Scuola Superiore di Studi Universitari e di Perfezionamento Sant'Anna

Statistical Learning and Large Data

a.a. 2024-2025

From Missing Data to Model Performance: A Full Pipeline for Diabetes Classification in Pima Indian Females

Gabriele Cini (Module 1-2), Flavio Di Lisio (Module 1-2),
Giorgia Di Santolo (Module 1)

Abstract

*Diabetes is a chronic progressive disease with growing numbers worldwide, it places a heavy burden both on individuals and on the healthcare systems. Its earlier manifestations are reversible, yet clinically silent; a diagnosis is often made only after irreversible complications have developed. Diabetic retinopathy remains a leading cause of adult blindness in the US [1] and diabetic foot ulcers frequently lead to lower limb amputations [2]. In this work, we aim to train a supervised classifier able to efficiently discriminate diabetes patients from healthy individual, using only non invasive, routinely collected measurements. We train and confront different models using the Pima Indians Diabetes Database obtained from Kaggle [3] (*768 *adult female records with pregnancies, diastolic blood pressure, triceps skinfold thickness, body mass index, diabetes pedigree function and age). After data cleaning and exploratory analysis, our methods of feature selection identified body mass index, diabetes pedigree function, number of pregnancies and glycemia as best predictors. We tuned decision tree, random forest, linear discriminant analysis, naïve Bayes, logistic regression, support vector machine and K-nearest neighbors models via stratified* 5*-fold cross-validation. We evaluated each on a held-out test set and computed* 95% *confidence intervals for accuracy, recall, precision and* $F1$*-score. Performances were similar across models: accuracy ranged from* 0.72 *to* 0.77*, recall from* 0.53 *to* 0.77*, precision from* 0.52 *to* 0.71 *and* $F1$*-score from* 0.61 *to* 0.69*. Logistic regression achieved the most balanced performance (accuracy* $= 0.77$*; recall* $= 0.71$*; precision* $= 0.65$*;* $F1$*-score* $= 0.67$*). Our findings demonstrate that machine learning models can be effectively trained to predict the presence of diabetes using only non-invasive clinical features.*

Contents

1 Introduction

Diabetes mellitus represents one of the most significant public health challenges of the 21st century, with an estimated 830 million people living with diabetes worldwide as of 2022 [4]. The disease is characterized by a state of chronic hyperglycemia resulting from defects in insulin secretion, insulin action, or both. This leads to disturbances in carbohydrate, fat, and protein metabolism.

Prolonged untreated hyperglycemia can lead to severe microvascular and macrovascular complications affecting multiple systems. Diabetic retinopathy remains the leading cause of preventable blindness among working-age adults [1], while diabetic nephropathy is the primary cause of end-stage renal disease in many developed countries [5]. Cardiovascular complications account for approximately 65% of mortality in diabetic patients, and diabetic neuropathy affects nearly 50% of individuals with long-standing diabetes, frequently leading to diabetic foot ulcers and lower limb amputations [2].

The Pima Indian population of Arizona has been extensively studied in diabetes research due to their extraordinarily high prevalence of type 2 diabetes, estimated at 38% among adults aged 20 years and older—one of the highest rates globally [6].

In this work, we aim to train classification algorithms and compare the results to identify the most robust classifier. In our dataset, we only use features routinely obtained in a clinical setting, thus enhancing the translational potential of our work.

2 Data

The dataset we used was sourced from Kaggle [3], and comes from the National Institute of Diabetes and Digestive and Kidney Diseases. It is a collection of 768 female's patients at least 21 years old of Pima Indian heritage. This community has a high incidence rate of diabetes, making it very interesting to analyze. For each patient a categorical output variable was assigned, called Outcome, indicating if the person had or had not manifested Diabetes, and eight medical continuous variables were measured:

- "Pregnancies": number of times pregnant,

- "Glucose": Plasma Glucose Concentration at 2 Hours in an Oral Glucose Tolerance Test (GTT),
- "BloodPressure": Diastolic Blood Pressure (mmHg),
- "SkinThickness": Triceps Skin Fold Thickness (mm),
- "Insulin": 2-Hour Serum Insulin (µU/ml),
- "BMI": Body Mass Index (Weight in kg / (Height in m)^2),
- "DiabetesPedigreeFunction": estimate of individual's genetic risk of developing diabetes by considering the presence of the disease in close relatives (parents, grandparents, siblings, aunts, uncles, and first cousins) and their degree of relatedness [7],
- "Age" (years).

The distribution of the continuous variables and the Bar Plot of the Outcome are displayed in **Figure 1**. First, examining the Bar Plot, we saw that the dataset is not balanced since about two thirds are not affected (500 to 268). Second, it is evident from the distributions that certain variables — "BloodPressure", "SkinThickness", "Insulin", "BMI", and "Glucose" — have a higher number of "0". Since the values of all these variables, with the exception of Insulin, could not be "0" from a medical perspective, we substituted a "NaN" for the "0". The statistics (mean, variance, lowest and maximum value, median, first and third quartile) related to each variable are collected in **Table 1**.

Table 1: Data before cleaning and scaling

	Observations	Mean	Variance	Min	1st qrt.	Median	3rd qrt.	Max	NaN	% NaN
Pregnancies	768	3.845	10.970	0	1	3	6	17	0	0.000%
Glucose	763	121.7	960.952	44.0	99.0	117.0	142.0	199.0	5	0.651%
Blood Pressure	733	72.41	151.542	24.0	64.0	72.0	80.0	122.0	35	4.557%
Skin Thickness	541	29.15	110.752	7.0	7.0	22.0	36.0	99.0	227	29.557%
Insulin	768	79.8	15130.859	0.0	0.0	30.5	127.2	846.0	0	0.000%
BMI	757	32.46	47.350	18.2	27.5	32.3	36.6	67.1	11	1.432%
Diabetes Pedigree function	768	0.4719	0.119	0.078	0.2437	0.3725	0.6262	2.42	0	0.000%
Age	768	33.24	115.812	21	24	29	41	81	0	0.000%

