Project 13 (Diabetes Prediction)
Assignment 1 - Data Analytics

Simone Tarenzi

April 12, 2024

1 Introduction

This assignment [1] required the analysis of a set of medical data of a number of people and successive prediction of diabetes on the subjects based on the information given.

In the following sections I will introduce the dataset, visualize and analyze the data using various functions and plots, inspect the correlation between the various features, explain how I prepared the data to create a prediction model, and finally use my findings to discuss some conclusions.

2 Data Analysis

After reading the Kaggle page's description of the project's dataset, I initially thought that the data would have been more modern and representative of today's health statistics, but looking at some of the entries of the dataset, I could instantly see that was not the case.

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0
768	0	123	77	0	1	36.3	0.252	55	1

Figure 1: First and last 5 entries.

	count	mean	std	min	25%	50%	75%	max
Pregnancies	769.0	3.840052	3.370237	0.000	1.000	3.000	6.000	17.00
Glucose	769.0	120.897269	31.951886	0.000	99.000	117.000	140.000	199.00
Blood Pressure	769.0	69.115735	19.345296	0.000	62.000	72.000	80.000	122.00
Skin Thickness	769.0	20.509753	15.959020	0.000	0.000	23.000	32.000	99.00
Insulin	769.0	79.697009	115.203999	0.000	0.000	29.000	127.000	846.00
BMI	769.0	31.998179	7.880557	0.000	27.300	32.000	36.600	67.10
Diabetes Pedigree Function	769.0	0.471590	0.331208	0.078	0.244	0.371	0.626	2.42
Age	769.0	33.269181	11.778737	21.000	24.000	29.000	41.000	81.00
Outcome	769.0	0.349805	0.477219	0.000	0.000	0.000	1.000	1.00

Figure 2: Overall statistics of the dataset.

First of all, the average number of pregnancies being 3.84 is unusually high for today's metrics, with only 30 countries in the world as of 2023 [2] surpassing it, and its maximum being 17 is also very interesting. The percentage of diabetic patients is also very high (35% compared to the world's average [3] of 10.5% as of 2021). However, by investigating more I found the actual source of this dataset, with more information about the data's origin, which explained all of my doubts but was somehow absent on the Kaggle's page given by the project's link.

2.1 Data Description

The dataset [4] is originally from the National Institute of Diabetes and Digestive and Kidney Diseases, and is a widely used dataset in machine learning and healthcare research [19]. It contains various attributes such as age, body mass index (BMI), blood pressure, and glucose levels, among others, of patients diagnosed with or at risk of diabetes. 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 near Phoenix, Arizona, and their medical data was collected through periodic standardized examinations every 2 years since 1965. The dataset is commonly employed for tasks like predictive modeling to determine the likelihood of an individual developing diabetes based on their characteristics. Researchers and data scientists utilize this dataset to develop and validate algorithms for early detection, risk assessment, and personalized treatment strategies for diabetes management.

For the following visual representations I decided to just remove all 0 value entries, but I had to handle them in another way for our prediction model. I will expand more on that subsequently.

Age: Expressed in years. As previously stated, all patients in the dataset are females of at least 21 years old, with the oldest being 81 years old. Most of the occurrences are in the 21 to 30 years old range, with an expected steady decrease the higher the age becomes.

Pregnancies: As mentioned earlier, the average number of pregnancies is unusually high, with 44.73% of patients having an amount over the median of 3. The origin of the data explains the reason for this, since the Pima is a tribe that had a very high birth rate.

Glucose: Plasma glucose concentration at 2 hours after eating in an oral glucose tolerance test, also known as two-hour postprandial glucose test, expressed in milligrams per decilitre (mg/dL). The histogram (Figure 3) shows an unusual distribution (Figure 4) of low and normal glucose levels ([5] and [6]) given the high amount of diabetic patients.

