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

Statistical Learning and Large Data

Gabriele Cini, Flavio Di Lisio, Giorgia Di Santolo

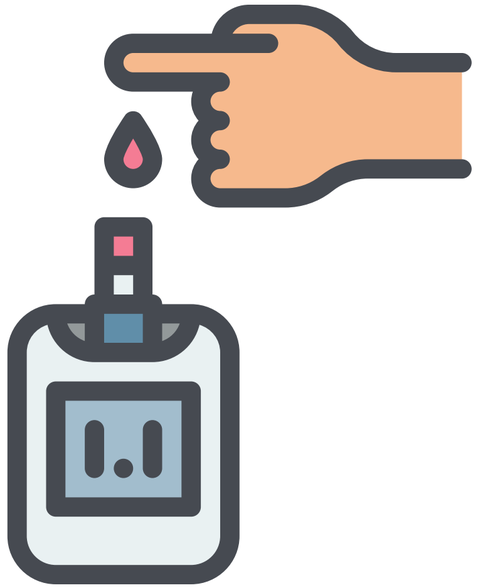
Sant'Anna
School of Advanced Studies – Pisa

OVERVIEW

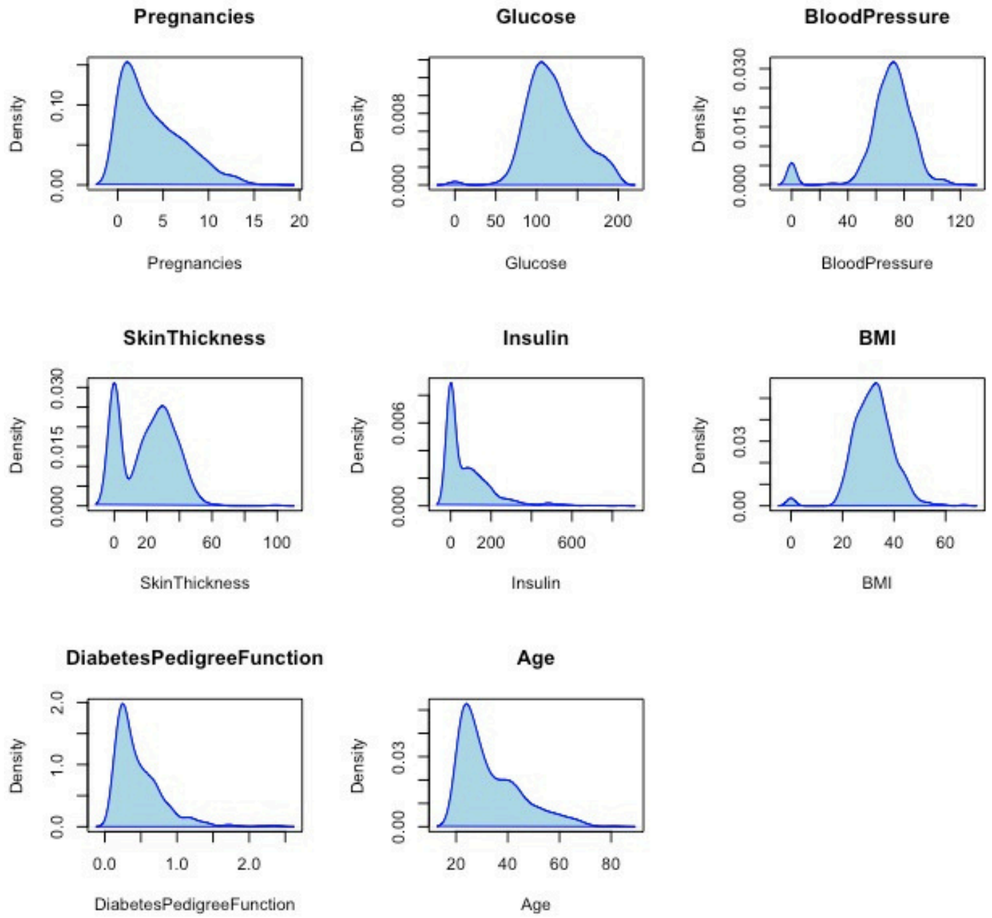
1. Introduction
2. EDA and data wrangling
3. PCA
4. Clustering
5. Feature selection
6. Classification
7. Closing remarks

INTRODUCTION

- Context and background: why diabetes?
- Our aim
- Medical considerations



DATA VISUALIZATION - I

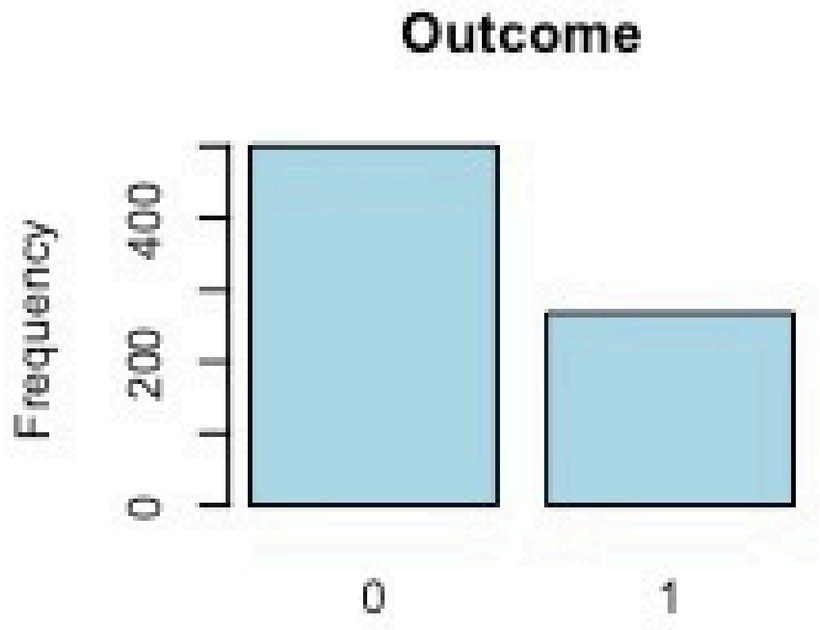


The Dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases, all patients (768) are females at least 21 years old of Pima Indian heritage.

In figure the distribution of the 8 continuous variables:

- Number of times pregnant
- Plasma Glucose Concentration at 2 Hours in an Oral Glucose Tolerance Test (GTT)
- Diastolic Blood Pressure (mmHg)
- Triceps Skin Fold Thickness (mm)
- 2-Hour Serum Insulin (μU/ml)
- Body Mass Index (Weight in kg / (Height in m)^2)
- Diabetes Pedigree Function: a synthesis of the diabetes mellitus history in relatives
- Age (years)

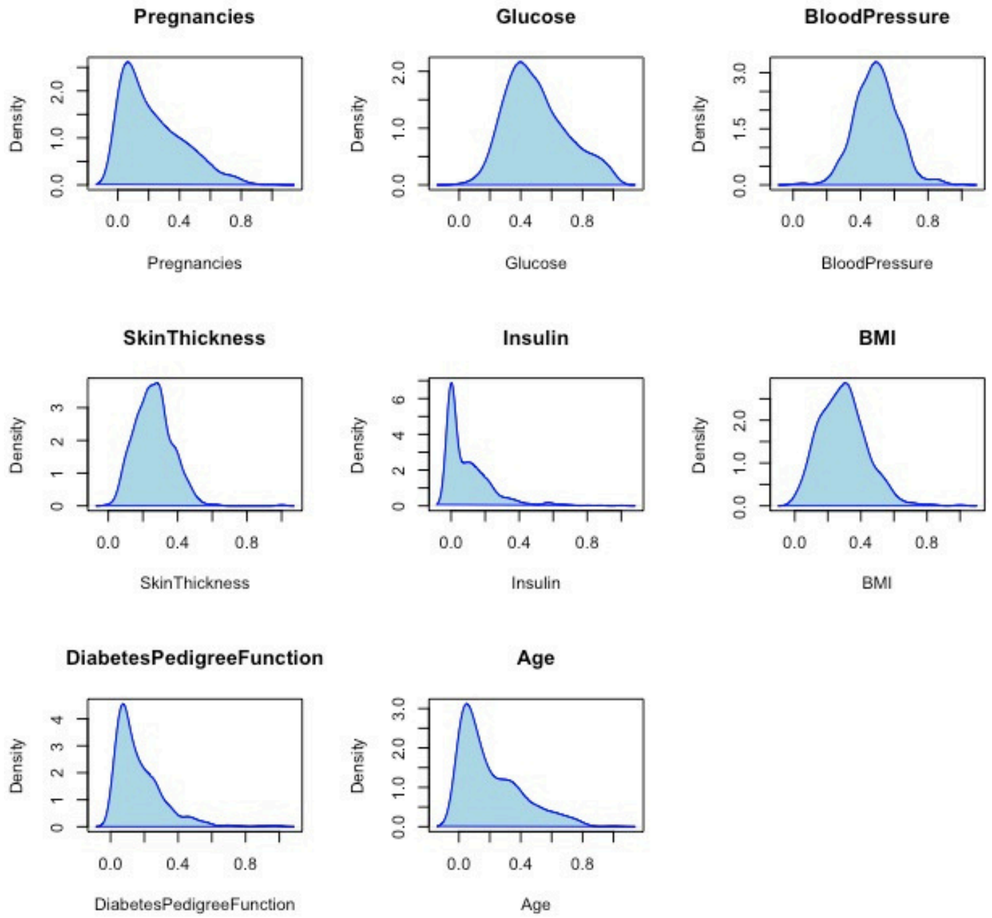
DATA VISUALIZATION - II



In figure the barplot of the categorical variable:

