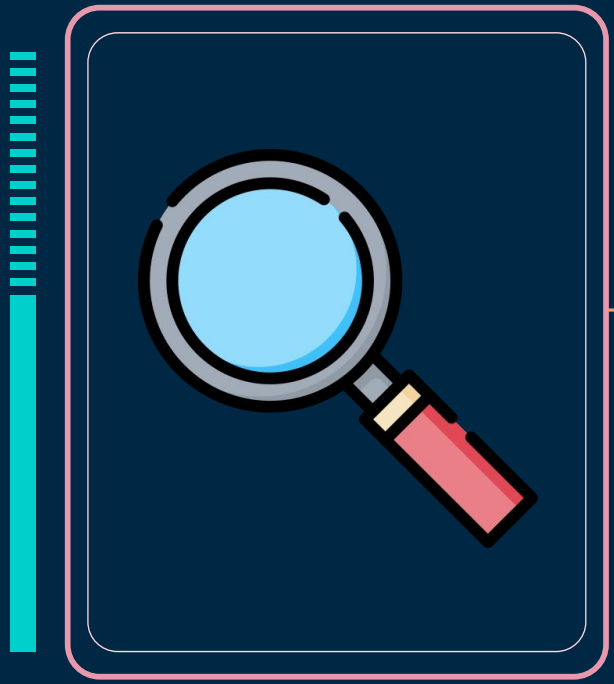
Diabetes Risk Factors Inference

Pima Indians diabetes dataset

The dataset is composed by 768 female individuals of at least 21 years old of Pima Indian heritage.

Each individual is characterized by 9 variables.



VARIABLES

Pregnancies	Numerical
Glucose	Numerical
BloodPressure	Numerical
SkinThickness	Numerical
Insulin	Numerical
BMI	Numerical
DiabetesPedigreeFunction	Numerical
Age	Numerical
Outcome	Categorical nominal, binary

TABLE OF CONTENTS

TABLE OF CONTENTS

DATA CLEANING

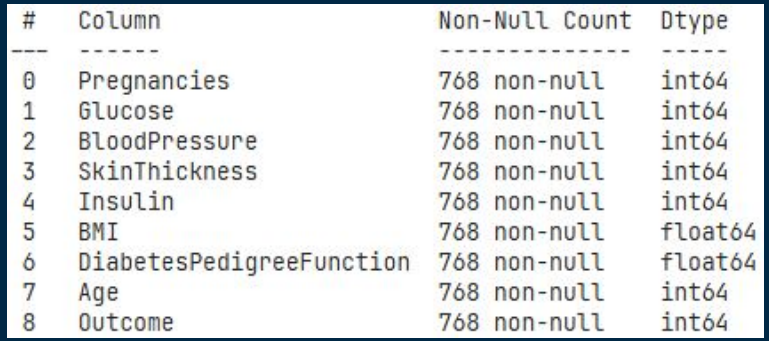
01

DATA CLEANING CHALLENGES

1. VARIABLES NAMES

2. UNEXPECTED ‘0’ VALUES

3. DATA IMBALANCE



#	Column	Non-Null Count	Dtype
0	Pregnancies	768 non-null	int64
1	Glucose	768 non-null	int64
2	BloodPressure	768 non-null	int64
3	SkinThickness	768 non-null	int64
4	Insulin	768 non-null	int64
5	BMI	768 non-null	float64
6	DiabetesPedigreeFunction	768 non-null	float64
7	Age	768 non-null	int64
8	Outcome	768 non-null	int64

VARIABLES NAMES

Variables have been renamed and styled with **snake case** to keep them consistent.

BloodPressure	→ blood_pressure
SkinThickness	→ skin_thickness
DiabetesPedigreeFunction	→ diabetes_pedigree
...	

UNEXPECTED ‘0’ VALUES

Several predictors (glucose, blood_pressure, skin_thickness, insulin, bmi) showed values equal to 0 which shouldn’t be possible values for such body vital parameters.

376 out of 768 were affected by this. Removing these individuals halves the size of the dataset but the total size is still big enough to be able to work.

	pregnancies	glucose	blood_pressure	skin_thickness	insulin	bmi	diabetes_pedigree	age	outcome
0	6	148	72	35	0	33.6	0.627	50	present
1	1	85	66	29	0	26.6	0.351	31	absent
2	8	183	64	0	0	23.3	0.672	32	present
5	5	116	74	0	0	25.6	0.201	30	absent
7	10	115	0	0	0	35.3	0.134	29	absent
...	...	...	...	...	...	...	...	...	...
761	9	170	74	31	0	44.0	0.403	43	present
762	9	89	62	0	0	22.5	0.142	33	absent
764	2	122	70	27	0	36.8	0.340	27	absent
766	1	126	60	0	0	30.1	0.349	47	present
767	1	93	70	31	0	30.4	0.315	23	absent

DATA IMBALANCE

The two classes of the dataset, diabetes being present or not, are affected by a mild imbalance, 34-66 split.

After removing the individuals with the unexpected 0 values the proportion still stays the same. In order to mitigate the situation, stratification is going to be employed with models.

```
df_raw['outcome'].value_counts()
```

```
outcome
0    500
1    268
Name: count, dtype: int64
```

```
df_raw['outcome'].value_counts()
```

```
outcome
absent     262
present    130
Name: count, dtype: int64
```

POST DATA CLEANING

HEAD

	pregnancies	glucose	blood_pressure	skin_thickness
0	1	89	66	23
1	0	137	40	35
2	3	78	50	32
3	2	197	70	45
4	1	189	60	23

insulin	bmi	diabetes_pedigree	age	outcome
94	28.1	0.167	21	absent
168	43.1	2.288	33	present
88	31.0	0.248	26	present
543	30.5	0.158	53	present
846	30.1	0.398	59	present

INFO

```
#   Column               Non-Null Count  Dtype
---  ------               --------------  -----
0   pregnancies          392 non-null    int64
1   glucose              392 non-null    int64
2   blood_pressure       392 non-null    int64
3   skin_thickness       392 non-null    int64
4   insulin              392 non-null    int64
5   bmi                  392 non-null    float64
6   diabetes_pedigree    392 non-null    float64
7   age                  392 non-null    int64
8   outcome              392 non-null    category
```

EXPLORATORY DATA ANALYSIS

02

EXPLORATORY DATA ANALYSIS

1. UNIVARIATE ANALYSIS

2. BIVARIATE ANALYSIS

NUMERICAL VARIABLES STATISTICS

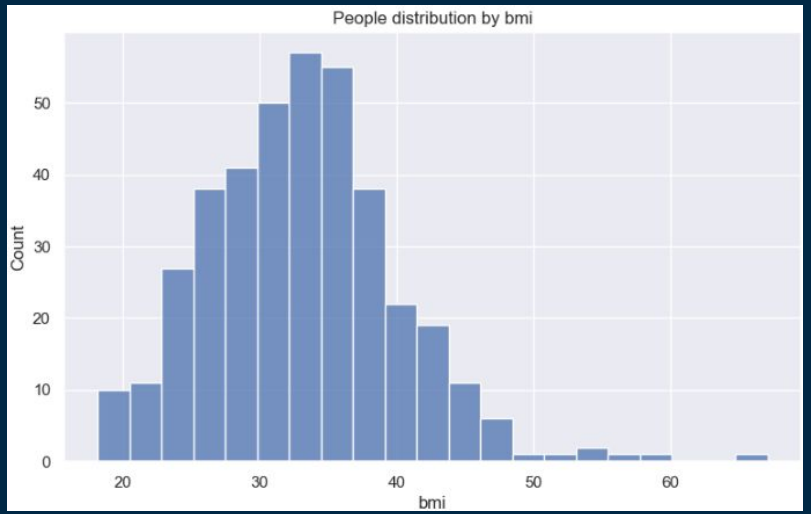
Interesting informations:

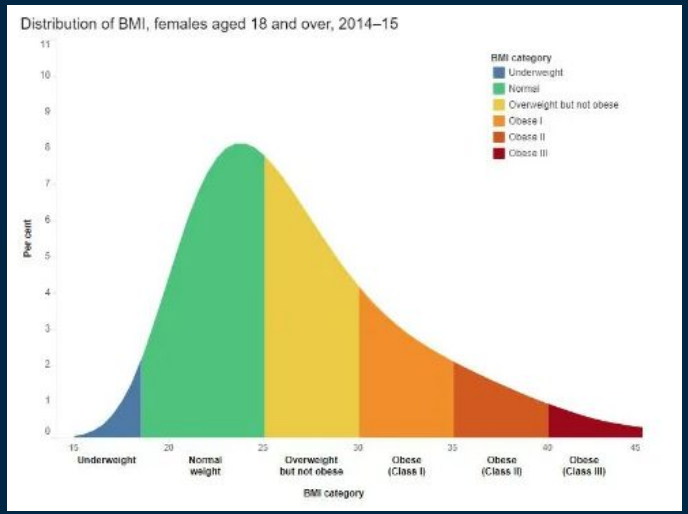
- The dataset individuals mean glucose level is ~122.
- The dataset individuals are mostly around type I obesity according to bmi.

	pregnancies	glucose	blood_pressure	skin_thickness	insulin	bmi	diabetes_pedigree	age
mean	3.301	122.628	70.663	29.145	156.056	33.086	0.523	30.865
std	3.211	30.861	12.496	10.516	118.842	7.028	0.345	10.201
var	10.313	952.388	156.152	110.595	14123.347	49.388	0.119	104.056
skew	1.336	0.518	-0.088	0.209	2.165	0.663	1.959	1.404
kurt	1.486	-0.483	0.795	-0.458	6.357	1.557	6.367	1.738
min	0.000	56.000	24.000	7.000	14.000	18.200	0.085	21.000
25%	1.000	99.000	62.000	21.000	76.750	28.400	0.270	23.000
50%	2.000	119.000	70.000	29.000	125.500	33.200	0.450	27.000
75%	5.000	143.000	78.000	37.000	190.000	37.100	0.687	36.000
max	17.000	198.000	110.000	63.000	846.000	67.100	2.420	81.000

BMI DISTRIBUTION

BMI distribution has a similar shape to the “expected” BMI distribution but it is more right shifted.

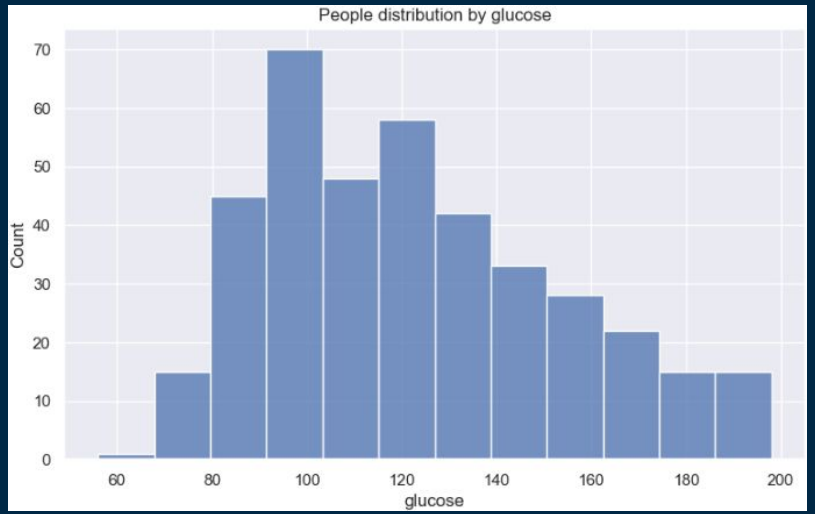


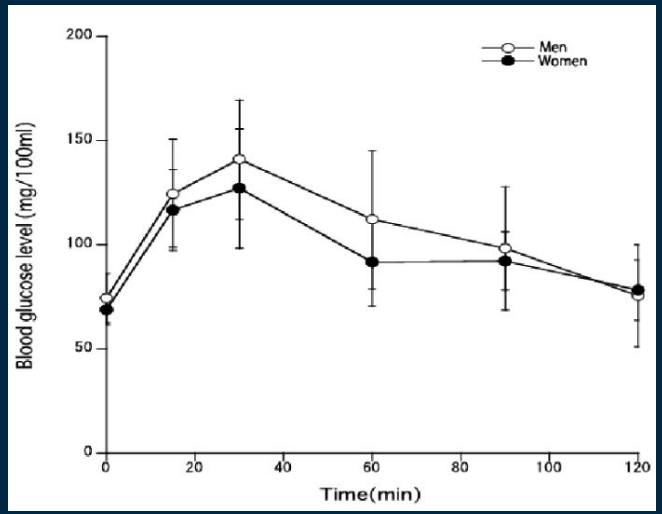


UNIVARIATE ANALYSIS

GLUCOSE LEVELS

Blood glucose levels are expected to be around ~90 after 2hrs for women while the individuals average appears to be around ~122.



