(Page count starts here & Minimum font size – 11; Maximum font size - 12)

Introduction

1. Data Collection

This report looks into the statistics and analysis of diabetes among the Native American Pima Indian People, specifically, females aged at least 21 years old.

The dataset contains Independent and dependent variables of integer and float types. The one dependent variable is classified as the Outcome in which—therein is represented by a "1" or "0" to indicate whether or not that person has diabetes. The independent variables are known as predictor variables as they have been medically considered to be of a possible link to being a diabetic, these independent variables form a co-dependency in which we hope to be able to link to the outcome to be able to explore and analyse the links within this dataset to better prevent, predict and combat diabetes in relation to the females of the Pima people.

The Independent variables:

1. Number of past pregnancies - Discrete
2. Blood glucose - Continuous
3. Diastolic blood pressure (mm Hg) - Continuous
4. Triceps skin fold thickness (mm) - Continuous
5. 2-Hour serum insulin (mm) - Continuous
6. Body mass index (kg) - Continuous
7. Diabetes pedigree function - Continuous
8. Age (years) - Discrete

The Dependent variable:

1. Outcome (Diabetic) - Discrete
    a. 0 = Nondiabetic
    b. 1 = Diabetic

2. Data Observation

Upon investigating this dataset, it was clear that the dataset contained data records in which the elements for that record were an entry error or missing value and replaced with a "0" during the data collection process.

Due to this dataset having originated from the National Institute of Diabetes and Digestive and Kidney Diseases and thus undergoing data collection and processing, via a third party to prepare the Data to specifically be constrained to the females of the Pima population 21 years of age and above. This has resulted in a lot of irregularities within the data records for

the columns of the specified scientific independent variable. These irregularities of our first findings within the data records are present heavily in the BMI, Insulin, Blood pressure and Skin Thickness columns.

Exploratory Data Analysis and Results

1. Data Cleaning

1.1. Handling missing values and errors

Handling missing data or null values is the most important part of data cleaning as it allows higher quality information for decision-making. For the dataset, these are the factors that would need to be considered for handling data cleaning:

- Dropping rows or columns
- Checking data type of data
- Replacing null values with 'Unknown' or 0
- Replacing missing values with predicted values
- Dealing with outliers
- Deleting duplicates

It was decided that instead of dropping rows or columns that had missing values or errors as a large percentage of the data would be deleted. Rather than dropping entire rows, we dropped specific elements that fit the NA / Outlier / '0' criteria and then these dropped elements were replaced with the best calculated predicted values depending on the Zscore and skewness of the data.

1.2. Check datatype of data

When checking each data in the dataset, data needs to be ensured that they have the right data type. For example, age must be type 'int' because a person's age can't be a fraction. From checking the data type of the dataset, it is clear to see that none of the data have obvious incorrect data type.

```
Original
----------
Pregnancies                 int64
Glucose                     int64
BloodPressure               int64
SkinThickness               int64
Insulin                     int64
BMI                       float64
DiabetesPedigreeFunction  float64
Age                         int64
Outcome                     int64
==================================
```

Figure 1 - Datatype of raw data

1.3. Filling in missing values

Instead of dropping the rows or columns that have missing value or errors, the mean or median replaces the columns with the missing values or errors. The mean would be used if the data is more centralized, while the median would be used if the data is

a skewed distribution as it is better than mean because it is not heavily impacted by large values.

