1|Introduction:

In this study, I explored the Pima Indians dataset, a community near Phoenix, Arizona in the United States of America who have exhibited a high risk of type 2 diabetes. The dataset consisted of two segments, containing information on several clinical measurements such as pregnancies (Npreg), glucose concentration (glu), blood pressure (Bp), skinfold thickness(skin), body mass index (bmi), pedigree function(ped) (a form of evaluation of the chances of diabetes based on family heritage) and finally the target variable(type) if the individual has diabetes. Our objective in this experiment will be to build and evaluate classification models to optimally predict the diabetes status of an individual.

This will be achieved by cleaning the data by checking for missing values and conducting explanatory data analysis, building models like Bayes classifier for linear or quadratic discriminant analysis, and building logistic regression with best subset selection. Later, the performance of these models will be compared using cross-validation- based on test error to determine the most accurate predictive model.

1.1 Data cleaning and pre-processing:

This initial step in our data-cleaning process is to check if there are any missing data in every row and this is a crucial step of our data-cleaning process for our models to work. We used the functions 'is.na()' and 'colSums()' to check and discovered that there were no missing values in our combined dataset. We also had to check all the columns to make sure they were numeric and we need to transform the categorical variable 'type' to numeric which was originally presented as 'Yes' or 'No' this is to imply that the type labeled "yes" was converted into the integer *1* and "No" into the integer *0*.

1.2 Exploratory Data Analysis:

To understand the distribution of our continuous variable plotted histograms to examine and determine if all the variables follow normal distribution. The visualization helped us understand that blood pressure, BMI, Glucose level, and skinfold thickness follow a normal distribution (Figure 1). It also revealed that most of the subjects in our study are between the ages of 20 -30 years. Pregnancies and pedigree function is skewed towards the right. To understand the numerical summary better, we obtained the mean, median and other metrics of all our variables. The values are presented in the table below –

Table 1. Descriptive statistics of all variables in our dataset

	Npreg	glu	Bp	Bmi	ped	skin	Age
Min	0.000	56.00	24.00	18.20	0.0850	7.00	21.00
1st Qu	1.000	98.75	64.00	27.88	0.2587	22.00	23.00
Median	2.000	115.00	72.00	32.80	0.4160	29.00	28.00
Mean	3.517	121.03	71.51	32.89	0.5030	29.18	31.61
3rd Qu	5.000	141.25	80.00	36.90	0.6585	36.00	38.00
Max	17.000	199.00	110.00	67.10	2.4200	99.00	81.00

The above table illustrates that there are people from 21 years old to 81 years old in our dataset and the average subject is *31.6* years old, and almost half the group is 28 or younger. Glucose levels, seems to be one of the indicators for diabetes, ranges from 56 to 199 and has an average of 121 for women. Additionally, blood pressure averages at 71.5 suggesting balanced heart health. The skinfold thickness varies from 7 to 99 with a mean thickness of 29.2. concerningly, the BMI varies from 32.9 to 36.9 which makes many subjects in our study obese. Finally using the 'table()' function I counted how many people in our dataset have diabetes, and it results in 177 individuals that are categorized to '1' which corresponds to yes have diabetes.

