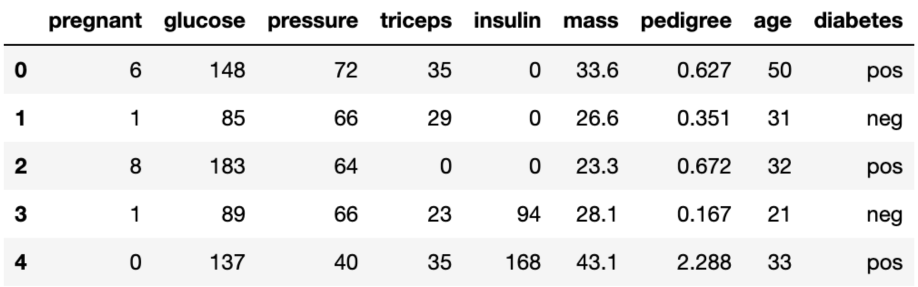
Arya Gupta - 918576066

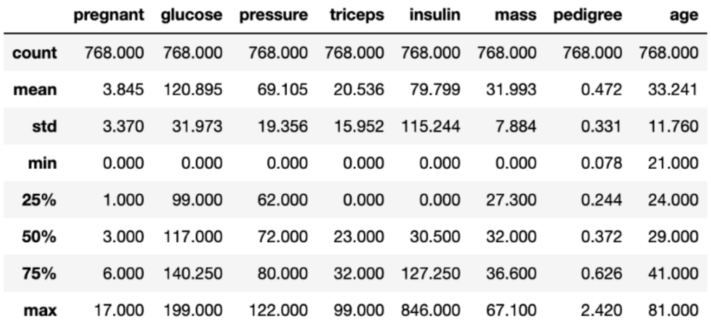
In this report, I aim to predict diabetes onset in female Pima Indians using historical medical data. The Pima Indians Diabetes dataset, sourced from the 'mlbench' package in R, comprises medical diagnostic measurements from women of Pima Indian descent. This dataset serves to investigate diabetes prevalence and its predictors within this high-risk group.



	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age	diabetes
0	6	148	72	35	0	33.6	0.627	50	pos
1	1	85	66	29	0	26.6	0.351	31	neg
2	8	183	64	0	0	23.3	0.672	32	pos
3	1	89	66	23	94	28.1	0.167	21	neg
4	0	137	40	35	168	43.1	2.288	33	pos

Figure 0: The first five rows of the diabetes dataset, featuring eight predictor variables and one target variable, 'diabetes'. Originally sourced from R, I transferred it into a CSV file for analysis in Python.

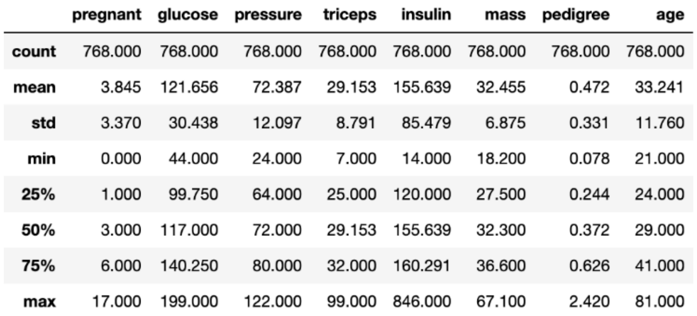
Before Imputation



	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age
count	768.000	768.000	768.000	768.000	768.000	768.000	768.000	768.000
mean	3.845	120.895	69.105	20.536	79.799	31.993	0.472	33.241
std	3.370	31.973	19.356	15.952	115.244	7.884	0.331	11.760
min	0.000	0.000	0.000	0.000	0.000	0.000	0.078	21.000
25%	1.000	99.000	62.000	0.000	0.000	27.300	0.244	24.000
50%	3.000	117.000	72.000	23.000	30.500	32.000	0.372	29.000
75%	6.000	140.250	80.000	32.000	127.250	36.600	0.626	41.000
max	17.000	199.000	122.000	99.000	846.000	67.100	2.420	81.000