- Outcome → unbalanced
  - 0: unaffected
  - 1: affected

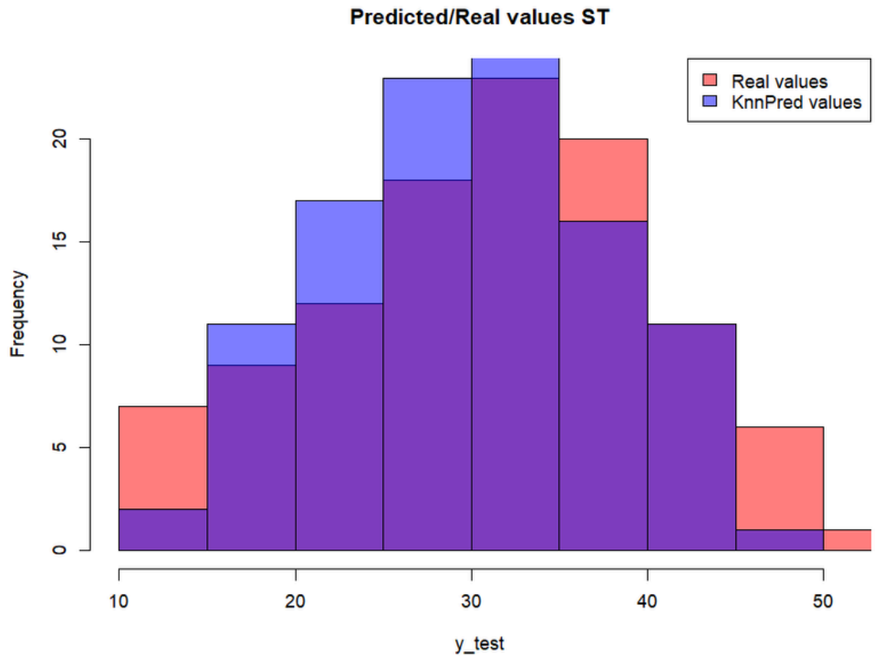
SCALING



- We decided to consider the “0” of Glucose, Blood Pressure and BMI variables as “NA”, and we removed those rows.

- We performed a Min-Max Scaling to the continuous variables.

SKIN THICKNESS PREDICTION



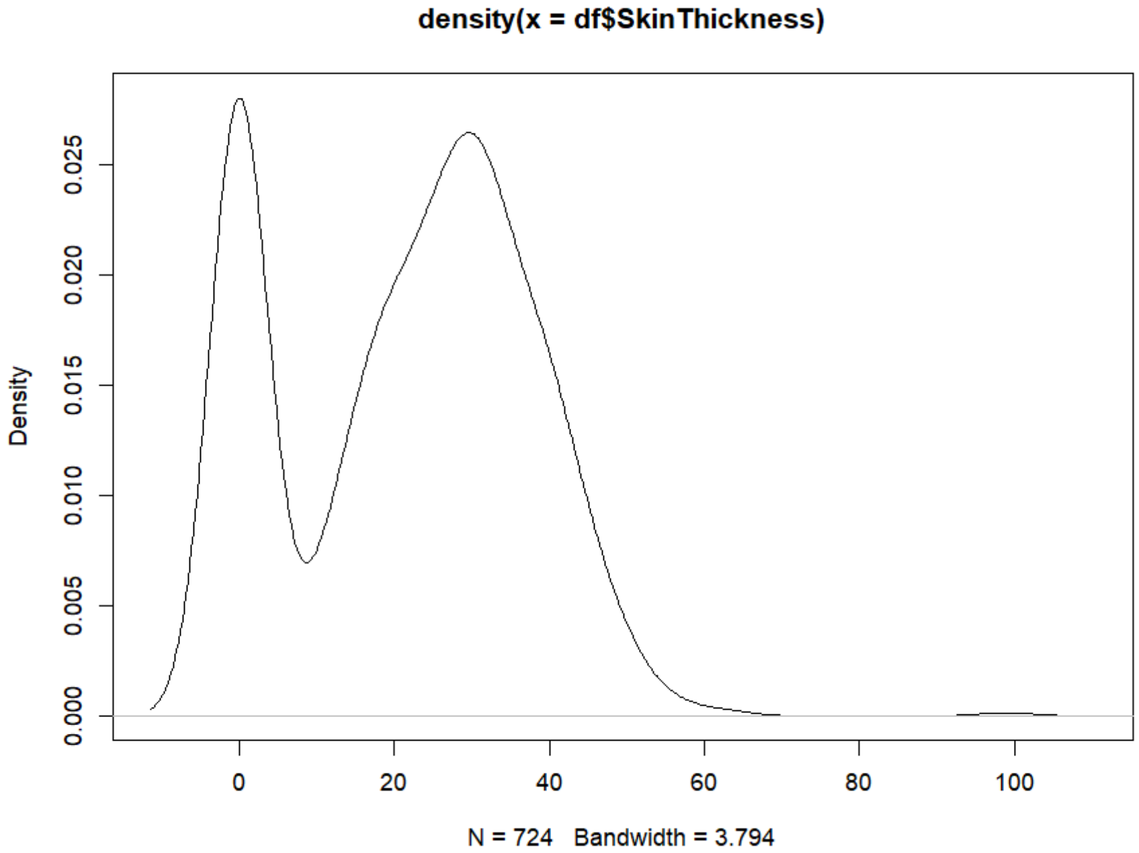
For the SkinThickness variable, we decided to estimate values rather than directly remove observations with NaN data.
This was decided due to the high number of not available observations (29.56%) and the high correlation between NaN SkinThickness and other NaN variables.

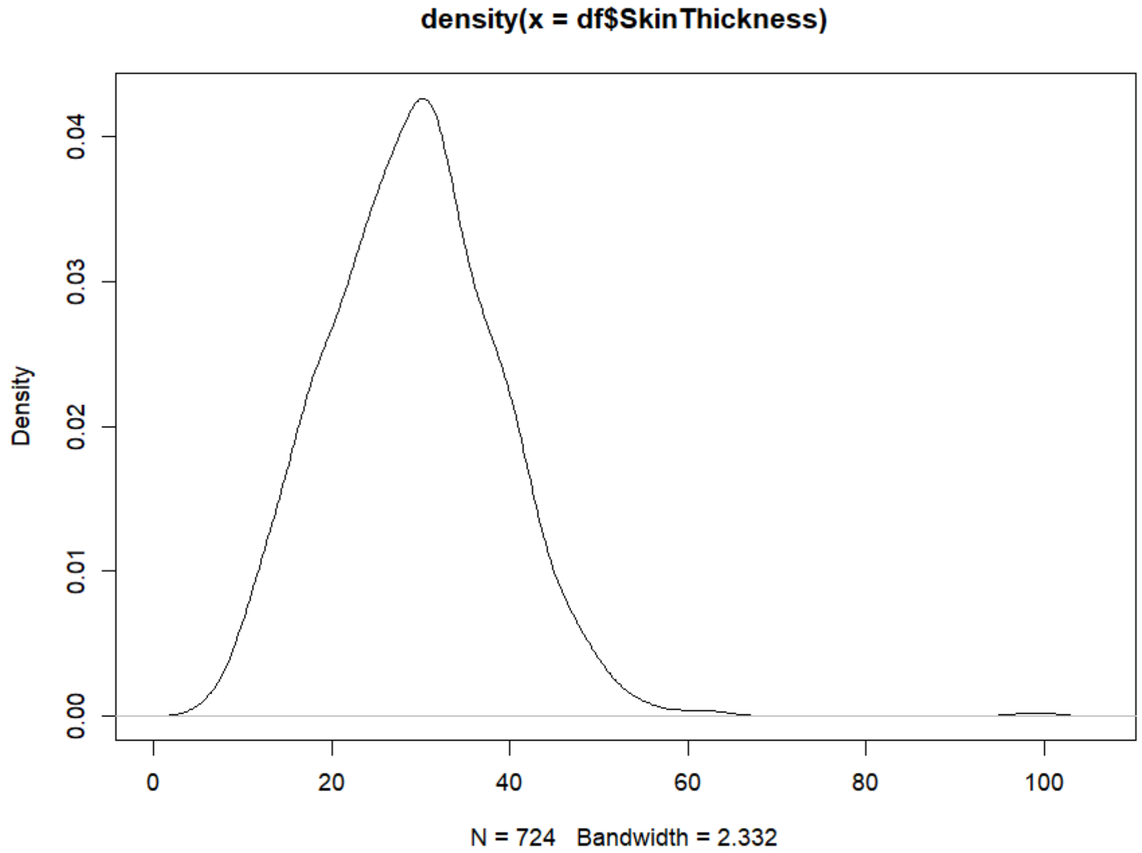
For the estimation we used a k-nearest neighbour model by assigning the mean value of the nearest k points to the interested observation.

We first had a test model run to observe the performance, then we applied the model to the entire dataset.

