Diabetes in Pima Indian Women: understanding the problem and searching for answers through Data Analysis

Diogo Silvério (45679)
Departamento de Informática
FCT/UNL
Caparica, Portugal
d.silverio@campus.fct.unl.pt

Marcin Konieczny (56551)
Departamento de Informática
FCT/UNL
Caparica, Portugal
marcin.l.konieczny@student.put.poznan.pl

***Abstract*—The purpose of this report is to try discover new information about the high rate of diabetes that occurred 50 years ago in the population of Pima Indian Women through data exploration and data analysis. The study is focused on 3 main Data Analysis techniques. We began by performing a Linear Regression analysis in hopes of finding correlation between two of the dataset variables; then moved onto a Principal Component Analysis in hopes of reducing the data dimension or scoring a hidden factor in the data; finally we ended our study by performing a Fuzzy Clustering with Anomalous Patterns analysis of the data.**

***Index Terms*—diabetes, Pima, data exploration, data analysis, linear regression, PCA, SVD, fuzzy clustering, anomalous patterns**

I. Introduction

The Pima people are a group of Native Americans living in an area consisting of central and southern Arizona. A few years ago, the San Carlos Irrigation project began blocking the flow of the water and diverging it to non-native farmers. It caused periods of drought and led to lengthy periods of famine that affected the Pima people. Once used to gathering, hunting and surviving from the land surrounding them, this population started to consume food with higher levels of carbohydrates that was provided to them by the US Government. This resulted in a general weight gain and blood sugar levels in the population resulting in a high rate of diabetes and obesity [1]. This curious yet tragic event motivated us to study it.

The data set consists of several medical predictor variables and one target variable, the Outcome. All the variables are numerical continuous, expect the age, number of pregnancies and outcome that are numeral discrete. The predictor variables include the number of pregnancies, the plasma glucose concentration, the age, the diastolic blood pressure ($mmHg$), the triceps skin fold thickness (mm), the insulin level (muU/ml), the body mass index (Kg/m^2), the diabetes pedigree function and the outcome class variable, (0 or 1) represent if the patient has diabetes or not. This dataset contains 768 entities, 9 features and missing data. It is originally from the National Institute of Diabetes and Digestive and Kidney Diseases and all patients are females at least 21 years old of Pima Indian heritage. The data set could for example be used to build a machine learning model to accurately predict whether or not the patients have diabetes.

We'll explore 3 types of Data Analysis: Linear Regression, Principal Component Analysis and Fuzzy Clustering with Anomalous Patterns. Regression modelling represents a powerful and elegant method for estimating the value of a continuous target variable through simple linear regression, where a straight line is used to approximate the relationship between a single continuous predictor variable and a single continuous response variable [2]. Principal Component Analysis is a statistical procedure that allows the removal of some features that may be redundant, allows to visualize high dimensional data and deal with the Curse of Dimensionality. This method also provides a way of finding a "hidden factor" in the data. This "hidden factor" can be used on our data to measure impact of each feature in the positive outcome of diabetes. Fuzzy Clustering is a type of clustering method that provides a way of dealing with uncertainty. Instead of using the conventional set theory that only permits a entity to belong or not to a set, the fuzzy set is used where each entity can belong to a set with a certain degree or grade of membership defined by a membership function.

II. Experimental Study

A. *Linear Regression*

Since the entire Linear Regression analysis is based on the existence of a linear relationship between a predictor and a response variable we must first find two features in our data set that might possess this quality. In order to facilitate our search we'll plot the Scatter Matrix of our data to check if any two features of the dataset have a "linear-like" scatter plot.

Analysing fig.21 we can see that the Skin Thickness and the Body Mass Index have the "linear-like" scatter plot that we're looking for. Therefore, the Skin Thickness will be our predictor variable while the BMI will be our response variable. Now that we have our variables selected let us build our linear regression and plot it.

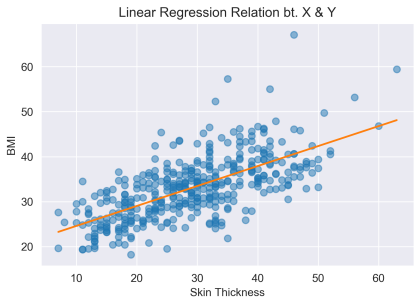

Fig. 1. Linear Regression relation between Skin Thickness and BMI and corresponding scatter plot.