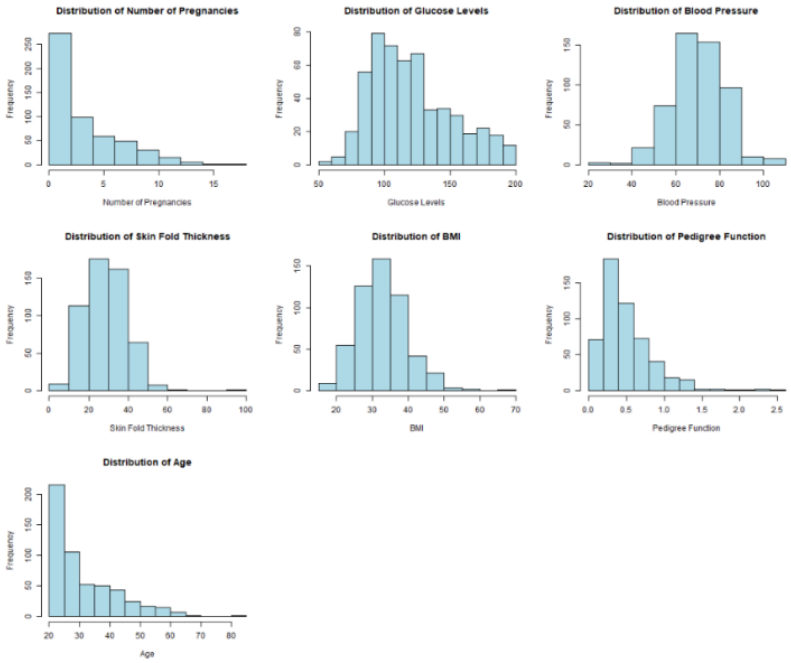

Figure 1. Distribution of independent variables

Moreover, to understand the relationships between these variables, we plotted a scatterplot matrix (Figure2). By using this method, one can visualize the correlation between each pair

of independent variables, independent-dependent pairs of independent variables, and independent-dependent variable pairs. The blue and red dots are very important to notice in the scatterplot matrix. The blue dots represent non–diabetic people and the red dots indicate diabetic patients (Figure 2). Through this color differentiation, patterns and relationships between these two types can be determined across a variety of variables. First, we observe a denser concentration of red dot in' glucose levels' implying that higher glucose levels can be associated with someone diagnosed with diabetes. Similarly, "number of pregnancies" and "age" show a positive correlation with both variables increasing together. Lastly, the "skin" and "bmi" seem to have a strong correlation implying that as skin fold thickness increases, BMi also increases. These findings are very beneficial for our analysis as it is beneficial for our model building and investigating the important variable further.

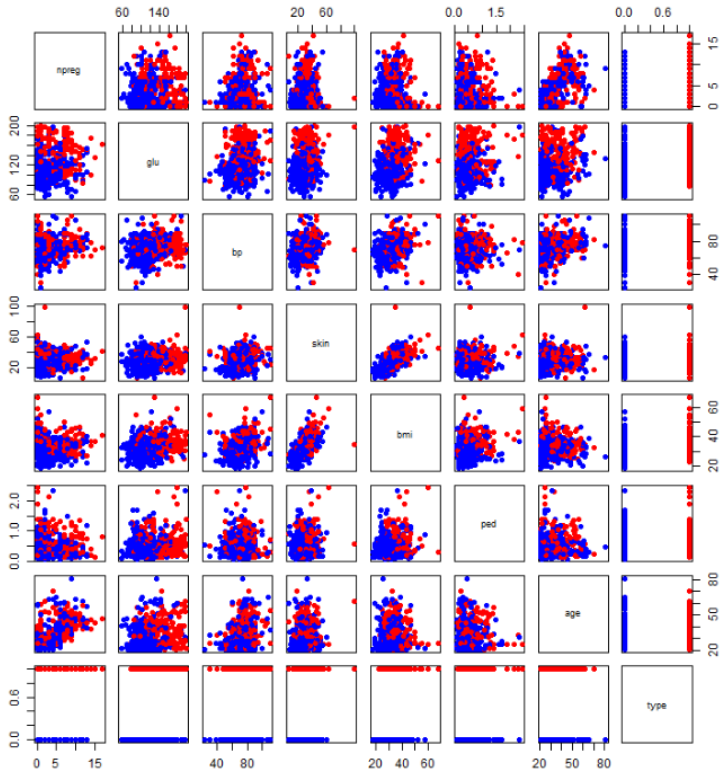

Figure 2. Scatterplot matrix of all independent variables and dependent variable.

