**A Study Adopting Statistical and Data Mining Approach to Draw Inferences on Diabetes Mellitus in Women**

Anthony Omowumi *(B00892305)*

*MSc Data Science, Ulster University* Magee Campus, Northern Ireland.

[anthony-o@uslter.ac.uk](mailto:anthony-o@uslter.ac.uk)

***Abstract- This study aimed at utilising statistical techniques to model the best predictors for Diabetes in women and classifying the outcome class based on those predictors - using a machine learning algorithm. The dataset used for this was sourced at Kaggle containing 768 records and 8 predictors/Independent variables and 1 Dependent variable with missing values recorded as NA, dealt with by using the missForest Algorithm in RStudio to impute the missing values. Appropriate statistical methods were done, and most attributes show rightly skewed and more asymmetrical than normal distribution around the mean. Two Logistics Regression Model was built to check statistically significant predictors and compared with each other using the ANOVA function. The results of the ANOVA chi-square test pr(>chi) equals 0.611, which shows the p-value for a chi-squared test comparing the two models Since the p-value is higher than 0.05 (the usual significance level), the null hypothesis cannot be rejected and it can be concluded that there is not enough evidence to suggest that the model 1l provides a significantly better fit to the data than model 2l. Furthermore, the dataset was split into a 70:30 ratio and a Two Random Forest model, first model was built with the 8 predictors and other with 4 statistically significant predictors while the first model with higher performance was chosen. Accuracy of 76%, Sensitivity of 80%, and Specificity of 66% were used as the metric to evaluate its performance, and the ROC curve was drawn having an AUC score of 82% which means the model is good at predicting the outcomes. Overall, the study was successful in achieving its goals.***

***Keywords: Diabetes, Data Mining, Statistical technique, Prediction, Data preprocessing, Logistics Regression, ANOVA, Random Forest***

1. **INTRODUCTION**

Diabetes is a global epidemic and a major cause of cardiovascular disease (CVD), chronic kidney disease, blindness, and amputation [1] and poses a substantial economic burden on individuals, communities, healthcare systems, and countries [2].

Diabetes especially Type 2 is more common in men than women [3]. Centers for Disease and Prevention (CDC) cited that women have more to manage, and women are at more serious complications and a greater risk chance of death. The most prevalent diabetic complication, heart disease, occurs nearly 4 times more frequently in women but only about 2 times in males, and women experience worse outcomes after a heart attack, even if it does not result in death. These factors make diabetes different for women than for men. Diabetes is not only different for women but also among women [4].

Additionally, women are at higher risk of other diabetes-related complications such as Blindness, Kidney disease, and vaginal dryness- which can make sexual intercourse uncomfortable and depression. Diabetes in women can affect their pregnancy (Gestational Diabetes) as high blood sugar can increase the risk of Preeclampsia (High Blood Pressure), Miscarriage, or stillbirth.

The aim of this paper is to utilise statistical techniques to draw inferences on the features that is more essential in predicting the likelihood of diabetes occurrence in Women based on the following life records - Glucose, Blood pressure, Family History of diabetes, Insulin, Body Mass Index, Age, and others as well as employing machine learning techniques in the classification of diabetes mellitus in women. This research aims not to eradicate medical diagnoses method like non-fasting blood tests (HbA1c), fasting blood tests, blood glucose tests, and others. However, in comparison to more conventional approaches that rely on human observation and judgment, this analysis can better help to know what to focus on when predicting Diabetes in women, and in the classification of Non-diabetic or Diabetes individuals, it would serve as a faster approach.

Statistics and Data mining can quickly and accurately handle enormous amounts of data while also finding patterns and relationships that humans would not immediately notice [5]. The precise algorithms and techniques utilised for data mining, as well as the quality and quantity of the data provided, will all affect how accurate the model is and an early diabetes diagnosis can result in more effective therapy to help several sectors utilised statistical Inferences and data mining techniques to predict diseases early [6].

