Diabetes EDA and Regression

1. Dataset Description

The ‘PimaDiabetes.csv’ dataset contains 750 observations of female members of the Akimel O’odham Indigenous community from the Gila River Indian Reservation in the USA (Radin, 2017). The dataset originates from the US National Institute of Diabetes and Digestive and Kidney Diseases (Smith et al., 1988).

‘PimaDiabetes.csv’ contains information about number of pregnancies, glucose (oral test, mg/dl), diastolic blood pressure (mm Hg), triceps skin thickness (mm), insulin, body mass index(BMI), diabetes pedigree score (measuring genetic risk), age, and diabetes outcome. While Smith et al. (1998) list insulin unit measurements as µU/ml, it is not possible to determine whether the scale has been appropriately reported. Hayashi et al. (2012) report an upper boundary value of 170.6(µU/ml) for insulin concentration 2 hours after an oral glucose test in their subjects. In comparison, the ‘PimaDiabetes.csv’ dataset has a maximum value of 843 on the same scale, which suggest a considerable data quality issue. Furthermore, the data has no encoded missing values, making it impossible to distinguish with certainty between logical zero values and missing data.

2. EDA

In order to perform any operations on the ‘PimaDiabetes.csv’ dataset, the missing values need to be cleaned. Zero values in Pregnancies (none) and Outcome (non-diabetic) are valid, and Age and DiabetesPedigree have no missing values. However, the null inputs in SkinThickness, BMI, BloodPressure, Insulin, and Glucose, are not logically valid. Thus, they were classified as missing, and their counts and percentages are shown in Table 1.

Table 1. Missing Values By Variable

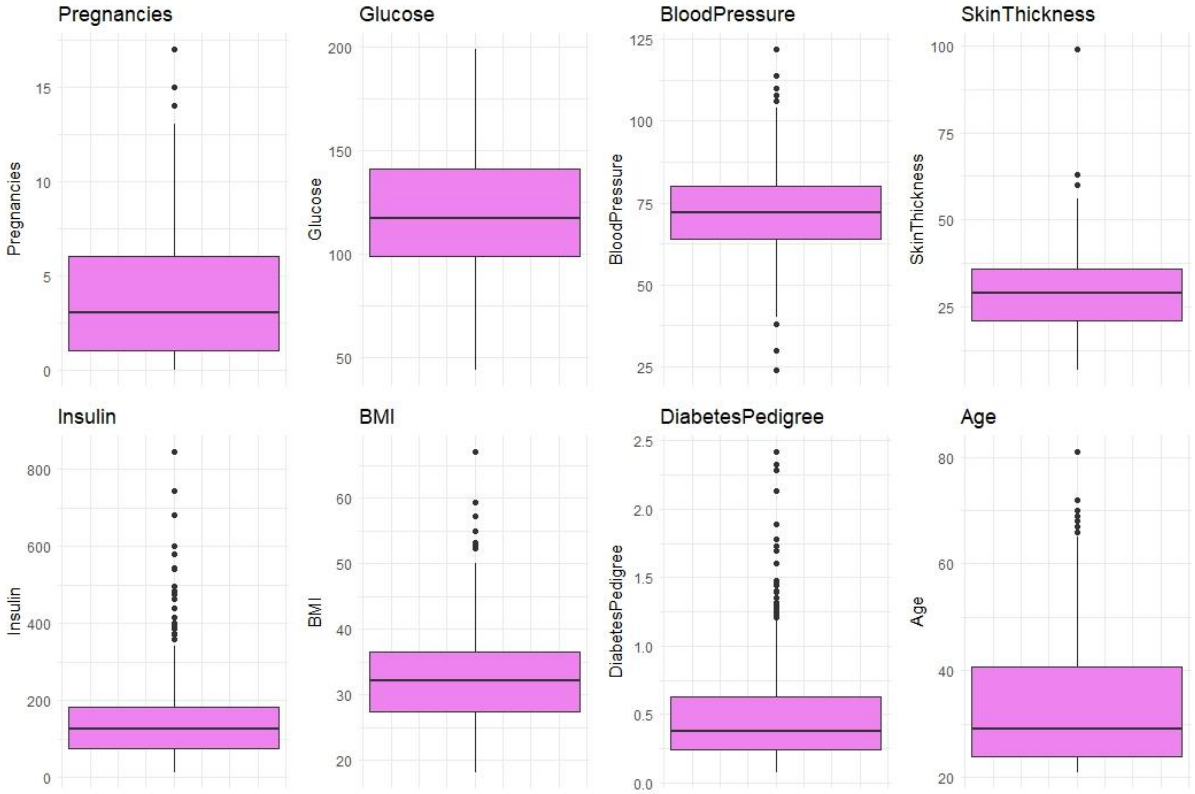
Variables	Missing Counts	Missing %
SkinThickness	221	29.5%
Glucose	5	0.7%
BloodPressure	35	4.6%
Insulin	362	48.2%
BMI	11	1.5%
Pregnancies, Age, Outcome, DiabetesPedigree	0	0%

The percentage of missing observations for Insulin (48.2%) and SkinThickness (29.5%) in Table 1., is much higher compared to the other variables in the dataset and could point to a systematic data quality issue. However, due to the lack of documentation available for the dataset, another method must be utilised to assess any patterns of missingness.