	npreg	glu	bp	skin	bmi	ped	age	type
npreg	1.000000000	0.1253296	0.204663421	0.09508511	0.008576282	0.007435104	0.64074687	0.2525855
glu	0.125329647	1.0000000	0.219177950	0.22659042	0.247079294	0.165817411	0.27890711	0.5036139
bp	0.204663421	0.2191779	1.000000000	0.22607244	0.307356904	0.008047249	0.34693872	0.1834319
skin	0.095085114	0.2265904	0.226072440	1.00000000	0.647422386	0.118635569	0.16133614	0.2548737
bmi	0.008576282	0.2470793	0.307356904	0.64742239	1.000000000	0.151107136	0.07343826	0.3009007
ped	0.007435104	0.1658174	0.008047249	0.11863557	0.151107136	1.000000000	0.07165413	0.2330739
age	0.640746866	0.2789071	0.346938723	0.16133614	0.073438257	0.071654133	1.00000000	0.3150968
type	0.252585511	0.5036139	0.183431874	0.25487371	0.300900748	0.233073898	0.31509683	1.0000000

>

Figure 3. Correlation matrix of all variables

Furthermore, the correlation matrix is also plotted to identify the linear relationships between the variables in our dataset. A Pearson correlation coefficient between two variables can be found in each cell in the matrix. A value that is closer to 1 implies a strong positive correlation implying if one variable increases the other variable also will increase. On the other hand, -1 implies a negative correlation implying if one variable increases, the other decreases. A value closer to 0 represents a weak correlation. Similarly, our scatterplot matrix (Figure 3) illustrates the same result. The following were discovered as positive correlations:

- Number of pregnancies and Age – 0.641
- Skin thickness and BMI – 0.647
- Glucose and Type – 0.504

Table 2. Diabetic vs Non – Diabetic Average

	Non- Diabetic (0)	**Diabetic (1)**
Number of Pregnancies	2.92	4.70`
Glucose levels	110.01	143.11
Blood Pressure	69.91	74.70
Skin Fold Thickness	27.29	32.97
BMI	31.42	35.81
Pedigree Function	0.44	0.61
Age	29.22	36.41

To conclude this statistical analysis, most of the findings seem logical. In terms of pregnancy history, diabetes has an average of 4.70 pregnancies, which is higher than the 2.93 average associated with non-diabetics. This suggests that a higher number of pregnancies can be related to an increased risk of diabetes, although we cannot confirm it based just on this observation. As expected from our initial analysis people who are diagnosed with diabetes have an elevated average glucose level of 143.12 which is relatively high compared to non-diabetic people with 110.02. This distinction is consistent in the medical context characterized by elevated blood sugar levels. People with diabetes also tend to have a slightly higher level of blood pressure than non-diabetic people. Additionally, BMI and pedigree functions which are core representatives of body fat are higher for people with Diabetes. This established the link between high body fat percentage leads to increased risk of diabetes. The pedigree function reveals a higher mean score among diabetic people, indicating a stronger hereditary inclination toward the disease. It is evident that this confirms the role of genetics in being prone to diabetes. Lastly, people in our study who have diabetes have a mean age of 36.41 years compared to 29.22 years for people without diabetes which can be recognized as diabetes risk increases with Age. Therefore, with these insights, we can conclude our EDA and proceed to the modeling stage of this analysis.

2| Modelling and subset selection

We will construct a simple logistic regression model first, for our data, setting the response variable to 'type' and using the rest of them as predictor variables. This is also done using <glm()> function to fit our logistic regression model and the results will produce the coefficients for each predictor and their significance. (Figure 4) below illustrates the model result –