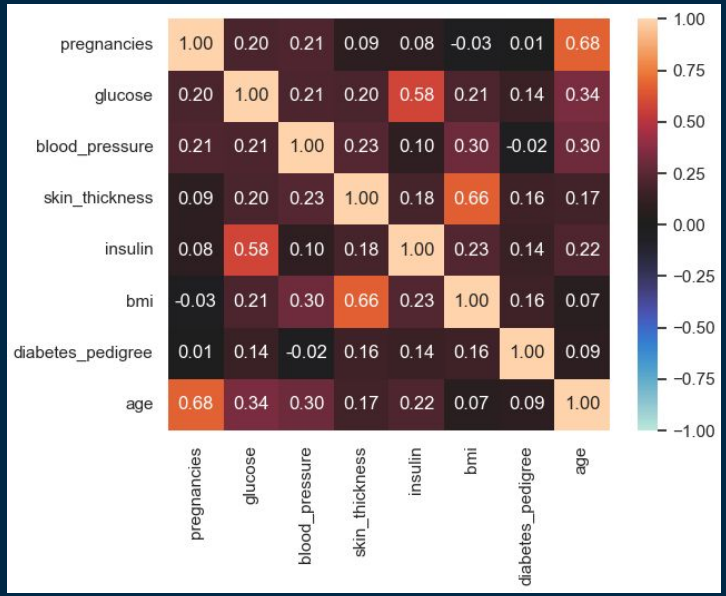


HEATMAP

Interesting informations:

There are several expected positive linear correlation due to the nature of the human body

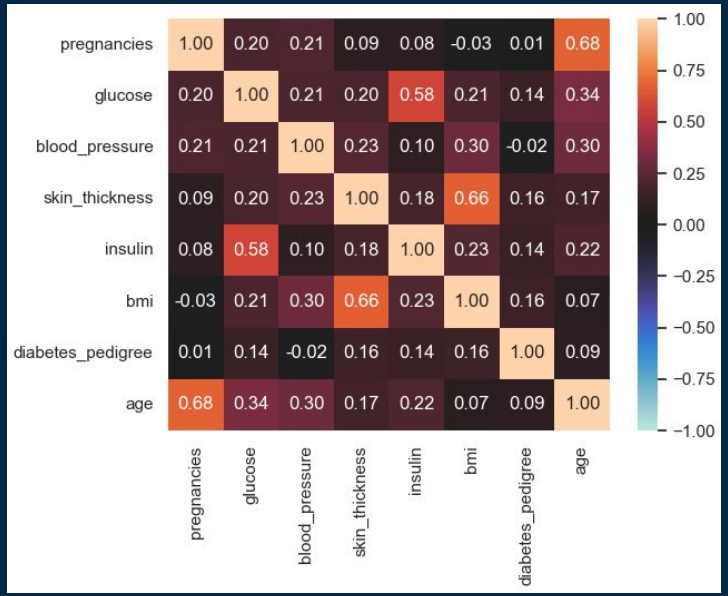
- **Age** and **Pregnancies** (0.68).
- **BMI** and **Skin thickness** (0.66).
- **Glucose** and **Insulin** (0.58).



HEATMAP

Interesting informations:

There are several absent linear correlation, especially regarding **Pregnancies, Age** and **Diabetes Pedigree** with other variables.

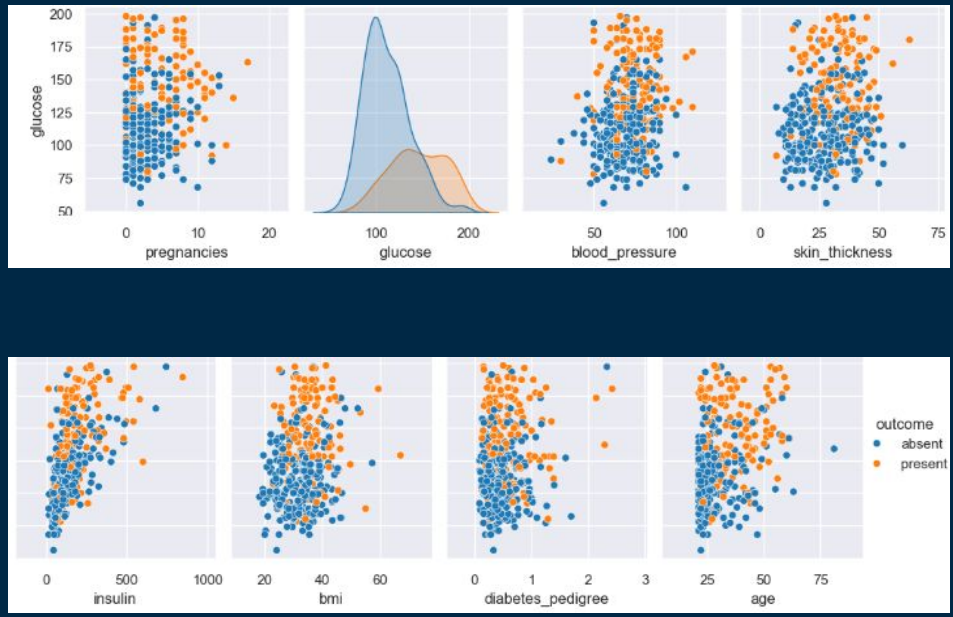


PAIRPLOT

Interesting informations:

- **Glucose** appears to be the predictor that best separates the two classes when plotted against any other predictor

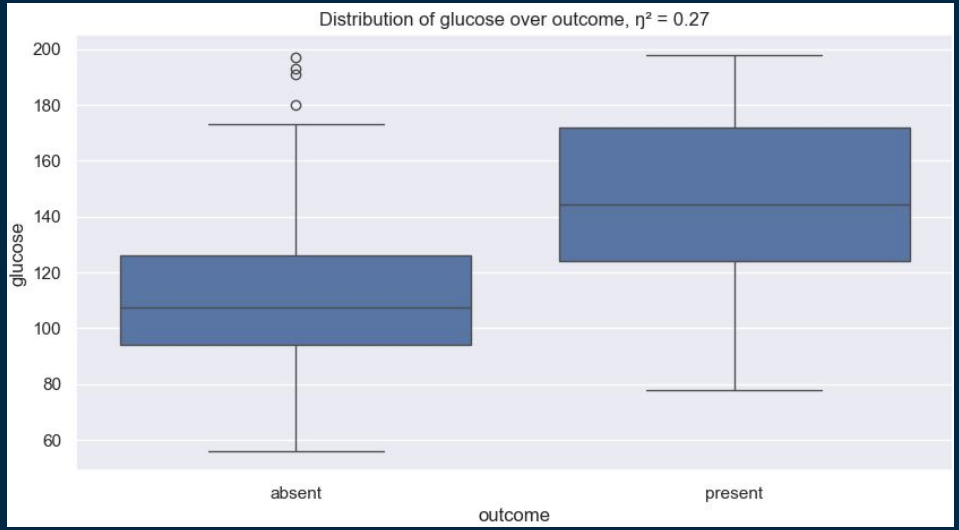
- The **Glucose-Glucose** plot showing the distribution-per-class appears to be the least overlapping compared to the other same-predictors-pairs plot.



OUTCOME-GLUCOSE DISTRIBUTION

Glucose being the most meaningful predictor can be further noticed by the eta squared coefficient value.

The IQR of people having diabetes includes people having higher levels of glucose.



PCA AND CLUSTERING

03

PCA AND CLUSTERING

1. ENCODING
2. SCALING
3. PCA
4. CLUSTERING (K-MEANS)

ENCODING AND SCALING

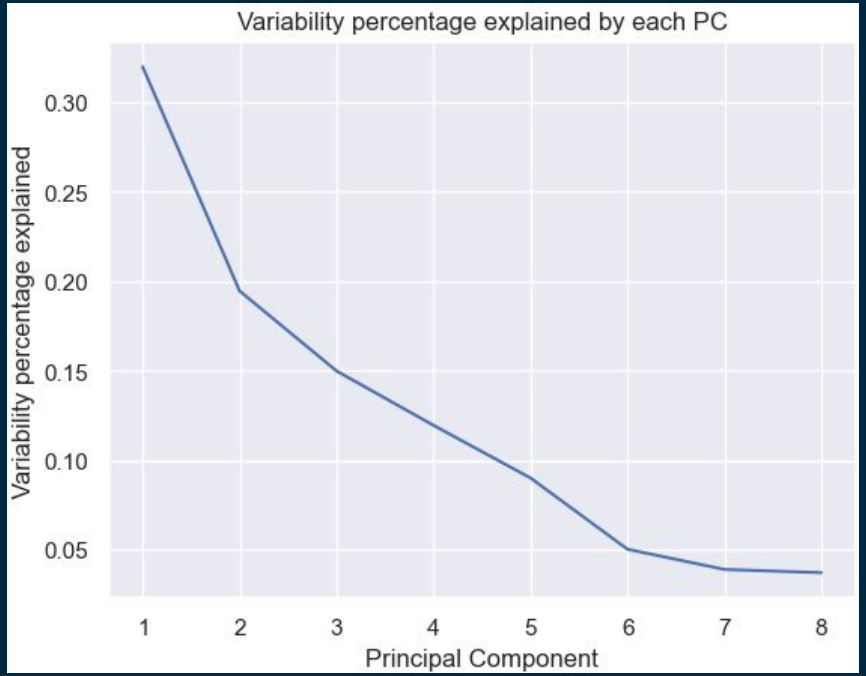
The only categorical variable present, outcome, is not taken into account for applying PCA since it is the response variable therefore no encoding is needed since all the remaining variables are numerical.

Scaling is a mandatory step before applying PCA so variables have been standardized.

PRINCIPAL COMPONENTS VARIABILITY

Interesting information:

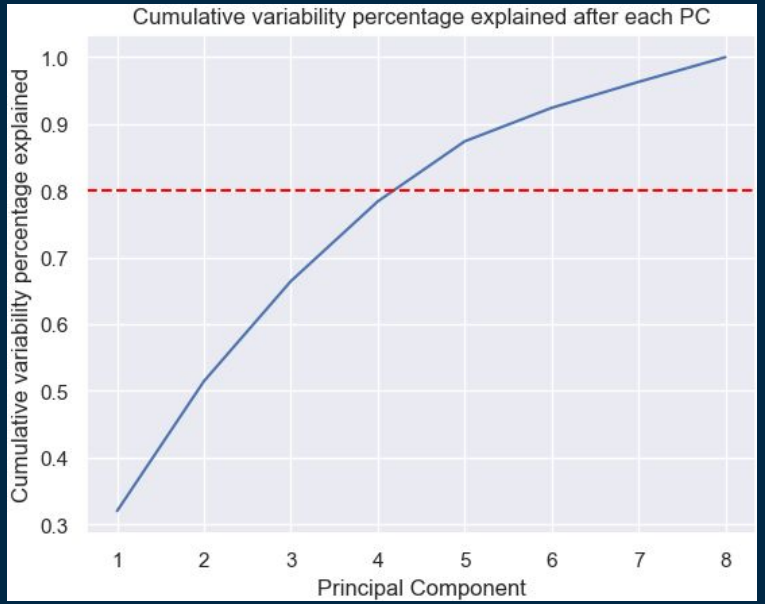
- An elbow is present at the second Principal Component.



