· U C ·
FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Pattern Recognition
2021-2022

dei25 departamento
de engenharia informática
1995 – 2020

Heart Classification Project

Beatriz Negromonte
Rafael Correia Molter
Yandra Vinturini Vieira Dantas

Subset of Dataset

- Dataset contains 318,958 entries.
- Changed labels 1.0 and 2.0 to 0 (negative) and 1 (positive), for better understanding and reading.
- Classification column was created, it takes the value of CoronaryHeartDisease.
- Classification column is changed for scenario C: has value 0 for healthy patients, value 1 for patients with heart conditions but no comorbidities and value 2 for patients with heart conditions and comorbidities. Comorbidities are considered either skin cancer or kidney disease diagnostic.

KS test

```
_BMI5 p-value:  0.0
Smoking p-value:  1.2131434371817858e-23
AlcoholDrinking p-value:  1.2131434371817858e-23
Stroke p-value:  1.2131434371817858e-23
PhysicalHealth p-value:  1.2131434371817858e-23
MentalHealth p-value:  1.2131434371817858e-23
DiffWalking p-value:  1.2131434371817858e-23
Sex p-value:  1.2131434371817858e-23
AgeCategory p-value:  1.1548262100906754e-120
Race p-value:  4.764860426058126e-80
Diabetic p-value:  3.603278870102406e-82
PhysicalActivity p-value:  2.0677757222974026e-31
GenHealth p-value:  4.764860426058126e-80
SleepTime p-value:  2.743841982214968e-200
Asthma p-value:  1.2131434371817858e-23
```