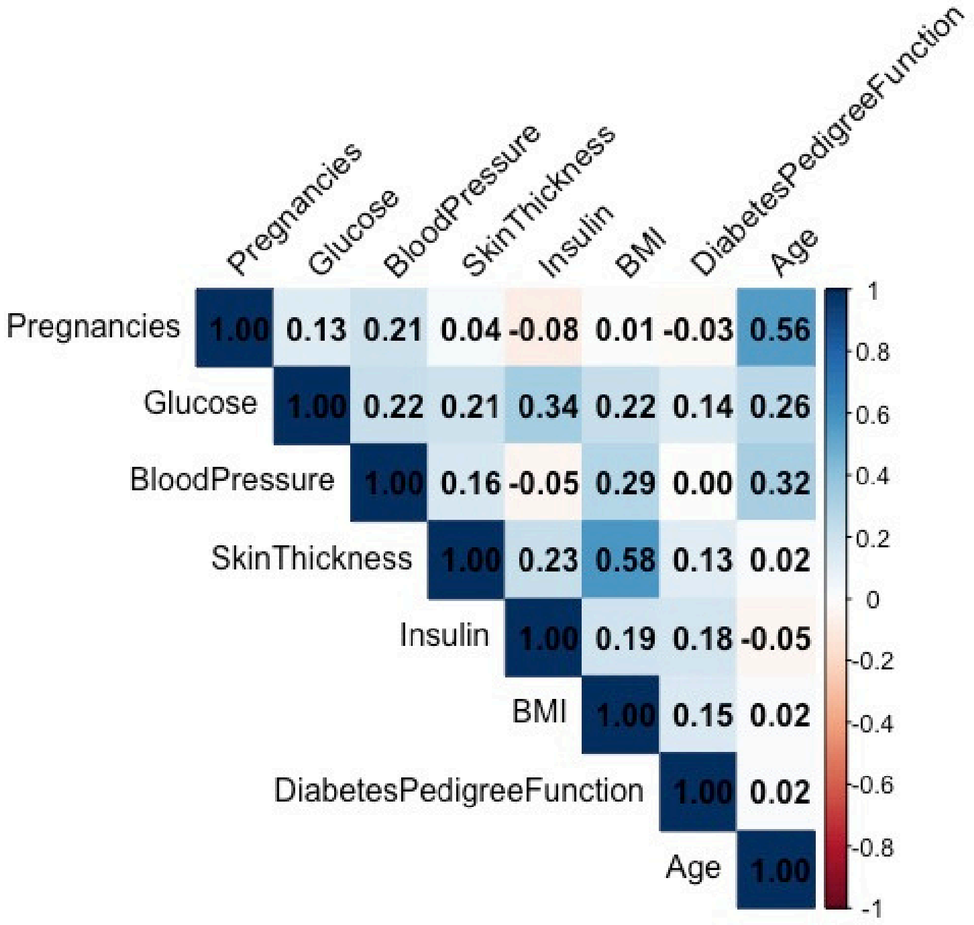
The model is quite good, as a Mean Absolute Percentage Error of about 10.57% is observed, a clearly better solution than using the average or median observed value (36% MAPE).

DATA VISUALIZATION FOR PREDICTED VALUES





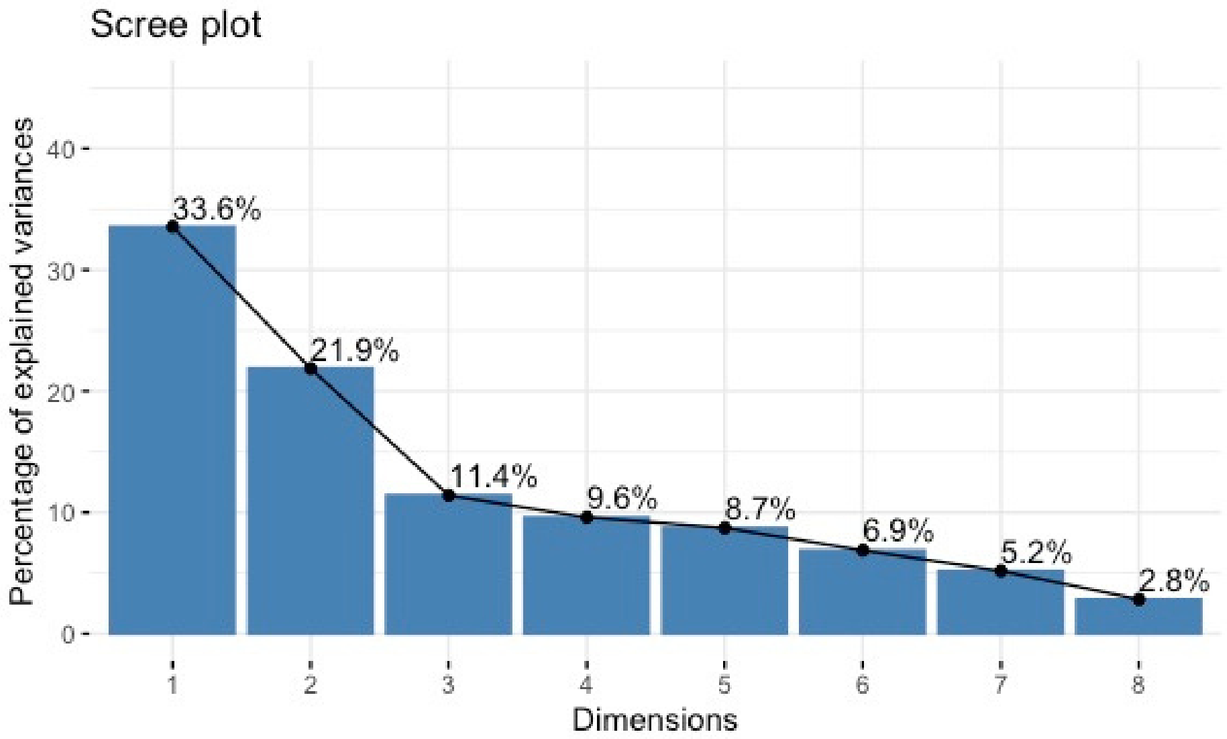
CORRELATION



We identified high correlation (> 0.30) between:

- BMI and SkinThickness
- Age and Pregnancies
- Insulin and Glucose
- Blood Pressure and Age

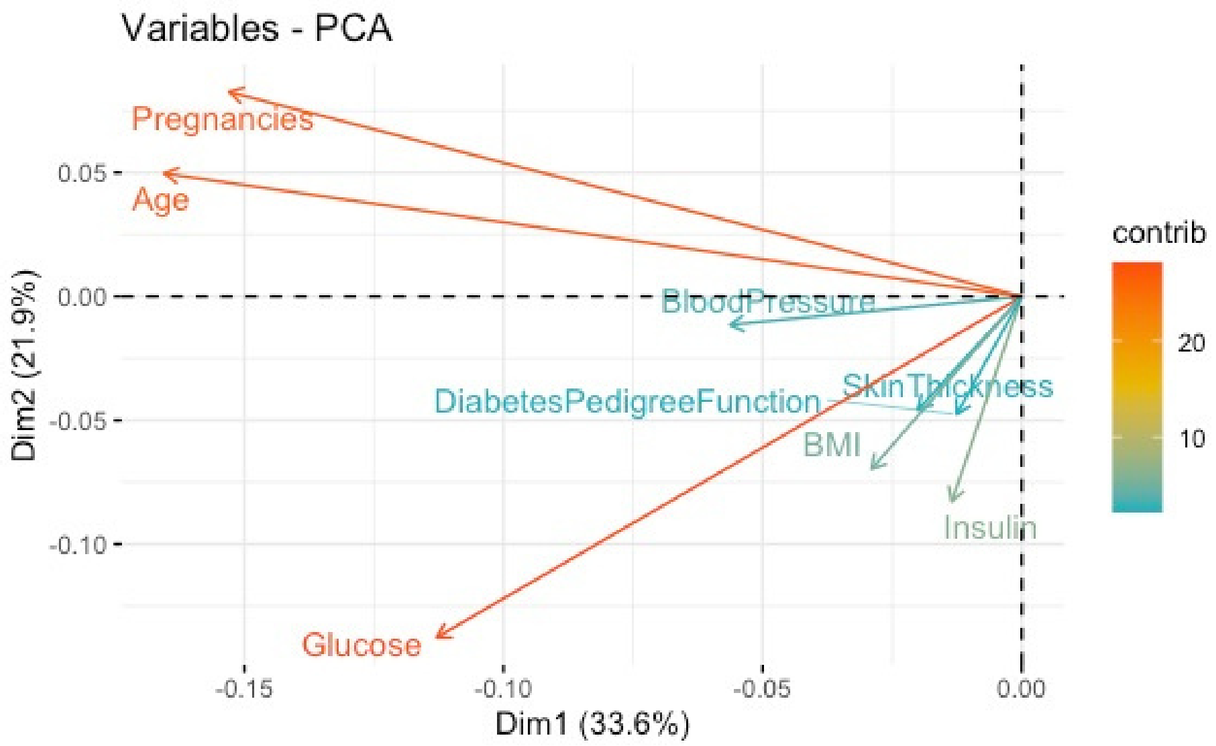
PRINCIPAL COMPONENT ANALYSIS (PCA) - I



We performed a PCA analysis, looking at the percentage of explained variances:

- Four elements are required in order to achieve 75% explainancy
- PC1 and PC2 explain 55.5% of the variance