PRINCIPAL COMPONENT CUMULATIVE VARIABILITY

Interesting information:

- The amount of Principal Component to reach ~80% variability explained is 4.



FIRST PRINCIPAL COMPONENT INTERPRETATION

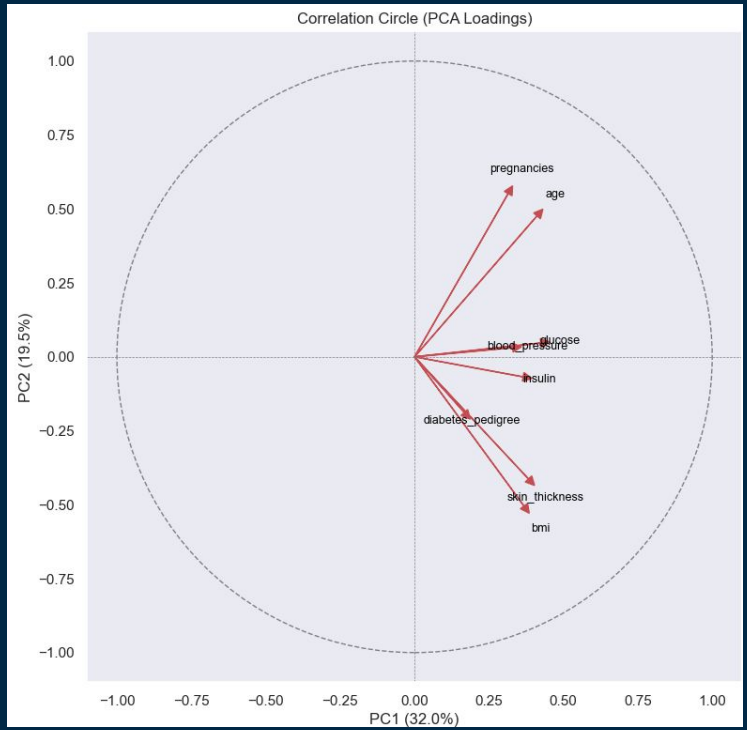
Diabetes factors

Positive values:

- Individuals with higher than average diabetes associated body parameters.

Negative values:

- Individuals with lower than average diabetes associated body parameters.



SECOND PRINCIPAL COMPONENT INTERPRETATION

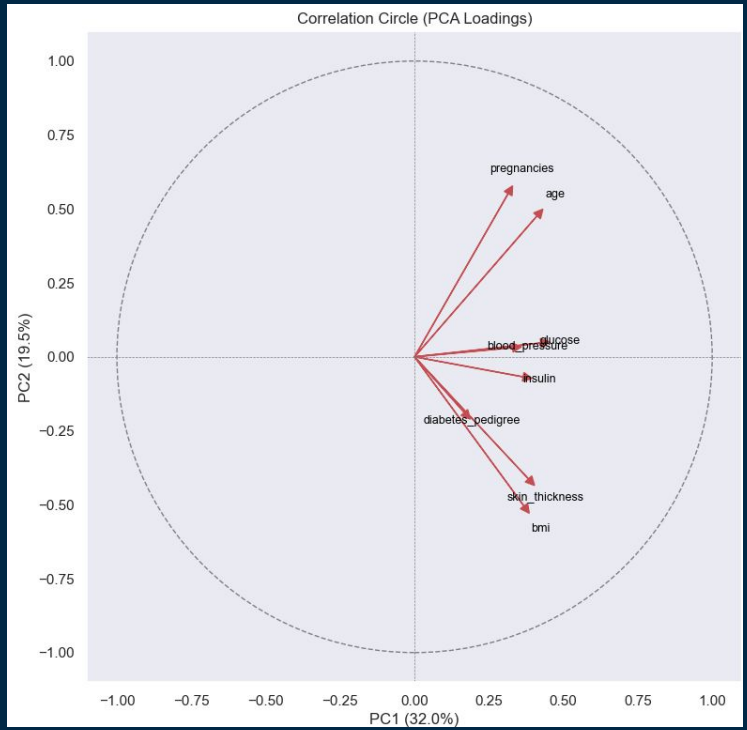
Life stage vs Body composition

Positive values:

- Older individuals with higher than average number of pregnancies

Negative values:

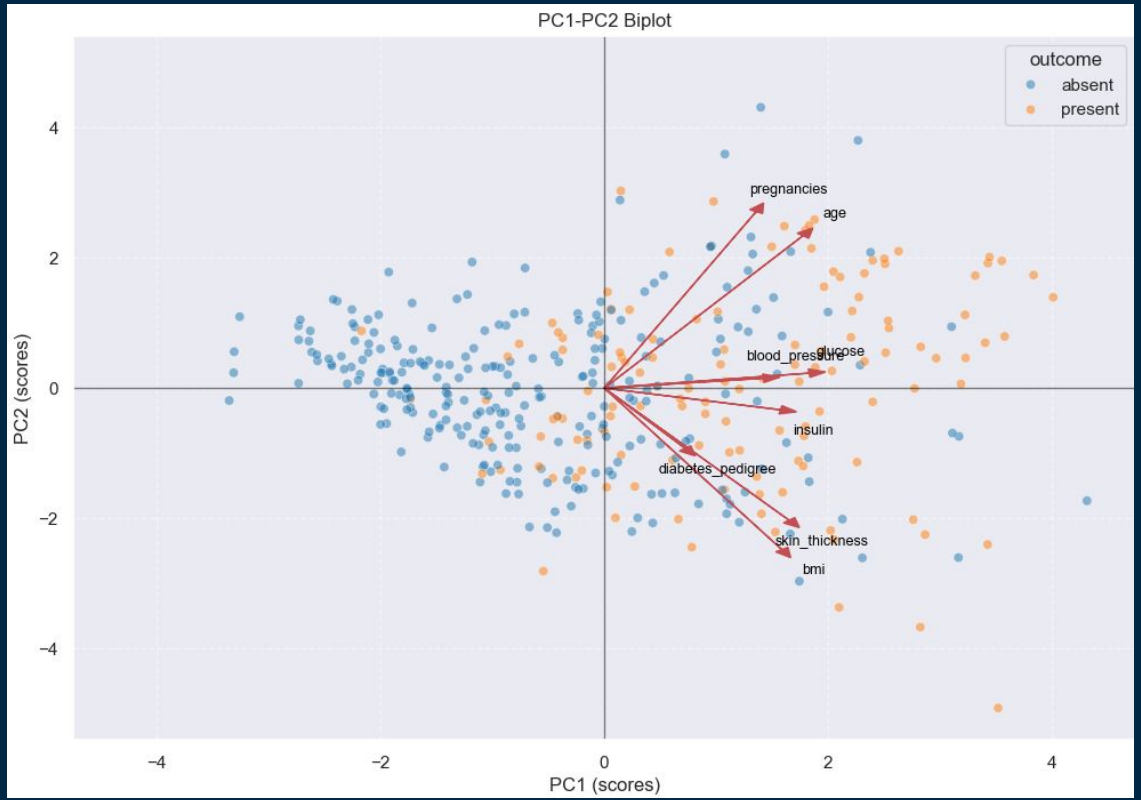
- Younger individuals with higher than average values of BMI and skin thickness



BIPLOT

Interesting informations:

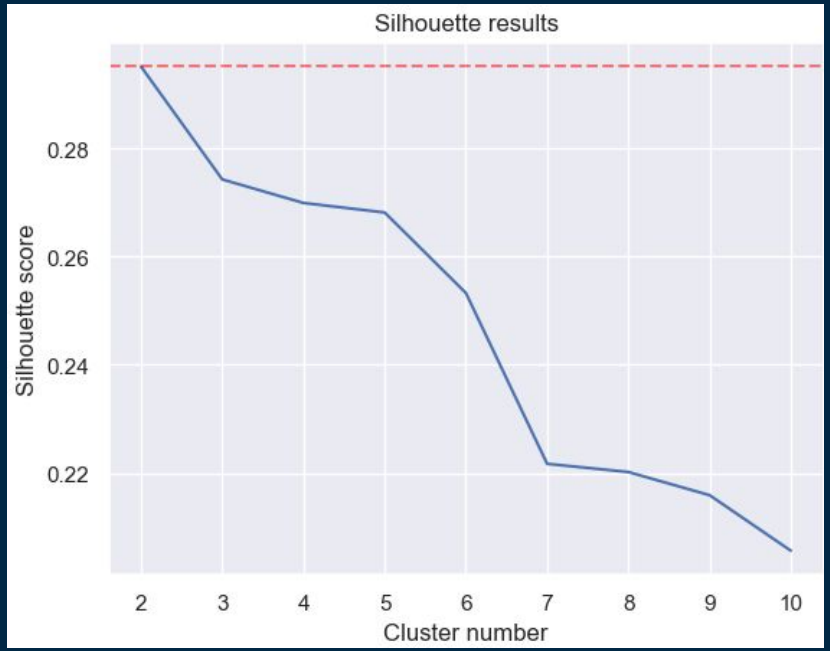
- Individuals affected by diabetes tend to have higher values of Glucose, Age and BMI.



SILHOUETTE SCORE

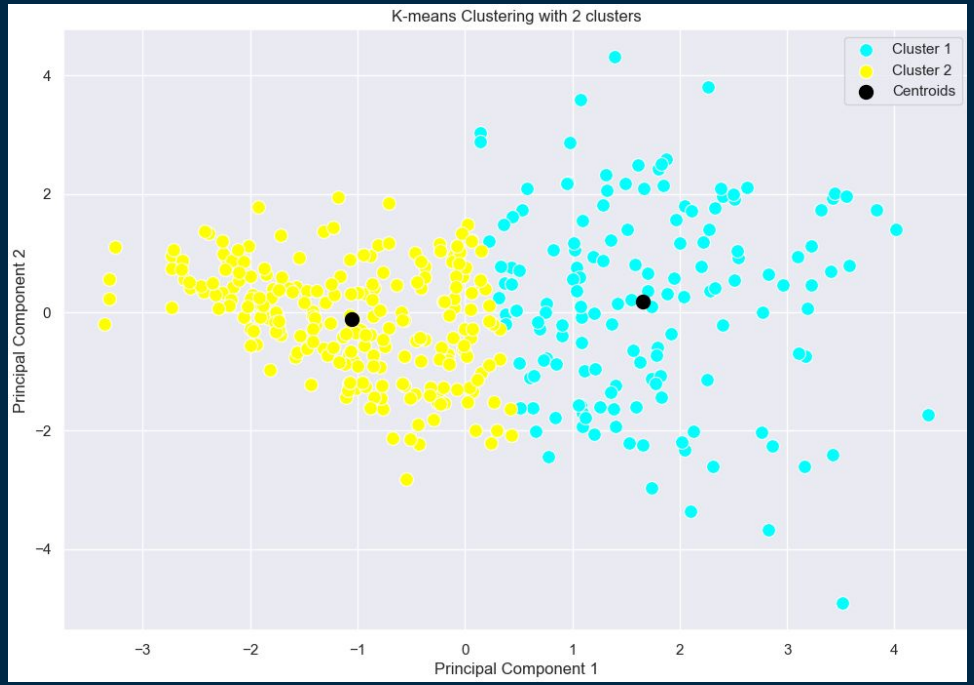
Interesting informations:

- Two clusters appear to be the best choice according to the Silhouette score.