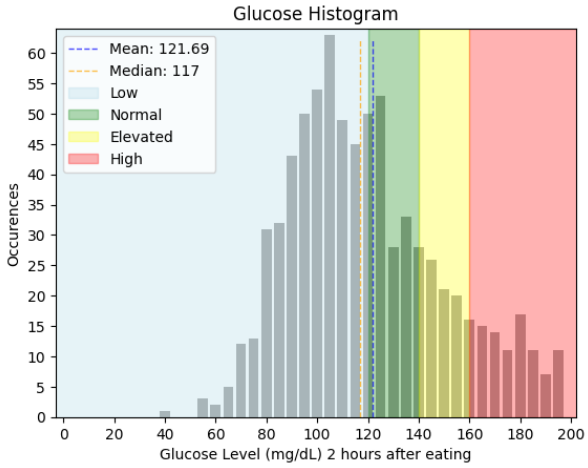

Figure 3: Histogram of glucose concentration occurrences.

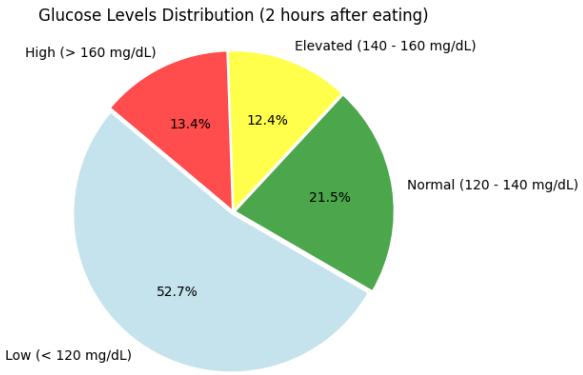

Figure 4: Pie chart of glucose concentration levels distribution.

Blood Pressure: Diastolic blood pressure, meaning the measurement is taken when the heart expands, expressed in millimeters of mercury (mmHg). The histogram (Figure 5) shows a more expected distribution (Figure 6) of blood pressure levels ([7], [8], [9] and [10]) given the outcome, since high blood pressure is known to be closely related with diabetes [11], with 46% of occurrences being within the elevated to hypertensive crisis range.

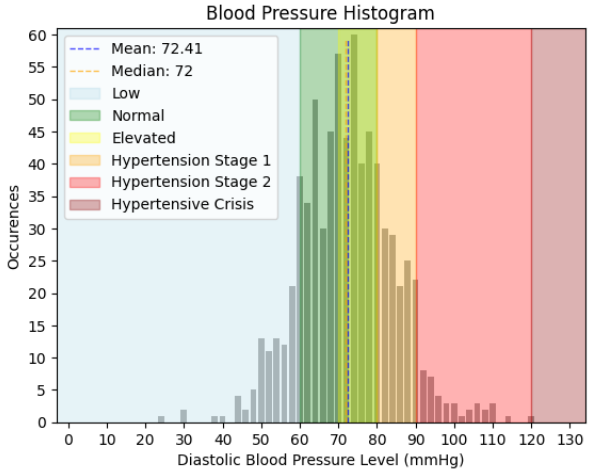

Figure 5: Histogram of blood pressure occurrences.

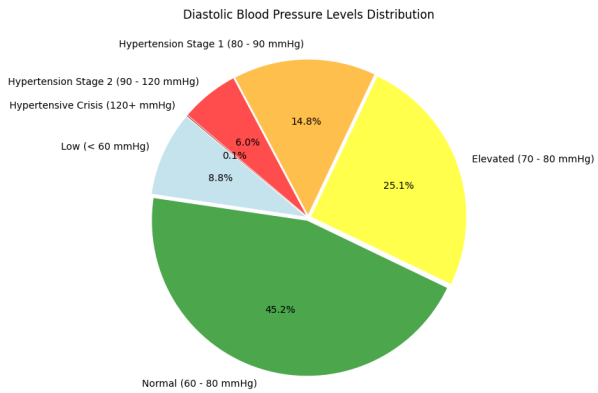

Figure 6: Pie chart of blood pressure levels distribution.

Skin Thickness: Triceps skin fold thickness, expressed in millimeters (mm). Nothing really interesting to say here, just that it (predictably) closely matches BMI's histogram which I will show subsequently.