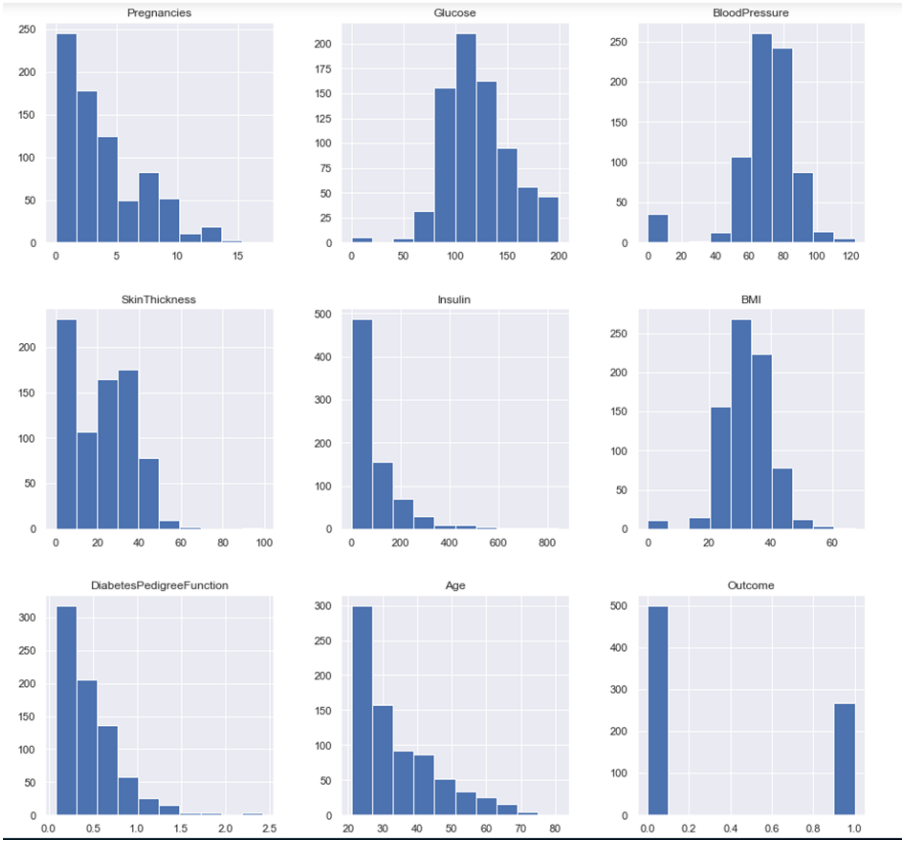


Figure 2 - Shape of data

For this dataset, 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', and 'BMI' are data that would be impossible for them to have the value 0. This can be solved by replacing those zeros with the mean or medium as mentioned before.

1.4. Checking for outliers

Checking outliers is important as it helps determine which values differ significantly from other observations or incorrect values within the dataset. The method that was used to define whether a value is an outlier or not was using z scores. If a z score value is less than -3 or greater than 3, that value would be an outlier. Those outliers then will be replaced with the mean or median of the data.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	0.639947	0.848324	0.149641	0.907270	-0.692891	0.204013	0.468492	1.425995	1.365896
1	-0.844885	-1.123396	-0.160546	0.530902	-0.692891	-0.684422	-0.365061	-0.190672	-0.732120
2	1.233880	1.943724	-0.263941	-1.288212	-0.692891	-1.103255	0.604397	-0.105584	1.365896
3	-0.844885	-0.998208	-0.160546	0.154533	0.123302	-0.494043	-0.920763	-1.041549	-0.732120
4	-1.141852	0.504055	-1.504687	0.907270	0.765836	1.409746	5.484909	-0.020496	1.365896

Figure 3 - Z Score of data

1.5. Others

As a result of replacing missing values or errors with mean or medium, the datatype of the dataset would be changed as the mean or medium will most likely be a decimal number.

```
Imputed
Pregnancies                 float64
Glucose                     float64
BloodPressure               float64
SkinThickness               float64
Insulin                     float64
BMI                         float64
DiabetesPedigreeFunction    float64
Age                         float64
Outcome                       int64
```