	Observations	Mean	0	1	NaN	% NaN
Outcome	768	0.349	500	268	0	0.000%

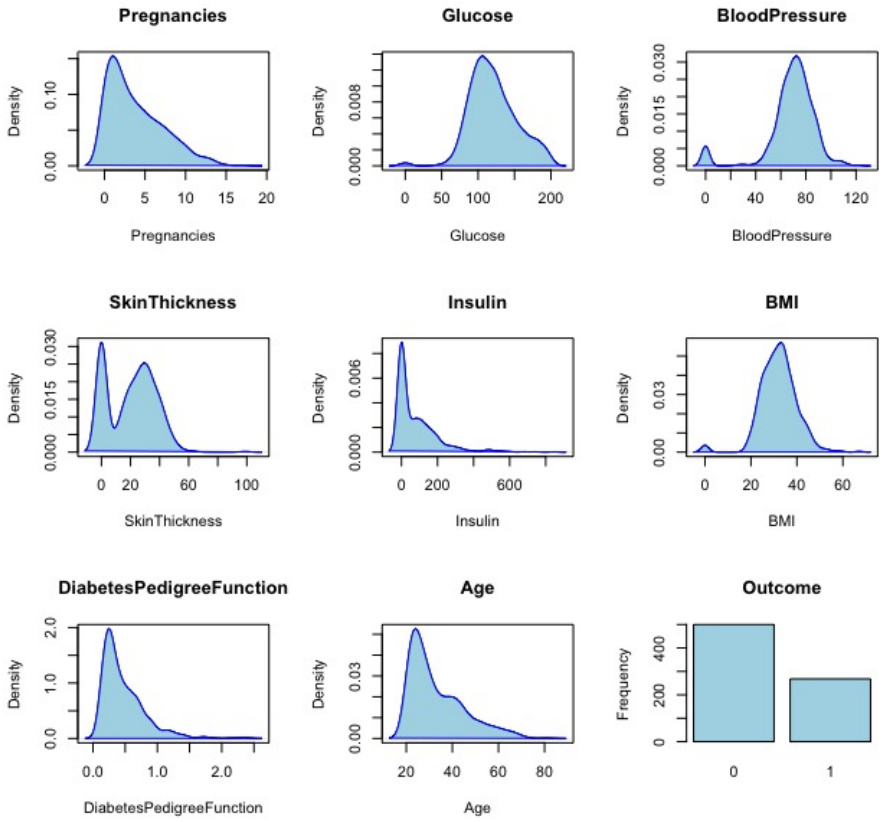


Figure 1: Density plot of the eight continuous variables prior to cleaning and scaling, along with a Bar plot of the target variable.

Then, in order to make our dataset clean, we eliminated any rows that had at least one "NaN" in "BloodPressure", "BMI", and "Glucose". This left us with 724 entries.

However, our approach in treating the "SkinThickness" variable was substantially different. Given the high number of missing values (measurements stating a 0 mm thickness of the triceps skin), which constituted almost a third of the dataset (29.56%), we decided that it was not possible to drop all these observations. Losing 1 out of 3 observations would have had detrimental effects on the amount of information at our disposal; it was then necessary to develop a predictive model capable of estimating our missing values for this variable.

In order to predict the missing "SkinThickness" values, we decided to develop a knn-model trained on the available information. This model would take the nearest k points in n-dimensional space calculated with Euclidean distance from our missing observation

(in other words, k points with the most similar values for variables other than "SkinThickness") and then would assign their arithmetic mean as the expected value for the variable "SkinThickness".

Firstly, we looked at the performance of this model on a training set containing all the complete case observations. We estimated an Absolute Mean Percentage Error (MAPE) of about 10% for a dozen of simulations, with better accuracy for middle values and less precisions for extreme values; however, the model performed much better than other standard data prediction approaches such as using the mean or the median observed value (around 35% MAPE each). Secondly, we substitute all the missing values in the dataset with our predicted values.

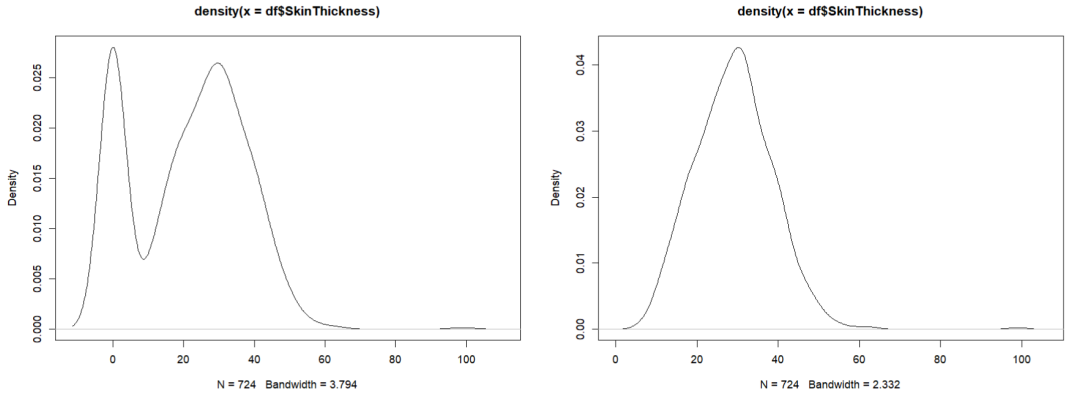


Figure 2: Density plot of the "SkinThickness" variable before and after applying our predictive model (before scaling).

As we can see in **Figure 2** the distribution of the "SkinThickness" variable drastically improved after adopting our knn predictive model, making it easier to work with such data.

Following predicting and cleaning, all of the updated statistics are compiled in **Table 2**.

We then used a Min-Max Scaling approach to get all of the variables to values between 0 and 1. We chose to use this method, which does not assume a normal distribution, because we do not have a normal distribution for all of the continuous variables. **Figure 3** displays the updated distributions of the continuous variables. From these we identified asymmetric distributions for the variables: "Pregnancies", "Insulin", "DiabetesPedigree-

Table 2: Cleaned data before scaling

	Observations	Mean	Variance	Min	1st qrt.	Median	3rd qrt.	Max	NaN	% NaN
Pregnancies	724	3.866	11.308	0	1	3	6	17	0	0
Glucose	724	121.88	945.564	44.0	99.8	117.0	142.0	199.0	0	0.000%
Blood Pressure	724	72.4	153.216	24.0	64.0	72.0	80.0	122.0	0	0.000%
Skin Thickness	724	29.13	93.531	7.0	22.3	29.0	35.3	99.0	0	0.000%
Insulin	724	84.49	13692.864	0.0	0.0	30.5	130.5	846.0	0	0.000%
BMI	724	32.47	47.458	18.2	27.5	32.4	36.6	67.1	0	0.000%
Diabetes Pedigree function	724	0.4719	0.11	0.078	0.2450	0.3790	0.6275	2.42	0	0.000%
Age	724	33.35	138.424	21	24	29	41	81	0	0.000%

	Observations	Mean	0	1	NaN	% NaN
Outcome	724	0.3439	475	249	0	0.000%

Function" and "Age". We expected these right skewed distributions: for "Pregnancies" because it highlights fewer women with more children; for "Insulin", which is caused by a greater number of patients with extremely low levels; "Age", which represents the distribution of the population. And the "DiabetesPedigreeFunction", which is a ratio of relatives who had diabetes to those who did not; it is reasonable to assume that the majority of subjects have values below 1, while a smaller percentage of patients have higher values.