Figure 1

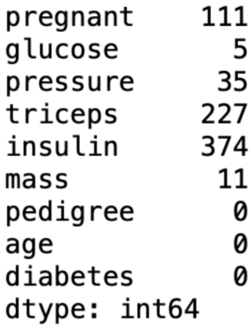
After Imputation



	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age
count	768.000	768.000	768.000	768.000	768.000	768.000	768.000	768.000
mean	3.845	121.656	72.387	29.153	155.639	32.455	0.472	33.241
std	3.370	30.438	12.097	8.791	85.479	6.875	0.331	11.760
min	0.000	44.000	24.000	7.000	14.000	18.200	0.078	21.000
25%	1.000	99.750	64.000	25.000	120.000	27.500	0.244	24.000
50%	3.000	117.000	72.000	29.153	155.639	32.300	0.372	29.000
75%	6.000	140.250	80.000	32.000	160.291	36.600	0.626	41.000
max	17.000	199.000	122.000	99.000	846.000	67.100	2.420	81.000

Figure 2

In my analysis, I discovered biologically impossible "zero" values in several columns (refer to Figure 3 for counts for each column). These 'zero' values, likely placeholders for missing data, were sparse in the 'glucose', 'mass', and 'pressure' columns, allowing me to simply replace them with the median values of their respective columns to preserve the data distribution. Conversely, the 'triceps' and 'insulin' columns exhibited a high frequency of zeros, prompting the use of iterative imputation. This method estimates the missing values using other features, offering a more accurate approximation of the extensive missing data. Figures 1 and 2 display descriptive statistics before and after imputation, showing noticeable increases in the means of 'insulin' and 'triceps'. Future visualizations will use this imputed dataset.



```
pregnant    111
glucose       5
pressure     35
triceps     227
insulin     374
mass         11
pedigree      0
age           0
diabetes      0
dtype: int64
```

Figure 3

Figure 4 presents a heatmap illustrating the correlations between various variables and the target, depicted by both magnitude and color intensity. Notably, 'glucose' and 'mass' show strong correlations with diabetes outcomes, making them key variables for my analysis. Additionally, Figure 5 highlights a somewhat imbalanced dataset, with a larger number of non-diabetic individuals compared to diabetic ones.