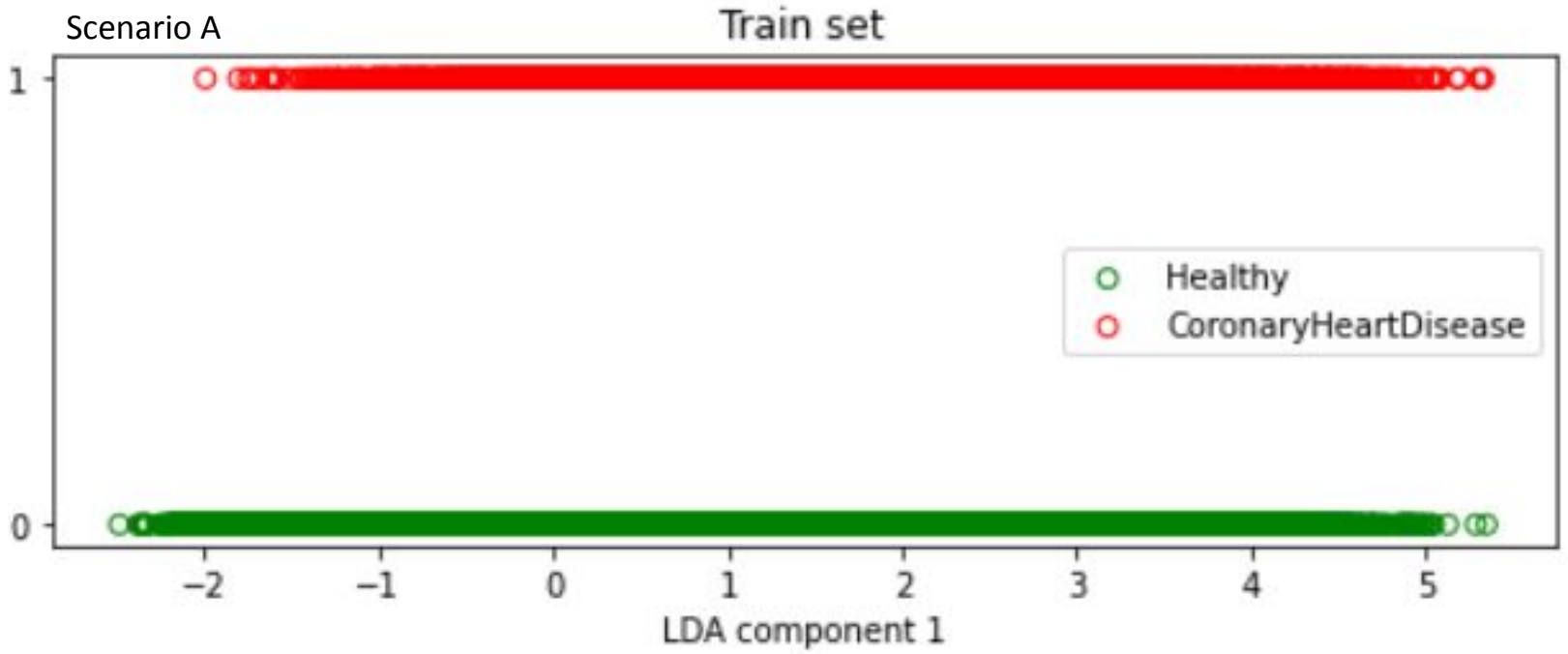
Feature selection

Kruskal Wallis test

	A	B	C
1º	GenHealth	BMI5	GenHealth
2º	AgeCategory	AgeCategory	AgeCategory
3º	DiffWalking	Stroke	DiffWalking
4º	Diabetic	SleepTime	Stroke
5º	PhysicalHealth	GenHealth	Diabetic
6º	Stroke	PhysicalHealth	PhysicalHealth
7º	BMI5	Smoking	BMI5
8º	Smoking	MentalHealth	Smoking
9º	PhysicalActivity	Race	PhysicalActivity
10º	SleepTime	Sex	SleepTime
11º	Sex	DiffWalking	Sex
12º	Race	Diabetic	MentalHealth
13º	MentalHealth	Asthma	Race
14º	Asthma	PhysicalActivity	Asthma
15º	AlcoholDrinking	AlcoholDrinking	AlcoholDrinking