Having built the Linear Regression model, let us plot the normal probability plot of residuals and the plot of standardized residuals against the predicted values to check if we have acceptable normality.

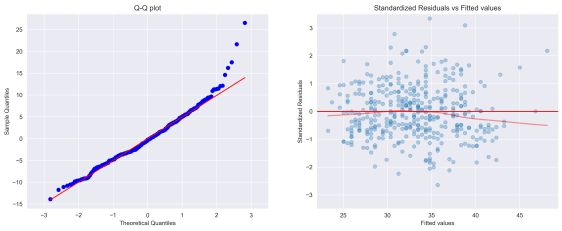

Fig. 2. Q-Q plot and Standardized Residuals vs Fitted Values

As we can see in fig.2, almost all of our data lines up nicely along the straight line, although the left tail deviates slightly and the right tail deviates a considerable amount. There are around 13 outliers and there's left skewness. It's not perfect but it's still quite a good result taking into account that this is real world data. Analysing the Standardized Residuals vs Fitted Values plot, we can see that it presents a "healthy" plot where no discernible patterns exist since all the data points form an overall rectangular shape. This means the Regression assumptions remain intact. Now we move on to feature engineering in hopes of improving the quality of our model. To do this we'll apply a log function to the response variable, the BMI.

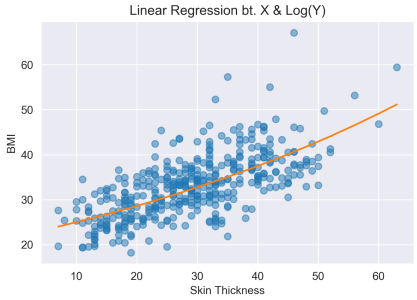

Fig. 3. New Regression model with transformation $log(y)$

Now like before, let us plot the normal probability plot of residuals and the plot of standardized residuals against the predicted values and check if we have still have acceptable normality and if our new model is improved.

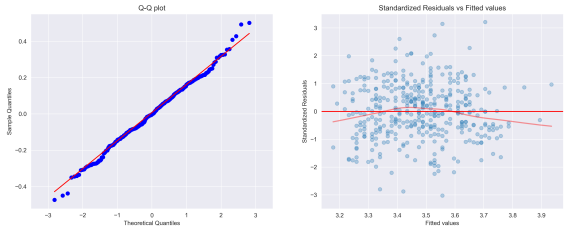

Fig. 4. Q-Q plot and Standardized Residuals vs Fitted Values

As we can see in fig.4, almost all of our data continues to line up nicely along the straight line, although both tails deviate slightly having 7 outliers. Analysing the Standardized Residuals vs Fitted Values plot, we can see that it still presents a "healthy" plot where no discernible patterns exist. This means the Regression assumptions remain intact and we obtained a slightly improved model. Because the improvement isn't very significant we'll continue the rest of the analysis with our basic model to simplify it. Having said that we can now write the population regression equation for our model:

$$BMI = 20.14686217 + 0.44395886 \cdot SkinThickness + \epsilon$$

This means that for each increase in one mm of Skin Thickness, a Pima Indian woman's Body Mass Index increases by 0.44395886 kg/m^2 plus 20.14686217 plus an error ϵ. For example a Pima Indian woman with a skin thickness of $30mm$ has an estimated BMI of 20.14686217 + 0.44395886(30) + ϵ = 33.46562797, roughly 33 kg/m^2.

We know that our variables have a linear relationship and already have our model built but we haven't questioned just much correlated they really are and how useful our regression line is for making predictions. In order to answer these questions we'll use the correlation and determinacy coefficients.

```
Coefficient of Determinacy: 0.44
Correlation: 0.67
```

Fig. 5. Determinacy and Correlation Coefficients

The Coefficient of Determinacy has a low score of 0.44 indicating that the regression line isn't useful for making predictions and the Correlation Coefficient (0.67) is positive indicating that the variables are positively correlated.

We've seen before that our model respects the Regression Assumptions but when inference or model building is performed these assumptions must be validated. In order to do this we use several inferential methods like the T-test and confidence intervals.