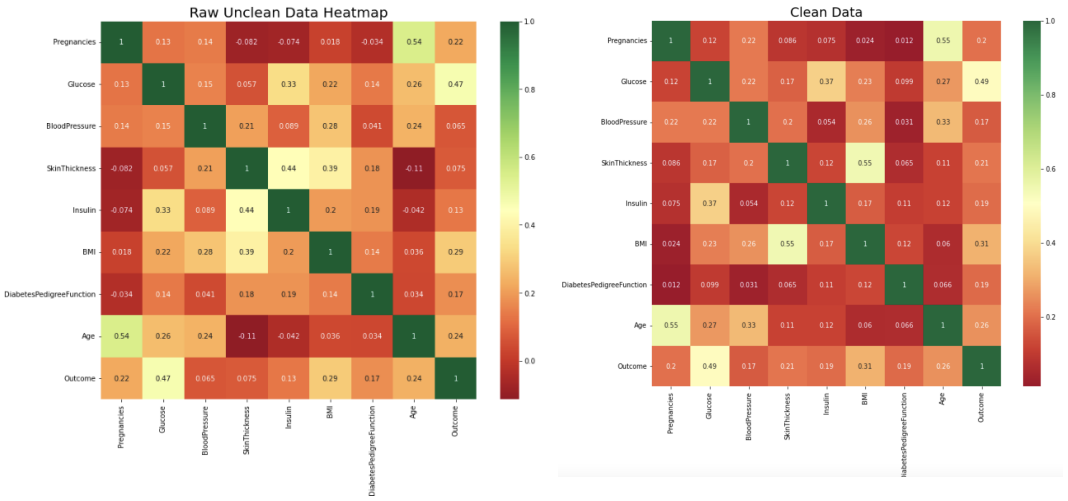
A solution to this problem is to round the mean or median to the nearest whole number.

2. Data Exploration Analysis

2.1. Principal Component Analysis (PCA)

Principal component analysis (PCA) is a fast and flexible unsupervised method for dimensionality reduction in data. PCA allowed us to merge all 8 variables into 2. Flattening the data simplifies the complexity in the high-dimensional data while retaining trends and patterns.

2.2. Correlation Heatmap



Interpreting the heatmap, we can see each variable correlates with one another to different degrees. The diagonal, red line across the centre, is a result of each feature matching up with itself, hence yielding 100% correlation and should be disregarded. After cleansing the data of outliers and imputing correctly averaged elements into the required field, we then checked the heatmap on the clean data and found that due to how we handled the data, the correlation between these highly correlated independent variables greatened or lessened, and thus the results show as for:

1. BMI and Skin Thickness = 0.55 (was 0.39 - increase of 0.11)
2. Insulin and Skin Thickness = 0.12 (was 0.44 - reduction of 0.32)
3. Age and Pregnancies = 0.55 (was 0.54 - increase of 0.01)
4. Glucose and Outcome = 0.49 (was 0.47 - increase of 0.02)
5. Insulin and Glucose = 0.37 (was 0.33 - increase of 0.04)

Correctly cleaning the data allowed us to obtain a stronger correlation of dependency within the variables of the dataset. The increase in correlation for Age / Pregnancies and Glucose / Outcome indicate that these independent variables have an important and heavily existent dependency on one another.

Independent variable pairs	Uncleaned correlation	Cleaned Correlation	Increase of correlation (%)
Glucose and Outcome	0.47	0.49	4.2%
Insulin and Outcome	0.13	0.19	46.1%
Insulin and Glucose	0.33	0.37	12.1%

From the above results from our cleaned data it is clear to say that Insulin / Glucose have a strong and increased correlation with one another which scientifically make sense, but that glucose (4.2% increase) and insulin (46.1% increase) also have a strong and increased correlation with the outcome of diabetes thus showing us that insulin there is a direct correlation between insulin/glucose while also there is strong correlation in those with higher levels resulting in diabetes.Most important takeaway from above is that insulin came to show it has a lot more correlation than first thought while also Glucose has the highest correlation although least increase with a correlation of 0.49.

Independent variable pairs	Uncleaned correlation	Cleaned Correlation	Increase of correlation (%)
BMI and Outcome	0.29	0.31	6.8%
Skin Thickness and Outcome	0.075	0.21	180%
BloodPressure and Outcome	0.065	0.17	161%