1. **LITERATURE REVIEW**

In the interest of predicting diabetes at an early stage by some group of researchers in Pakistan. A study was done evaluating the effectiveness of three classification algorithms: Random Forest (RF), Artificial Neutral Network (ANN), and K means clustering algorithm. The research used a variety of methods, and the Apriori method for association mining was used to identify Glucose and BMI as having a substantial correlation in the dataset. The three models—ANN, RF, and K means clustering—used in the analysis produced results with high prediction accuracy, with ANN achieving the best accuracy of 75.7%, which is helpful to medical practitioners in making treatment decisions [6]. However, the focus of this research is to build a logistics regression model to statistically identify the best predictors that should be used in model building for the prediction of diabetes and as well as build a Random Forest classifier for the classification of diabetes (yes/no)

1. **METHODS AND MATERIALS**

The dataset used for this study was sourced from the Kaggle dataset Repository. The dataset was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. The purpose of the data collection was to use the diagnostic metrics included in the dataset to estimate a patient's likelihood of having diabetes or not. The dataset was chosen from a bigger database, and it was restricted to all female patients in PIMA India who were at least 21 years old. Several Independent (predictors) variables and a single target dependent variable make up the dataset. The result is a kind of binary categorization, where 1 indicates that a person has diabetes and 0 indicates that they don't.

The dataset was analysed in this study using the R programming language. R is a popular statistical and data mining application that is utilized by both data scientists and researchers. All required libraries were imported for the analysis [7].

1. ***Dataset Description***

The diabetes dataset is multivariate with 9 attributes and 768 records with missing values recorded as zero values. This data was loaded into an Excel worksheet to convert the zero values to NA (652 in total). Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI (Body Mass Index), Diabetes Pedigree Function, Age, and the "Outcome" classification column make up the features. There are just two categories: 1 or 0 in the classification column. Below is a brief description of the dataset*“Table 1”.*

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset  Characteristics | Multivariate | Number of  records | 768 |
| Features  Characteristics | Real | Number of  Features | 9 |
| Associated  Tasks | Classification, Regression | Missing  Values | Yes |

*TABLE 1: Diabetes Dataset Description*

***Pregnancies:*** expressed as the number of times a participant has been pregnant

***Glucose:***expressed as the plasma glucose concentration in 2h oral glucose tolerance test

***Blood Pressure:***expressed as diastolic blood pressure which occurs as blood pumps into the arteries between the heart (mm Hg)

***Skin Thickness:***expressed as the triceps skinfold thickness (mm). The collage content provides a conclusion.

***Insulin:***expressed as 2h serum insulin (mu U/ml)

***BMI:***weight (kg) divided by height in (m2 ) express as the Body Mass Index

***DiabetesPedigreeFunction:***express the likelihood of a participant having diabetes based on family history.

***Age:***express as the age of participants in the study

***Outcome:***express the class variable for diabetes which is either yes/1 (Diabetes) or no/0 (non-diabetic)

1. ***Data Preprocessing and Transformation***

Data preprocessing and transformation are essential steps in the data mining process since raw data is frequently not in a format that can be evaluated. Several methods were employed in this study to enhance the quality of the data by identifying errors, outliers, and missing values. These techniques are essential since the quality of the data has a direct impact on how accurate the results are. Methods were taken to ensure that the data used was consistent.

*I)* ***Data Cleaning:***After importing the import libraries for data preprocessing and data mining, the data was loaded into R studio and stored in another data frame to ensure data provenance. The dataset's schema was then queried to examine the data types and structure of each column and dimension. Errors, special characters like "Inf" and "NaN," as well as non-negative values, were checked for; this was done by developing a check rule that was used to verify each data point and make sure it adhered to the rule and contained non-negative values. The presence of outliers was visualized using boxplot “Fig 1”. The plot showed that there is a mild outlier in most of the data variables which will have a minimal significant effect on the skewing result. However, Insulin has a higher Outliers presence which will significantly affect its skewness. The Insulin outliers were left because according to some researchers, the higher values may be due to errors or deregulation [8]. However, high blood sugar higher than 600mg/dL is very serious and at risk of higher dehydration, coma, or death which can be a result of high Insulin resistance [9].