First let us begin by testing the statistical hypothesis for determining whether a linear relationship exists between the chosen variables.

T-test for Relationship between x and y:

- H0: asserts $\beta_1 = 0$ (no linear relationship exists)
- Ha: asserts $\beta_1 \neq 0$ (linear relationship exists)

T-test rejects H0 when the p-value is small and 0.05 is used as rejection threshold.

```
Test statistc for coefficient estimate  0 : 25.712888384584428 | P-value: 0.0
Test statistc for coefficient estimate  1 : 17.553724658425693 | P-value: 0.0
```

Fig. 6. T-test statistic for both coefficient estimates

As we can see in fig.6 because the p-value is 0.0, we reject H0. This indicates that a linear relationship exists between Skin Thickness and Body Mass Index.

Now we'll use confidence intervals. We'll start by constructing and interpreting a 95% confidence interval for the unknown true slope of the regression line.

```
β0 = [18.6064 21.6873]
β1 = [0.3942 0.4937]
```

Fig. 7. Computed confidence intervals for the unknown true slope and the y-intercept

According to fig.7 the confidence interval for β_1 is [0.3942, 0.4937]. This means we're 95% confident that the true slope of regression line lies between 0.3942 and 0.4937. For each additional millimetre of Skin Thickness, the Body Mass Index (Kg/m^2) increases between 0.3942 and 0.4937 points. Since β_1 = 0 is not contained in [0.3942, 0.4937], we are 95% confident of significance in linear relationship between Skin Thickness and BMI.

Constructing and interpreting a 95% confidence interval for the mean of the y-variable at a fixed value of our choice of the other variable ($x_{p=105}$).

```
For xp = 30.0 yp = 30.0, ȳ = [2.7499 4.0125]
```

Fig. 8. Computed confidence interval for the mean of the y variable at $x = 30$

We're 95% confident that the mean Body Mass Index of all Pima women with a Skin Thickness of $30mm$, lies between 32.727 and 33.9896 kg/m^2.

```
For xp = 30.0 with yp = 30.0, ȳ = [-6.878  13.8419]
```

Fig. 9. Computed confidence interval for the value of the y variable at $x = 30$

We're 95% confident that the Body Mass Index of a randomly selected Pima Woman with a Skin Thickness of $30mm$, lies between 23.0991 and 43.8191 kg/m^2.

This is a very wide interval containing values that represent very different body types ranging from the underweight range to the obese range [3]. For this reason we consider that this prediction interval is not useful.

B. Principal Component Analysis

To start the Principal Component Analysis we must choose 3 features to analyse. We'll analyse the same two features used in the Regression Analysis, the Skin Thickness and the Body Mass Index, and add the Insulin level as well because we believe we might be able to perform effective dimensionality reduction with these 3 features or score a hidden factor.

First we'll start by visualizing the data over these features in 2D/3D PC plane using normalization by range or by standard deviation.

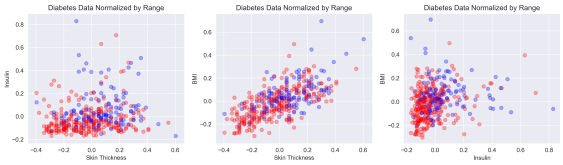

Fig. 10. Visualization of the data over the chosen features in 2D PC plane using normalization by range

Fig. 11. Visualization of the data over the chosen features in 2D PC plane using normalization by Z-Score

After both normalizations Insulin level has high correlation between the Skin Thickness and the Body Mass Index (what is shown on fig.10 and fig.11) what leads to the hypothesis that we can summarize this feature in effective way with PCA.

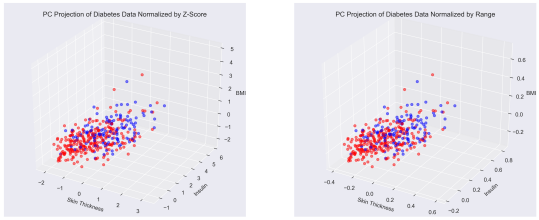

Fig. 12. Visualization of the data over the chosen features in 3D PC plane using normalization by Z-Score and Range

We'll continue the visualization of our data using the Single Value Decomposition with different data normalizations and choose one that we find best.