PRINCIPAL COMPONENT ANALYSIS (PCA) - II



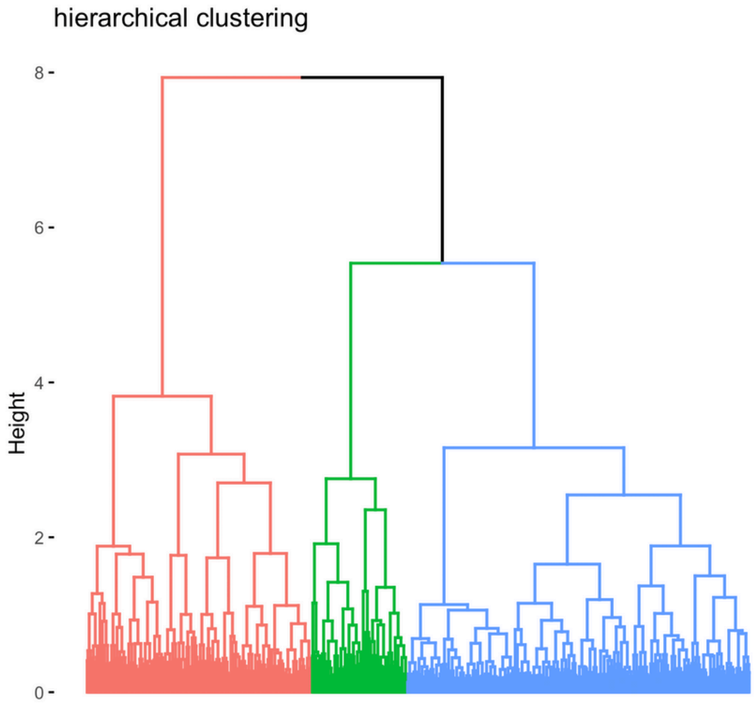
The loadings in the PCA - biplot:

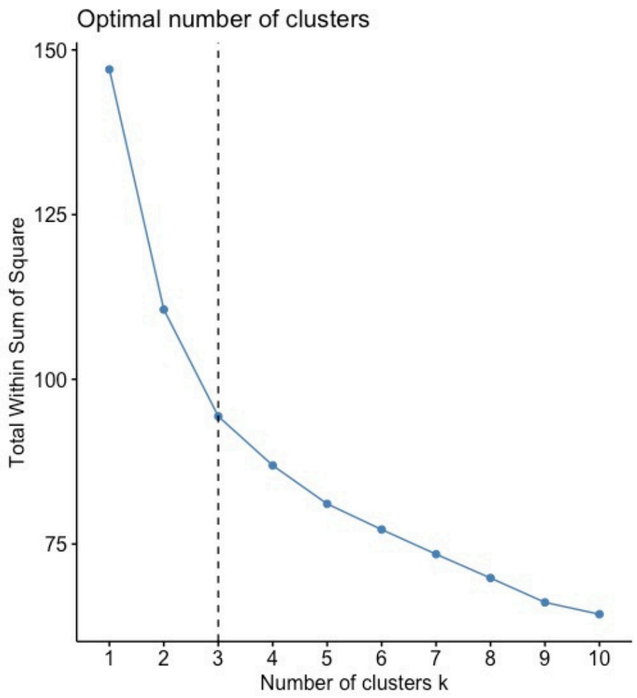
- PC1 shows a predominantly negative correlation with the variables Pregnancies and Age.
- Both PC1 and PC2 are negatively correlated with the variable Glucose.

CLUSTERING - I

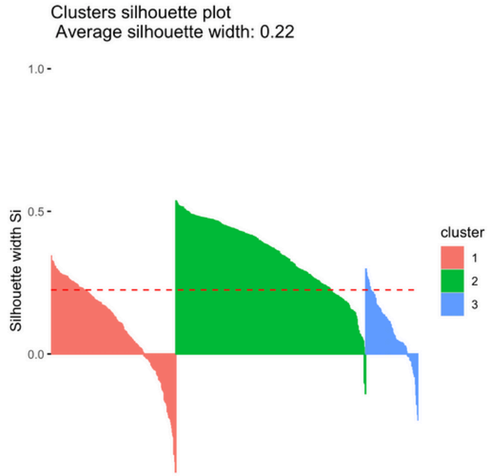
We indentified the optimal number of clusters:

- for the Hierarchical Clustering method: **3**
- for the K-Means clustering method using the Within cluster dissimilarity/distance strategy: **3**



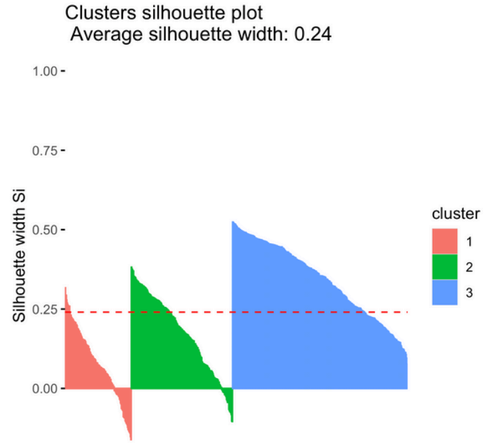


CLUSTERING - II




```
[1] "Average Silhouette Score: 0.2247"
  cluster size ave.sil.width
1       1  246          0.10
2       2  374          0.35
3       3  104          0.08
```

- **Hierarchical clustering**
  - Adjusted Rand Index (ARI) = 0.175

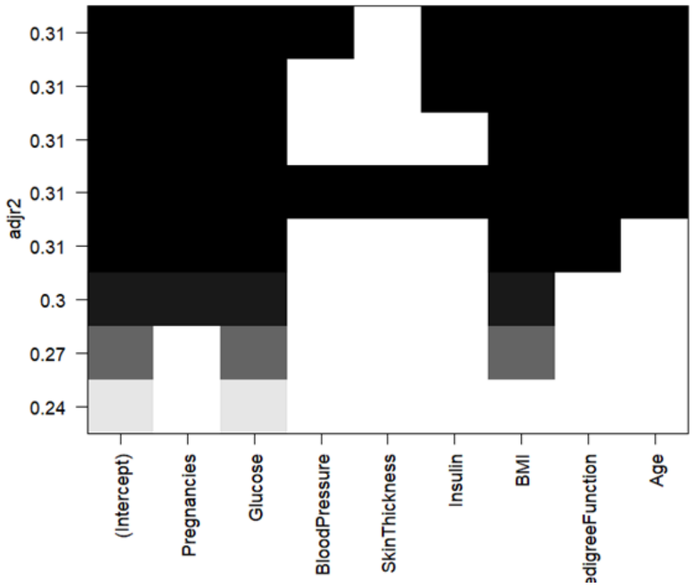


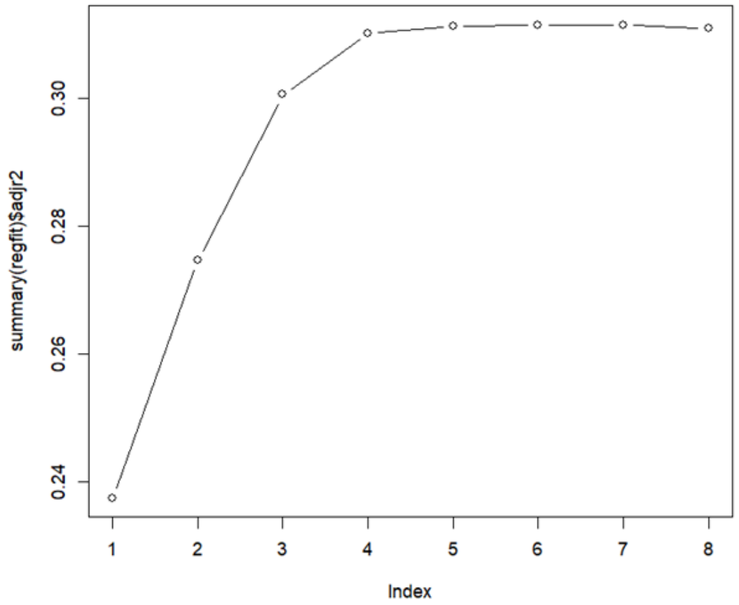

```
Average Silhouette Score: 0.2403
  cluster size ave.sil.width
1       1  140          0.08
2       2  214          0.17
3       3  370          0.34
```

- **K-Means clustering**
  - Adjusted Rand Index (ARI) = 0.207

FEATURE SELECTION

In doing feature selection we adopt the Best Subset approach, giving the low number of variables and thus models to test.
Our results suggests the best model adopts 4 variables, these being “Glucose”, “BMI”, “Pregnancies”, “DiabetesPedigreeFunction”