The Missing Completely at Random (MCAR) test was performed on the dataset (p-value<0.001), giving us evidence to suggest that the data are not missing at random.

This is a crucial implication for missing values imputation, as the majority of them operate under the assumption of MCAR. MICE (Multiple Imputation by Chained Equations) can handle missing not at random data, which makes it more suitable for the purposes of this analysis (Vink and van Buuren, 2013). However, imputation methods are sensitive to outliers (Quintano, Castellano, and Rocca, 2010). The boxplots in Figure 1. demonstrate that most of the variables in ‘PimaDiabetes.csv’ have outliers, which poses a limitation on the analysis.

Figure 1. Box Plots of the Variables (Excluding Outcome)



Therefore, to prepare the data for imputation, outliers need to be addressed. The treatment of outliers in 'PimaDiabetes.csv' was performed based on lower limit=first quartile–1.5*IQR and upper limit=third quartile+1.5*IQR (Johansen and Christensen, 2018). Subsequently, the missing data was imputed with MICE, utilising 15 cycles to stabilise the results (standard cycle range from 10-20) (White, Royston and Wood, 2011). The summary statistics for the imputed dataset are displayed in Table 2., rounded to 2 decimal places. The counts for non-diabetics and diabetics in Outcome are shown in Figure 2.

Table 2. Summary Statistics of Variables

Metric	Variable							
	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree	Age
Min	0.00	44.00	40.00	7.00	14.00	18.20	0.07	21.00
1st Quartile	1.00	99.00	64.00	20.00	78.00	27.50	0.24	24.00
Median	3.00	117.00	72.00	28.00	120.00	32.20	0.38	29.00
Mean	3.83	121.50	72.10	28.50	142.50	32.40	0.46	33.12
3rd Quartile	6.00	141.00	80.00	36.00	183.00	36.60	0.62	40.75
Max	13.00	199.00	104.00	58.00	360.00	50.33	1.20	66.00

Figure 2. Bar Chart of Diabetes Outcome

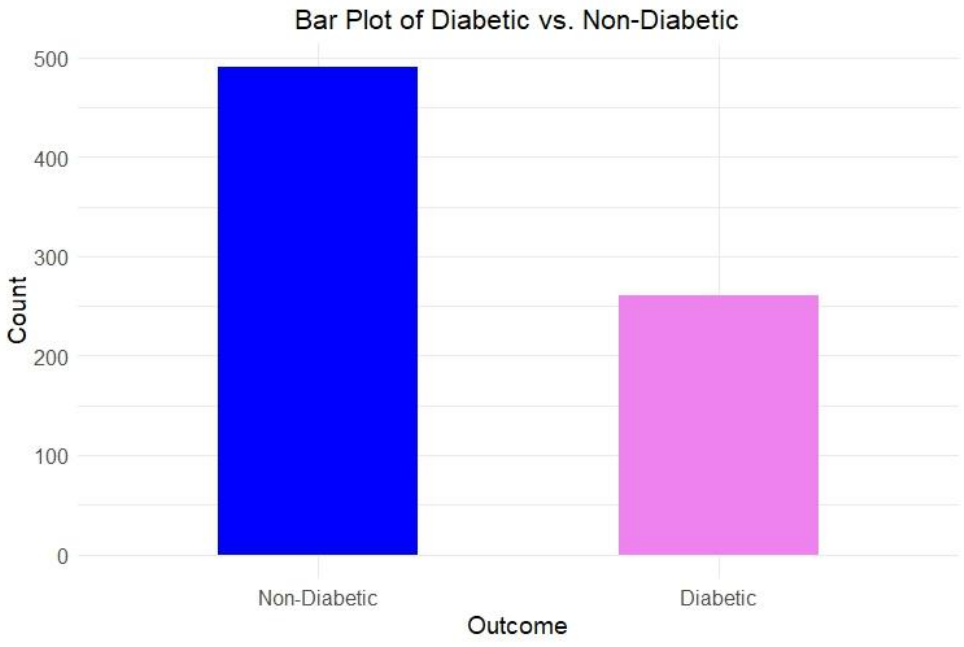


Figure 2. demonstrates that Outcome of diabetes is imbalanced, which could affect the construction of regression models.