K-MEANS

Clustering through K-means provides a similar result also noticeable from the Biplot, not adding much information.



SUPERVISED ANALYSIS

04

SUPERVISED ANALYSIS

1. SVC
2. SVM WITH RBF KERNEL
3. LOGISTIC REGRESSION
4. K-NN

SUPERVISED ANALYSIS

The following models have been chosen for the analysis:

- SVC
- SVM with RBF kernel
- Logistic regression
- K-NN

Recall is going to be the chosen metric since we're interested in not having false negatives being patients classified as not having diabetes when they have diabetes.

The classes are imbalanced (138 patients having diabetes while 238 patients don't, with a 36-64 split) therefore **Accuracy** might not be the ideal metric.

Data has been divided into training and testing sets with a 70-30 split and **stratified** .

PCA vs no PCA

PCA is a useful tool for the purpose of dimensionality reduction and maximizing variability explained through fewer predictors, employing in this case only the four Principal Components to explained ~80% of variability.

At the same time, based on the scenario, interpretability might be heavily lost due to the predictors transformation.

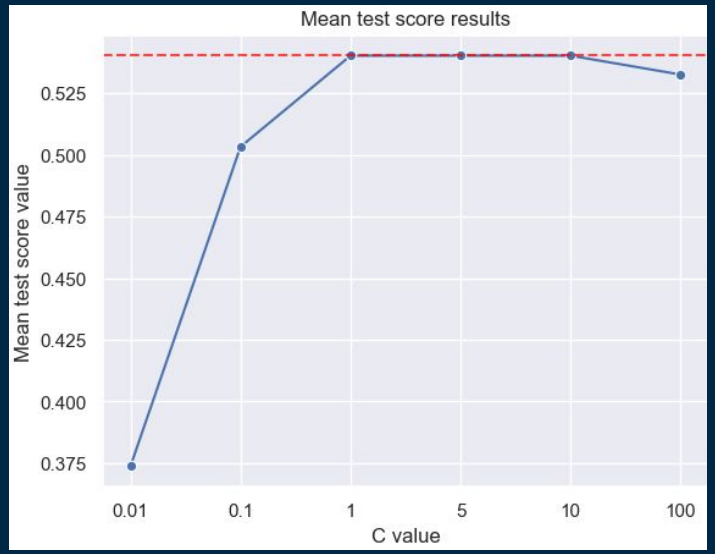
It has been decided to compare the models performance on PCA vs no PCA in order to verify if the lost interpretability due to working on Principal Components instead of the starting predictors is a worth trade-off.


SVC

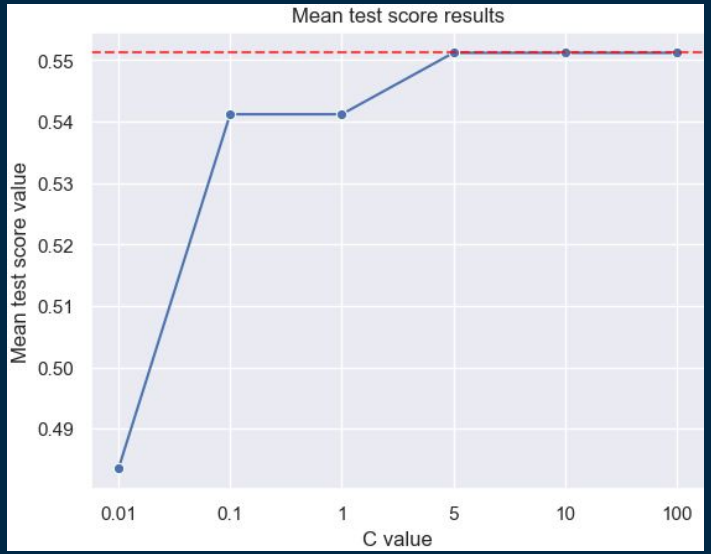
C VALUE SELECTION

Employing cross-validation, the selected C value is 1 for the PCA scenario and 5 for No PCA scenario.

PCA



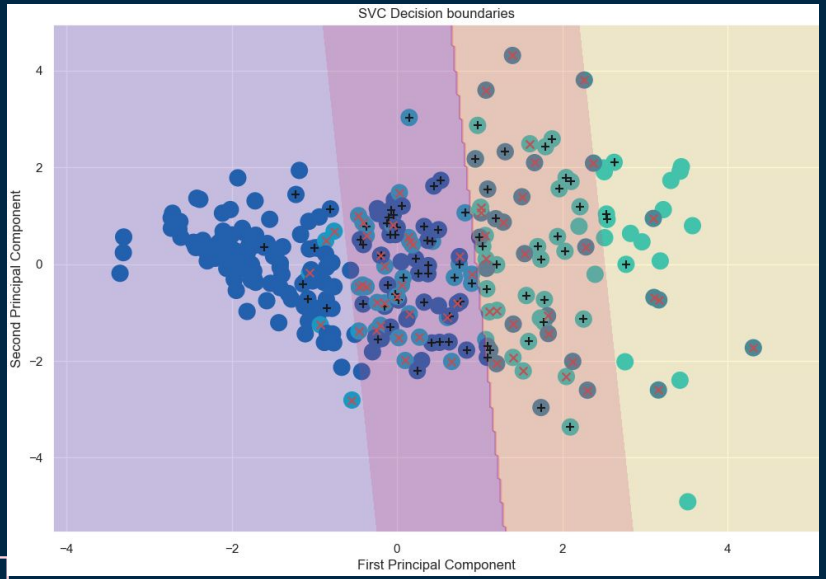
No PCA



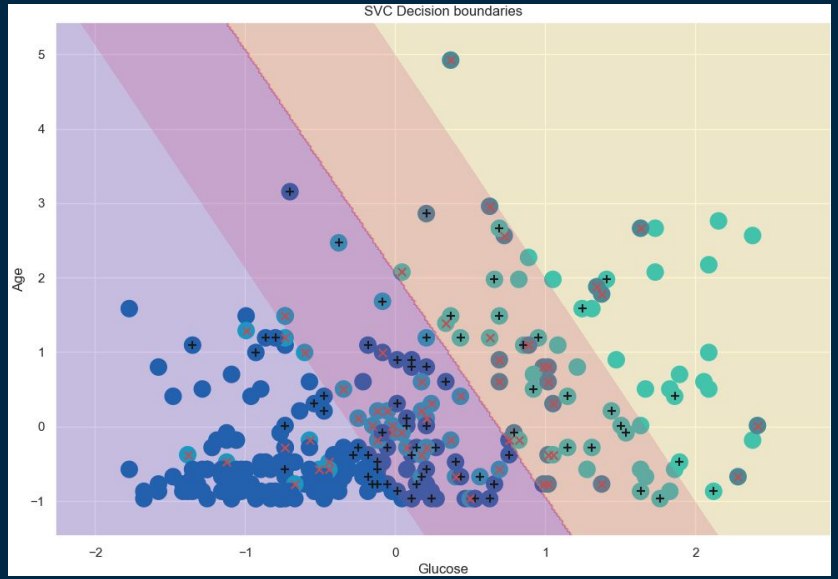
SVC

DECISION BOUNDARIES

PCA



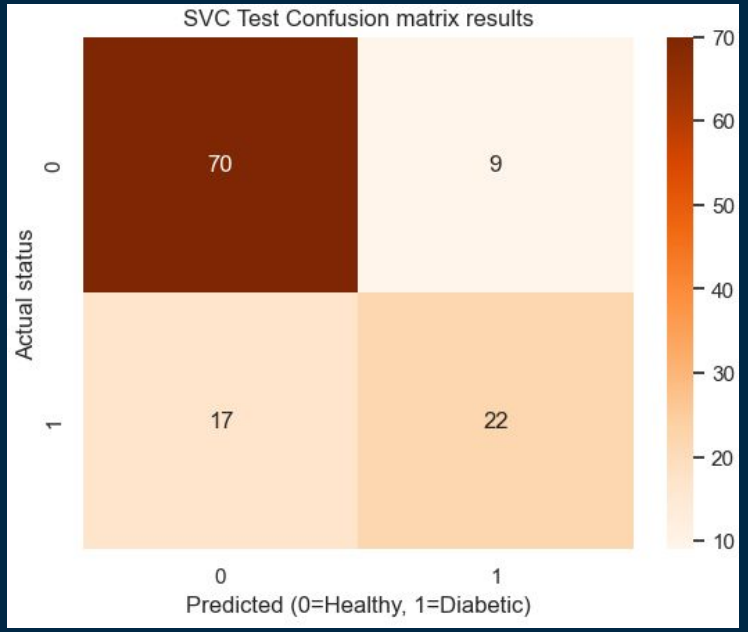
No PCA



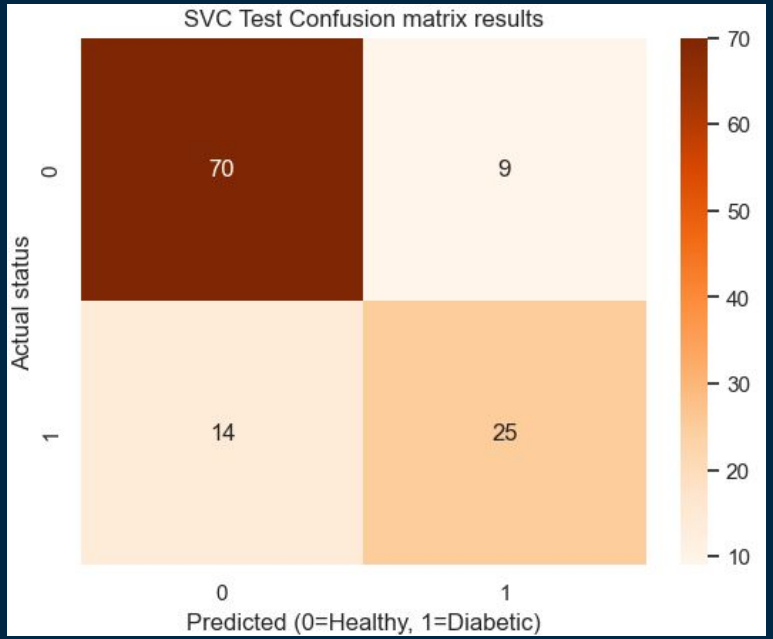
SVC

CONFUSION MATRIX

PCA



No PCA



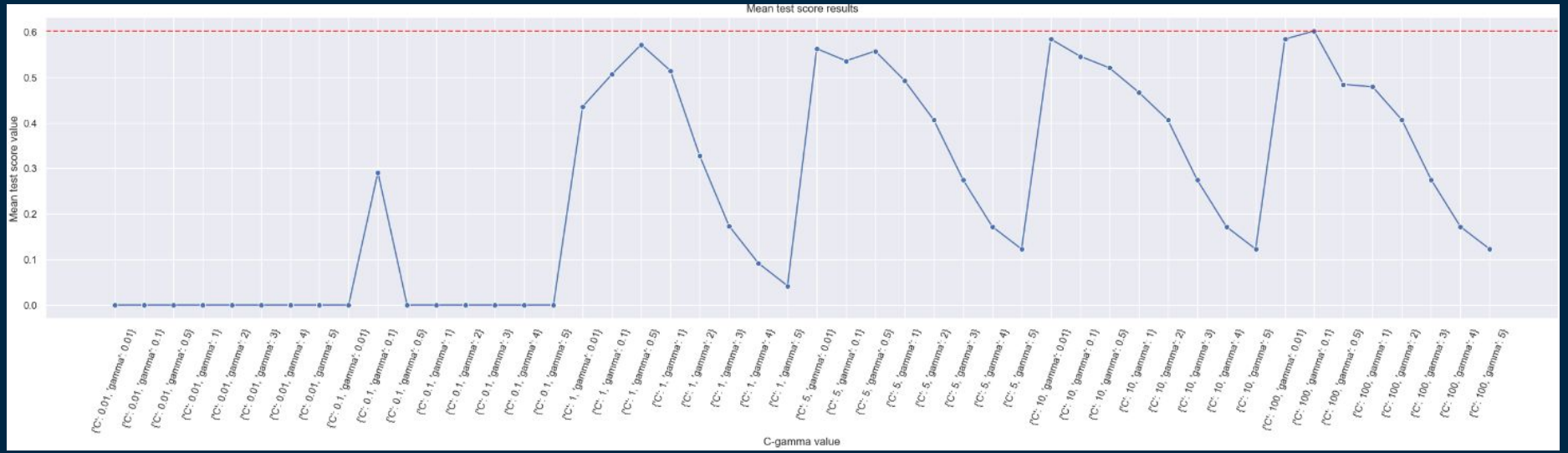
SVC

SVM

C AND GAMMA VALUE SELECTION

Employing cross-validation, the selected C value is 100 and the Gamma value is 0.1 in the PCA scenario