Chart, box and whisker chart

Description automatically generated

*Fig 1:* Boxplot of data attributes in the Diabetes dataset

*II)* ***Missing Values Imputation:***missForest is a non-parametric missing values imputation using Random Forest techniques. This can be applied to mixed data types, including numerical and categorical data, and because it is resistant to noise, outliers, and used non-linear data, making it more effective than KNN or other techniques. Along with the computation, an out-of-bag (OOB) Error is produced. The imputation efficiency is assessed using this error [10] An out-of-bag (OOB) imputation estimate error for a continuous variable was 0.5877423 NMRSE (Normalized Mean Squared Error) after 4 iterations. This is quite high in this analysis because, according to the general rule for Miss Forest, datasets are effectively imputed after 4 to 5 iterations. However, this depends on the size and quantity of missing data, and since the dataset has 652 missing values, this likely influence the outcome [11].

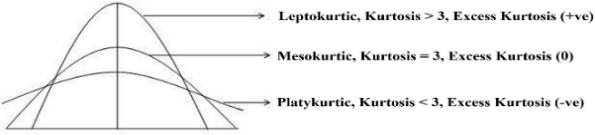
III***) Descriptive and Graphical Analysis after Missing Value Imputation*:** various aspects of the data were inspected and analysed to draw an inference statistically.

**Data Distribution**: Histogram is used to identify the shape of a data distribution. Histogram plots of each variable were visualized to check for data distribution. The plots show unimodal with a single peak and are mostly right skewed looking at the mean and median values (mean > median) [12]. Furthermore, a normality test using Shapiro Wilk’s test, a non-parametric statistical test was used to evaluate if each data variable distribution deviates from a normal distribution. Setting a threshold of 0.05, the p-value of the data features has a p-value lower than the threshold, hence the hypothesis of normality is rejected [20].

**Skewness** is the measure of the symmetric/asymmetric of a variable. Farther away from zero indicates it is more asymmetric than symmetric. It considers the data distribution's direction and relative magnitude [13]. From the analysis in order of asymmetric, Blood Pressure ~0.14 < Glucose ~0.53 < BMI ~0.61 < Skin Thickness ~0.70 < Pregnancies ~0.89 < Age ~1.13 < DiabetesPedigreeFunction ~1.92 < Insulin ~2.12, this shows that Insulin is more highly asymmetric than all other variables and Blood Pressure is closely symmetric.

**Kurtosis** measures the tailedness of a data distribution while tailedness refers to how often an outlier occurs in a distribution. A normal distribution has a Kurtosis of 3, therefore, to measure for excess kurtosis which indicates the presence and seriousness of outliers [13]. “*Equation 1*” was used for the evaluation and the result is Insulin is closely related to Leptokurtic having a higher amount of excess kurtosis, same as DiabetesPedigreeFunction and Skin Thickness. Glucose is Platykurtic while other variables fall slightly within Mesokurtic.

***Equation 1****: Excess Kurtosis = kurtosis – 3 (1)*



*Fig 3: Kurtosis Explained*

**Correlation and Variance Inflation Factor**: The presence of multicollinearity between the features in "Fig. 4" was examined using a correlation plot of the nine attributes. Multicollinearity occurs when two or more independent variables, sometimes referred to as predictors, have a strong correlation with one another in a regression model. Although multicollinearity might not significantly affect the model's accuracy, it might make it more challenging to understand the effects of independent characteristics on the dependent feature, which could make the model's conclusions [14].

Table

Description automatically generated

*Fig4:Spearman Correlation plot showing relationships between the features.*

The Spearman correlation plot shows a moderately positive between Age and Pregnancies, a strong correlation between (Insulin and Glucose), and (BMI and Skin Thickness) [19]. This is further studied by using the Variance Inflation Factor (VIF) to confirm the presence of multicollinearity. VIF determines the strength of multicollinearity between independent variables in a regression model. A VIF equal to 1 means no correlation between predictors while VIF < 5 means low multicollinearity or moderately correlated and VIF > 5 or 10 means high multicollinearity between predictors [14]. None of the 8 features had VIF values larger than 5, as shown by the computed VIF values, implying that none of the features have high multicollinearity with one or more of the variables in a regression model. “Fig 5“.