Table 3 displays the dataset's statistics following all of our processing.

Table 3: Cleaned and scaled data

	Observations	Mean	Variance	Min	1st qrt.	Median	3rd qrt.	Max	NaN	% NaN
Pregnancies	724	0.227	0.0391	0	0.059	0.176	0.353	1	0	0
Glucose	724	0.502	0.0394	0	0.360	0.471	0.632	1	0	0.000%
Blood Pressure	724	0.494	0.0160	0	0.408	0.490	0.571	1	0	0.000%
Skin Thickness	724	0.241	0.0110	0	0.166	0.239	0.308	1	0	0.000%
Insulin	724	0.100	0.0160	0	0.000	0.057	0.154	1	0	0.000%
BMI	724	0.292	0.0198	0	0.190	0.290	0.376	1	0	0.000%
Diabetes Pedigree function	724	0.169	0.0201	0	0.071	0.129	0.235	1	0	0.000%
Age	724	0.206	0.0385	0	0.050	0.133	0.333	1	0	0.000%

	Observations	Mean	0	1	NaN	% NaN
Outcome	724	0.3439	475	249	0	0.000%

To determine if some variables exhibit strong correlation values, we produced a correlation matrix, which is shown in **Figure 4**. This was not the case, as the highest value, between "BMI" and "SkinThickness", was 0.58, as we expected considering that it makes sense to assume that high BMI values are associated with high skin thickness levels. Following this, there are also weak logical correlation pairings between "Age" and "Pregnancies" (0.56), "Insulin" and "Glucose" (0.34), and "BloodPressure" and "Age" (0.32).

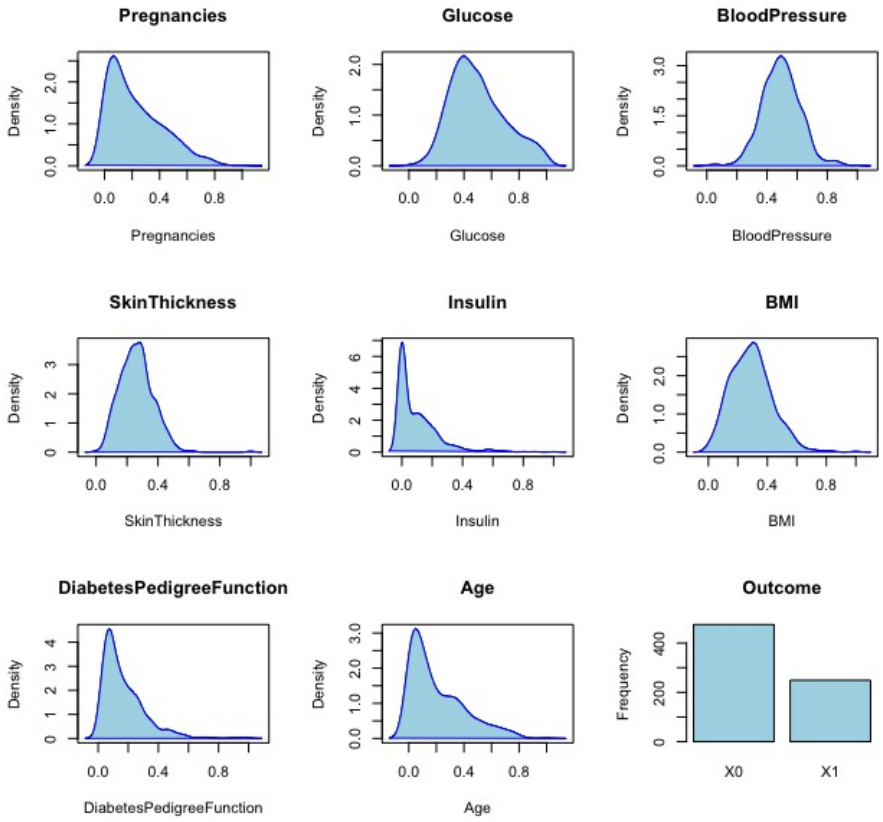


Figure 3: Density plot of the eight continuous variables following cleaning and scaling, along with a Bar plot of the target variable.

3 Statistical Methods

3.1 Dimensionality reduction

We used Principal Components Analysis (PCA), an unsupervised linear dimension reduction technique that maintains as much variability as possible, on the scaled dataset to examine the variables. Regarding the percentage of explained variance, four components are needed to reach 75% explainancy; nevertheless, PC1 and PC2 alone account for a remarkable 55.5% of the variance. Plotting the loading vectors on a PC1 and PC2 biplot, as shown in **Figure 6**, reveals that all the variables negatively correlate with the Principal Components. However, it appears that all of the variables contribute to the Principal Components rather than just a few, as indicated by the fact that the maximum loading vectors values for "Glucose", "Age", and "Pregnancies" are approximately 0.15.

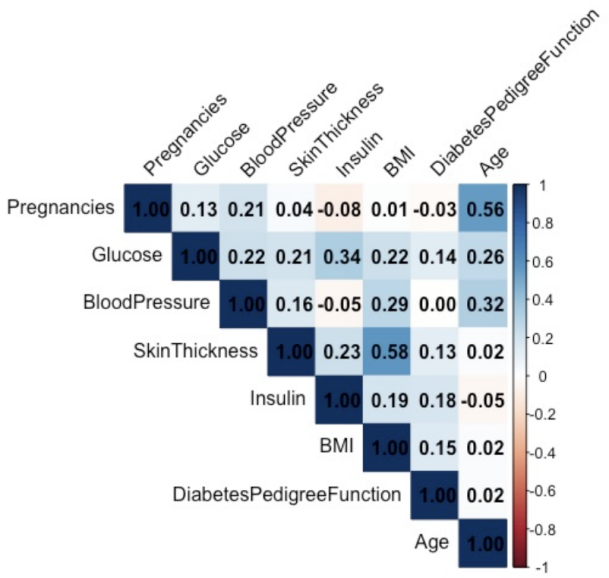


Figure 4: Matrix of correlations between continuous variables.

3.2 Clustering

We were interested in finding out additional details regarding the observation, specifically whether our dataset contains subgroups and whether these clusters match our dataset's Outcome. We decided to perform Hierarchical Clustering using the Ward.D2 linkage method, which tends to produce compact clusters and minimizes the squared distances between the points inside a cluster. It is evident from the dendrogram in **Figure 6** that there are three clusters.