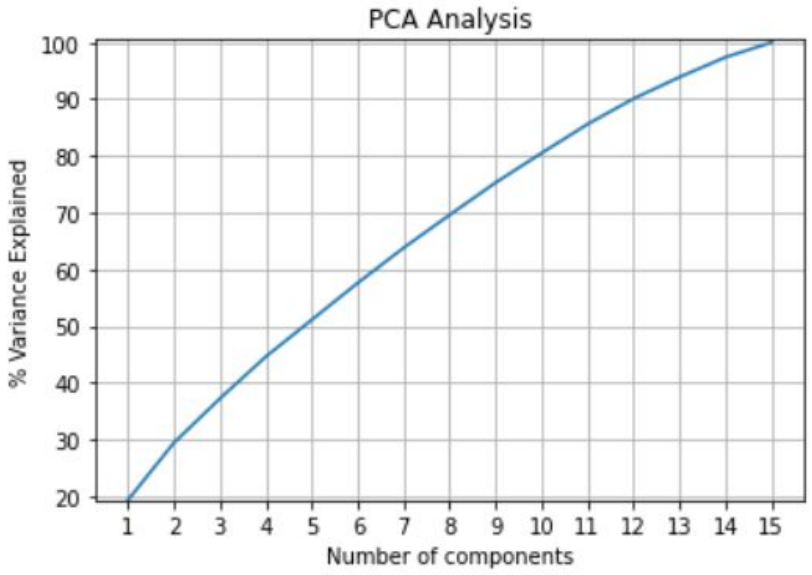
Dimensionality reduction - LDA

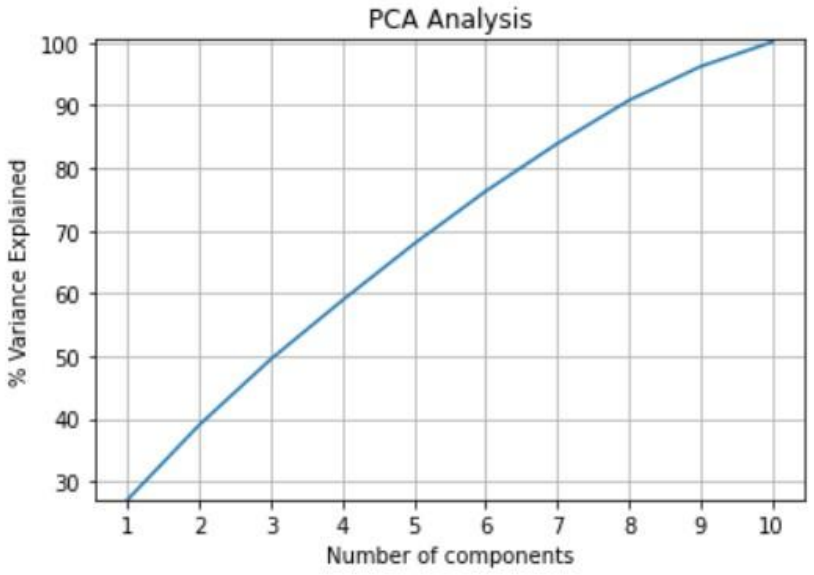
Scenario A

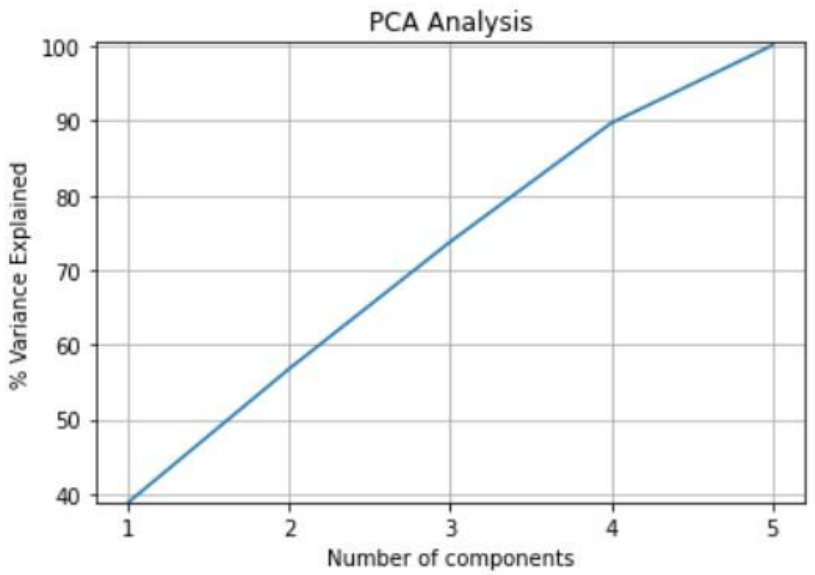


Dimensionality reduction - PCA

Variance explained by the components