PCA

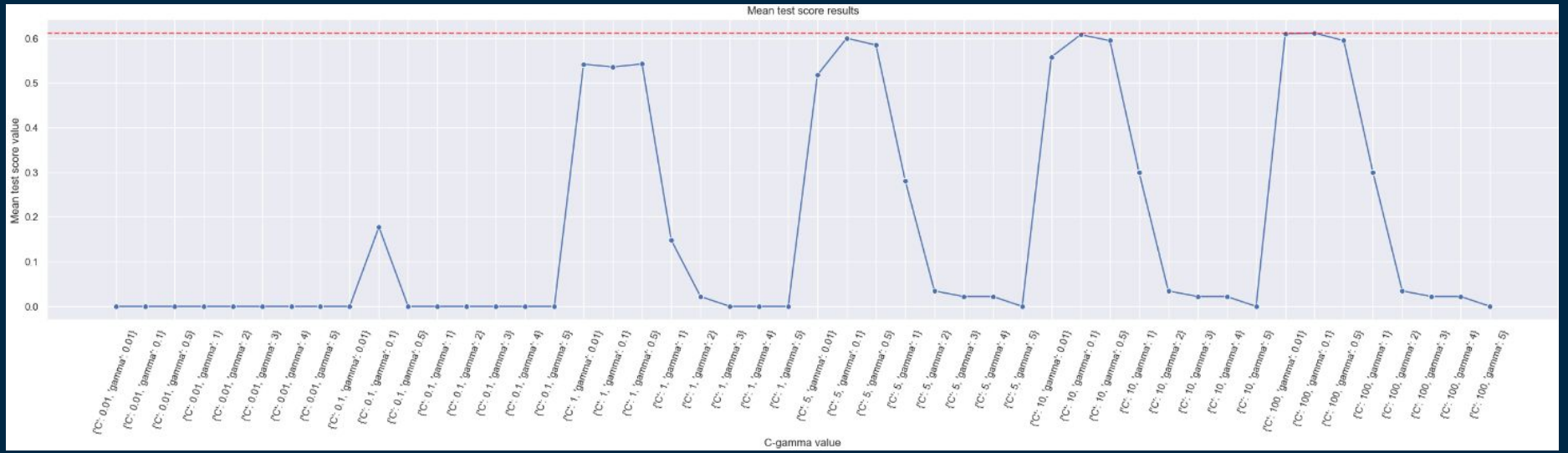


SVM

C AND GAMMA VALUE SELECTION

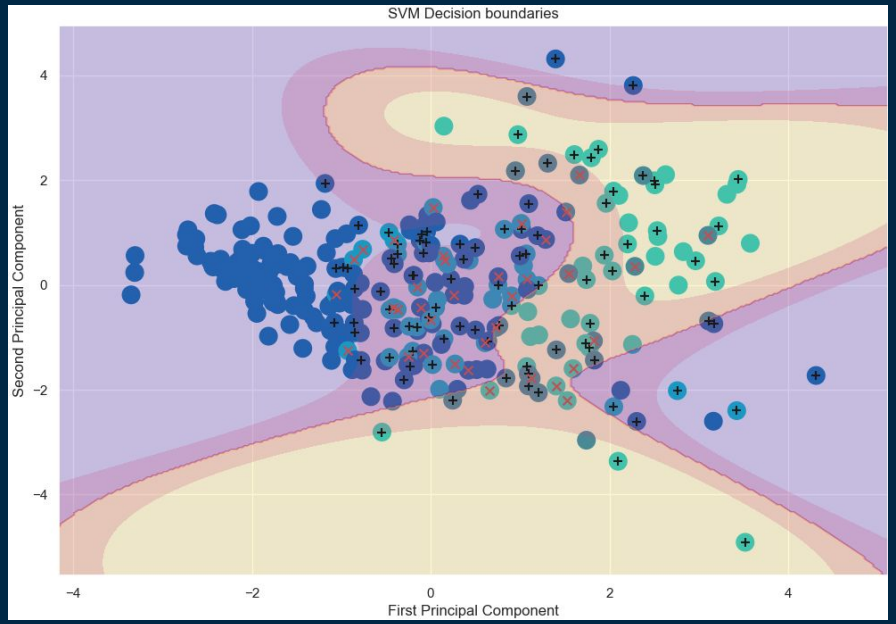
Employing cross-validation, the selected C value is 100 and the Gamma value is 0.01 in the No PCA scenario

