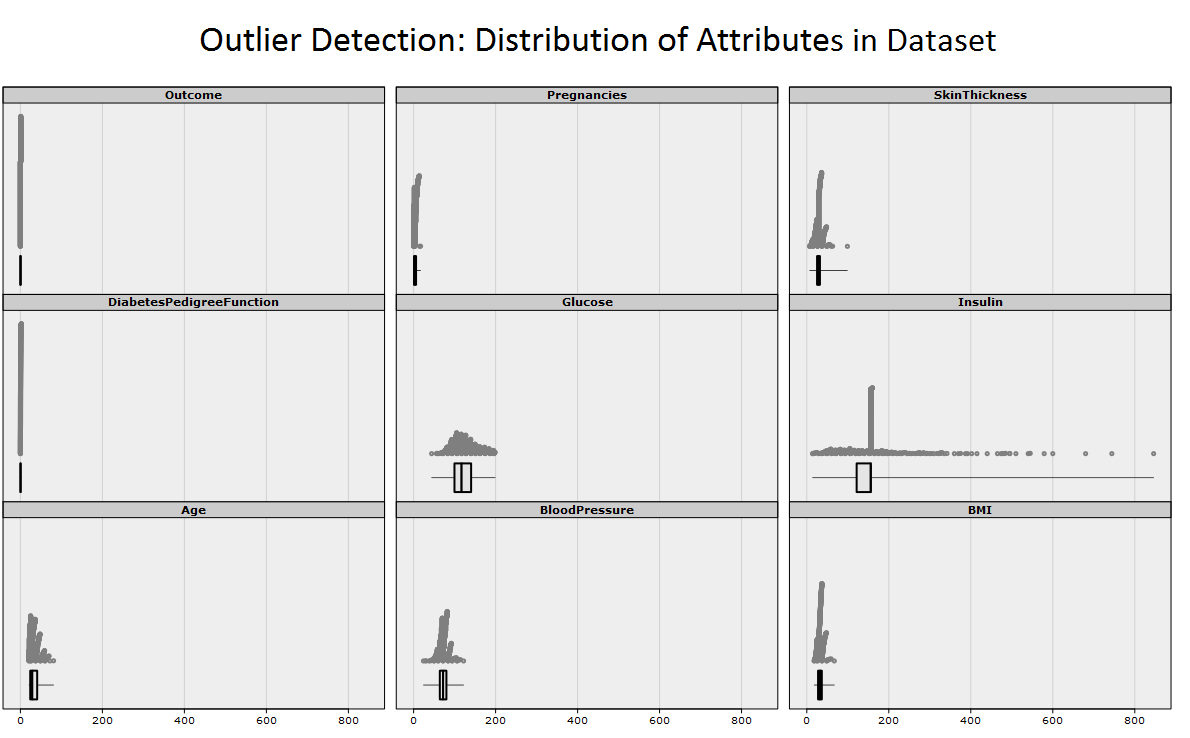
7 Pedigree Function: Diabetes pedigree function  
 8 Age: Age (years)  
 9 Outcome: Class variable (0 or 1)"

**2.4 Data Preparation**:

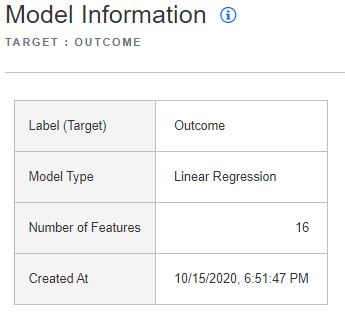
Analyzed 0 value records across all attributes. There were several attributes where a 0 value did not make sense and I think were used for missing data. (There were 0's for Blood Pressure, Skin Thickness, Insulin, BMI, Glucose). I used mean substitution for these cases. 0 Value Analysis in Dataset.



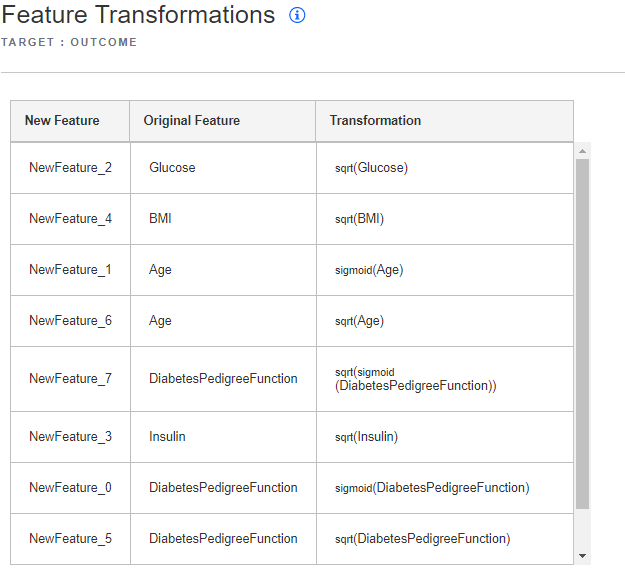
There were not too many outliers in the data. None of the data was removed. The below fig shows the distribution of the data by each attribute after a few of the attributes were adjusted with mean substitution Attribute Distribution using iNZight