Insulin: Serum insulin level 2 hours after eating, expressed in micro units per millilitre (μU/ml). The histogram (Figure 7) shows a very peculiar distribution (Figure 8), with 34.4% of patients having abnormal ([12] and [13]) levels, either lower than 16 μU/ml (with one patient having an absurdly low insulin level of 1 μU/ml) or higher than 166 μU/ml (one patient having an absurdly high insulin level of 846 μU/ml). These strange measurements might be simply explained by user-error, but this is data which was continuously updated by the National Institute of Diabetes and Digestive and Kidney Diseases [19], so the presence of this kind of mistake seems very unusual. Nonetheless, I decided to keep this data which did not seem to affect our prediction model that much.

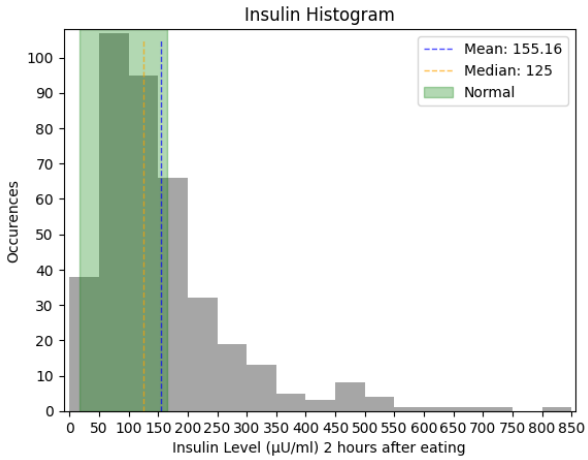

Figure 7: Histogram of insulin level occurrences.

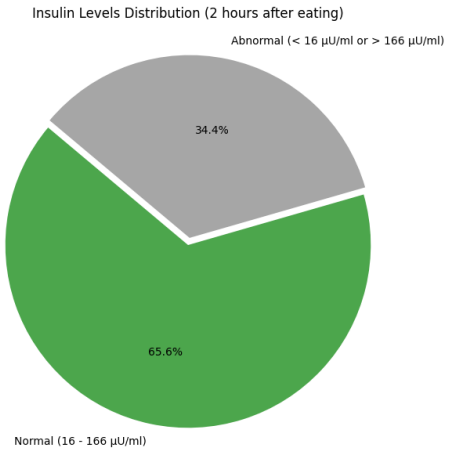

Figure 8: Pie chart of insulin levels distribution.

BMI: Body mass index, expressed in weight divided by height in squared meters (kg/m^2). The histogram (Figure 9) shows a very high distribution (Figure 10) of patients within the overweight to obesity class 3 range ([14] and [15]). Understandable and expected given the population's traditional diet which is high in carbohydrates [18].

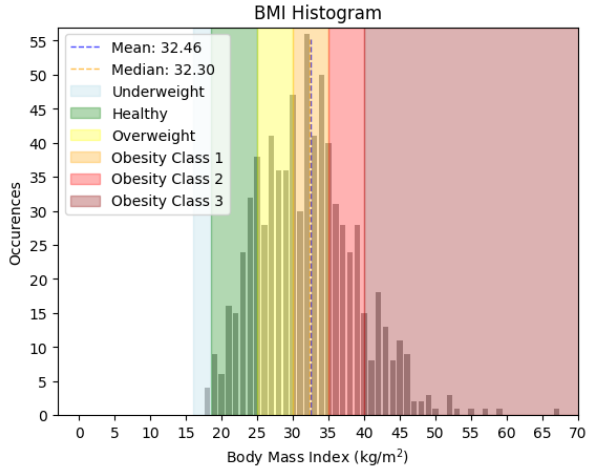

Figure 9: Histogram of BMI measurement occurrences.