Figure 3. Combined Density and Histogram Plots

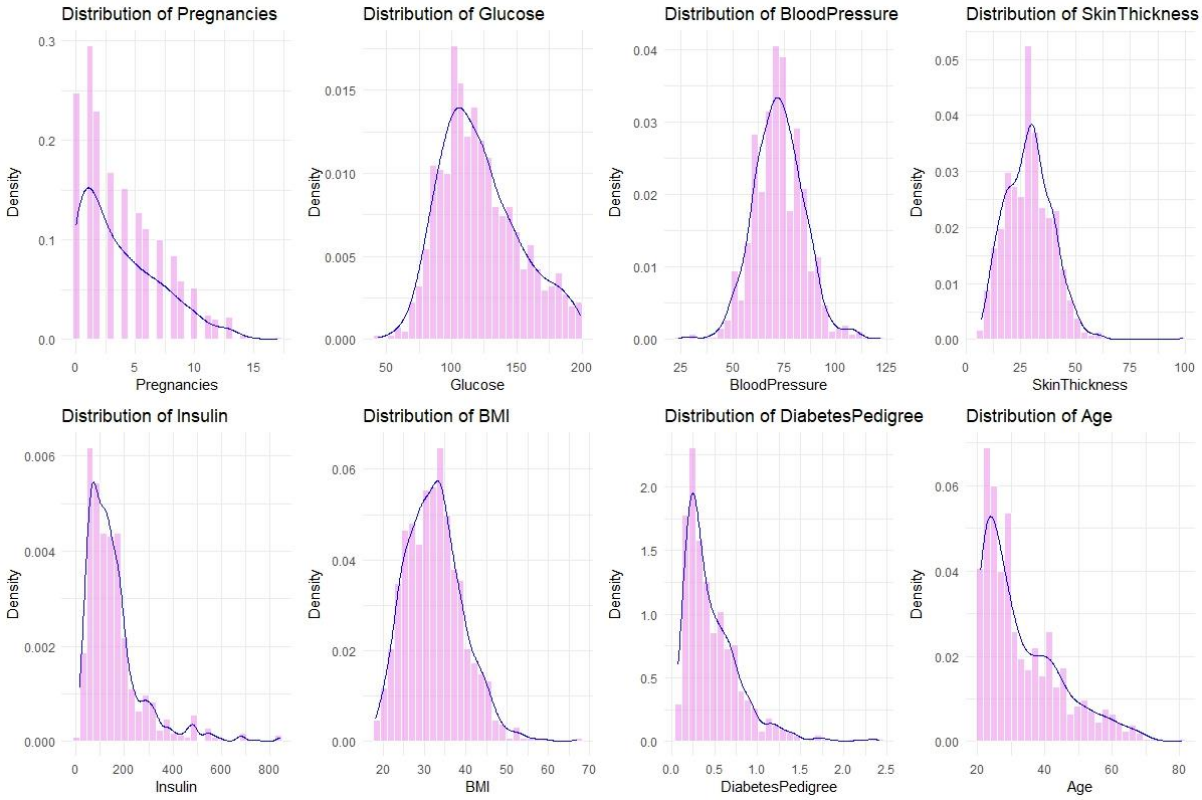


Figure 3. illustrates the combined density and histogram plots for all variables except Outcome. BloodPressure has a normal distribution, while Glucose is nearly normally distributed. However, the rest of the variables are skewed which could impact results.

3.Probability of Diabetes Outcome Based on Number of Pregnancies

A new column was created in the dataset – if a woman has had 7 or more pregnancies, the value is 1, and for 6 or less – 0. In order to calculate the probability of being diagnosed with diabetes based on the number of pregnancies (more or less than 7), a fitted logistic regression model was used, where Outcome is the

dependent variable and SevenOrMorePregnancies is the predictor. The following formula was used to derive the probabilities for being diabetic given the value of SevenOrMorePregnancies :

p(x) = 1/ 1+e-(β0 + β1x), where β0 is the intercept and β1 is the slope.

The glm() function in R automatically uses the positive outcome, and the slope corresponds to the coefficient for SevenOrMorePregnancies in the model.

Figure 4. Calculations in RStudio

```
intercept <- coef(model)[1]
coefficient_SevenOrMorePregnancies <- coef(model)[2]
logistic_function <- function(x) {
  1 / (1 + exp(-(intercept + coefficient_SevenOrMorePregnancies * x)))
}
#Probability for 7 or more pregnancies (1)
prob_seven_or_more <- logistic_function(1) # 0.5670732
# Probability for 6 or fewer pregnancies (0)
prob_six_or_fewer<- logistic_function(0) # 0.2849829
```

- Pdiabetes given 7 or more pregnancies: 0.5670732
- Pdiabetes given 6 or fewer pregnancies: 0.2849829

4. Predicting Diabetes Outcome for Observations in 'ToPredict.csv'

Before building a regression model, possibilities of multicollinearity need to be checked for our predictor(feature) variables. Computing the Variance Inflation Factor(VIF) of a logistic regression model with all the predictors provides information about the level of multicollinearity for each predictor, the results for which is presented in Table 3.