**2.5 Modeling & Evaluation:** I used linear regression Algorithm with 16 features for best RMSE scores.

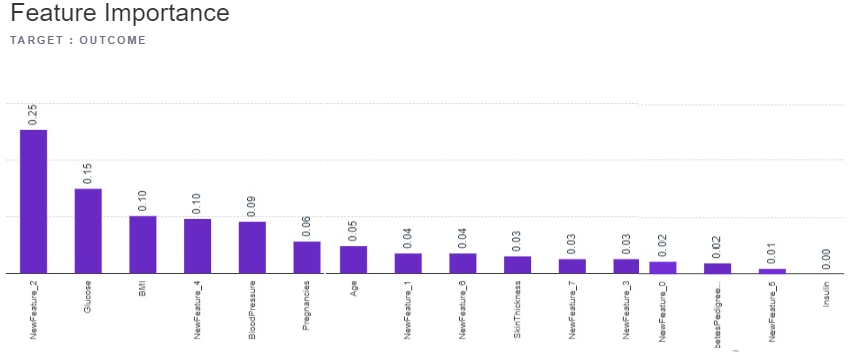


2.5.1 Table: Model information

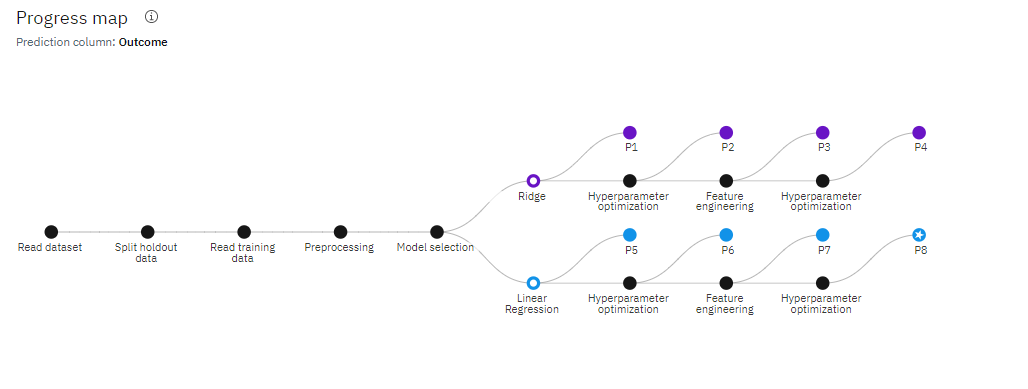


2.5.2.Feature Transformation

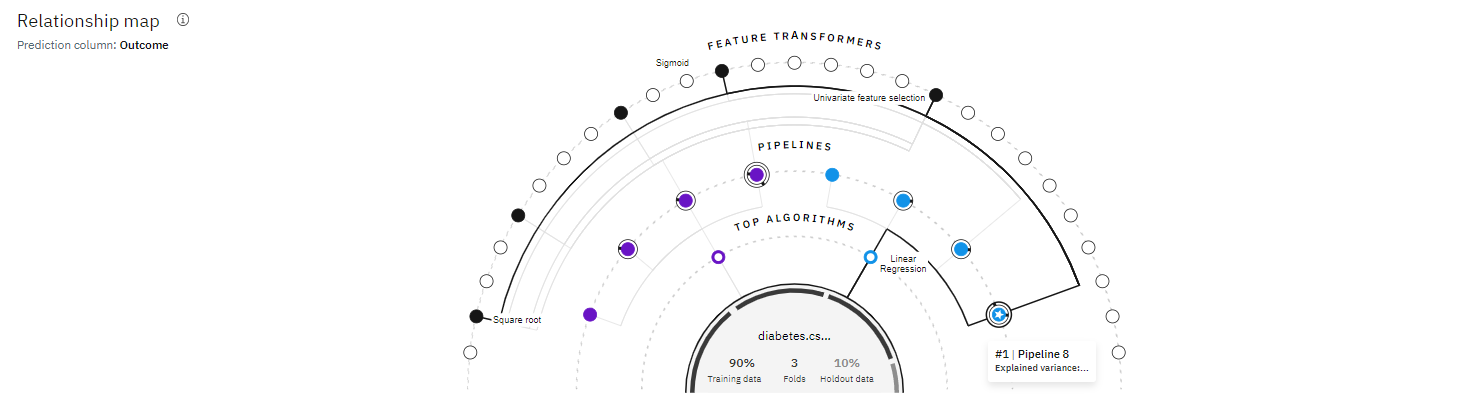
**3.Flowcharts:**



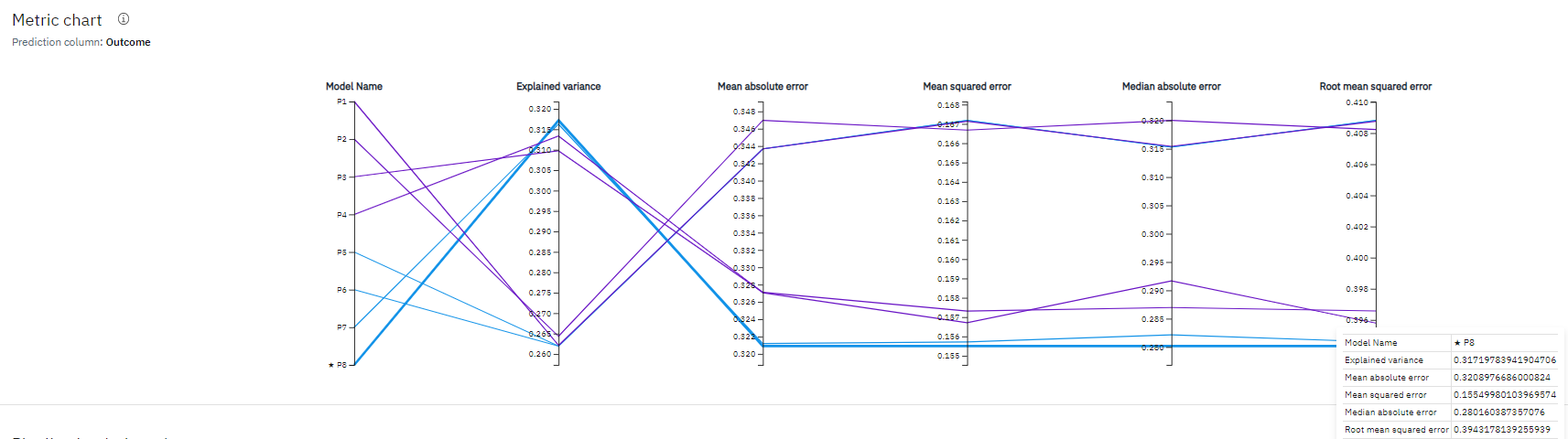
**3.1.1 Feature Importance Chart**



**3.1.2 Progress Map Chart**

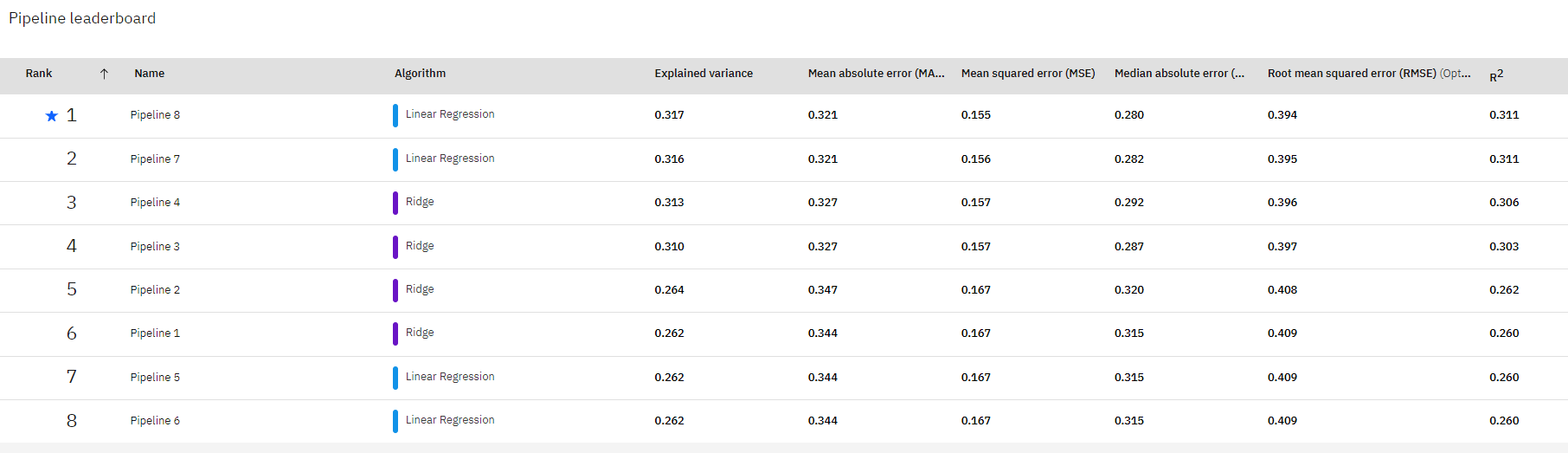
****

**3.1.3 Relation Ship Map**

****

**3.1.4 Metric Chart**

**4.Table Values:**



**4.1.1 Table: Pipeline Leaderboard**

**5. Results:**

**Cross Validation:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Enhancements** | **Name** | **Explained Variance** | **MAE** | **MSE** | **MedAE** | **RMSE** | **R^2** |
| Linear Regression | HP0-1,FE,  HPO-2 | Pipeline8 | 0.317 | 0.321 | 0.155 | 0.28 | 0.394 | 0.311 |
| Ridge | HP0-1,FE,  HPO-2 | Pipeline 4 | 0.313 | 0.327 | 0.157 | 0.292 | 0.396 | 0.306 |
| Linear Regression | HP0-1,FE | Pipeline7 | 0.316 | 0.321 | 0.156 | 0.282 | 0.395 | 0.311 |