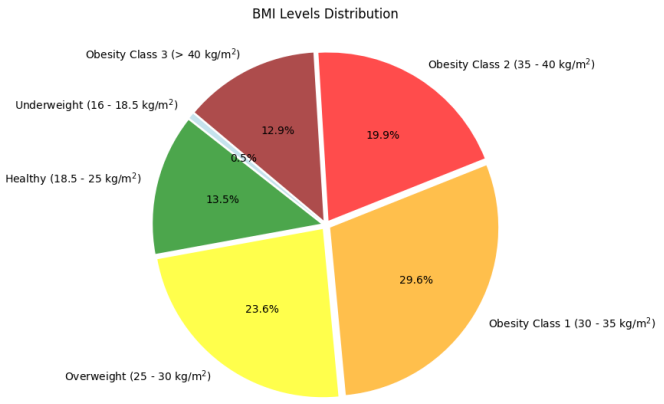

Figure 10: Pie chart of BMI levels distribution.

Diabetes Pedigree Function: Calculates diabetes likelihood depending on the subject's age and on diabetic family history, with values going from 0.078 to 2.42. I will expand more on this later.

Outcome: Expressed as 1 for diabetic patients and as 0 for non-diabetic patients. 500 subjects (65%) were healthy, while 269 (35%) of them were diagnosed with diabetes. As mentioned earlier, this is a very high percentage of diabetic people, but understandable given the history of the Pima Indian population [16].

2.2 Correlation Analysis

For the analysis of the correlation between features of the dataset, I used a pair plot (Figure 11), and a correlation matrix (Figure 12). I suggest to look at the Jupyter Notebook [1] to see the images in full resolution.

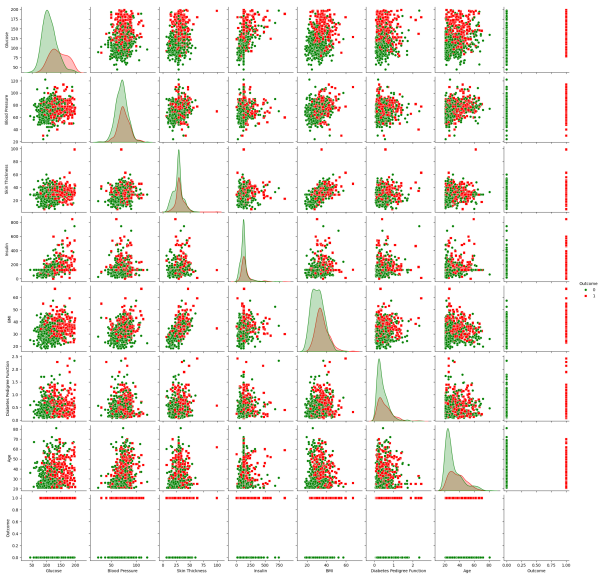

Figure 11: Pair plot of the features of the dataset.

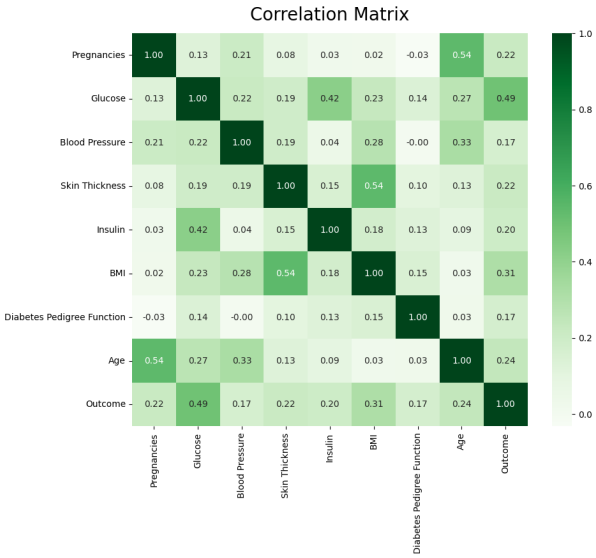

Figure 12: Correlation matrix of the features of the dataset.

The correlation efficient between features is mostly expected, for example it has a value of 0.54 between BMI and Skin Thickness, 0.42 between Glucose and Insulin, or 0.54 between Age and Pregnancies, but the most interesting one is between the Diabetes Pedigree Function and Age/Outcome. As it is explained in the dataset, this function is calculated depending on the subject's age and on history of diabetes in their family, so it is very peculiar for its correlation efficient to be 0.03 and 0.17 for Age and Outcome, respectively. However, as I will show next, the Diabetes Pedigree Function is an important feature to enable in order to accurately predict diabetes.