No PCA

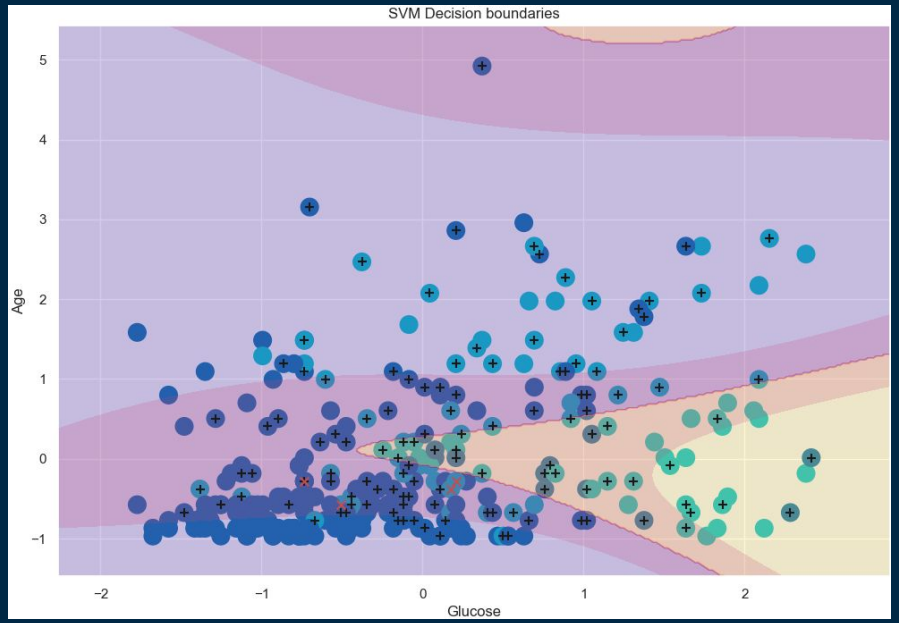


DECISION BOUNDARIES

PCA

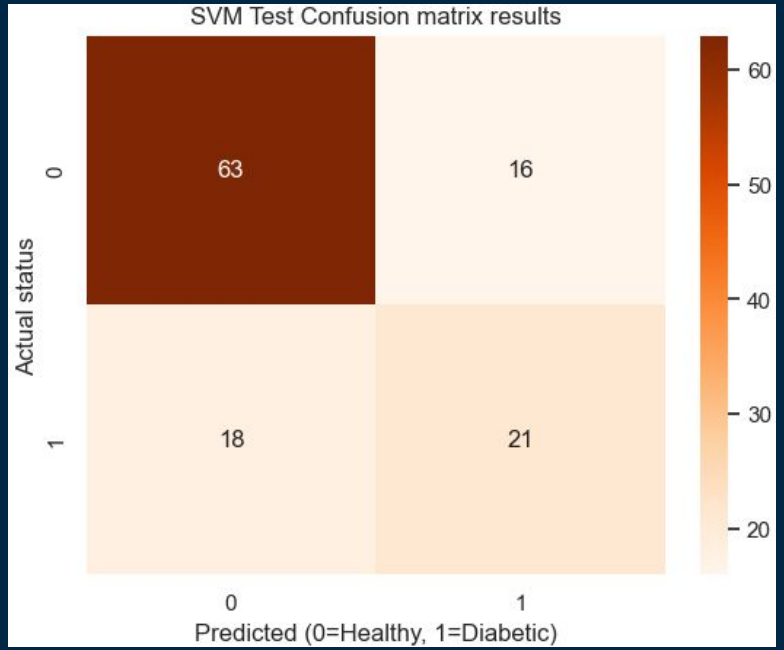


No PCA

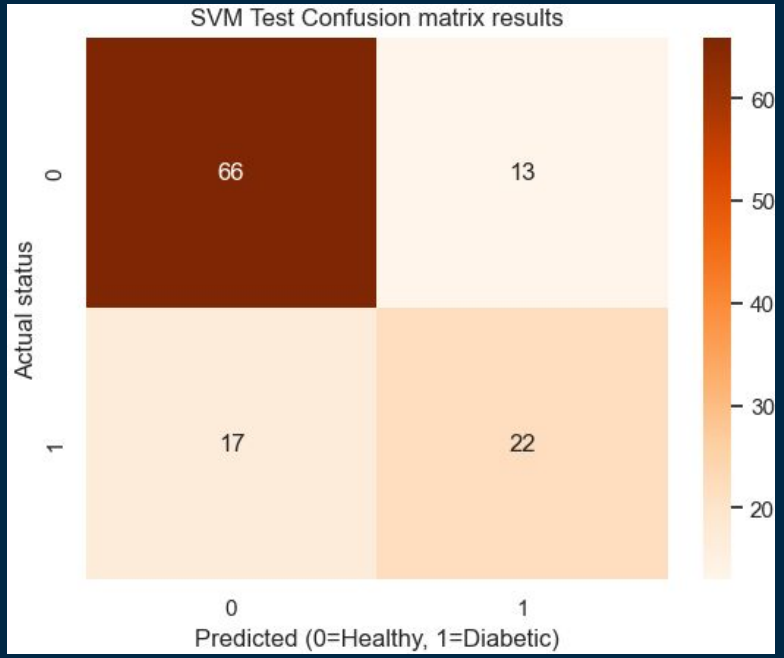


CONFUSION MATRIX

PCA



No PCA

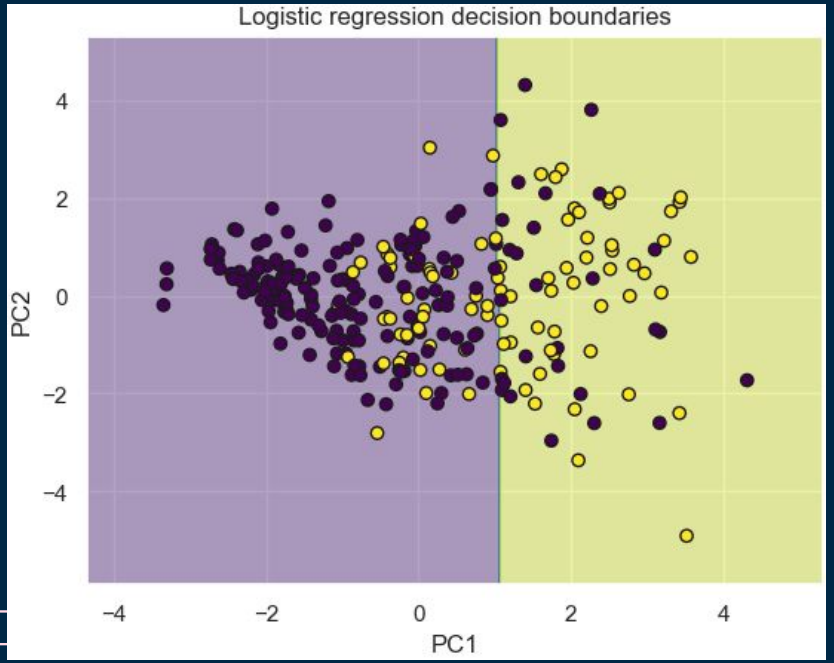


SVM

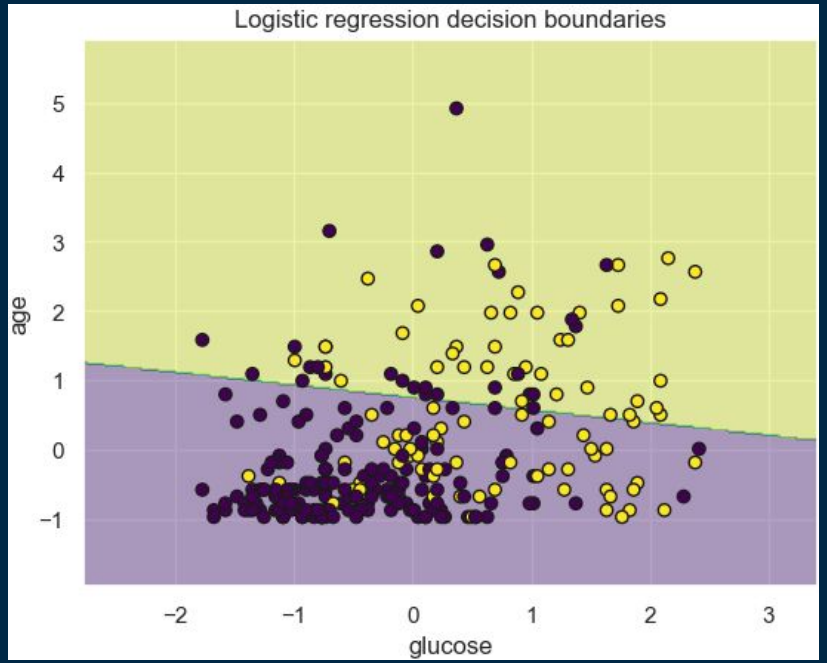
LOGISTIC REGRESSION

DECISION BOUNDARIES

PCA

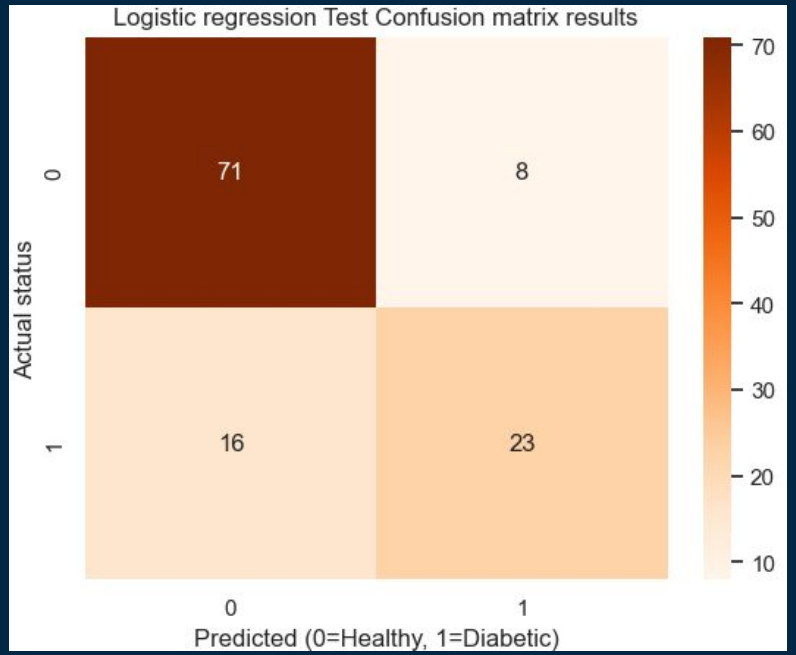


No PCA

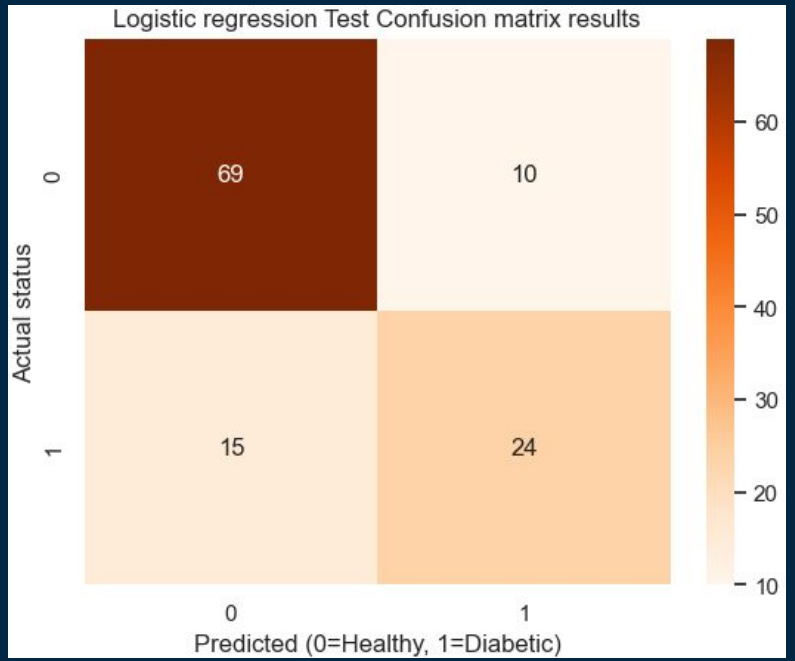


CONFUSION MATRIX

PCA



No PCA



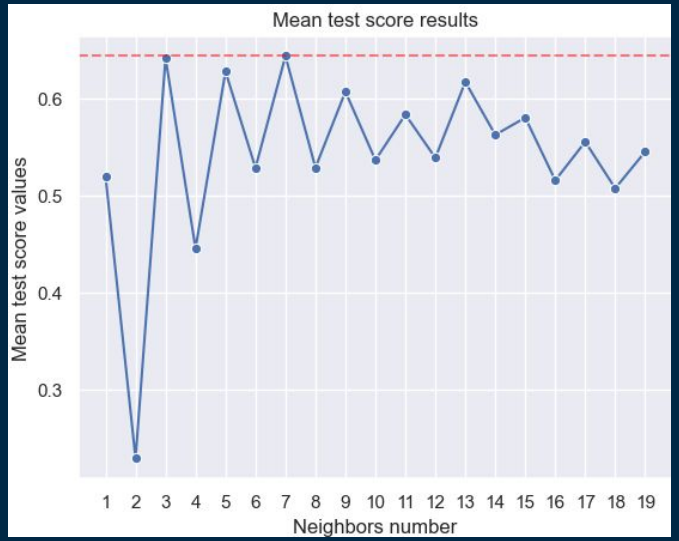
LOGISTIC REGRESSION

K-NN

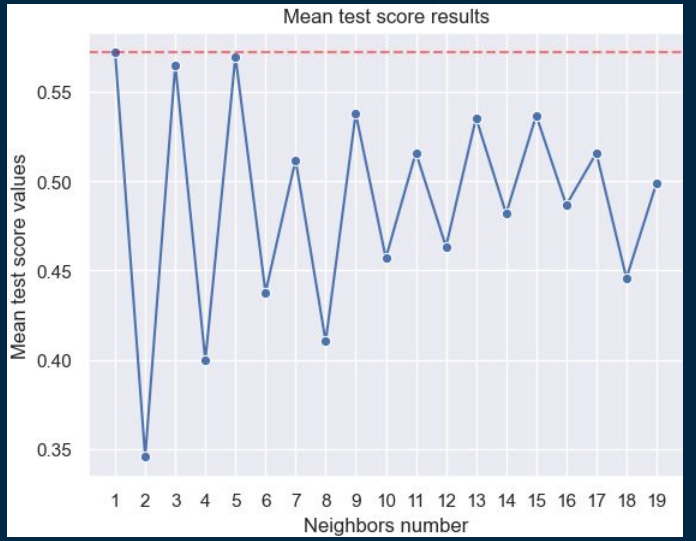
NEIGHBORS NUMBER SELECTION

Employing cross-validation, the selected K value is 7 for the PCA scenario and 1 for the No PCA scenario