| Ridge | HP0-1,FE | Pipeline3 | 0.31 | 0.327 | 0.157 | 0.287 | 0.397 | 0.303 |
| Liner Regression | HP0-1 | Pipeline6 | 0.262 | 0.344 | 0.167 | 0.315 | 0.409 | 0.26 |
| Ridge | HP0-1 | Pipeline2 | 0.264 | 0.347 | 0.167 | 0.32 | 0.408 | 0.262 |
| Linear Regression | None | Pipeline5 | 0.262 | 0.344 | 0.167 | 0.315 | 0.409 | 0.26 |
| Ridge | None | Pipeline1 | 0.262 | 0.344 | 0.167 | 0.315 | 0.409 | 0.26 |

**5.1.1 Table: Cross Validation Value for Linear Ridge Algorithms**

**Hold out Score:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Enhancements** | **Name** | **Explained Variance** | **MAE** | **MSE** | **MedAE** | **RMSE** | **R^2** |
| Linear Regression | HP0-1,FE,HPO-2 | Pipeline8 | 0.363 | 0.322 | 0.151 | 0.304 | 0.388 | 0.359 |
| Ridge | HP0-1,FE,HPO-2 | Pipeline 4 | 0.352 | 0.328 | 0.153 | 0.313 | 0.391 | 0.348 |
| Linear Regression | HP0-1,FE | Pipeline7 | 0.362 | 0.322 | 0.151 | 0.306 | 0.388 | 0.358 |
| Ridge | HP0-1,FE | Pipeline3 | 0.355 | 0.329 | 0.152 | 0.294 | 0.39 | 0.352 |
| Liner Regression | HP0-1 | Pipeline6 | 0.341 | 0.388 | 0.155 | 0.32 | 0.394 | 0.34 |
| Ridge | HP0-1 | Pipeline2 | 0.334 | 0.347 | 0.156 | 0.321 | 0.396 | 0.33 |
| Linear Regression | None | Pipeline5 | 0.341 | 0.338 | 0.155 | 0.32 | 0.394 | 0.34 |
| Ridge | None | Pipeline1 | 0.34 | 0.338 | 0.155 | 0.319 | 0.394 | 0.339 |

**5.1.2 Table: Hold Out Score Value for Linear Ridge Algorithms**

The offspring of women who had diabetes during pregnancy, on average, were more obese and had higher glucose concentrations and more diabetes than the offspring of women who developed diabetes after pregnancy or who remained non diabetic. Although no new analyses were attempted, several of the older publications were updated by repeating the analyses on later, expanded data sets.

**6.Conclusions:**

The diabetic pregnancy, in addition to its effects on the newborn, has effects on the subsequent growth and glucose metabolism of the offspring. These effects are in addition to genetically determined traits.