3 Prediction Model

3.1 Data Preparation

Before using the model, I had to prepare the data by substituting all the 0 values mentioned before, otherwise it would have too little data to work on. To decide how to substitute the values, I calculated each feature skewness: with a skewness higher than 0.5 or lower than -0.5, it is good practice to substitute the missing values with the median value, while otherwise, it is better to substitute them with the mean [17]. This method is called imputation and it greatly helps when using prediction models with few data entries.

3.2 SVC (Support Vector Classifier)

The model I chose is the Support Vector Classifier (SVC), which is a pretty simple model but perfect for data with binary classification such as this. I decided to use it with a 80-20% data split and iteratively reduce the number of selected features one by one, ordered by their correlation efficient with the outcome. For simplicity purposes, I only include the confusion matrices for the model with all features enabled (Figure 13), and the one with all but Blood Pressure enabled (Figure 14) in this report.

As it can be seen from the plot (Figure 15), and from the ROC curve of the best model (Figure 16), the best scores were obtained when using all features but Blood Pressure, which got an overall accuracy score of 82% (all scores available in the Jupyter Notebook [1]). Even though blood pressure and diabetes are common comorbidities [11], it is curious that removing the Blood Pressure feature from the model improved the accuracy. This is a very interesting outcome that shows that having more data to work with not always results in better predictions, at least in this specific dataset. Perhaps this is the Bonferroni's principle at work.

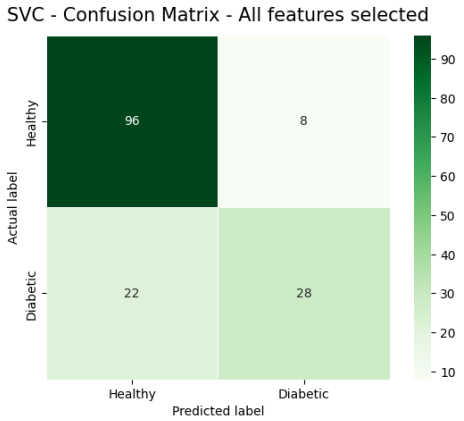

Figure 13: Confusion matrix of the SVC model with all features enabled.

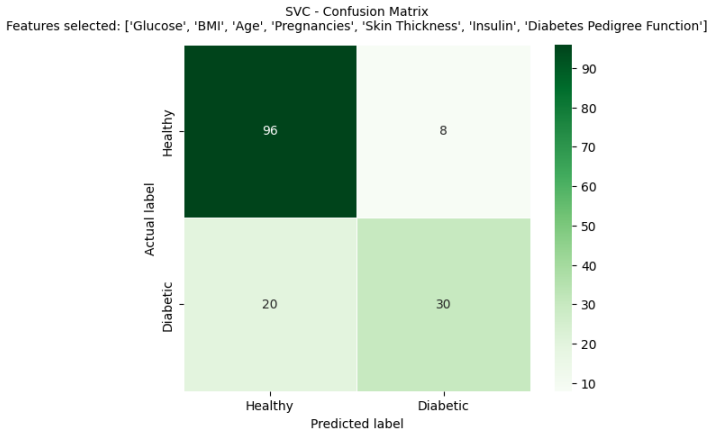

Figure 14: Confusion matrix of the SVC model with Glucose, BMI, Age, Pregnancies, Skin Thickness, Insulin and Diabetes Pedigree Function features enabled.

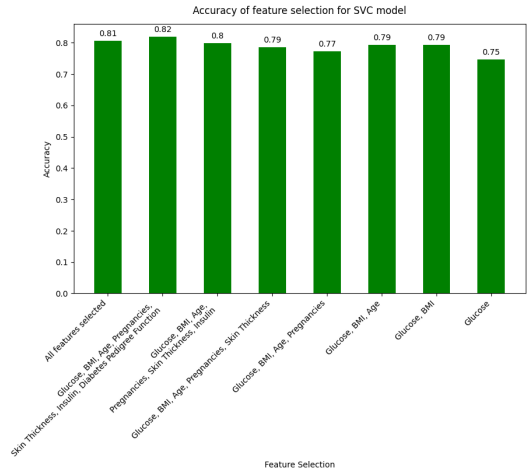

Figure 15: Plot of feature selection accuracy scores.