Second, we tested the k-means clustering. To determine the number of clusters, we used the Within cluster dissimilarity/distance technique. Based on the generated graph, we were able to spot a tiny elbow at three clusters.

Based on these findings, we evaluated our clusterings using silhouette widths; the results are displayed in **Table 4**, and the silhouette plots are shown in **Figure 7**. It seems obvious that some observations are misclassified, because there are some values below 0, and the values in general are not very high. However, one group in both methods have an average silhouette width of 0.34 and 0.35, which is much greater than the other two groups and represents the majority of the patients who were not affected by diabetes.

Since our goal was to determine whether our clustering aligned with the Outcome

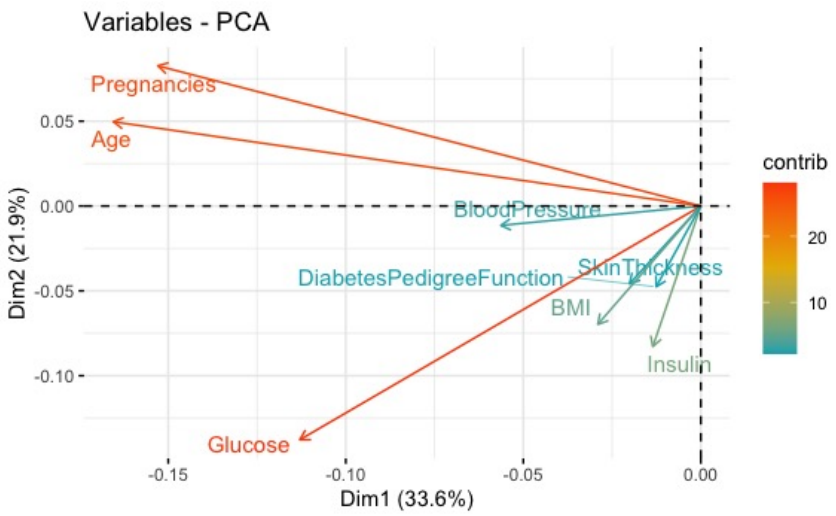


Figure 5: Biplot of the first two Principal Components. The loading vectors are indicated by arrows that are colored according to contribution.

variable, we computed the Adjusted Rand Index (ARI), which is shown in **Table 4**. The results for both approaches were approximately 0.2, which is a relatively low value.

Table 4: Clustering statistics

Clustering method	Average Silhouette Score	Cluster Number	Cluster Size	Average Silhouette Width	Adjusted Rand Index
Hierarchical	0.2247	1	246	0.10	0.175
		2	374	0.35	
		3	104	0.08	
k-means	0.2403	1	140	0.08	0.207
		2	214	0.17	
		3	370	0.34	

3.3 Feature Selection

Before developing proper classification models we decided to try and extrapolate the best subset of features in order to obtain the best possible classification with the least amount of information. To do so we employed a Best Subset Selection approach.

The Best Subset selection is a feature selection approach which aims at finding the best subsect of predictors by examining all the possible combination of them. This means fitting all the p models with exactly one predictor, all the $\binom{p}{2}$ models with two predictors and so forth. We then look at the resulting models and identify the best one.

The main problem with the Best Subset Selection approach is the computational bur-

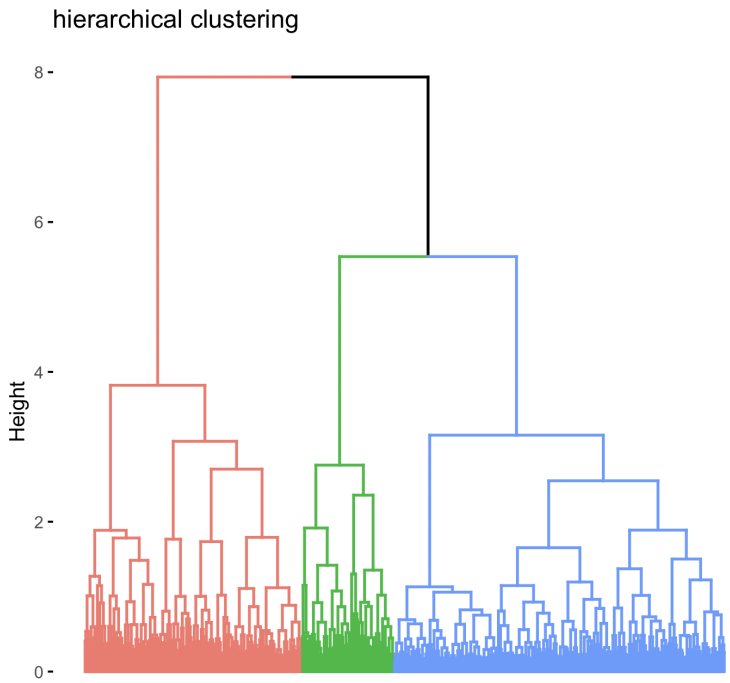


Figure 6: Dendrogram of the Hierarchical Clustering using the Ward.D2 linkage.

den which may derive from testing 2^p different models; however, in our specific case, this issue does not arise given the low number of models to test ($2^8 = 256$ different models). Moreover, it is important to remember that such approach works only for least squared linear regressions.

The results of our feature selection are visible in **Figure 8**. The optimal number of predictors appears to be four, these being "Pregnancies", "Glucose", "BMI" and "DiabetesPedigreeFunction". We will thus proceed in examining the performance of our classification models both by using the whole set of variables and by using only the subset of four best variables.

4 Classification Results

We obtained a training and a test dataset via a stratified 0.8/0.2 split of the processed dataset. The classifiers we evaluated include: decision tree, random forest, linear discriminant analysis, naïve Bayes, logistic regression, support vector machine and K-nearest neighbors. We selected the best hyperparameters for each model via grid search and a

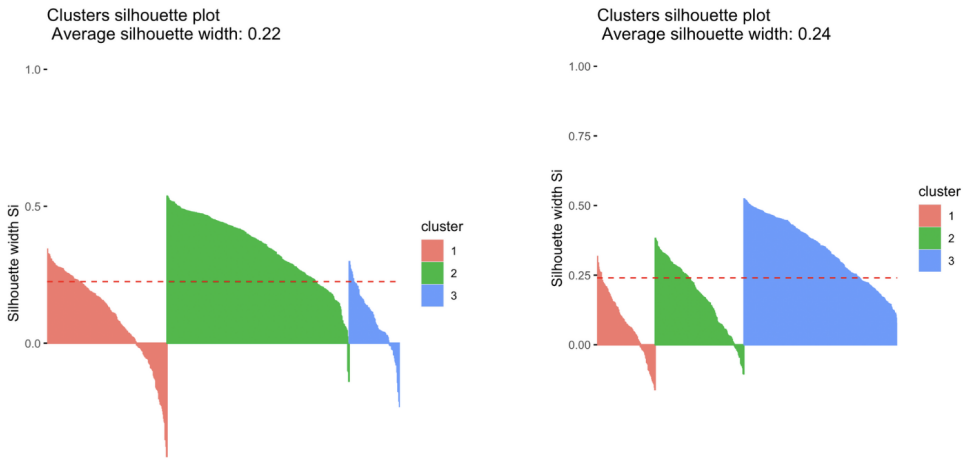


Figure 7: Silhouette Plots of Hierarchical Clustering (left) and K-means Clustering (right).