Fig. 13. SVD projection and SVD US of the data normalized by Range

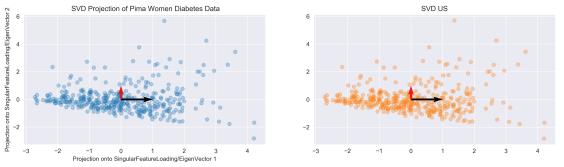

Fig. 14. SVD projection and SVD US of the data normalized by Z-Score

As we can see both visualizations (fig.13 and fig.14) are very similar despite the different normalizations. This being said, we think that the Z-Score normalization is better because it satisfies the intuitive equality principle – all features contribute to the data scatter equally after they have been divided by their standard deviations [4].

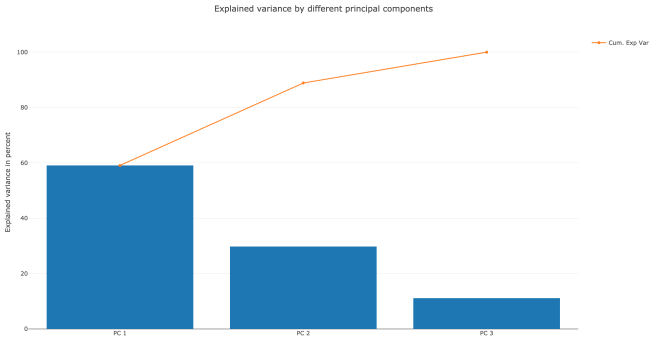

Fig. 15. Explained Variance by different principal components

In the next visualization, the data was divided into two groups. One for the diabetic women and another for the non diabetic women. The diabetic women are represented by the orange circles while the healthy women are represented by the blue plus signs. We chose these two groups because they're already present in our data.

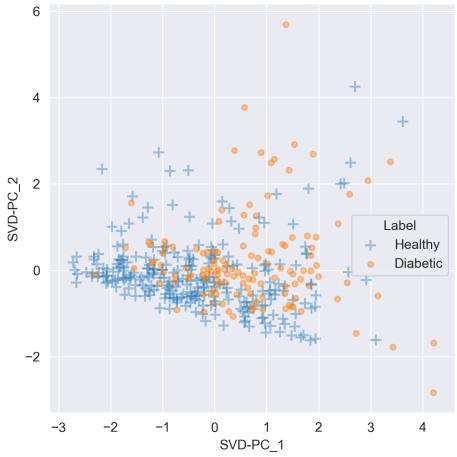

Fig. 16. SVD Visualization with a pre-specified group, the diabetic women and the non diabetic women

Below we present the graphical presentation of the "quality" of the PC projection of the data.

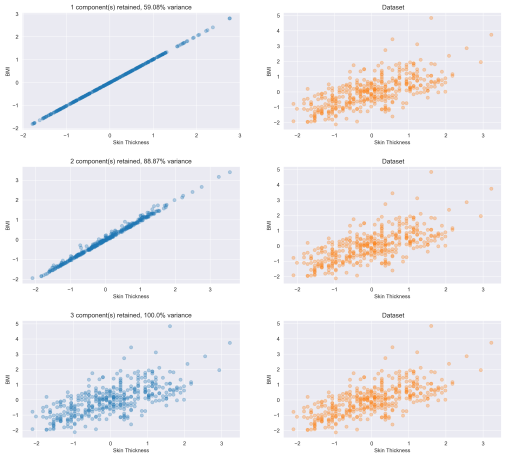

Fig. 17. Retained Principal Components and the Cumulative Explained Variance

We can see that the quality of the PC projection is quite good since with 2 components it's possible to cover 88.89% of the variance.

The last step was to compute a hidden factor that can be used on our data to measure the impact of each feature in the positive outcome of diabetes. To do this we applied the Single Value Decomposition to find the first singular triplet. Then we used the equation $z = Xc\alpha$ to rescale z to a scale of 0-100, obtaining the following equation:

$$Z = 0.09 * SkinThickness + 0.78 * Insulin + 0.11 * BMI$$

Compared to the mean, the weight of the Skin Thickness and the BMI greatly decreased while the weight of the Insulin greatly increased. This means, as we know, that the level of Insulin has a very big impact in the outcome of having diabetes.