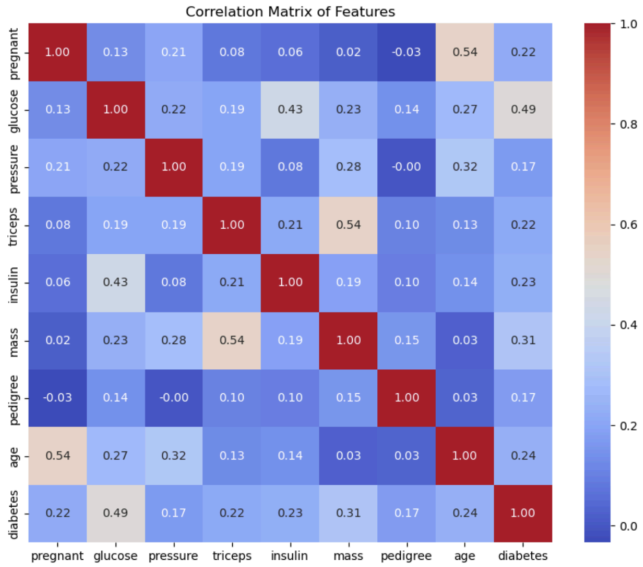


Figure 4

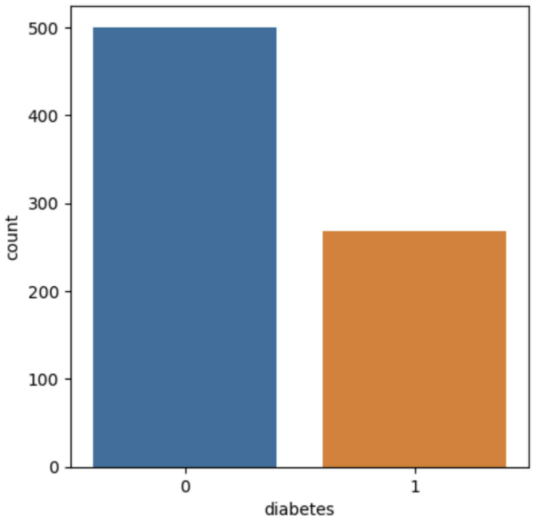


Figure 5

Moving forward, Figure 6 visualizes the distributions of 'glucose' and 'mass,' displaying roughly normal curves with a slight right skew. The box plots in Figure 7 demonstrate significantly higher glucose levels in diabetic individuals compared to non-diabetic ones. Likewise, it shows a higher BMI in diabetic individuals compared to non-diabetic ones.

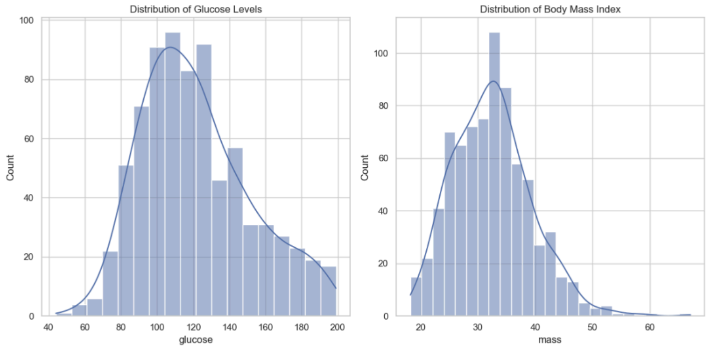


Figure 6

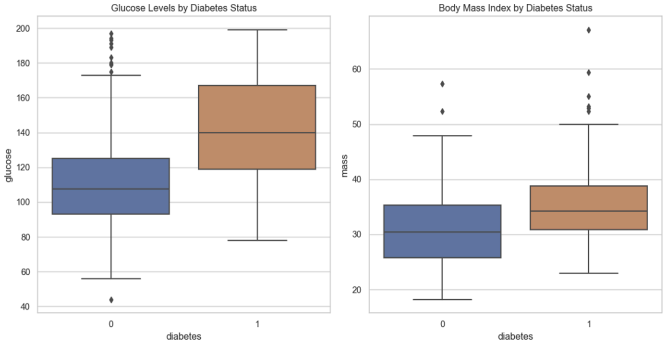


Figure 7

Methodology - PCA: Principal Component Analysis (PCA) is a statistical technique for dimensionality reduction that aims to preserve data variability. It transforms original variables into Principal Components (PCs), which are orthogonal linear combinations of the original variables ranked by the amount of variance they capture. This process simplifies the model while minimizing information loss.

The PCA process began by standardizing the features to ensure that the analysis was not biased towards variables with greater variance. I then implemented PCA using the scikit-learn PCA module. To make an informed decision on the number of components to retain, I adopted a three-pronged approach: analyzing cumulative explained variance, utilizing a scree plot, and conducting parallel analysis.

Cumulative explained variance quantifies the variance contribution of each component, helping me decide how many components are necessary to capture roughly 80% of the total variance. A scree plot is a visual tool which plots the eigenvalues in descending order against their indices, illustrating the variance captured by each component. This plot is essential for identifying the "elbow point," where the addition of more components provides minimal increase in explained variance. Lastly, parallel analysis compares the eigenvalues with those generated from random data of the same size and scale. This analysis is vital to determine if the components retained represent true underlying patterns rather than noise.