stratified 5-fold cross-validation on the training dataset. To address the class imbalance, we applied the Synthetic Minority Over-sampling Technique (SMOTE)[8] exclusively on the training folds within each cross-validation iteration, thus avoiding data leakage from the validation sets. After selecting the best hyperparameters, each model was retrained on the whole training dataset. We calculated the point estimates of the accuracy, the recall (with class 1—diabetes-positive—as the positive class), the precision and the F1-score of each model on the test dataset. We trained and tested our models first by training them on the full dataset and then training them using the best predictive features, obtaining very similar results as shown in **Table 5** and in **Table 6**.

Table 5: Performance metrics for each classifier using all features.

Model	**Accuracy**	**Recall**	**Precision**	**F1-score**
Decision Tree	0.72	0.67	0.58	0.62
Random Forest	0.75	0.65	0.63	0.64
LDA	0.75	0.62	0.71	0.66
Naïve Bayes	0.76	0.62	0.76	0.68
Logistic Regression	0.76	0.63	0.74	0.68
SVM	0.76	0.57	0.53	0.60
KNN	0.73	0.58	0.77	0.66

To quantify the uncertainty of our estimates, we applied the bias-corrected and accelerated (BCa)[9] bootstrap method with R = 1000 to compute the 95% confidence interval of the calculated metrics for the models that used only the selected features. The results are

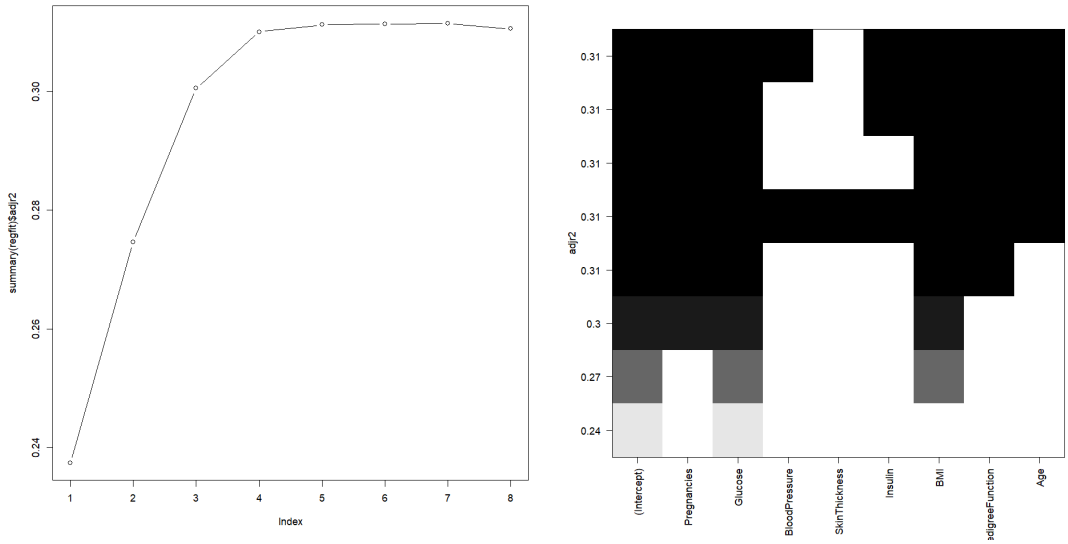


Figure 8: Feature selection plots
On the left plot for the value of the adj. R^2 for an inceasing number of variables. On the right the stability selection plot for the best features to chose for each model.

Table 6: Performance metrics for each classifier using selected features.

Model	**Accuracy**	**Recall**	**Precision**	**F1-score**
Decision Tree	0.72	0.77	0.57	0.65
Random Forest	0.75	0.69	0.62	0.66
LDA	0.77	0.72	0.65	0.68
Naïve Bayes	0.76	0.77	0.62	0.69
Logistic Regression	0.77	0.71	0.65	0.67
SVM	0.77	0.57	0.71	0.63
KNN	0.72	0.77	0.57	0.65

displayed in **Table 7**. For our analysis, we utilized the `caret` package [10] in R statistical software[11].

5 Conclusion

The models trained using only the features found with our techniques for feature selection exhibited very similar performances to those of the ones trained using all available features. “Glucose”, “DiabetesPedigreeFunction”, “Pregnancies” and “BMI” are thus sufficient to build a robust classifier. From our experiments, the highest accuracy (0.77 as shown in

Table 7: Performance metrics for each classifier with 95% confidence intervals.

Model	**Accuracy**	**Recall**	**Precision**	**F1-score**
Decision Tree	0.72 (0.69, 0.83)	0.77 (0.53, 0.79)	0.57 (0.51, 0.78)	0.65 (0.53, 0.76)
Random Forest	0.75 (0.67, 0.81)	0.69 (0.67, 0.73)	0.62 (0.5, 0.77)	0.66 (0.52, 0.69)
LDA	0.77 (0.67, 0.82)	0.72 (0.58, 0.83)	0.65 (0.48, 0.75)	0.68 (0.55, 0.76)
Naïve Bayes	0.76 (0.61, 0.79)	0.77 (0.52, 0.81)	0.62 (0.41, 0.67)	0.69 (0.48, 0.72)
Logistic Regression	0.77 (0.67, 0.82)	0.71 (0.56, 0.81)	0.65 (0.48, 0.75)	0.67 (0.53, 0.75)
SVM	0.77 (0.67, 0.81)	0.57 (0.34, 0.63)	0.71 (0.47, 0.8)	0.63 (0.42, 0.68)
KNN	0.72 (0.59, 0.79)	0.77 (0.52, 0.86)	0.57 (0.44, 0.68)	0.65 (0.54, 0.76)

Table 6) is achieved by logistic regression, support vector machine and linear discriminant analysis. The highest recall (with class 1—diabetes-positive—as the positive class) is equal to 0.77 and it is achieved by decision tree, naïve bayes and k nearest neighbor classifier. Support vector machines achieves the worst recall (0.57) and the highest precision (0.71). We observe a convergence of results across diverse modeling approaches, this further strengthens the credibility of our conclusions and suggests that the identified patterns are genuine rather than artifacts of any particular modeling technique. The similarity in performance metrics, as shown in **Table 7**, particularly in accuracy (ranging from 0.72 to 0.77) and F1-score (ranging from 0.63 to 0.69), indicates that the signal in our data is being captured consistently regardless of the classification approach employed.