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.554651   0.994217  -9.610  < 2e-16 ***
npreg        0.122517   0.043743   2.801 0.005097 **
glu          0.035321   0.004244   8.322  < 2e-16 ***
bp          -0.007695   0.010314  -0.746 0.455602
skin         0.006774   0.014759   0.459 0.646242
bmi          0.082678   0.023334   3.543 0.000395 ***
ped          1.308708   0.364040   3.595 0.000324 ***
age          0.026375   0.014000   1.884 0.059581 .
```

Figure 4. The output of the logistic regression model

From the model output (the entire results can be found in the appendix), the intercept is estimated to be -9.554651 which can be interpreted as the log odds of having diabetes when all the predictors are set to zero. However, it is important to note that this scenario is not practical or realistic due to the nature of our variables. The glucose level(glu) coefficient was determined to be extremely significant ($p<0.001$) with an estimate of 0.035. This means that the log odds of getting diabetes rise by roughly 0.035 for every unit increase in the glucose level. This was consistent with our EDA findings, where we discovered that glucose levels differed between the two groups and showed a positive link with diabetes status. Number of pregnancies(npreg) was also shown to be a significant predictor ($p=0.005$). its coefficient of 0.122 indicates that for each consecutive pregnancy, the log odds of being diabetic increase by around 0.123. Once again, this result supports the findings from EDA that diabetic people had a greater mean number of pregnancies than non-diabetic people. BMI's coefficient (bmi) was estimated at 0.0826 and was statistically significant ($p=0.0003$). This indicates that for every unit increase in BMI, the log odds of developing diabetes are increased by 0.083. This also favors our EDA result. The Diabetes Pedigree Function (ped) has a coefficient of 1.308 which actually shows a strong influence on the likelihood of diabetes. The p-value of 0.0003, which was also statistically significant. Although we did not emphasize this in our EDA and its inclusion in this regression model stresses the importance of this variable. Age's coefficient was 0.026 which implies that there with advancing age, there is a modest rise in the likelihood of developing diabetes. The p-value was very borderline ($p= 0.059$) implying that age may have a role in diabetes prediction but it is not the best predictor in our model. Lastly, blood pressure(bp) and skin fold thickness('skin') did not have great p-values which implies that they may not be the best predictors for predicting diabetes.
Finally, the AIC value of the model was 482.32 and the residual deviation on 524 degrees of freedom was 466.32 indicating a good model fit to the data.

2.1 Best subset Analysis –

To determine the ideal number of independent variables for our mode, and to ensure minimal error and maximum predictive accuracy, we are conducting a Best Subset Analysis. However, as the number of predictors increases, there is a risk of overfitting, which occurs when the model fits the training data well but performs badly on fresh unknown data. To combat this challenge, and select the best model, we employ assessment features like as Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). This strategy encompassed three distinct factors of model error which are AIC, BIC, and k- fold Cross Validation.

In figure.5 There are three separate graphs, each showcasing a unique error metric for the optimal variable selection at k values. By analysing AIC and BIBC against the number of predictors(K), it is evident the way it affects model selection criteria. AIC selected 5 predictors which are “npreg”, “glu”, “bmi”, “ped” and “age”. On the other hand, adding new predictors incurs a higher cost in BIC. The best model with BIC employs 4 predictors that are “npreg”, “glu”, “bmi”, “ped”. It is worth noting that both of them chose only a limited number of predictors out of 7 implying that only a few variables are important for explaining the variance in the dependent variable. The intercept for the AIC model is -1.079 and for the BIC model, it is -1.018.

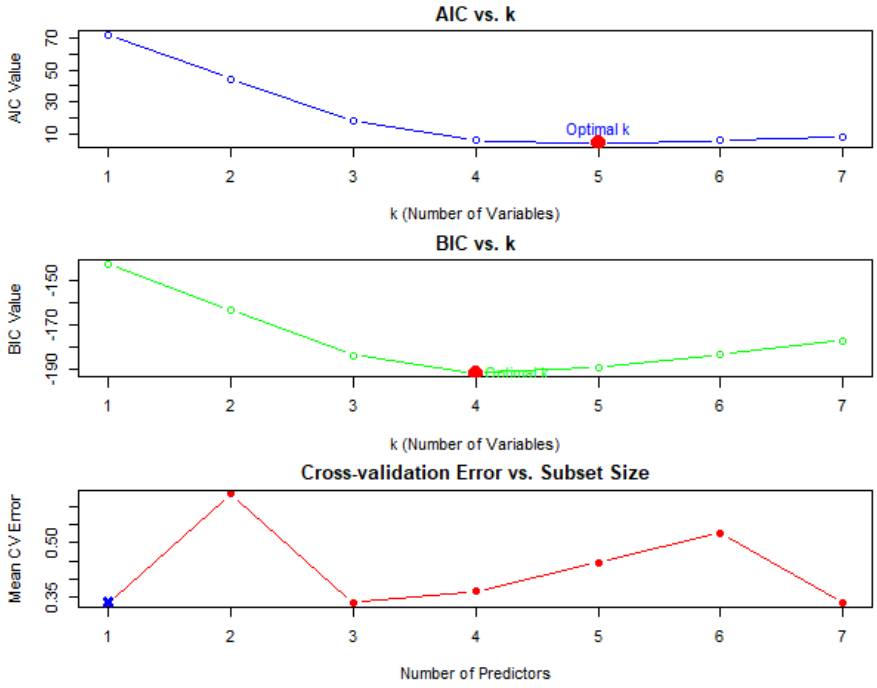

Figure 5. Best subset Analysis plots for each error measure for each number of variables with ideal k value measure indicated in red.