PCA

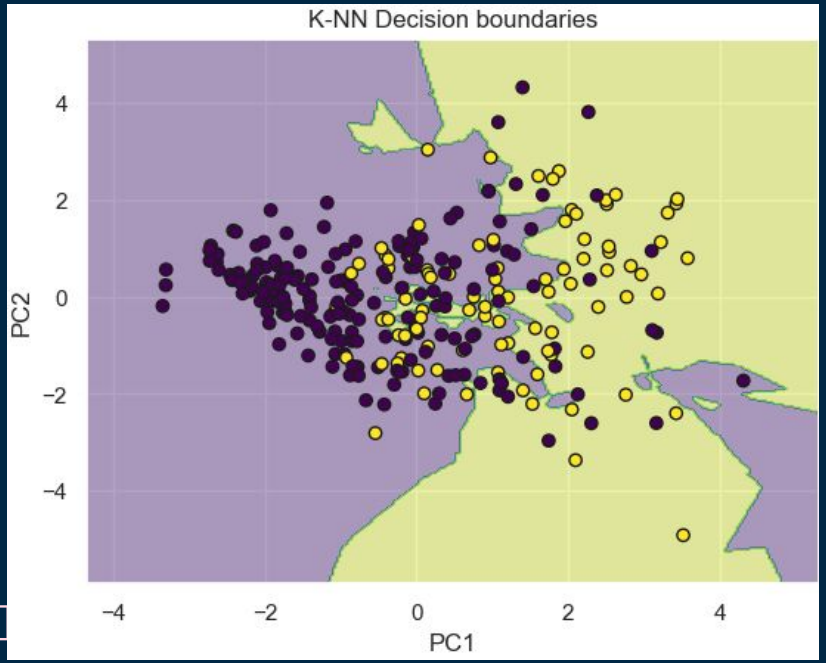


No PCA

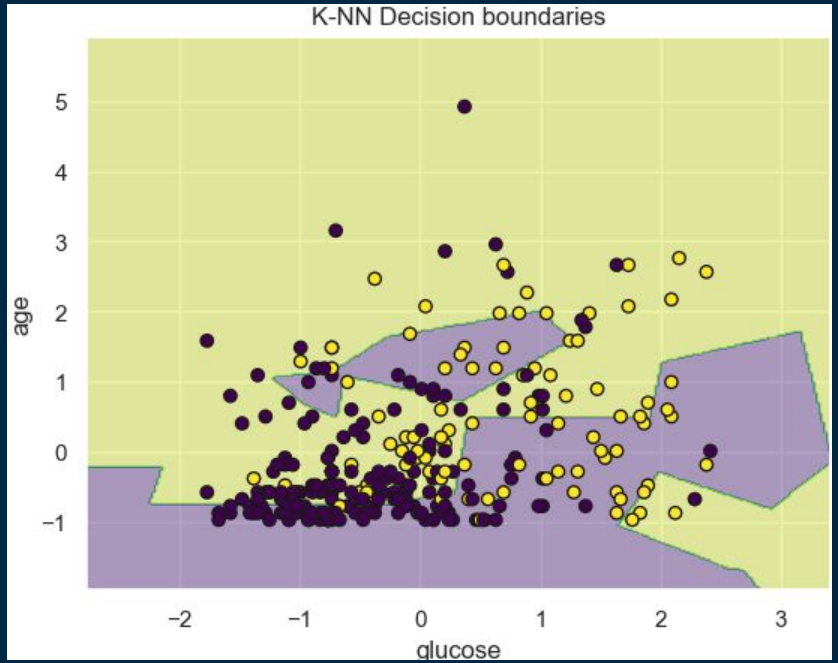


DECISION BOUNDARIES

PCA

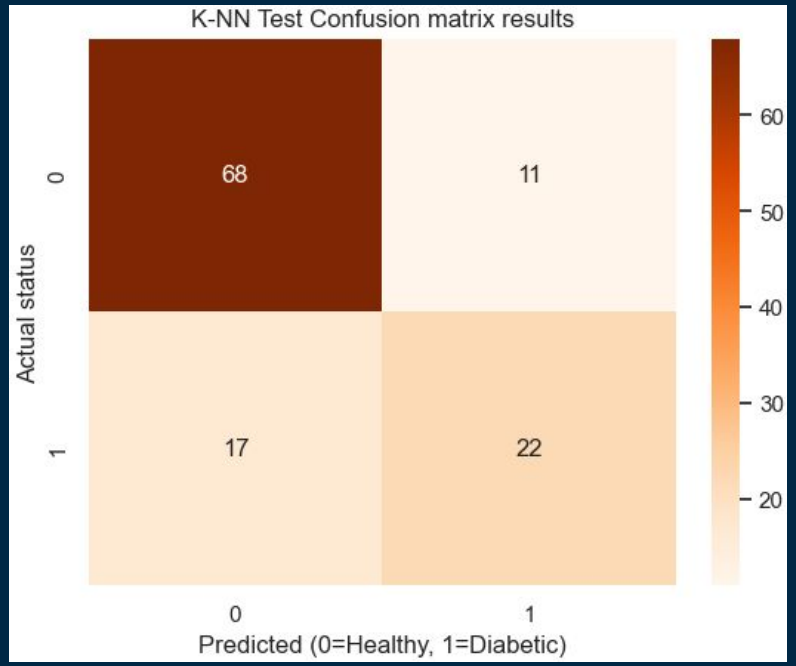


No PCA

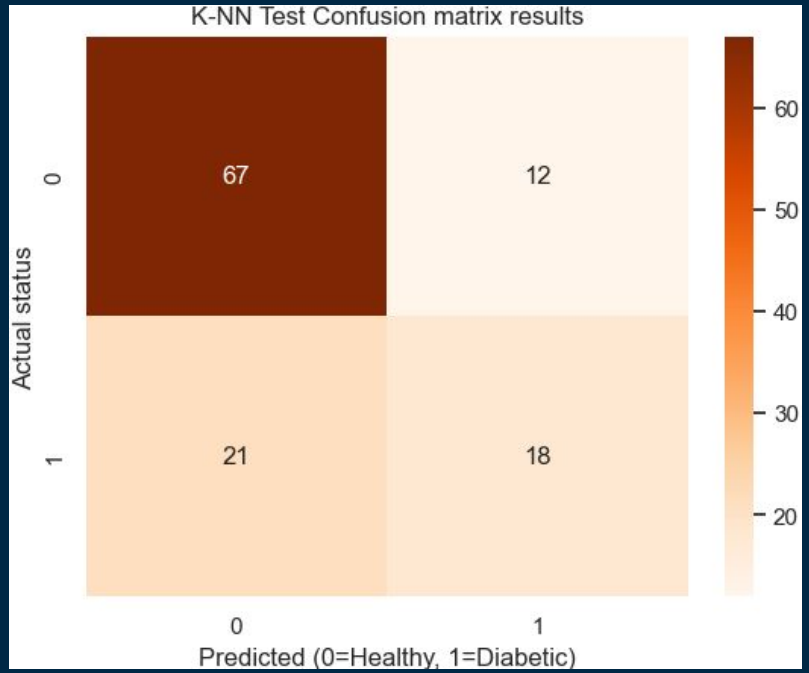


CONFUSION MATRIX

PCA



No PCA



K-NN

MODEL PERFORMANCE AND INTERPRETATION

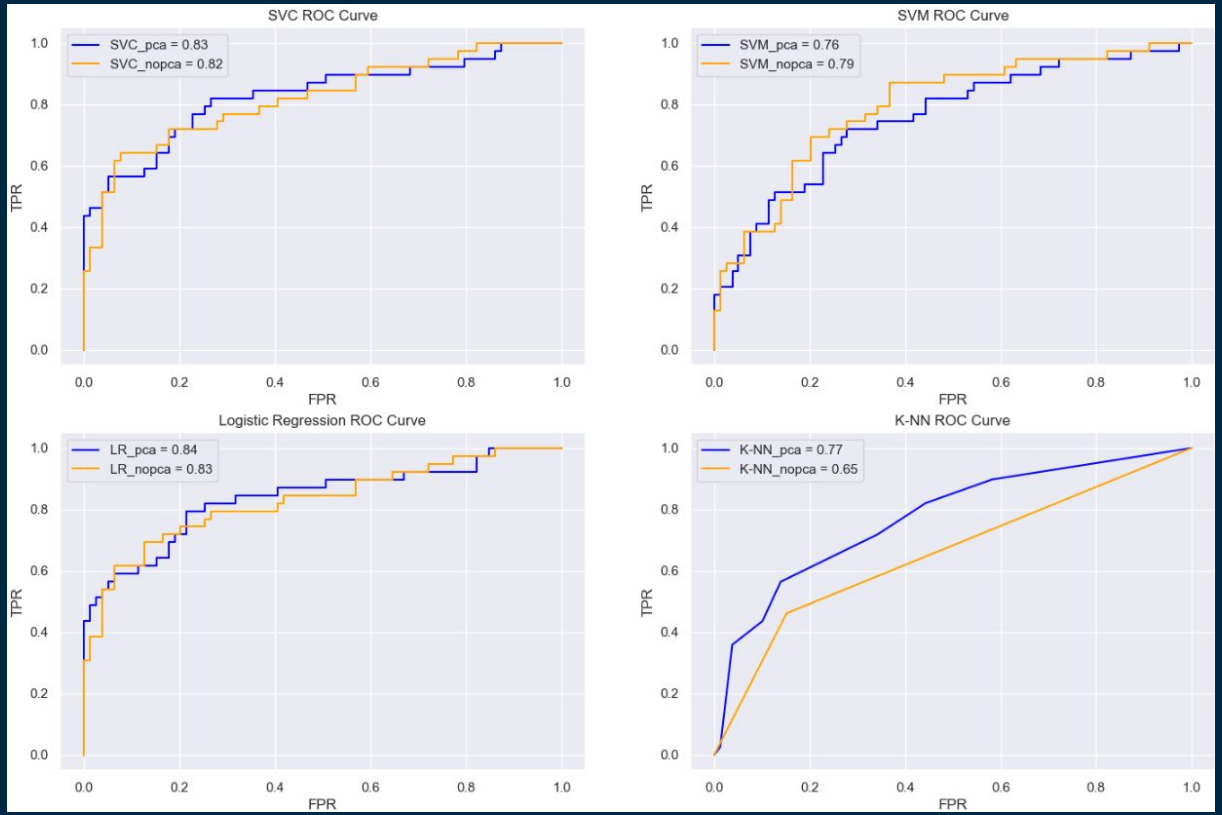
05

MODEL PERFORMANCE AND INTERPRETATION

1. MODEL PERFORMANCE

2. MODEL INTERPRETATION

ROC CURVE



MODEL COMPARISON

TRAINING vs TESTING, RECALL

SVC and **SVM**

SVC testing performances were better and more consistent with training performances compared to the SVM whose training performances were way higher compared to testing.

This could be due SVM overfitting because of an overly complex decision boundaries that ended up being too performing on training data and not as performing on unseen data.

	Training	Testing
SVC_PCA	0.56	0.56
SVC_noPCA	0.57	0.64
SVM_PCA	0.76	0.54
SVM_noPCA	0.96	0.56
LGS_PCA	0.56	0.59
LGS_noPCA	0.53	0.62
KNN_PCA	0.73	0.56
KNN_noPCA	1.00	0.46

TRAINING vs TESTING, RECALL

Logistic regression

Logistic regression appeared to be, together with SVC, the most consistent in performances between training and testing, both in the PCA and noPCA results.

	Training	Testing
SVC_PCA	0.56	0.56
SVC_noPCA	0.57	0.64
SVM_PCA	0.76	0.54
SVM_noPCA	0.96	0.56
LGS_PCA	0.56	0.59
LGS_noPCA	0.53	0.62
KNN_PCA	0.73	0.56
KNN_noPCA	1.00	0.46

TRAINING vs TESTING, RECALL

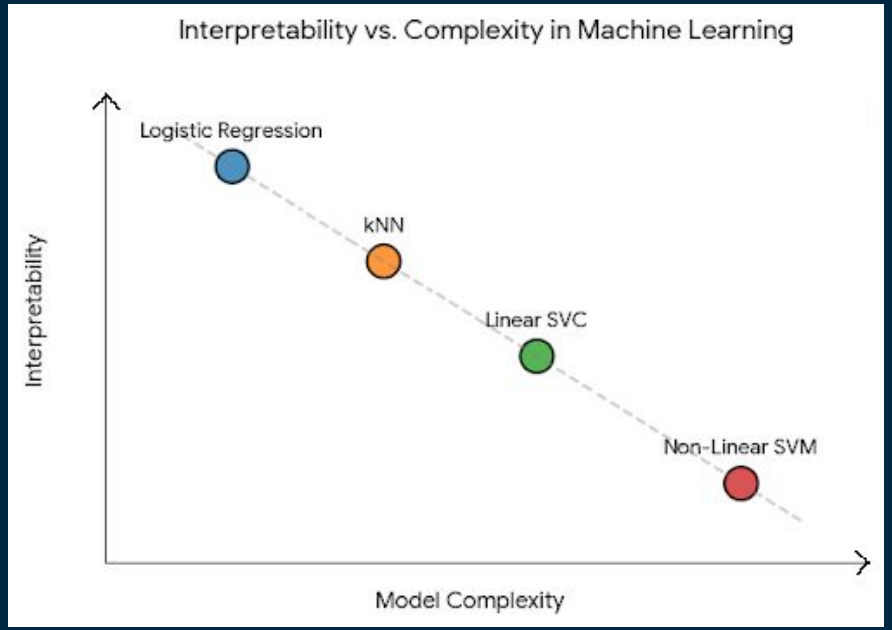
K-NN

K-nearest neighbors ended up overfitting way too much, both in the PCA and especially in the noPCA results.