References

1. For Disease Control, C. & Prevention. *About Common Eye Disorders and Diseases* 2024. `https://www.cdc.gov/vision-health/about-eye-disorders/index.html`.

2. Armstrong, D. G., Tan, T.-W., Boulton, A. J. & Bus, S. A. Diabetic foot ulcers: a review. *Jama* **330,** 62–75 (2023).

3. *Pima Indians Diabetes Database* `https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database`.

4. Organization, W. H. *WHO Diabetes* `https://www.who.int/news-room/fact-sheets/detail/diabetes`.

5. For Disease Control, C. & Prevention. *CDC Diabetic Nephropathy* https://www.cdc.gov/diabetes/diabetes-complications/diabetes-and-chronic-kidney-disease.html.

6. Bogardus, C. & Lillioja, S. Pima Indians as a model to study the genetics of NIDDM. *Journal of cellular biochemistry* **48,** 337–343 (1992).

7. Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C. & Johannes, R. S. *Using the ADAP learning algorithm to forecast the onset of diabetes mellitus* in *Proceedings of the annual symposium on computer application in medical care* (1988), 261.

8. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16,** 321–357 (2002).

9. Efron, B. Better bootstrap confidence intervals. *Journal of the American statistical Association* **82,** 171–185 (1987).

10. Kuhn & Max. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28,** 1–26. https://www.jstatsoft.org/index.php/jss/article/view/v028i05 (2008).

11. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2024). https://www.R-project.org/.

12. James, G., Witten, D., Hastie, T., Tibshirani, R., *et al. An introduction to statistical learning* **1** (Springer, 2013).