Chart

Description automatically generated

*Fig 5: Showing the VIF values level for the features.*

IV) ***Feature Selection***: There are several effective methods for reducing the number of dimensions; one such method is principal component analysis (PCA). PCA is useful for several reasons. Firstly, employing variance explained by the data variables reduces the number of dimensions in the data, thus lowering the computation complexity of a machine learning model. It can also be used to visualise a dataset by reducing the number of dimensions in the data to two or three. [15]. PCA allows you to find correlations and patterns in a data collection so that it can be compressed in size without losing any crucial data by assessing the variances of the original dataset stored as a Principal component. This requires scaling/normalizing of features because interest is in component that maximise the variance hence comparism done between the features has to be on the same scale for effective result. The data features were normalize. Chart, line chart

Description automatically generated

*Fig 6. The Variation in the dataset explained in Principal Component*

The plot explains the variation and clearly shows that PC1 accounts for ~33% of the total variation. However, the cumulative proportion variance computed in R studio reveals that to explain 95% of the variation, PC1, PC2, PC3, PC4, PC5, PC6, and PC7 must be retained.

1. **STATISTICAL ANALYSIS OF THE INDEPENDENT VARIABLES USING LOGISTICS REGRESSION (LR):**

LR is a statistical model that is mostly used forpredictive analytics and classification. It is widely used to estimate the probability of an event occurring based on independent variables (either categorical or continuous)**.** The dependent variable is binary(dichotomous). Using the logit formula in logistic regression, the odds, or likelihood of success divided by the probability of failure is transformed. This is also sometimes referred to as the log odds or the natural logarithm of odds, with many of the model's coefficient estimations being calculated using the Maximum Likelihood Estimate method (MLE) [16]. The normalized dataset was used to build this model

**Interpretation of the Logistic Regression Model Built:**

1. The model was built by entering all 8 independent variables and computing the statistically significant independent variables in the model using a p-value of significance level alpha equal to 0.05. Null hypothesis is rejected if the p-value is < a significance level (alpha = 0.05), whereas a p-value > 0.05 indicates that the deviation from the null hypothesis is not statistically significant and thus the null hypothesis is not rejected. Based on this hypothesis, the model built with the 8 independent variables has Pregnancies, Glucose, BMI and DiabetesPedigreeFunction having a p-value < 0.05 means it has a statistically significant relationship with the response variable in the model. To determine this, the odds ratio and its Confidence Interval were calculated.
2. The deviance residual result is Min is 2.8257, Max is 2.4368, 1Q/3Q is 0.7228 and 0.7236 with a median of 0.3975 taking the absolute value of the results. This shows a median a bit closer to Zero, and min/max, 1Q/3Q showing about the same distribution (evenly distributed) the smaller values show that the model probably fitted well.
3. The Coefficients Estimates show the average changes in the log odds of Outcome (diabetes =1/Yes, Non-diabetic = 0/NO). Inspecting the BMI attribute having a value of 4.2126 which is positive, which means with a one unit increase in BMI, there is an increase in the Log odds of the outcome occurring to be diabetes.
4. The Odds ratio (OR) and Confidence Interval (CI) were computed to ascertain the statistically significant inference drawn. Pregnancies have an OR of ~8.372 which is greater than 1, this shows that at a one-unit increase in Pregnancy values, the odds of having diabetes increase by a factor of ~8.372, a likelihood of Diabetes outcome occurring with a 25% to 95% CI of ~2.872, ~25.013. Glucose also has a strong effect on Diabetes occurring as the OR is ~240.247 and a 25%, 95% CI of ~67.460, ~910.21. However, Blood pressure has an OR of ~0.4633 which is less than 1, this means the Blood pressure variable is associated with lower odds of diabetes occurring, for a one-unit increase in blood pressure, the odds of having diabetes decrease by a factor of ~0.4633 holding all other variable constants. Skin Thickness has an OR greater than 1, ~1.8608 however, the 25%, 95% CI is ~0.1511, ~24.30 which includes 1 but the expected true OR may be above or below 1, so it is uncertain whether the odds of having diabetes or not is happening with a specific level of confidence. This interpretation was used for all other independent variables (Insulin, BMI, DiabetesPedigreeFunction, and Age. This confirms the p-values of the z-test indicating their statistically significant or not.
5. Another logistic regression model was built using statistically significant independent variables (Pregnancies, Glucose, BMI, and DiabetesPedigreeFunction). The ANOVA function using the Chi-square test was run to compare the two models. The result obtained showed that comparing the first model with 8 Independent variables and the second model with 4 statistically significant independent variables. The results of the ANOVA chi-square test pr(>chi) equals 0.611, this shows the p-value for a chi-squared test comparing the two models. Since the p-value is higher than 0.05 (the usual significance level), the null hypothesis cannot be rejected and it can be concluded that there is not enough evidence to suggest that the full model provides a significantly better fit to the data than the reduced model. Alternatively, it means Model 2, which includes only 4 predictor variables, is sufficient to explain the variability in the response variable *(The null hypothesis is that the condensed model (Model 2) fits as well as the complete model (Model 1)).*
6. **MODEL TRAINING AND TESTING USING RANDOM FOREST ALGORITHM:**