C. *Fuzzy Clustering with Anomalous Patterns*

We'll start by clustering the data with the Fuzzy C-Means algorithm. The number of clusters obtained is a parameter chosen by us. In order to choose it we'll apply the algorithm to our data with several different c values to find which number of clusters better fits our data. The classification metric used for this purpose was the fuzzy partition coefficient that can be define by:

$$F_c(\underset{\sim}{U}) = \frac{tr(\underset{\sim}{U} * \underset{\sim}{U}^T)}{n}$$

where $\underset{\sim}{U}$ is the fuzzy partition matrix being segregated into c classes (partitions), n is the number of data sets, and the operation "*" is standard matrix multiplication. The product $\underset{\sim}{U} * \underset{\sim}{U}^T$ is a matrix of size c × c [5].

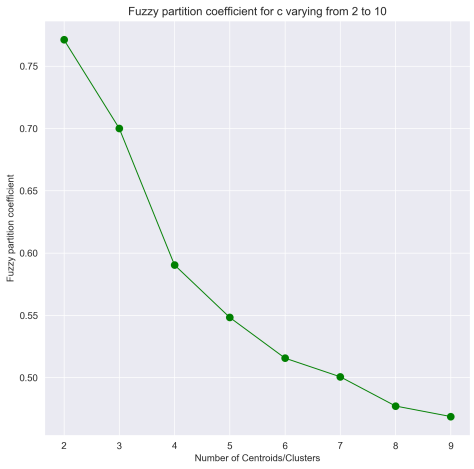

Fig. 18. Graphic of the Fuzzy Partition Coefficient vs the Number of Clusters

As $F_c(\underset{\sim}{U})$ increases, the decomposition of the data sets into the classes chosen is more successful. Looking at the plot we can conclude that the best number of clusters is 2 because it's the c value with the highest fuzzy partition coefficient value.

The Fuzzy C-Means algorithm has another parameter which is an initial fuzzy c-partitioned matrix. When this matrix isn't provided by the user, then it is randomly computed. We'll implement the Anomalous Patterns algorithm and use the resulting clusters as the initial fuzzy c-partitioned matrix for the FCM. This is the Anomalous-Patterns Fuzzy C-Means.

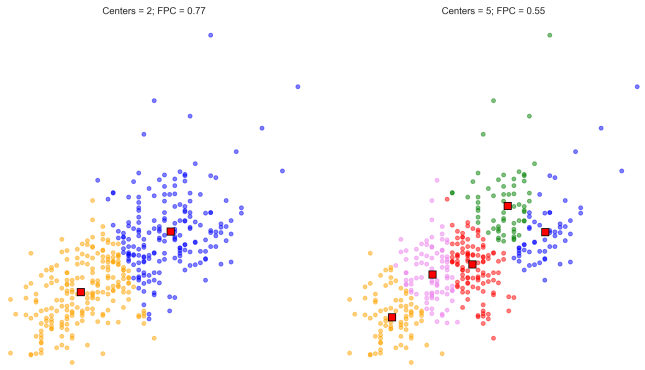

Fig. 19. Graphic of the clusters found by the Fuzzy C-Means and the Anomalous-Patterns Fuzzy C-Means

The initial prototypes are far away from the reference point with some being closer to it as expected with the evolution of the algorithm. There are a few near the center given the natural shape of our data because as we can see above most of the points are all concentrated in one particular area without many groups of points presenting an anomalous behaviour and being far from the grand mean. The number of clusters was chosen based on a threshold applied to a cluster's cardinality, in this case we decided that every cluster must have at least cardinality $N_k > 20$.