CLASSIFICATION PIPELINE

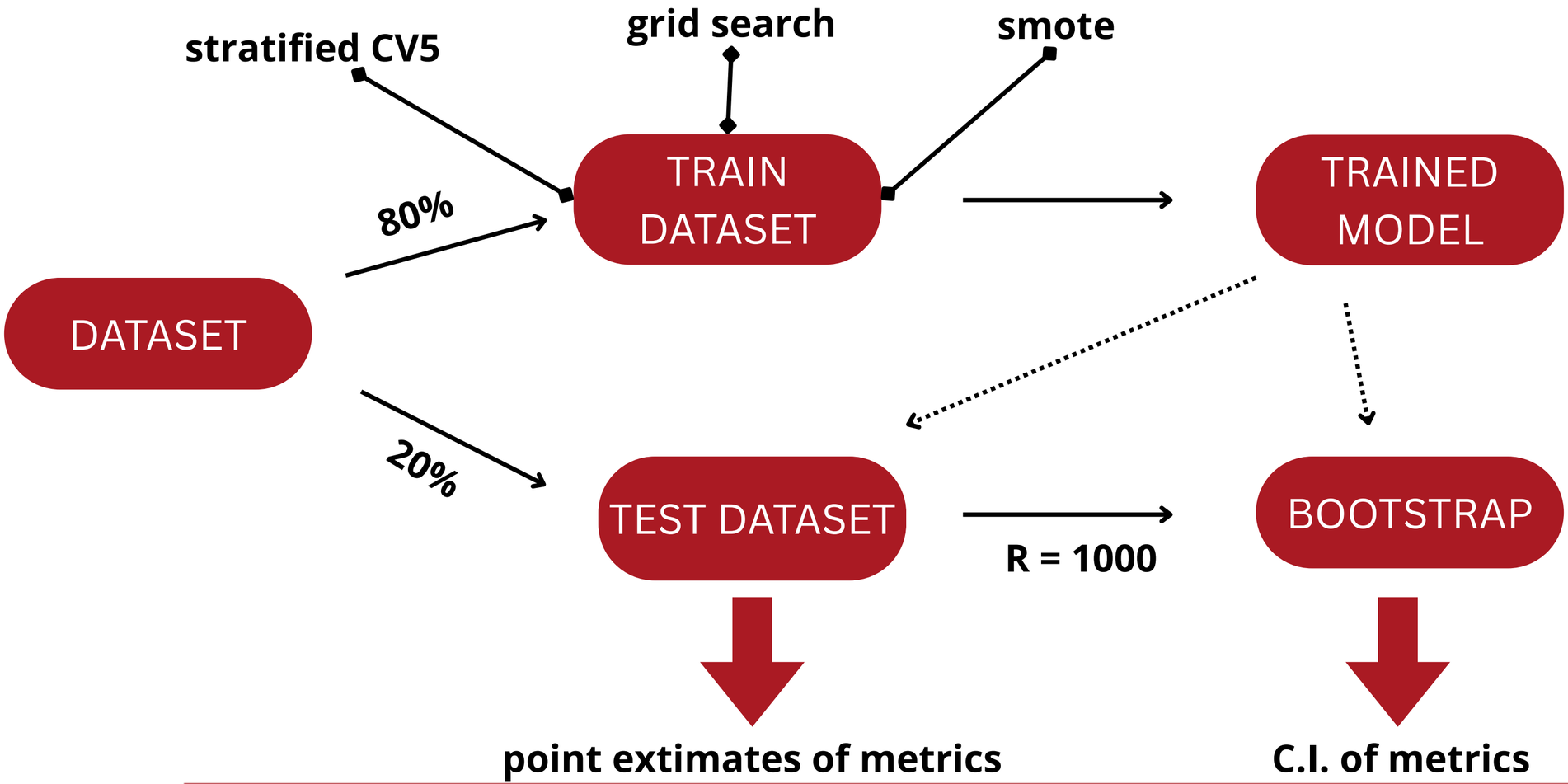
Algorithms tested:

- K nearest neighbors
- Logistic regression
- LDA
- Naive Bayes Classifier
- Decision Tree
- Random Forest
- SVM

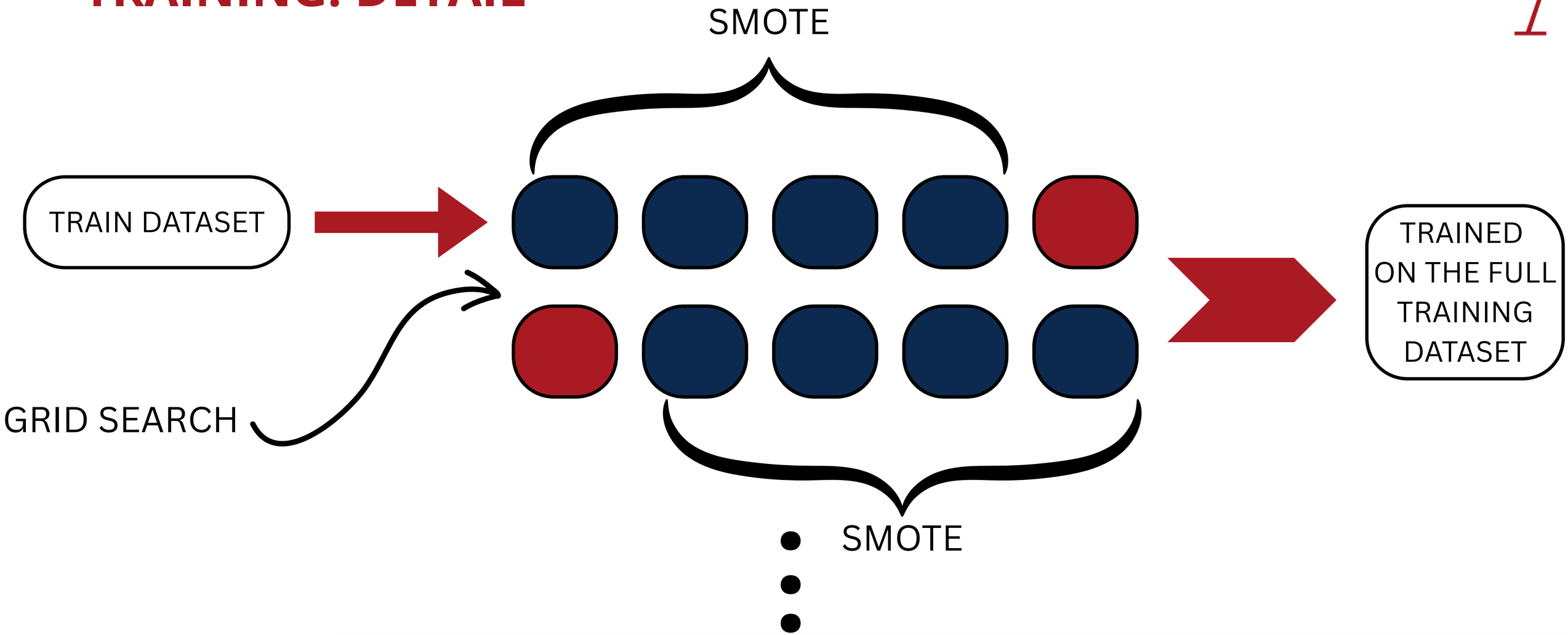
- Grid search stratified CV5 with SMOTE
- Training on entire dataset

CI of metrics estimated with bootstrap with B=1000

CLASSIFICATION PIPELINE: DETAIL



TRAINING: DETAIL



SUMMARY

ALL FEATURES

model	accuracy	recall	precision	f1
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
naive bayes	0.76	0.62	0.76	0.68
logistic regression	0.76	0.63	0.74	0.68
svm	0.76	0.7	0.53	0.6
knn	0.73	0.58	0.77	0.66

SELECTED FEATURES

model	accuracy	recall	precision	f1
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
naive 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

SIMILAR PERFORMANCES!

CONFIDENCE INTERVAL OF METRICS

model	accuracy	recall	precision	f1
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 (0.47, 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)
naive 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 (0.43, 0.68)
knn	0.72 (0.63, 0.79)	0.77 (0.62, 0.86)	0.57 (0.44, 0.69)	0.65 (0.54, 0.75)

SUMMARY

- The dataset is unbalanced
- We performed Exploratory Data analysis
- We filled the missing values using a k-nearest neighbour model
- We scaled the dataset
- We performed a Principal Component Analysis and a Clustering Analysis
- We selected the most relevant Features
- We tested several classification techniques
- What's next …

ESSENTIAL BIBLIOGRAPHY

- The dataset: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data
- Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.
- Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2023). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications, 35(22), 16157-16173.
- Nitesh, V. C. (2002). SMOTE: synthetic minority over-sampling technique. J Artif Intell Res, 16(1), 321.
- McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Thank you for your attention

Gabriele Cini, Flavio Di Lisio, Giorgia Di Santolo

Sant'Anna
School of Advanced Studies – Pisa

DESCRIPTIVE STATISTICS OF THE UNSCALED DATASET - I

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction
Min. : 0.000	Min. : 0.0	Min. : 0.00	Min. : 0.00	Min. : 0.0	Min. : 0.00	Min. :0.0780
1st Qu.: 1.000	1st Qu.: 99.0	1st Qu.: 62.00	1st Qu.: 0.00	1st Qu.: 0.0	1st Qu.:27.30	1st Qu.:0.2437
Median : 3.000	Median :117.0	Median : 72.00	Median :23.00	Median : 30.5	Median :32.00	Median :0.3725
Mean : 3.845	Mean :120.9	Mean : 69.11	Mean :20.54	Mean : 79.8	Mean :31.99	Mean :0.4719
3rd Qu.: 6.000	3rd Qu.:140.2	3rd Qu.: 80.00	3rd Qu.:32.00	3rd Qu.:127.2	3rd Qu.:36.60	3rd Qu.:0.6262
Max. :17.000	Max. :199.0	Max. :122.00	Max. :99.00	Max. :846.0	Max. :67.10	Max. :2.4200

Age	Outcome
Min. :21.00	Min. :0.000
1st Qu.:24.00	1st Qu.:0.000
Median :29.00	Median :0.000
Mean :33.24	Mean :0.349
3rd Qu.:41.00	3rd Qu.:1.000
Max. :81.00	Max. :1.000