Class Distribution for the outcome column was analysed and found to be 65% (NO/1) and 35%(YES/1), this shows a moderate imbalance class. The data was then split into 70% (538) and 30% (230) records without scaling the features because Random Forest (RF) is known for its high accuracy performance, robustness to outliers, ability to handle non-linear data, and inability to overfit with many features. Given that it utilised a tree-based model and does not require scaling of features. It is an ensemble learning technique that works well for both classification (data labels that are discrete) and regression (data labels that are continuous). For these reasons, this classifier was chosen. [14]. Two RF was built, the first with 8 predictors variable and the second model with the 4 statistically significant variables (Glucose, BMI, Pregnancies and DiabetesPedigreeFunction).

1. **RESULTS AND DISCUSSION**

Using the test data, the evaluation of the Trained RF model's performance was conducted. The first model with the 8 predictors was chosen because it has a higher accuracy, sensitivity, specificity and F1 score.

Utilising the metrics Accuracy, Sensitivity, Specificity, and F1, the model's performance was evaluated. The non-parametric classifier RF model yields accuracy, sensitivity, and specificity values of 76%, 80%, and 66% respectively. The area under the curve (AUC) for the model's performance according to the ROC curve is 82%.

Inspecting the metric of the model, Accuracy was less considered because the data is considered moderately imbalanced data with 0/non-diabetic having higher distribution. The sensitivity of 80% shows that if the classifier predicts that one does not have diabetes, it is 80% probably certain they do not have diabetes, while the Specificity of 66% on the other hand shows that if the classifier predicts that one has diabetes, there is a high probability that they are healthy without diabetes (False Positive rate).

Chart, line chart

Description automatically generated

*Fig 7: The ROC curve showing the performance of the model.*

1. **CONCLUSION**

It can be inferred from the outcomes of the various statistical analyses performed on the dataset indicates that all independent variables (8 features) from the original dataset are crucial for the prediction of Diabetes in women. Although It is suggested that further analysis should be conducted to ascertain the 4 statistically significant importance probably by using statistical controls and other methods. Also, the analysis's use of statistical and data mining tools was successful in reaching the study's objective as the conclusion was reached based on this study that the 8 features are necessary for the prediction of Diabetes Mellitus in Women. The RF classifier was quite good at detecting a category. Future data-collecting efforts are advised to concentrate on gathering minimal error-free data. This will help prevent having a lot of missing values entered as Zero values. Further analysis can also be carried out on the dataset by balancing the dataset because it was moderately imbalanced data. ML balancing data techniques like Random over-sampling, under-sampling, or Synthetic Minority Oversampling Technique (SMOTE) and much more may be considered to balance the class distribution which might improve the performance of the data. This approach can as well be utilised for men’s diabetes classification and other types of diseases of diagnosis.