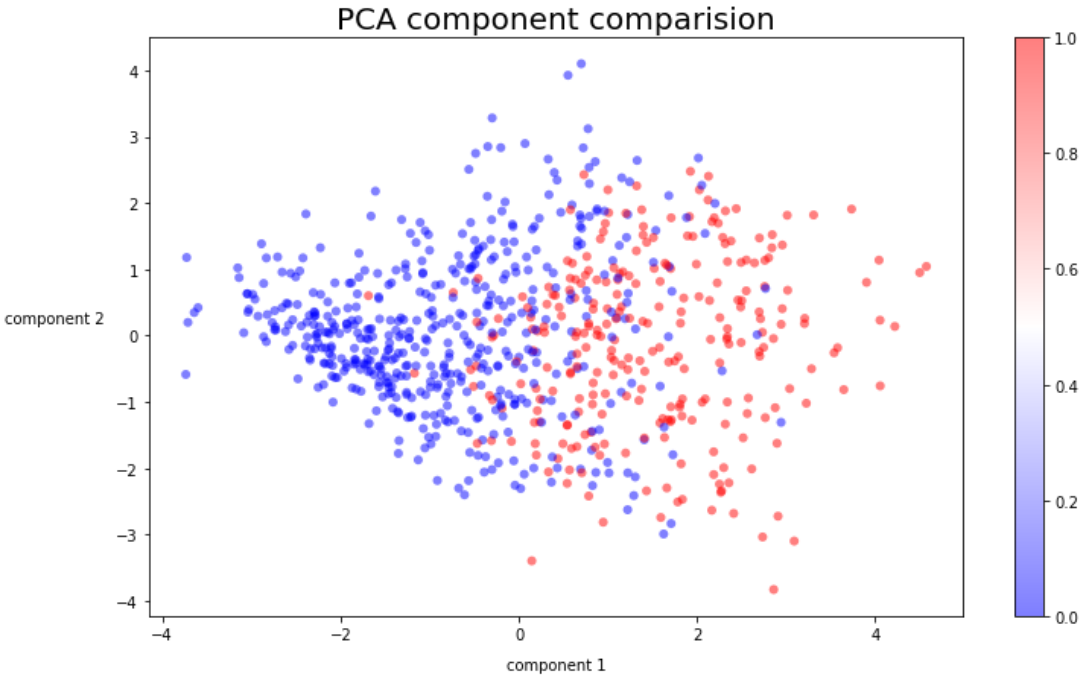
The results emphasize that BMI, skin thickness, and blood pressure not only correlate with each other, but also play a clear role in their direct correlation with diabetes results. Exercise to maintain "weight" focuses attention on BMI and its associated skin thickness and blood pressure, and directly reduces the three scientifically strongly increased independent variables This is important because the resulting reduction in pregnancy indicates that this is not just due to pregnancy, but may be due to a lifestyle around having so many pregnancies than cannot be seen in the dataset. Adjusted data shows that pregnancy is 55% correlated with age and BMI is 55% correlated with skin thickness.
As pregnancy increases, age tends to increase. Similarly, women's BMI correlates with skin thickness. Overweight (high BMI) women tend to achieve higher values on skin thickness tests and vice versa. Skin thickness readings tend to increase with increasing BMI.

2.3. PCA Component Analysis

Moreover, looking at the PCA component comparison. Which compares the two flattened variables on a scatter plot. Once we have reduced the dimensions from the top table to the one underneath, via fitting and transforming the data, we are left with two principal components.

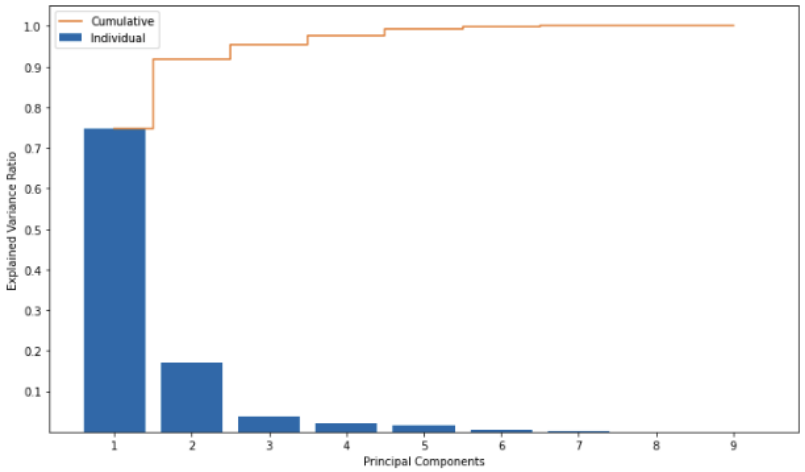
Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes pedigree Function	Age