In particular, the extreme overfitting showed in the noPCA example proved to be extremely impactful on testing where the Recall was even lower than 0.5

	Training	Testing
SVC_PCA	0.56	0.56
SVC_noPCA	0.57	0.64
SVM_PCA	0.76	0.54
SVM_noPCA	0.96	0.56
LGS_PCA	0.56	0.59
LGS_noPCA	0.53	0.62
KNN_PCA	0.73	0.56
KNN_noPCA	1.00	0.46

MODEL INTERPRETATION

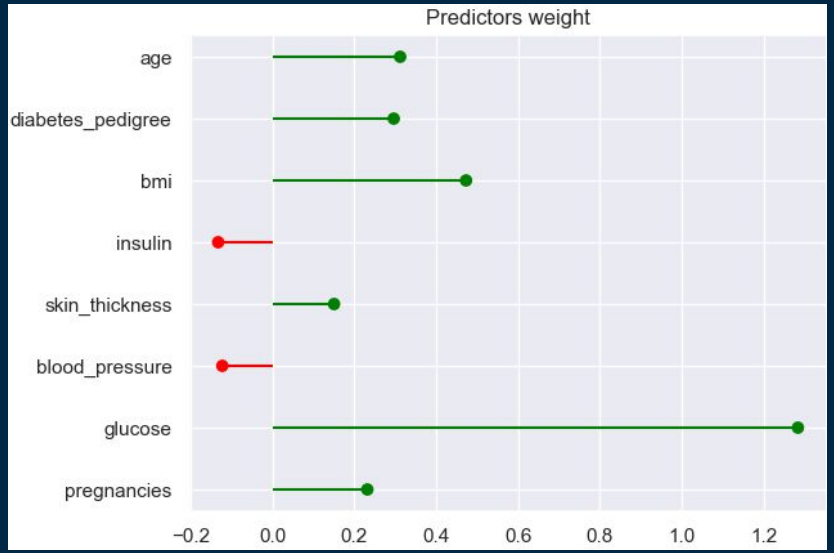


LOGISTIC REGRESSION

Logistic regression provides the highest interpretability in the set of chosen models.

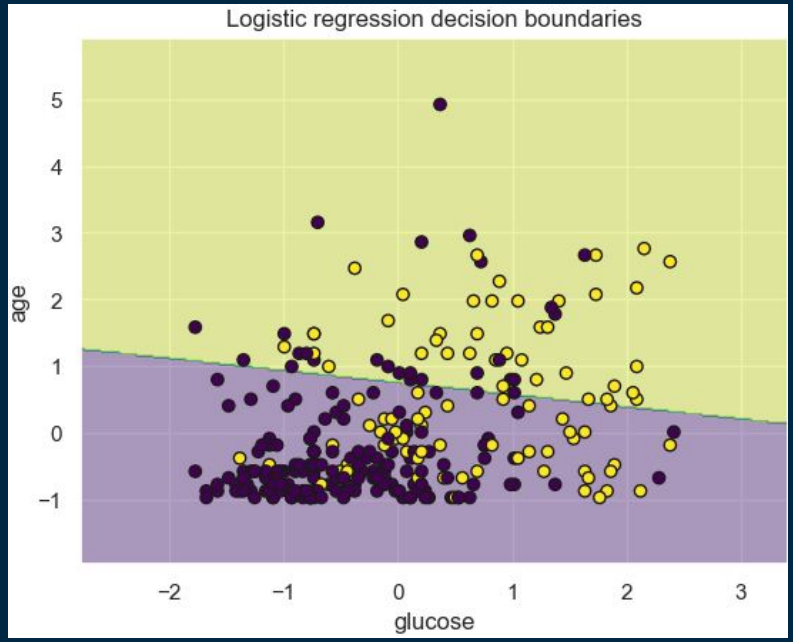
It is possible to understand the weight of each predictors through their coefficient thanks to log odds.

For example, Glucose variable weighs roughly ~3 times BMI.



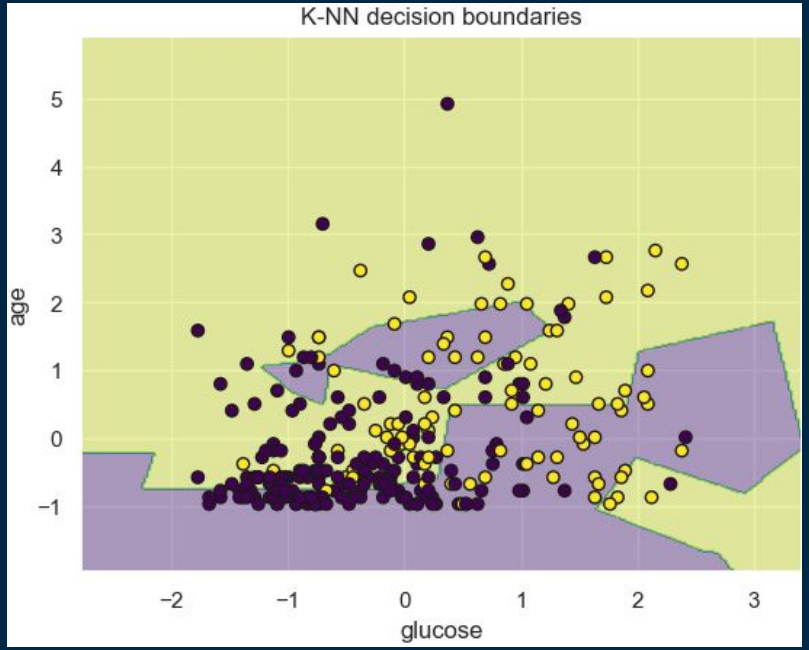
LOGISTIC REGRESSION

Logistic regression also allows for the creation of decision boundaries.



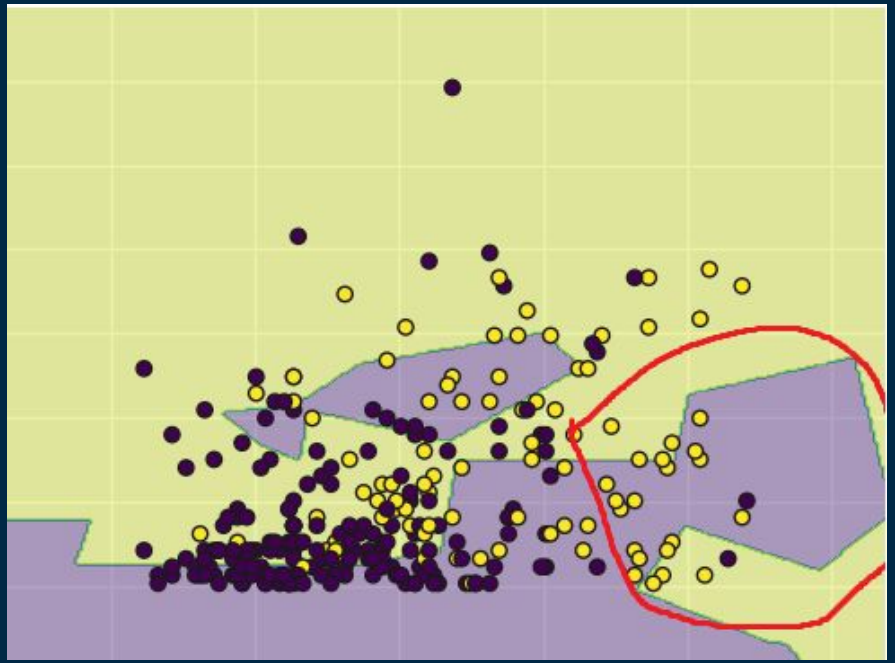
K-NN

K-NN allows for the creation of decision boundaries.



K-NN

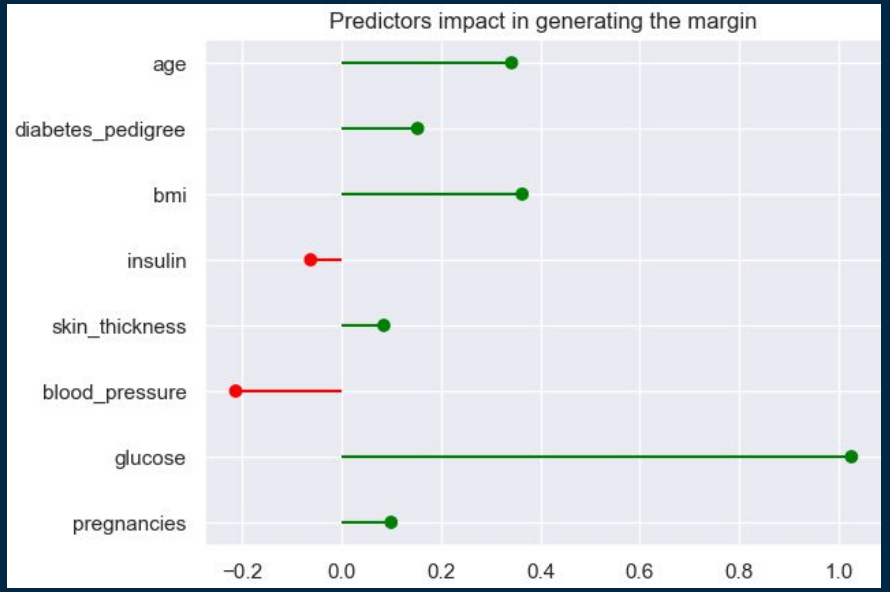
Interpretability can be decreased when data dimensionality starts to increase.



SVC AND SVM

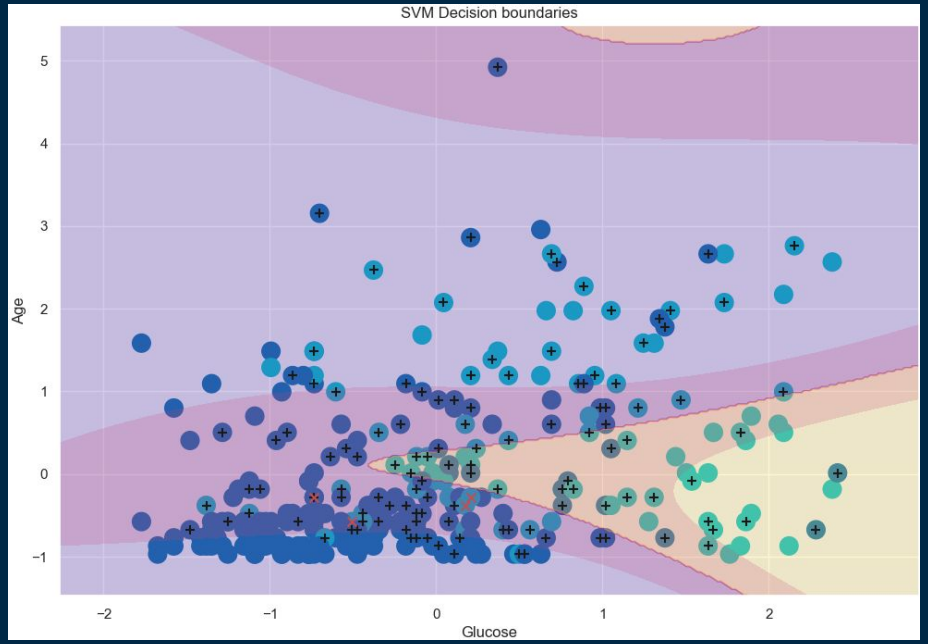
Similarly to K-NN, SVC and SVM also provides decision boundaries.

In the SVC case, thanks to libraries like scikit-learn it is also possible to have coefficients that highlights the importance of certain predictors in defining the hyperplane.



SVC AND SVM

In the SVM case, such coefficients are not available and the only interpretability is linked to decision boundaries that end up being overly complex.



CONCLUSIONS

Data cleaning:

- Unexpected 0 values

Data exploration:

- Glucose in Pairplot. Glucose and Age association with Outcome through eta squared coeff.

Unsupervised analysis:

- Glucose, BMI, Age weight towards the response variable.

Supervised analysis:

- SVC and Logistic regression Glucose, BMI and Age weights.

Further improvements:

- Data size and data imbalance.