**REFERENCES**

[1]de Ritter, R. et al "Sex difference in the risk of vascular disease associated with diabetes," *Biology of sex difference,* vol. 11, no. 1, pp.1-11, 2020.

[2] Facts & Figures - "International Diabetes Federation," 9 December 2021. [Online]. Available: https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures. [Accessed 24 February 2023].

[3] Diabetes affects Men and Women. "Medical News Today," 2 March 2021. [Online]. Available: https://www.medicalnewstoday.com/articles/diabetes-affects-men-women. [Accessed 24 February 2023].

[4] Diabetes and Women, "Center for Diseases Control and Prevention," Department of Health and Human Services, 20 June 2022. [Online]. Available: https://www.cdc.gov/diabetes/library/features/diabetes-and-women.html. [Accessed 24 February 2023].

[5] Goh, Y.C. et al. (2020) “Evaluating human versus machine learning performance in classifying research abstracts,” Scientometrics, 125(2), pp. 1197–1212. Available: https://doi.org/10.1007/s11192-020-03614-2. [Accessed 25 February 2023]

[6] Alam, T.M. et al. (2019) “A model for early prediction of diabetes”. *Informatics in Medicine vol. 16*, p.100204, 2019. (Science direct.com)

[7] Python vs. R: What’s the Difference? 2021. [Online] Available: https://www.ibm.com/cloud/blog/python-vs-r [Accessed 28 February 2023].

[8] Wu, H., Yang, S., Huang, Z., He, J. and Wang, X., “Type 2 diabetes mellitus prediction model based on data mining”. *Informatics in Medicine vol*, *10*, pp.100-107, 2018

[9] WebMD High Blood Sugar: Complications that can happen. [Online] Available: https://www.webmd.com/diabetes/uncontrolled-blood-sugar-risks [Accessed 14 March 2023]

[10] R package “*Imputing missing values in R*”. 4 March 2016. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/ [Accessed 6 March 2023]

[11] Towards Data Science Miss Forest. *“The best-missing data imputation algorithm”* 31 August 2020 [Online]. Available: https://towardsdatascience.com/missforest-the-best-missing-data-imputation-algorithm [Accessed 13 March 2023]

[12] Standard Deviation- National Institute of Health," *National Library of Medicine”,* [Online]. Available: https://www.nlm.nih.gov/nichsr/stats\_tutorial/section2/mod8\_sd.html. [Accessed 14 March 2023].

[13] Skewness and Kurtosis, Analytics Vidhya, 2 May 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/shape-of-data-skewness-and-kurtosis/. [Accessed 14 March 2023].

[14] Multicollinearity, Analytics Vidhya *“Multicollinearity in Data Science Models”*, 19 March 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/03/multicollinearity-in-data-science. [Accessed 14 March 2023].

[15] "Principal Component Analysis," Wikipedia, 2022. [Online]. Available: https://en.wikipedia.org/wiki/Principal\_component\_analysis. [Accessed 14 March 2023].

[16] What is Logistics Regression, IBM, [Online]. Available: https://www.ibm.com/uk-en/topics/logistic-regression. [Accessed 22 March 2023].

[17]Random Forest Algorithm in Machine Learning- An Overview (2022) Available at: https://www.mygreatlearning.com/blog/random-forest-algorithm (Accessed: 12 April 2023)

[18]Correlation Coefficients-Andrew University (2005)Available at : https://www.andrews.edu/~calkins/math/edrm611/edrm05.htm#:~:text=Correlation%20coefficients%20whose%20magnitude%20are%20between%200.5%20and%200.7%20indicate,which%20have%20a%20low%20correlation (Accessed on: 24 April 2023)

[19] Spearman’s Rank Calculator. April 2023 [Online] Available at :https://geographyfieldwork.com/SpearmansRankCalculator.html (Accessed on: 26 April 2023)

[20]Shapiro Wilk’s Test 2010 [Online] Available at: https://www.sciencedirect.com/topics/psychology/shapiro-wilk-test (Accessed on: 28 April 2023)

[21] What is Akaike Information Criterion? November 2022 Available at: https://builtin.com/data-science/what-is-aic Accessed on: 26 April 2023)