Table 3. VIF results

Predictor	VIF value
SkinThickness	1.613837
BMI	1.842475
Pregnancies	1.441725
Age	1.576488
Glucose	1.485819
BloodPressure	1.221940
DiabetesPedigree	1.026673
Insulin	1.483329

While all the VIF values are below 10, therefore not indicating a level of problematic multicollinearity, the statistic for DiabetesPedigree is very close to 1. In this case, it is more reasonable exclude it from the logistic regression model.

Since we want to predict a new dataset (‘ToPredict.csv), it is appropriate to construct a logistic regression model by dividing the ‘PimaDiabetes.csv’ into a test and train set and see how the model performs when encountered with new observations. The test/train split ratio used for the model was 2:1 due to its high reliability evidenced in academic sources (Dobbin and Simon, 2011). Subsequently, a K-Fold cross-validation was performed to ensure consistent performance of the model, with k = 10 (Brownlee, 2023). The process of cross validating, predicting Outcome for the test data, and displaying classifier performance measures was done in a loop in R, which iterates over the k different train splits. The first logistic model utilized all feature

variables except DiabetesPedigree, while the second one included only BMI, Age, Glucose and Pregnancies as predictors.

Table 4. Model Comparison

	Model 1	**Model 2 (chosen model)**
Features	BMI, Age, Pregnancies, Glucose, Insulin, SkinThickness, BloodPressure	BMI, Age, Glucose, Pregnancies
Average Accuracy	0.7693333	0.8133333
Average Precision	0.794279	0.8571429
Average Recall	0.8754517	0.8571429
Average F1 Score	0.8308366	0.8571429

For this data, accuracy might not be the best measure due to the class imbalance in the Outcome variable (Hicks et al., 2022). The recall in this dataset it could be more important than accuracy due to the danger of missing a positive case of diabetes (Hicks et al., 2022). Precision provides information about the correctly assigned outcomes within the sample of outcomes (Hicks et al., 2022)., in which Model 2 performs better. The harmonized mean of precision and recall, F1 Score (Hicks et al., 2022), is better for Model 2 but so we could argue that it has superior overall performance. Thus, Model 2 was picked for the prediction of the diabetes outcome in the ‘ToPredict.csv’ The updated dataset is provided in Table. 5.

However, should be noted that both models are limited due to the quality of the dataset, which is why we should be cautious when making assumptions about their performance.

Table 5. Updated 'ToPredict.csv' dataset

Pregnancies	Glucose	BP	Skin Th.	Insulin	BMI	Diabetes Pedigree	Age	Pred. Outcome	Pred. Probability
4	136	70	0	0	31.2	1.182	22	0	0.36
1	121	78	39	74	39.0	0.261	28	0	0.35
3	108	62	24	0	26.0	0.223	25	0	0.09
0	181	88	44	510	43.3	0.222	26	1	0.89
8	154	78	32	0	32.4	0.443	45	1	0.73

References

Brownlee, J. (2023). A Gentle Introduction to k-fold Cross-Validation. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/k-fold-cross-validation/.

Dobbin, K.K. and Simon, R.M.. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics*, 4(1), p.31. [online]. Available from: https://dx.doi.org/10.1186/1755-8794-4-31.

Hayashi, T. et al. (2012). 'Patterns of Insulin Concentration During the OGTT Predict the Risk of Type 2 Diabetes in Japanese Americans', *Diabetes Care*, 36(5), pp.1229–1235. doi:https://doi.org/10.2337/dc12-0246.

Hicks, S.A. et al. (2022). On evaluation metrics for medical applications of artificial intelligence. Scientific Reports, [online] 12(1), p.5979. doi:https://doi.org/10.1038/s41598-022-09954-8.

Johansen, M. and Christensen, P. (2018). A simple transformation independent method for outlier definition. *Clinical Chemistry and Laboratory Medicine (CCLM)*, Vol. 56 (Issue 9), pp. 1524-1532. https://doi.org/10.1515/cclm-2018-0025

Quintano, C., Castellano, R. and Rocca, A. (2010). Influence of outliers on some multiple imputation methods. *Advances in Methodology and Statistics*, 7(1).

Radin, J. (2017). "Digital Natives": How Medical and Indigenous Histories Matter for Big Data. *Osiris.*, 32(1), pp.43–64.

Smith, J.W. et al. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings - Annual Symposium on Computer Applications in Medical Care*. pp. 261–265.

Vink, G. and van Buuren, S. (2013). Multiple Imputation of Squared Terms. *Sociological methods & research*, 42(4), pp.598–607.

White, I.R., Royston, P. and Wood, A.M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine*, 30(4), pp.377–399.