Appendix

Figure 9: Pairwise scatterplots coloured by Outcome

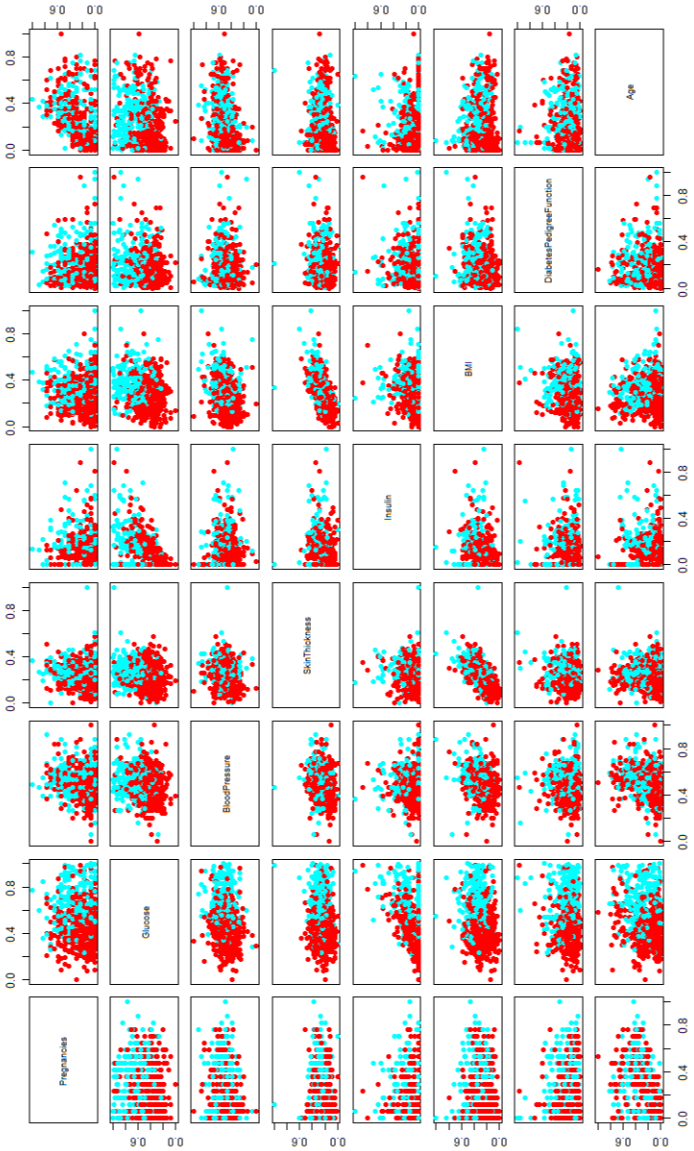


Figure 10: Knn performance in predicting SkinThickness values

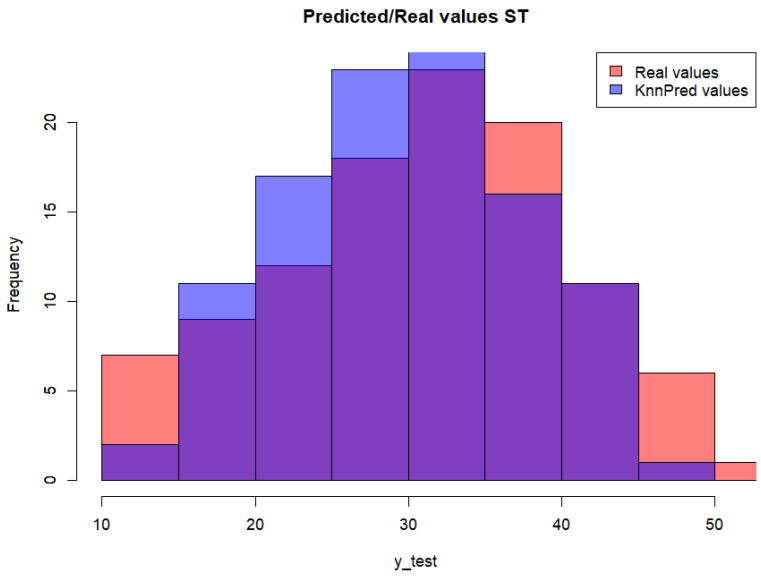


Figure 11: PCA biplot for the first two variables

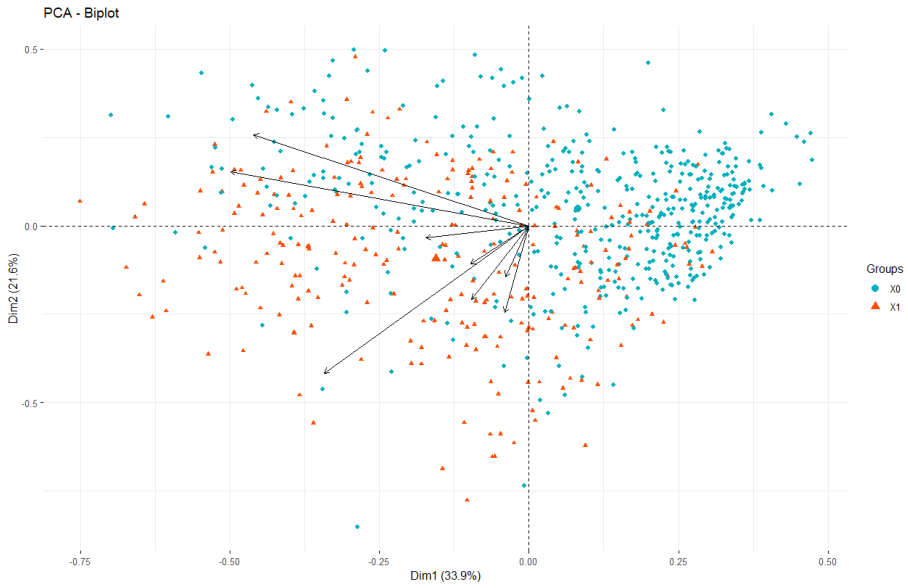


Figure 12: PCA scree plot

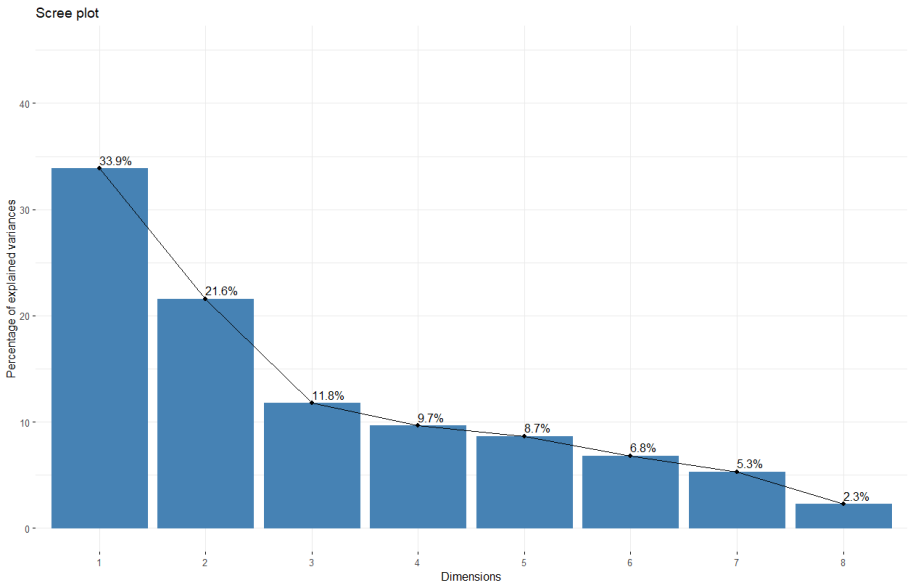


Figure 13: PCA elbow plot

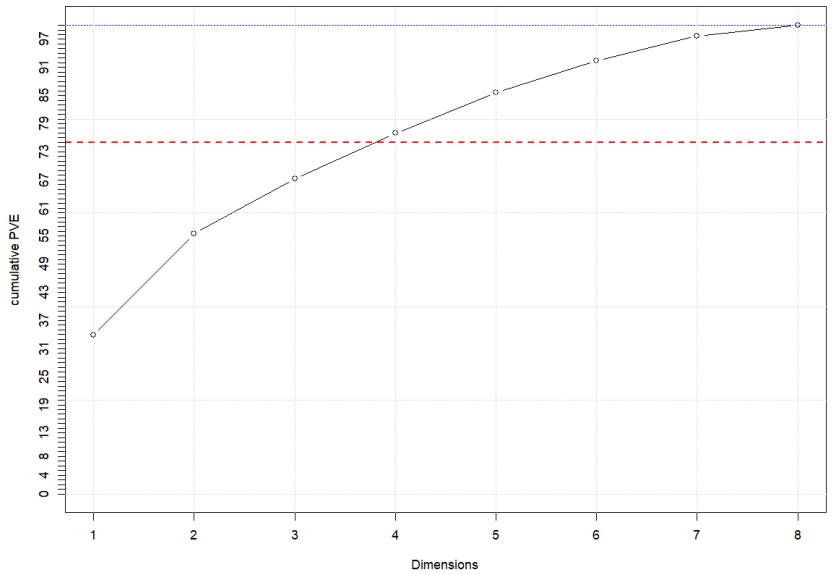


Figure 14: K-means Within cluster dissimilarity/distance

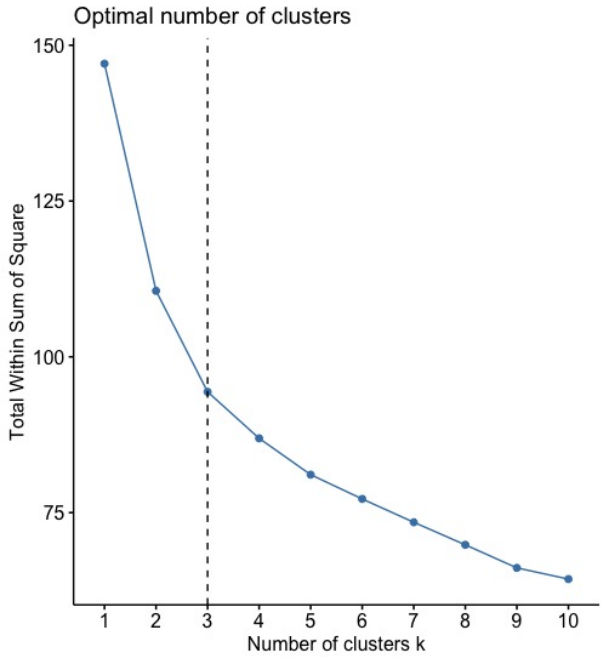


Figure 15: Feature selection plots for various metrics (R^2, adj.R^2, BIC, AIC)