To perform clustering validation two internal validation indexes, the Adjusted Rand Index and the Xie-Beni Index were used. The ARI has a range of $[0, 1]$ with 1 being when the cluster is close to the true labels. This means we want to maximize the ARI score and minimize the XB score.

```
Fuzzy C-Means Adjusted Rand Index score: 0.0736
Fuzzy C-Means Xie-Beni score: 0.1377
Anomalous Pattern Fuzzy C-Means Adjusted Rand Index score: 0.022
Anomalous Pattern Fuzzy C-Means Xie-Beni score: 0.2566
```

Fig. 20. Adjusted Rand Index and Xie-Beni

As we can see in fig.20, both clusters didn't achieve very good scores in either index but the cluster with 2 centers found by the Fuzzy C-Means did slightly better in both scores achieving 0.0736 on the Ajusted Rand Index and 0.1377 on the Xie-Beni index against the cluster with 5 centers found by the AP-FCM that scored 0.022 on the ARI and 0.2566 on XBI. These scores can be explained by the fact that the FCM is biased towards globular shaped clusters and our data is one big globular like shape. Any number of clusters above 1 will deteriorate their quality as seen in fig.18. Because the AP algorithm found 5 clusters and we used them to initialize the FCM the results will be affected from the beginning.

The cluster found by the Anomalous Patterns Fuzzy C-Means wasn't very good. As discussed before the natural shape of our data doesn't present any anomalous patterns, having most of the points all concentrated in one particular area near the grand mean. The only anomalous patterns visible in our data are a few outsiders that aren't enough to build a cluster on their one and that get cut once a threshold is applied.

So it's expected for the AP-FCM algorithm to have similar performance to the original FCM because there's no benefit to looking for Anomalous Patterns in a dataset where there are none present.

III. Conclusion

Overall the Data Analysis techniques used in this project revealed interesting information about our data. We were able to conclude that there is a Liner Regression relationship between the Skin Thickness and the Body Mass Index of the Pima Indian Women and that we're able to perform inference on this regression model. We were also able to find there's a strong correlation between the Skin Thickness, BMI and Insulin, allowing us to only use 2 principal components and retain around 88.80% of the total explained variance. Last, we performed a clustering analysis on our data using the Anomalous Patterns and the Fuzzy C-Means algorithms and unfortunately didn't obtain very useful results but were able to understand why these algorithms didn't perform well given the nature of our data set.

IV. Appendix

Heads of the developed functions for Anomalous Cluster Algorithm

anomalous(X, me, rang, D) - Anomalous, iterative extraction of anomalous clusters based on the algorithm referred to as Separate/Conquer (Mirkin, 1999, Machine Learning). Input: X - data matrix, me - grand mean, range - normalizing values, D - normalised data scatter. Output: ancl[ind, 1] list of indices in cluster, ancl[ind, 2] standardised center, ancl[ind, 3] contribution to the data scatter.

anpat(X, remains, rang, centroid, me) -Iterative soubroutine in anomalous based on the algorithm 'Separate/Conquer' (Mirkin, 1999, Machine Learning). Input: X - full data matrix, remains - set of its row indices, range - normalizing values, centroid - initial center of the anomalous cluster, me - vector to shift to the 0 (origin). Output: cluster - set of row indices in the anomalous cluster, centrod - center of the cluster.

center(X, cluster) - Finding centroid to a cluster.

separc(X0, remains, rang, a, b) - Separating a cluster around 'a' from that around 'b'

distm(X,remains,rang,a) - Finding normalized distances in 'remains' to point 'a'. Input: X - the original data matrix, remains - set of X-row indices under consideration, range - normalizing vector, a - point the distances relate to. Output: distan - column of distances from a to remains.

dist(X, remains, stand, me) - Finding normalized distances in 'remains' to point 'a'. Input: X - data matrix, remains - data still to cluster, stand - values to standardize, me - array point to calculate dist. Output: distan - column of distances from a to remains.

post_process_ANCL(ancl, threshold=10) - Post-processing of anomalous clusters. Removal of small anomalous clusters. Input: ancl - anomalous clusters, threshold - the maximum cardinality of a small cluster. Output: new_ancl - new anomalous clusters.

References

[1] Wikipedia contributors. Pima people. Wikipedia, The Free Encyclopedia. March 25, 2019, 18:47 UTC. Available at: https://en.wikipedia.org/w/index.php?title=Pima_people&oldid=889443834. [Accessed April 27, 2019]

[2] Larose, T. Larose, C. (2015). Data Mining and Predictive Analytics, Wiley Series on Methods and Applications in Data Mining, Wiley, Chapter 8

[3] What is the body mass index (BMI)?. NHS-UK. Page last reviewed: 12/07/2016. Available at: https://www.nhs.uk/common-health-questions/lifestyle/what-is-the-body-mass-index-bmi/ [Accessed April 27, 2019]

[4] Boris Mirkin. Core Concepts in Data Analysis: Summarization, Correlation and Visualization, Undergraduate Topics in Computer Science, Springer-Verlag London

[5] Timothy J. Ross. Fuzzy Logic with Engineering Applications, Third Edition, Wiley, 2010, 3rd edition, page 358.