DESCRIPTIVE STATISTICS OF THE UNSCALED DATASET - II

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Pregnancies	1	768	3.85	3.37	3.00	3.46	2.97	0.00	17.00	17.00	0.90	0.14	0.12
Glucose	2	768	120.89	31.97	117.00	119.38	29.65	0.00	199.00	199.00	0.17	0.62	1.15
BloodPressure	3	768	69.11	19.36	72.00	71.36	11.86	0.00	122.00	122.00	-1.84	5.12	0.70
SkinThickness	4	768	20.54	15.95	23.00	19.94	17.79	0.00	99.00	99.00	0.11	-0.53	0.58
Insulin	5	768	79.80	115.24	30.50	56.75	45.22	0.00	846.00	846.00	2.26	7.13	4.16
BMI	6	768	31.99	7.88	32.00	31.96	6.82	0.00	67.10	67.10	-0.43	3.24	0.28
DiabetesPedigreeFunction	7	768	0.47	0.33	0.37	0.42	0.25	0.08	2.42	2.34	1.91	5.53	0.01
Age	8	768	33.24	11.76	29.00	31.54	10.38	21.00	81.00	60.00	1.13	0.62	0.42
Outcome	9	768	0.35	0.48	0.00	0.31	0.00	0.00	1.00	1.00	0.63	-1.60	0.02

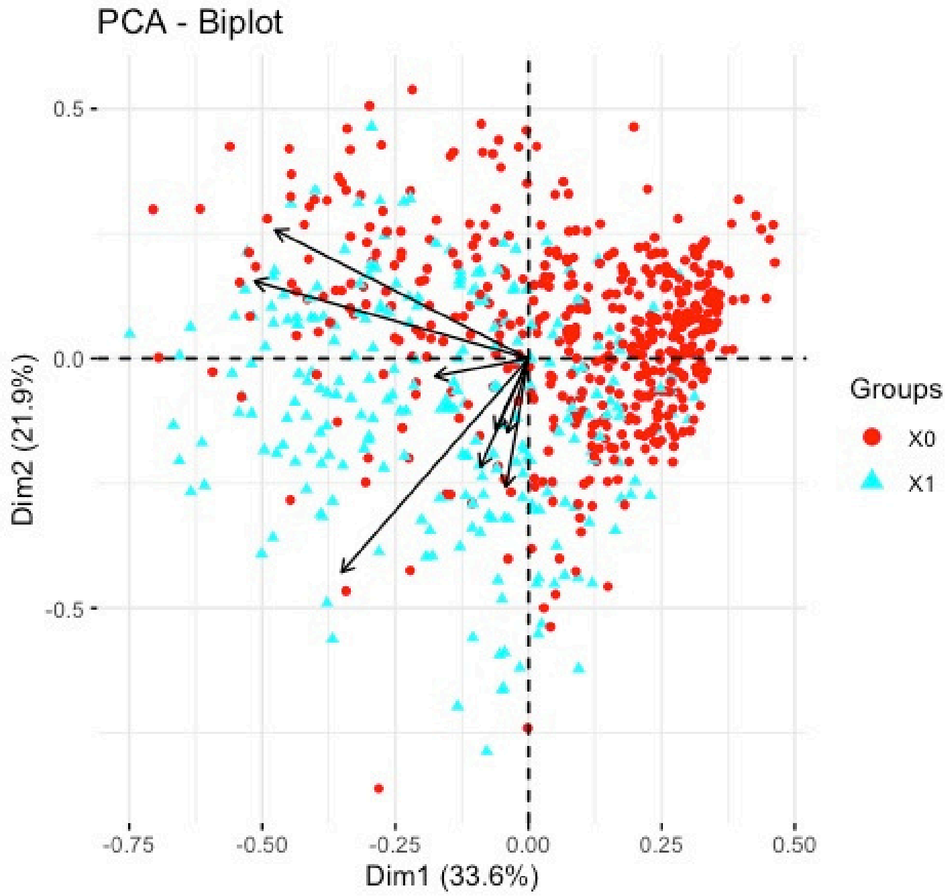
DESCRIPTIVE STATISTICS OF THE SCALED DATASET - I

```
  Pregnancies        Glucose       BloodPressure    SkinThickness      Insulin           BMI
Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
1st Qu.:0.05882   1st Qu.:0.3597   1st Qu.:0.4082   1st Qu.:0.1658   1st Qu.:0.00000   1st Qu.:0.1902
Median :0.17647   Median :0.4710   Median :0.4898   Median :0.2391   Median :0.05674   Median :0.2904
Mean   :0.22741   Mean   :0.5025   Mean   :0.4939   Mean   :0.2405   Mean   :0.09988   Mean   :0.2918
3rd Qu.:0.35294   3rd Qu.:0.6323   3rd Qu.:0.5714   3rd Qu.:0.3080   3rd Qu.:0.15426   3rd Qu.:0.3763
Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
 DiabetesPedigreeFunction      Age          Outcome
Min.   :0.00000           Min.   :0.0000   X0:475
1st Qu.:0.07131           1st Qu.:0.0500   X1:249
Median :0.12852           Median :0.1333
Mean   :0.16941           Mean   :0.2058
3rd Qu.:0.23463           3rd Qu.:0.3333
Max.   :1.00000           Max.   :1.0000
```

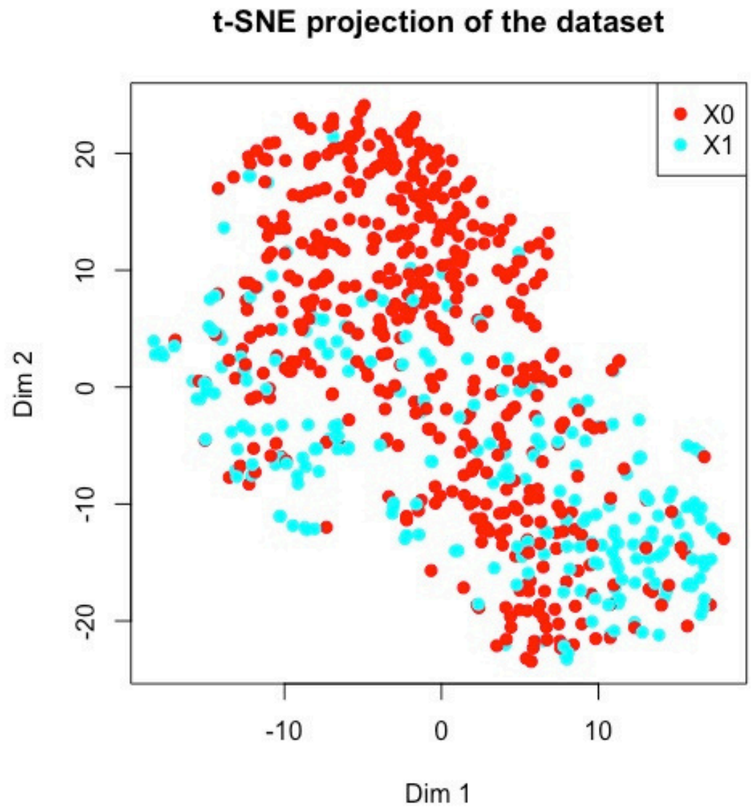
DESCRIPTIVE STATISTICS OF THE SCALED DATASET - II

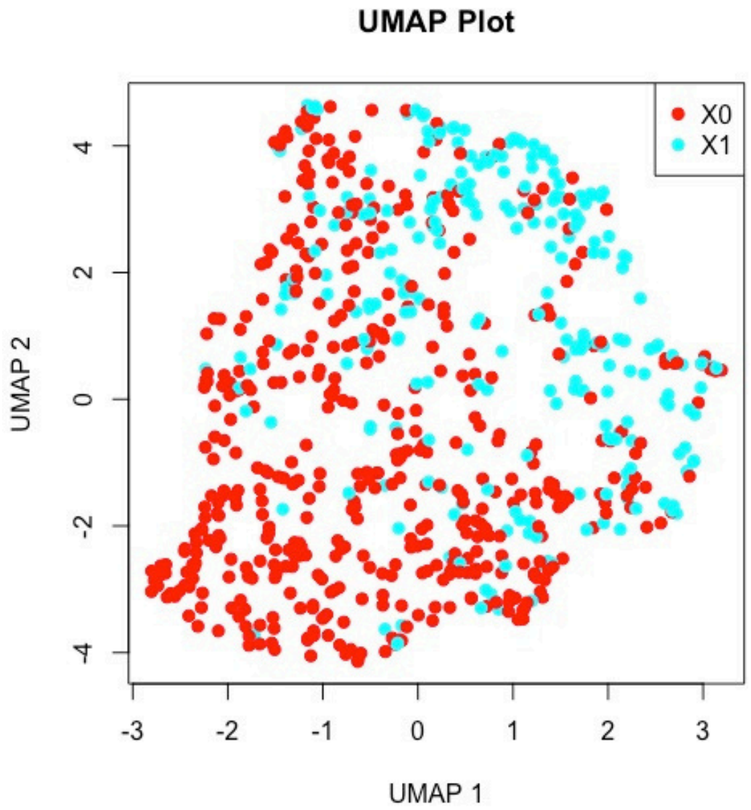
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Pregnancies	1	724	0.23	0.20	0.18	0.21	0.17	0	1	1	0.90	0.16	0.01
Glucose	2	724	0.50	0.20	0.47	0.49	0.19	0	1	1	0.52	-0.32	0.01
BloodPressure	3	724	0.49	0.13	0.49	0.49	0.12	0	1	1	0.14	0.90	0.00
SkinThickness	4	724	0.24	0.11	0.24	0.24	0.10	0	1	1	0.65	3.22	0.00
Insulin	5	724	0.10	0.14	0.06	0.07	0.08	0	1	1	2.19	6.74	0.01
BMI	6	724	0.29	0.14	0.29	0.28	0.14	0	1	1	0.60	0.91	0.01
DiabetesPedigreeFunction	7	724	0.17	0.14	0.13	0.15	0.11	0	1	1	1.91	5.56	0.01
Age	8	724	0.21	0.20	0.13	0.18	0.17	0	1	1	1.08	0.48	0.01
Outcome*	9	724	1.34	0.48	1.00	1.31	0.00	1	2	1	0.66	-1.57	0.02