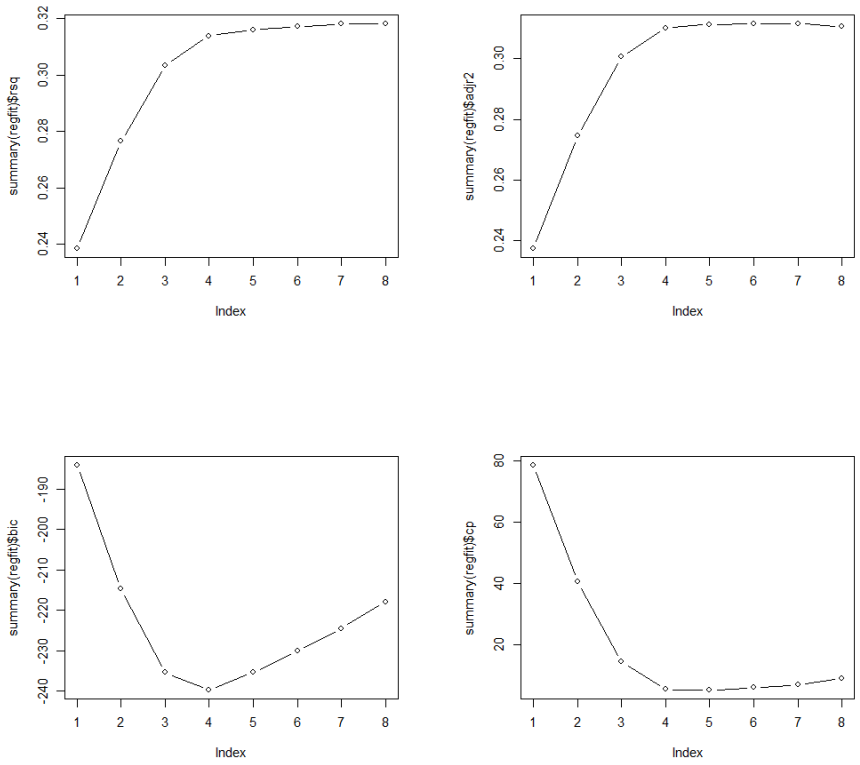


Figure 16: Feature selection stability selection plots for adj.R^2 and BIC

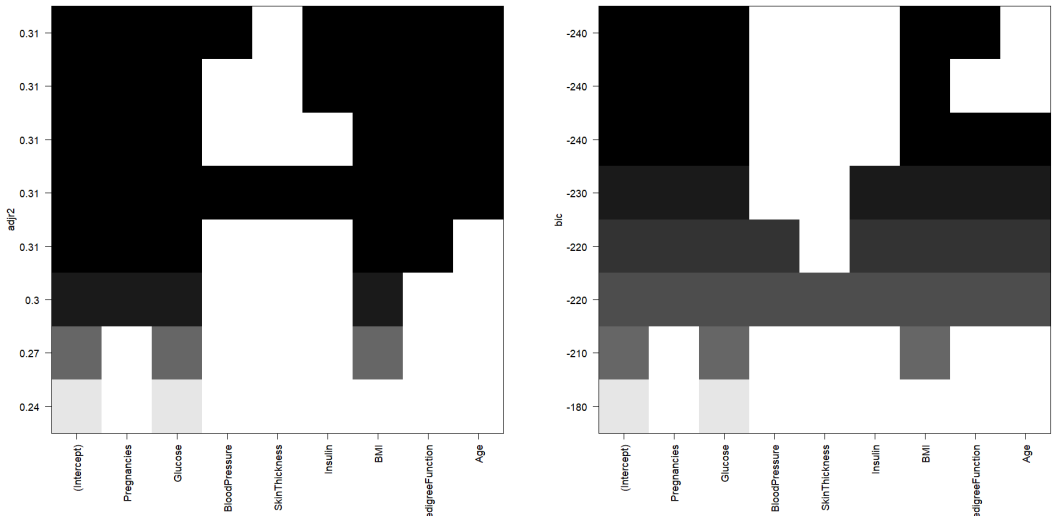
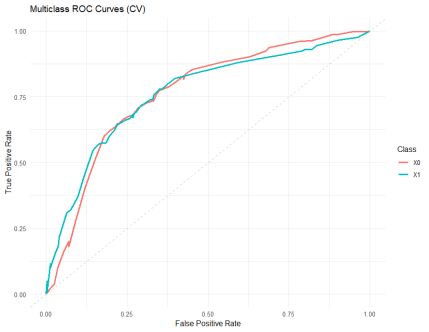
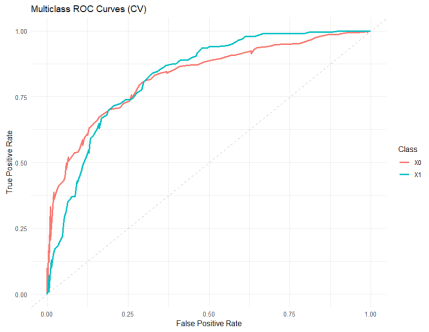


Figure 17: ROC Curves for each classification model

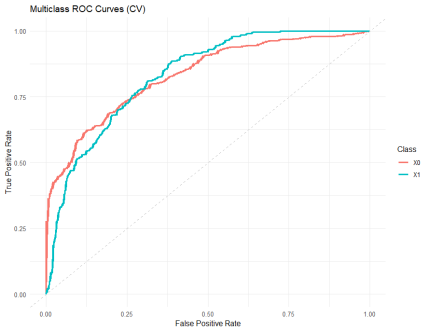
(a) Decision Tree



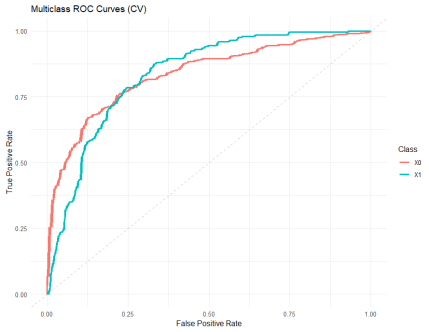
(b) Random Forest



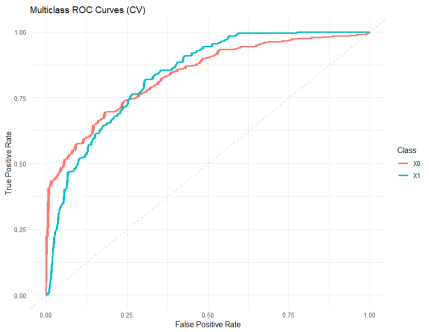
(c) LDA



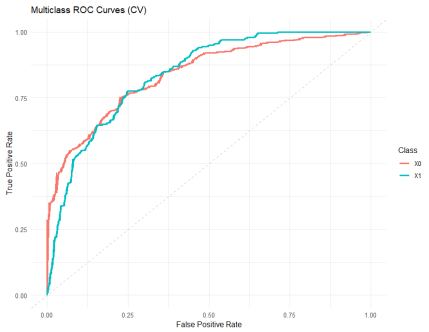
(d) Naïve Bayes



(e) Logistic regression



(f) SVM



(g) Knn