Methodology - LDA/QDA:

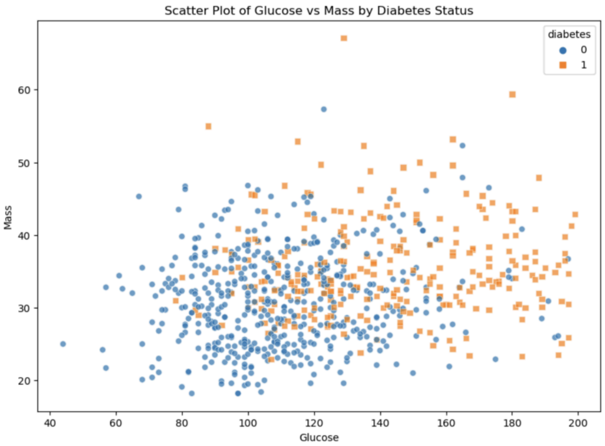


Figure 8

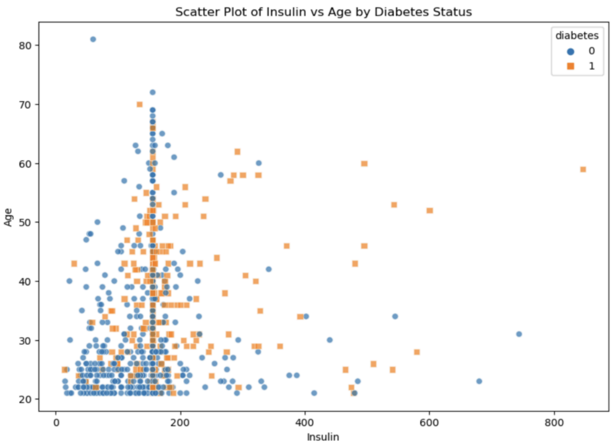
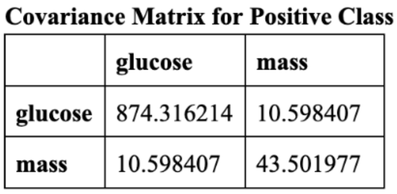


Figure 9

My next step was to proceed with discriminant analysis techniques on the retained principal components, distinguishing between Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA).

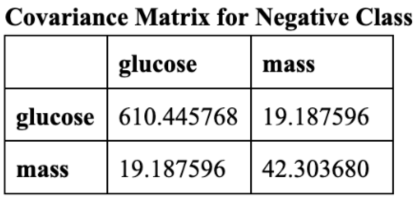
LDA is a classification method designed to identify a linear combination of features that separates two or more classes. It is most effective when the data distributions for each class share the same covariance matrix and follow a normal distribution. This method excels in situations where the separation between classes is linear.



Covariance Matrix for Positive Class

	glucose	mass
glucose	874.316214	10.598407
mass	10.598407	43.501977

Figure 10



Covariance Matrix for Negative Class

	glucose	mass
glucose	610.445768	19.187596
mass	19.187596	42.303680

Figure 11

On the other hand, QDA permits each class to have its own covariance matrix, which accommodates nonlinear class boundaries. This flexibility is particularly advantageous for our dataset, as evidenced by scatter plots of feature pairs (shown in Figures 8 and 9). These plots reveal complex, non-linear interactions, suggesting quadratic decision boundaries are more appropriate due to significant overlap in the middle ranges of these features. Further examination of the covariance matrices for diabetic and non-diabetic classes (illustrated in Figures 10 and 11) showed pronounced differences, especially in variables like glucose and mass, which, as shown previously, are highly correlated with diabetes outcomes. These variations in variances and covariances highlighted the suitability of QDA, which can more effectively model the distinct characteristics of each class compared to LDA. With these in mind, I decided to move forward with QDA.

For model training and evaluation, I used the selected principal components as inputs for the QDA model. I divided the dataset into training and testing subsets, allocating 70% of the data for training and the remaining 30% for testing. This split helps ensure that the model generalizes well to new data and allows me to detect potential overfitting. I evaluated the model's predictive accuracy using standard metrics such as a confusion matrix and accuracy score. Additionally, I employed a stratified 5-fold cross-validation approach to further enhance the model's stability and ensure its generalizability across different data subsets.