Based on cross-validation error, an optimal model only contains one predictor which is glucose (“glu”). This predictor seems to be the most important in determining whether or not a person has diabetes which seems logical. With each incremental unit increase in “glu” the log odds of the specific type witness an increase of 0.0077. The value of the intercept is -0.5946 and the log odds of type are -0.5946 when “glu” is 0. After careful consideration, I chose BIC as my leading metric since it strikes a delicate balance, while it evaluates a model’s

fit to observable data, it also incorporates a higher penalty for model complexity compared to AIC. Also, its properties to favor a simpler model keeps us away from overfitting especially with large datasets. Moreover, K-fold cross-validation only indicated a model with one predictor which simplifies the underlying data patterns. Given that our goal is to obtain a model that is both accurate and efficient, BIC preference for simpler models aligns with our overall objectives. Therefore, the BIC subset was used to construct and test out the linear discriminant analysis (LDA) model and the Quadratic discriminant analysis (QDA) model.

In the context of LDA model, the average value for several predictors within each class reveals information about the typical predictor values for those classes. The “glu” predictor has mean values of 110.016 and 143.11 for the two groups implying a possible link between higher “glu” values and the group corresponding to 1 (people with diabetes). Similarly, the QDA model group means to indicate the average predictor values for each class. However, because QDA allows for various covariance structures for each class, it allows for a more flexible boundary between both classes. For example, the group means for BMI, 31.429, and 35.819 show us how the values of this predictor differ across the two groups. Hence, these parameters are critical in understanding the complex interaction between predictors and responses. They not only show which predictors are important, but they also provide information on the direction and degree of these relationships.

3| Model performance and comparison-

Now that we have built our models, we are going to compare their performance to determine which is the best to predict outcomes on raw-unseen data. We will be evaluating Logistic regression, LDA, and QDA and examine their test errors through cross-validation. The 10-fold cross-validation method was used consistently for all our models to provide a fair comparison. This method splits the dataset randomly into ten subgroups, training the model on nine of them and testing it on the tenth, iterating until each subset has served as a test set.

First, comparing the accuracies, logistic regression proved to have 79.14% during the testing phase. It misclassified about 20.11% of the training set aligning to its test error 20.86%. The LDA model had a test error of 21.99% and its training error was slightly higher at 21.05%. Although, LDA functions are based on the assumption that predictors, regardless of class, have a common covariance matrix. However, it may not be consistent with the properties of every dataset. QDA had a test error similar to LDA 21.99% and avoid the assumption that was indicated earlier. The unusual flexibility of this model was clear as it had the largest training error 22.18%.

It was critical to approach a model that neither oversimplified the underlying relationships which would result in underfitting nor become extremely sensitive to the training data. Our training errors for logistic regression and QDA are evident for this strategy. The lower training error suggests that it was a strong fit for the training data but this also has to be considered alongside test error to determine the model’s ability y generalize. The use of Bic

for feature selection demonstrated the idea of using features strategically. Our primary objective was to select simpler models that matched the data and make statistically relevant.

4| Conclusion –

Overall, all the models we built and tested today had their unique strengths and weaknesses, our training errors of logistic regression (0.201), LDA (0.211), and QDA(0.222) gave us a unique perspective on how each model performed in this dataset. Logistic regression was the most suitable model and narrowly outperformed with 79.14% accuracy and performed well on the test data. LDA and QDA models both had a similar accuracy of 78.01%. finally, expanding our feature selection strategy beyond might have helped in forming an ideal predictor combination.