Principle Component 1	Principle Component 2

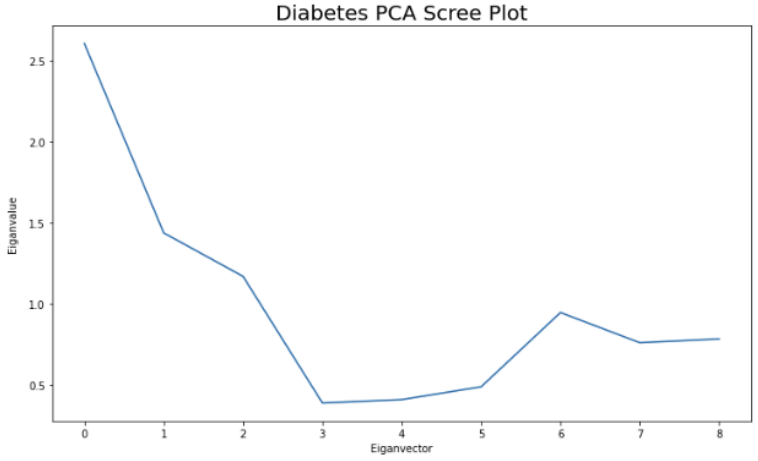


Using a scatter plot, we can see that the non-diabetic group tends to cluster when both components 1 is negative and to a lesser degree when component 2 is negative. Conversely, the diabetic group shows a trend of occurring frequently when

component 1 is positive. As this is an unsupervised learning environment, it shows potential that a model could predict the outcome of diabetes accurately.



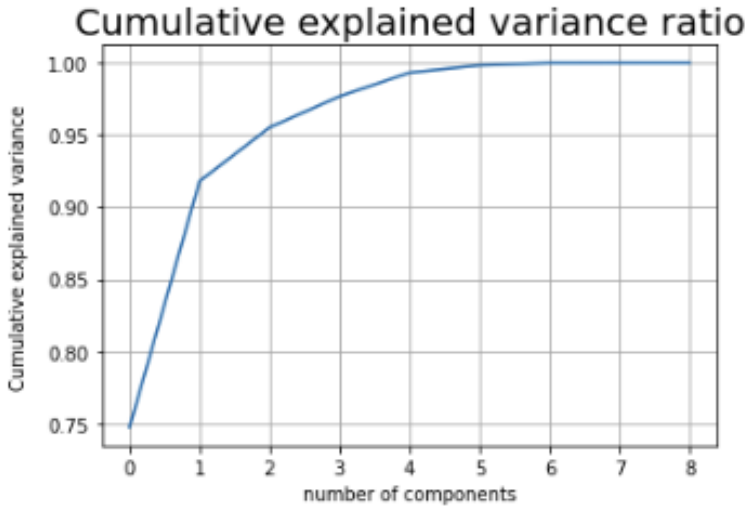
Eigenvectors and eigenvalues of a covariance matrix represent the basis of a PCA. The eigenvectors or principal components determine the directions of the new feature space, and the eigenvalues determine their magnitude. In other words, the eigenvalues explain the variance of the data. In the image below we can see the matrix of eigenvectors - as PCA reduces our dimensionality from 7 variables to 2.



The above plot is a scree plot, which takes the eigenvalues of principal components in an analysis. The plot is used to determine the number of principal components to keep in a PCA analysis. This graph is simply a visualisation of the eigenvector table above. The y-axis is the eigenvalues and the x-axis is the number of components or eigenvectors being used.

2.4. Cumulative Explained Variance Ratio

Now we will plot the components versus the cumulative variance. This will tell us how much impact each of the components has in the outcome.



One component alone counts for about 92% of the multi-dimensional variance. Four only being of marginal significance and the last three components being completely redundant.