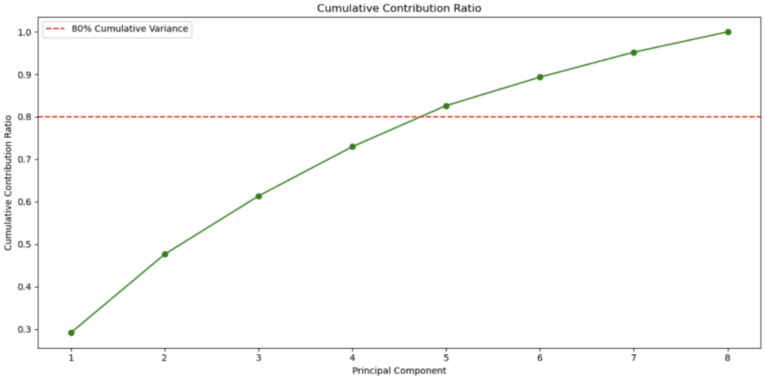


Figure 12

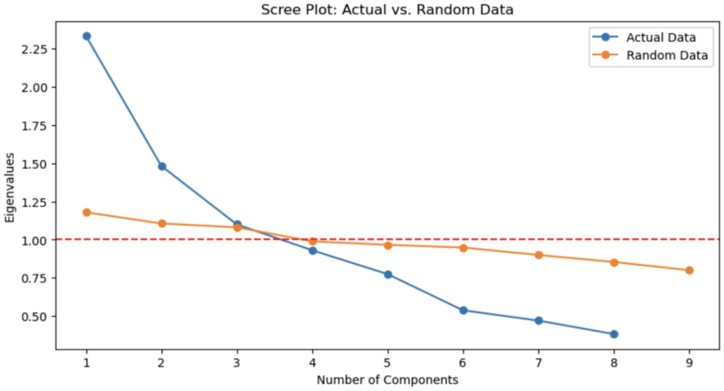
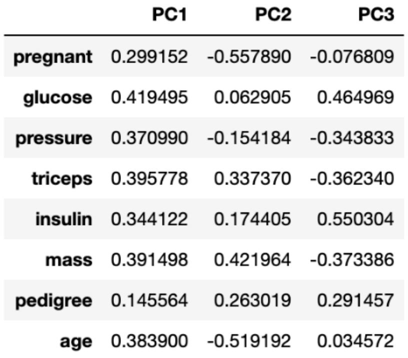


Figure 13

PCA Results: As shown in Figures 15 and 16 below, my PCA analysis found that the first five principal components explained over 80% of the variance, meeting my initial target. Figure 12's Cumulative Contribution Ratio plot confirms this, justifying the initial selection of these components based on their significant variance contribution.

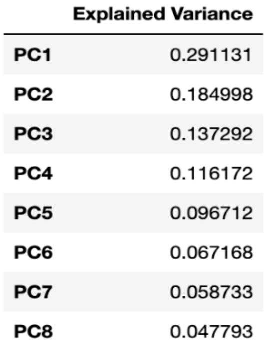
However, despite the quantitative appeal of retaining five components, the scree plot and parallel analysis suggests a more conservative approach. The scree plot in Figure 13 visually demonstrates an "elbow" point around the third component. Additionally, the Kaiser criterion, which suggests dropping all components with eigenvalues under 1.0, is satisfied after the third component. Parallel analysis in the same figure further supports this, indicating that only the first three components have eigenvalues significantly higher than those from a random dataset, capturing genuine, non-random patterns.

Therefore, despite accounting for only 61% of the explained variance, focusing on the first three principal components helps me balance between maximizing the explained variance and ensuring the components retained are statistically robust and not influenced by noise. The data frame in Figure 14 shows the loadings for each of these first three components. Interpreting these loadings help me gain insights into how different variables contribute to major variance factors within the data, and may also assist in creating new features for predictive modeling.