DIMENSIONAL REDUCTION - PCA



DIMENSIONAL REDUCTION - t-SNE and UMAP





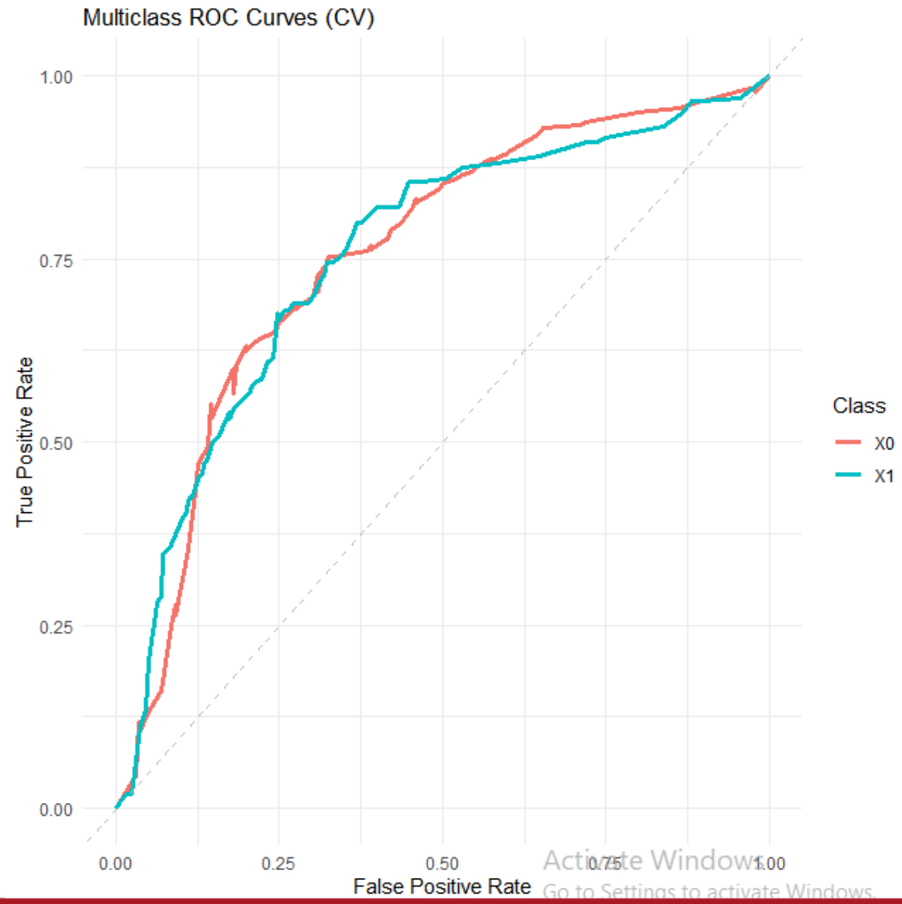
CLASSIFICATION - DECISION TREE

Metrics:

- Accuracy: 0.72 (0.69, 0.83)
- Recall: 0.77 (0.53, 0.79)
- Precision: 0.57 (0.51, 0.78)
- F1: 0.65 (0.53, 0.76)

CI 95% estimated with R=1000, BCa method

cp = 0.001

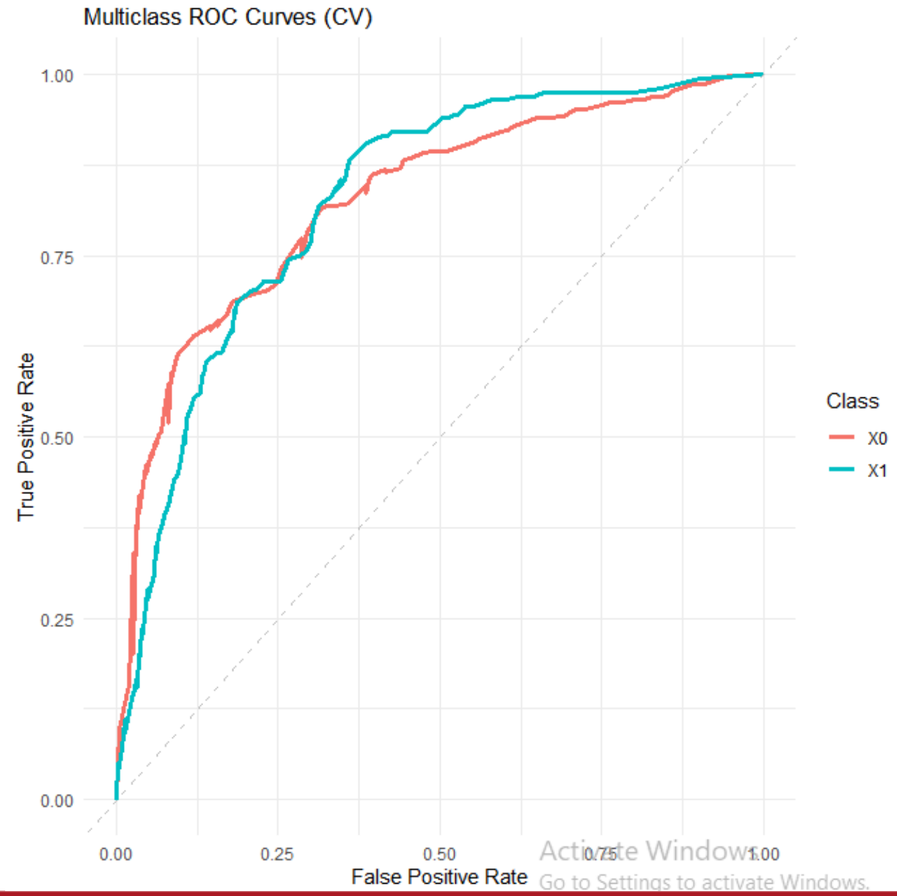


CLASSIFICATION - RANDOM FOREST

mtry = 2

Metrics:

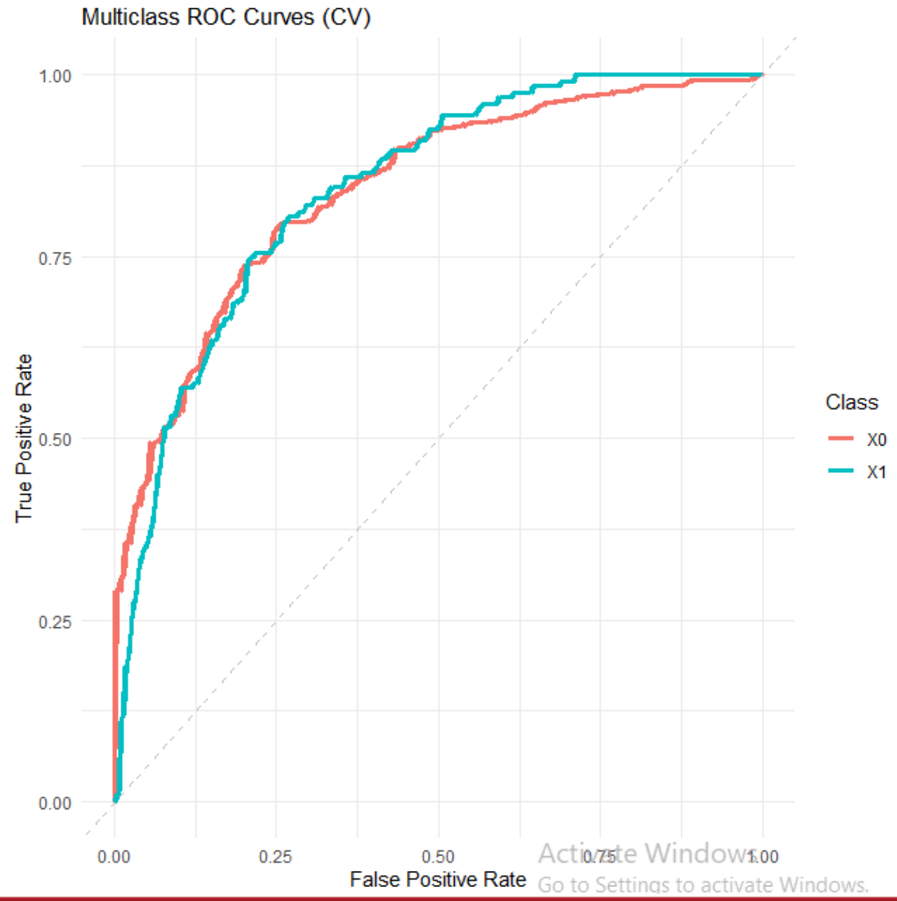
- Accuracy: 0.75 (0.67,0.81)
- Recall: 0.69 (0.47, 0.73)
- Precision: 0.62 (0.5, 0.77)
- F1: 0.66 (0.52, 0.69)