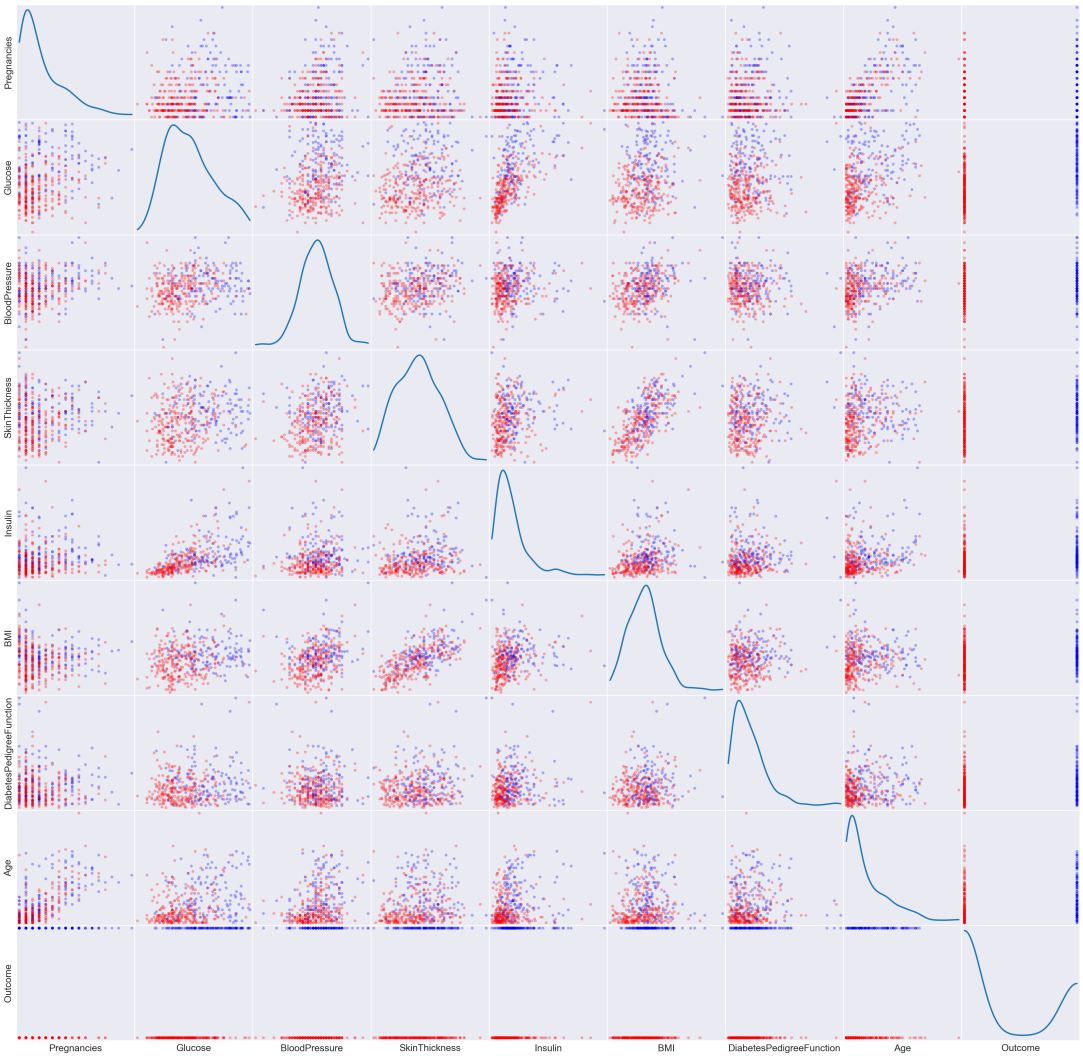

Fig. 21. Scatter Matrix of the dataset