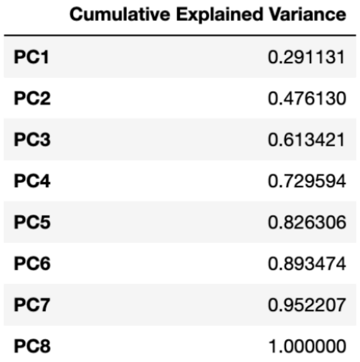
	PC1	PC2	PC3
pregnant	0.299152	-0.557890	-0.076809
glucose	0.419495	0.062905	0.464969
pressure	0.370990	-0.154184	-0.343833
triceps	0.395778	0.337370	-0.362340
insulin	0.344122	0.174405	0.550304
mass	0.391498	0.421964	-0.373386
pedigree	0.145564	0.263019	0.291457
age	0.383900	-0.519192	0.034572

Figure 14



	Explained Variance
PC1	0.291131
PC2	0.184998
PC3	0.137292
PC4	0.116172
PC5	0.096712
PC6	0.067168
PC7	0.058733
PC8	0.047793

Figure 15



	Cumulative Explained Variance
PC1	0.291131
PC2	0.476130
PC3	0.613421
PC4	0.729594
PC5	0.826306
PC6	0.893474
PC7	0.952207
PC8	1.000000

Figure 16

QDA Results: The QDA model achieved a predictive accuracy of 73.16% on the testing dataset, effectively distinguishing between diabetic and non-diabetic cases. Cross-validation showed a consistent average accuracy of 73.5% with a standard deviation of 1.4%, indicating no overfitting to the training set. The confusion matrix (Figure 17) indicates that the model correctly identified 123 non-diabetic and 46 diabetic cases but misclassified 28 non-diabetic as diabetic and missed 34 diabetic cases. This resulted in a Sensitivity (True Positive Rate) of 57.5% and a Specificity (True Negative Rate) of 81.46%.

The higher number of False Negatives relative to False Positives indicates a tendency of the model to err on the side of caution. However, this pattern could be problematic where failing to diagnose diabetes has more severe consequences than false alerts. These results highlight the benefits of using QDA to address non-linear relationships in complex medical data. Further analysis should focus on optimizing the sensitivity-specificity trade-off and investigating whether including more or different principal components could improve diagnostic accuracy.

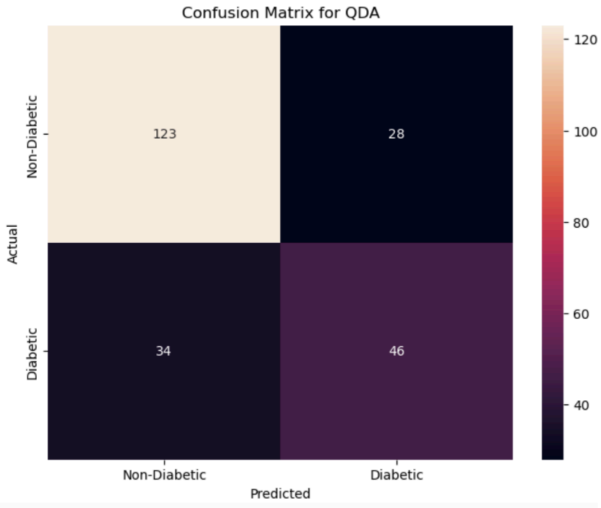


Figure 17

Conclusion: My study on the Pima Indians Diabetes dataset integrated PCA and QDA to enhance the understanding and prediction of diabetes. This approach allowed me to effectively reduce the dimensionality of the dataset while retaining the most significant features that encapsulate the primary variance within the data. Retaining three principal components was justified by comprehensive visualizations and parallel analysis, ensuring that they were statistically significant and relevant in capturing the underlying patterns in the data. QDA took advantage of its capacity to model non-linear boundaries through class-specific covariance matrices. The QDA model achieved a predictive accuracy of 73%, with substantial success in identifying true negatives, although it also pointed out the challenges of managing false negatives within the model.