CLASSIFICATION - LDA

Metrics:

- Accuracy: 0.77 (0.67, 0.82)
- Precision: 0.65 (0.48, 0.75)
- Recall: 0.75 (0.58, 0.83)
- F1: 0.68 (0.55, 0.76)

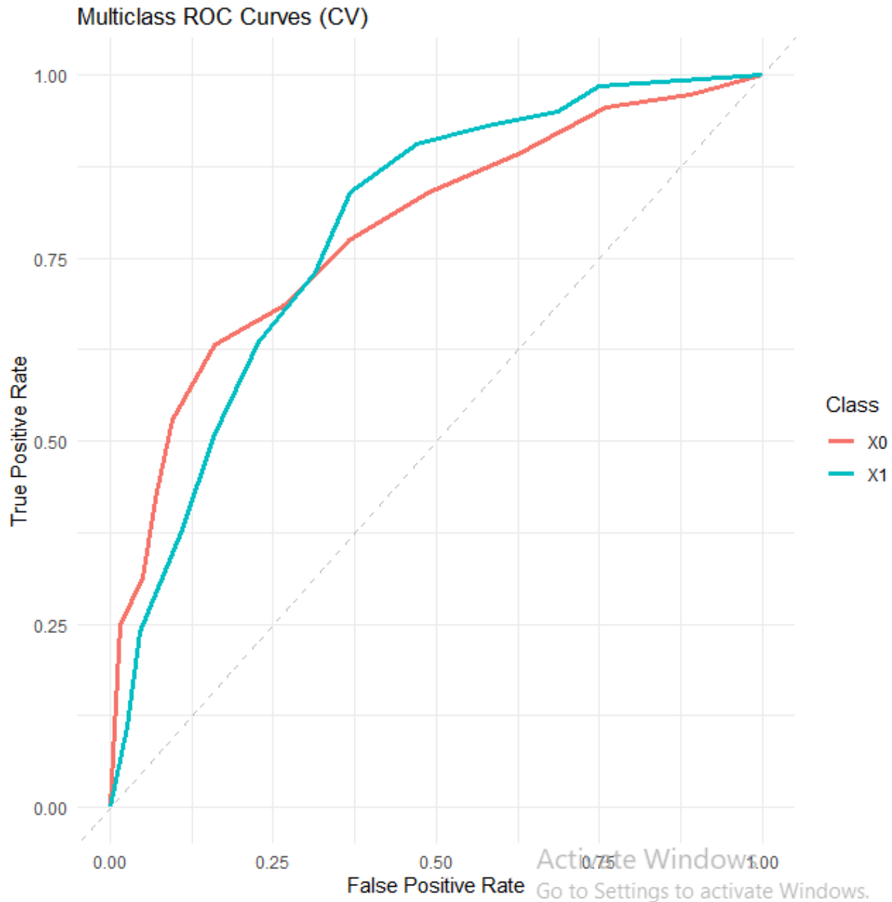


CLASSIFICATION - KNN

k= 11

Metrics:

- Accuracy: 0.72 (0.63, 0.79)
- Precision: 0.57 (0.44, 0.69)
- Recall: 0.77 (0.62, 0.86)
- F1: 0.65 (0.54, 0.75)

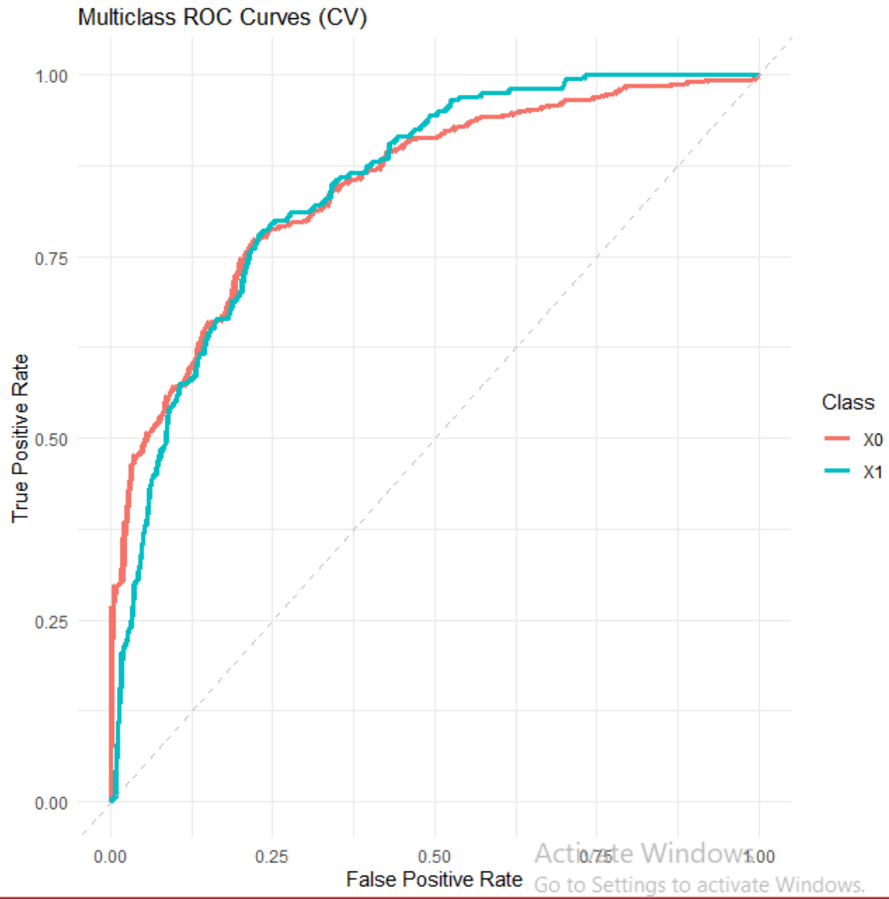


CLASSIFICATION - LOGISTIC REGRESSION

alpha=0.5, lambda =0

Metrics:

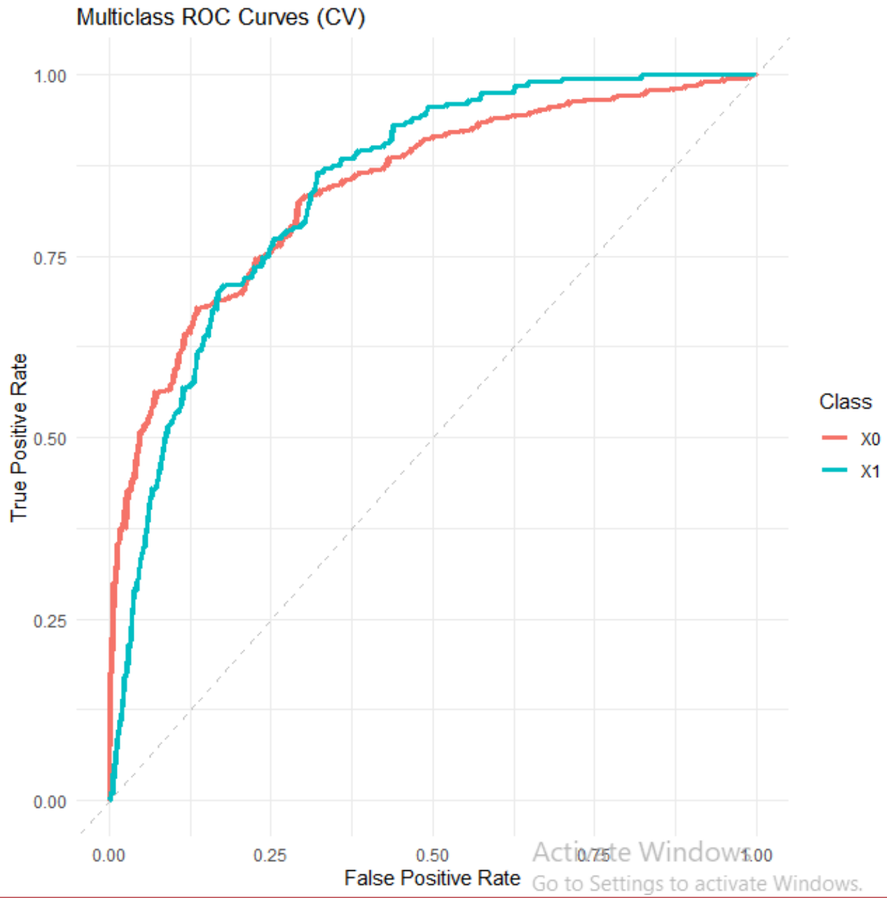
- Accuracy: 0.77 (0.67, 0.82)
- Precision: 0.65 (0.48, 0.75)
- Recall: 0.71 (0.56, 0.81)
- F1: 0.67 (0.53, 0.75)



CLASSIFICATION - NAIVE BAYES

Metrics:

- Accuracy: 0.76 (0.61, 0.79)
- Precision: 0.62 (0.41, 0.67)
- Recall: 0.77 (0.52, 0.81)
- F1: 0.69 (0.48, 0.72)



CLASSIFICATION - SVM

kernel = ‘rbf’, sigma=0.01, C=1

Metrics:

- Accuracy: 0.77 (0.67, 0.81)
- Precision: 0.71 (0.47, 0.8)
- Recall: 0.57 (0.34, 0.63)
- F1: 0.63 (0.43, 0.68)