2.5. Decision Tree

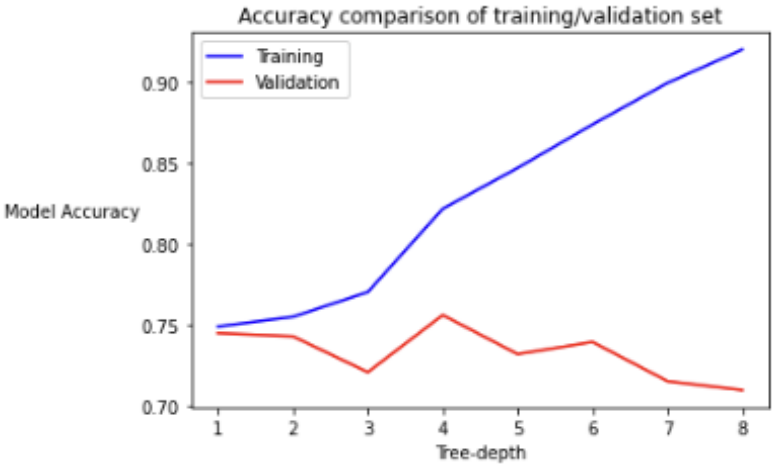
Moreover, we will look into implementing a machine learning model that can reasonably predict the probability of developing diabetes. We will split our data, 70% will be used for training and 30% will be used for testing. We will also set the maximum depth of the tree to 8.

Our training accuracy is 90.88% and our testing accuracy is 73.16%. We will implement 10-fold cross validation to find the average accuracy.

The average accuracy for each fold is

1	2	3	4	5	6	7	8	9	10
75.93%	77.78%	64.81%	61.11%	75.93%	66.67%	66.67%	67.92%	75.47%	77.36%

Now we will examine the tree size's effect on accuracy. We'll build 8 decision trees, with 8 different maximum depths. This is to reflect the 8 different variables we are using. Which excludes the outcome variable.
Mean accuracy - 70.96%



The highest accuracy is observed when the tree-depth is four. From there on, the model overfits the data. This is notable because earlier we saw that exactly four factors accounted for 100% of the cumulative variance with four factors being redundant. After implementing these factors and running again. Our Final prediction model could then 73.2% accurately predict diabetes based on four features.

Conclusions

1. Main purpose

The purpose of the experiment was to investigate and build a machine learning model against the Pima Indians diabetes dataset, which looks into the statistics and analysis of diabetes among the Native American Pima Indian People, specifically, females aged at least 21 years old. By applying the knowledge learnt throughout the class activities from weeks 1-7, we are equipped with the knowledge to train a machine learning model and report on our findings.

2. Observations

Most of the main observations we had were during the data cleaning and manipulations stages. One of the main observations we had was when we decided that instead of dropping rows or columns that had missing values or errors as a large percentage of the data would be deleted. Rather than dropping, areas where there were missing values or errors were replaced with predicted values. This allowed us to have the most concise data that would work perfectly with our models when we train them. This ended up working very well. As when we went to train our prediction models, all of our data was completely clear, and we removed most of the outliers, so we did not have any crazy values being returned in any of our models.

3. Major Findings

One of the major findings that we discovered whilst creating our model, was the correlation of Glucose, compared to the other variables. Glucose presented a 49% correlation. Following along is BMI which accounts for 31%. Quite a significant decrease from Glucose.

4. Outcome

After our training concluded, we fitted 70% of the data to testing and 30% to training. This resulted in our model producing a 74.46% accuracy. Without using Stratify we produced a result of 76.57%, then once stratify was introduced, this value was increased to 77.73%. Showing that our model in its best situation, performed 77.73% correctly. We looked into why this value was not higher, and we found the limits of the dataset. If it only has so many values to train a model on, there are aspects of the training in which it will not be able complete

References

Scribbr. (n.d.). *What's the best measure of central tendency to use?* [online] Available at: https://www.scribbr.com/frequently-asked-questions/whats-the-best-measure-of-central-tendency-to-use/

Healthline (29 June 2018) Do I Have Prediabetes or Diabetes?, Healthline, accessed 30 March 2022. https://www.healthline.com/health/type-2-diabetes/a1c-fpgt-ogtt-tests/

IYYER, S., 2021. *Step by Step Diabetes Classification-KNN-detailed*. [online] Kaggle.com. Available at: <https://www.kaggle.com/code/shrutimechlearn/step-by-step-diabetes-classification-knn-detailed> [Accessed 9 March 2022].