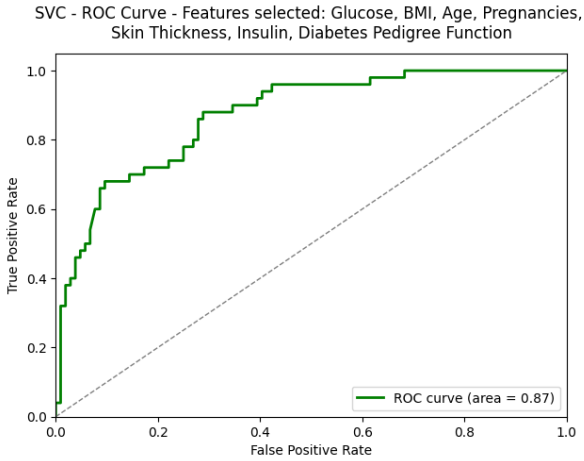

Figure 16: ROC curve of best model.

4 Conclusions

- The dataset came from a very interesting medical study of the Pima Indian population of Arizona, which is a fascinating research into the dietary preferences of this Native American tribe.
- The dataset's composition, which includes a variety of health attributes such as BMI, blood pressure, and glucose levels, provides a comprehensive view of the health status of the Pima Indian population. This makes it an invaluable resource for studying the complex interplay of these factors in the context of diabetes.
- High (and low) correlation scores do not always directly translate to better (or worse) feature selection for prediction models, as I saw with the Diabetes Pedigree function.
- The Support Vector Classifier (SVC) was an effective model for this binary classification task, given its more than respectable 82% accuracy score and ROC AUC score of 0.87 when selecting the right features.
- Even though more data usually helps with predictions, too many features can have an adverse effect on the accuracy of the model. However, the same can be said about selecting too few, so the right balance must be reached.

References

[1] https://github.com/Bluxen/DA-Project1.

[2] https://en.wikipedia.org/wiki/List_of_countries_by_total_fertility_rate.

[3] https://idf.org/about-diabetes/diabetes-facts-figures/.

[4] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[5] https://www.lark.com/resources/blood-sugar-chart.

[6] https://www.verywellhealth.com/blood-sugar-levels-after-eating-5118330.

[7] https://www.cdc.gov/bloodpressure/about.htm.

[8] https://www.singlecare.com/blog/blood-pressure-levels/.

[9] https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings.

[10] https://www.webmd.com/hypertension-high-blood-pressure/diastolic-and-systolic-blood-pressure-know-your-numbers.

[11] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5953551/.

[12] https://pharmeasy.in/diagnostics/tests/insulin-pp-162.

[13] https://emedicine.medscape.com/article/2089224-overview?form=fpf.

[14] https://www.cdc.gov/obesity/basics/adult-defining.html.

[15] https://www.researchgate.net/figure/International-classification-of-adult-underweight-overweight-and-obesity-according-to_tbl5_323009432.

[16] https://diabetesjournals.org/diabetes/article/64/12/3993/34762/Dissecting-the-Etiology-of-Type-2-Diabetes-in-the.

[17] https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/#:~:text=Mean%20imputation%20is%20often%20used,to%20outliers%20than%20the%20mean.

[18] Boyce, V. L., and Swinburn., B. A. The traditional pima indian diet. composition and adaptation for use in a dietary intervention study. *Diabetes care vol. 16,1* (1993), 369–71.

[19] Smith, J., Everhart, J., Dickson, W., Knowler, W., and Johannes, R. Using the adap learning algorithm to forcast the onset of diabetes mellitus. *Proceedings - Annual Symposium on Computer Applications in Medical Care 10* (11 1988).