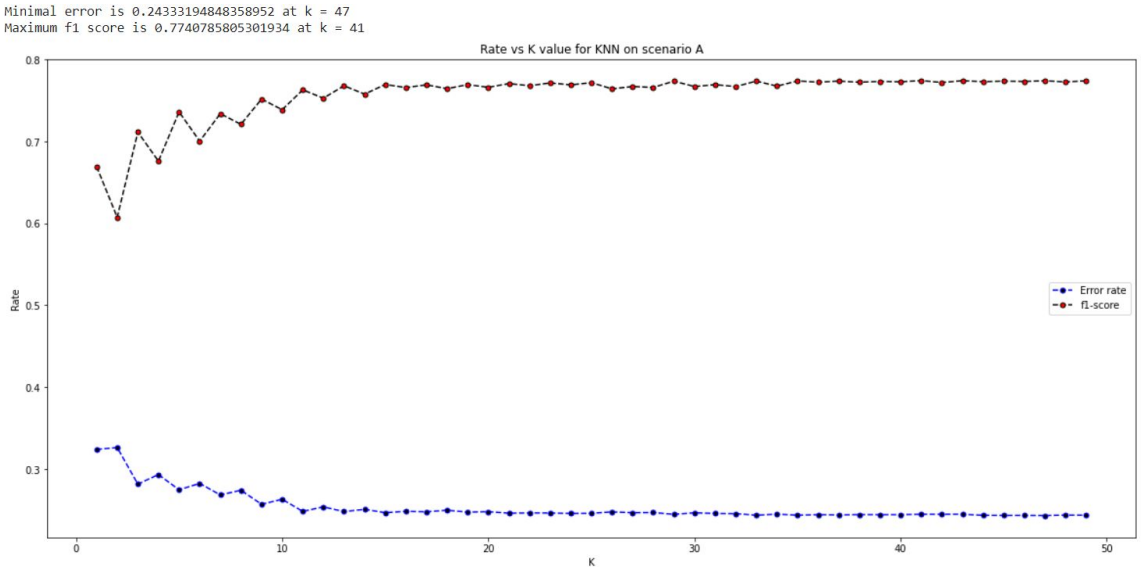




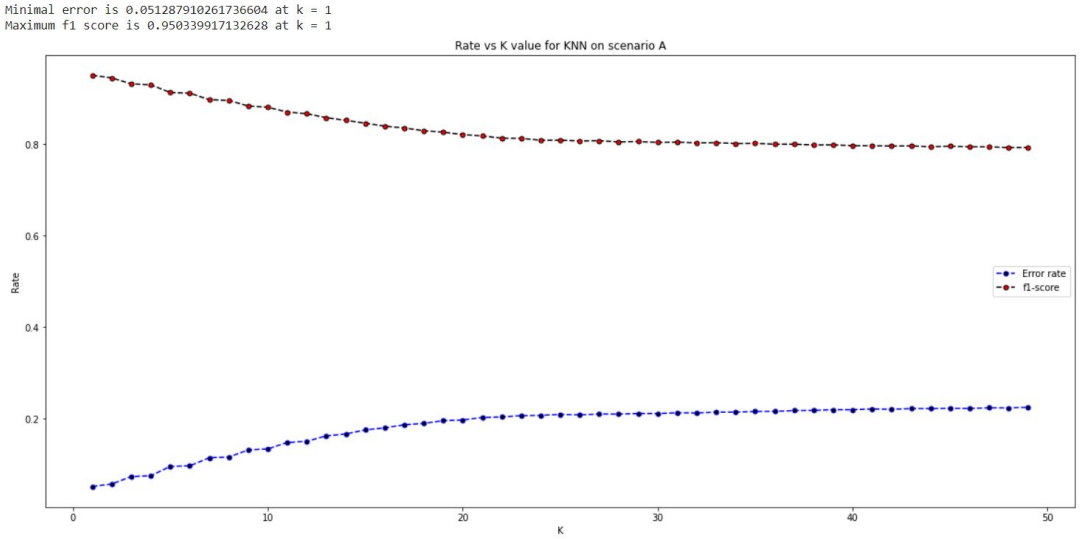
PCA for Scenario C

KNN classifier

Choice of best value of K



5 features and 4 components

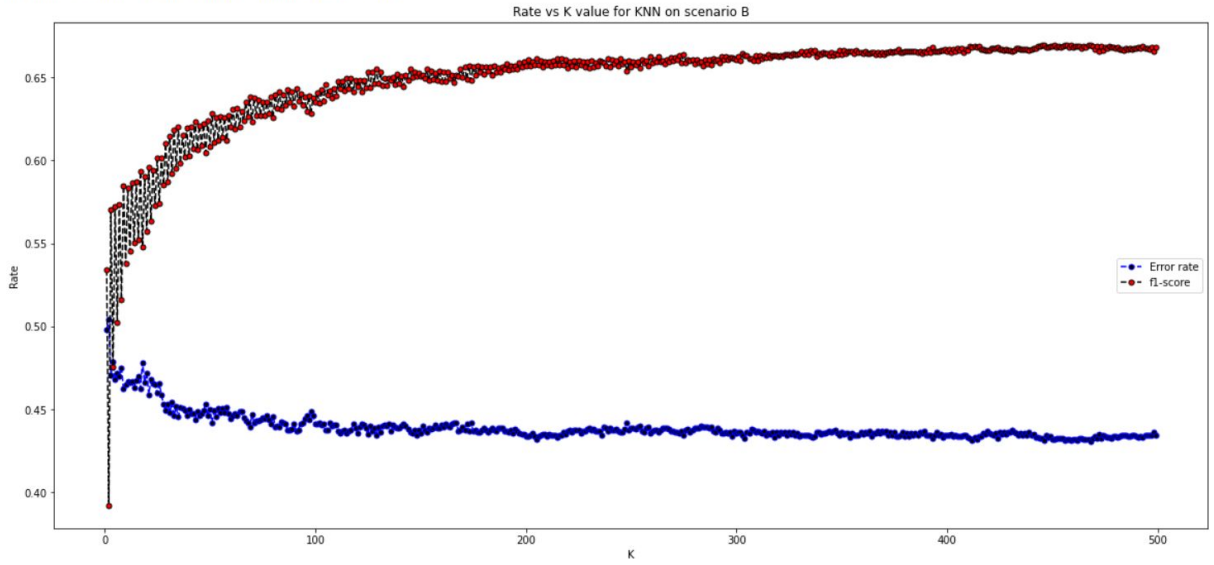


10 features and 8 components

KNN classifier

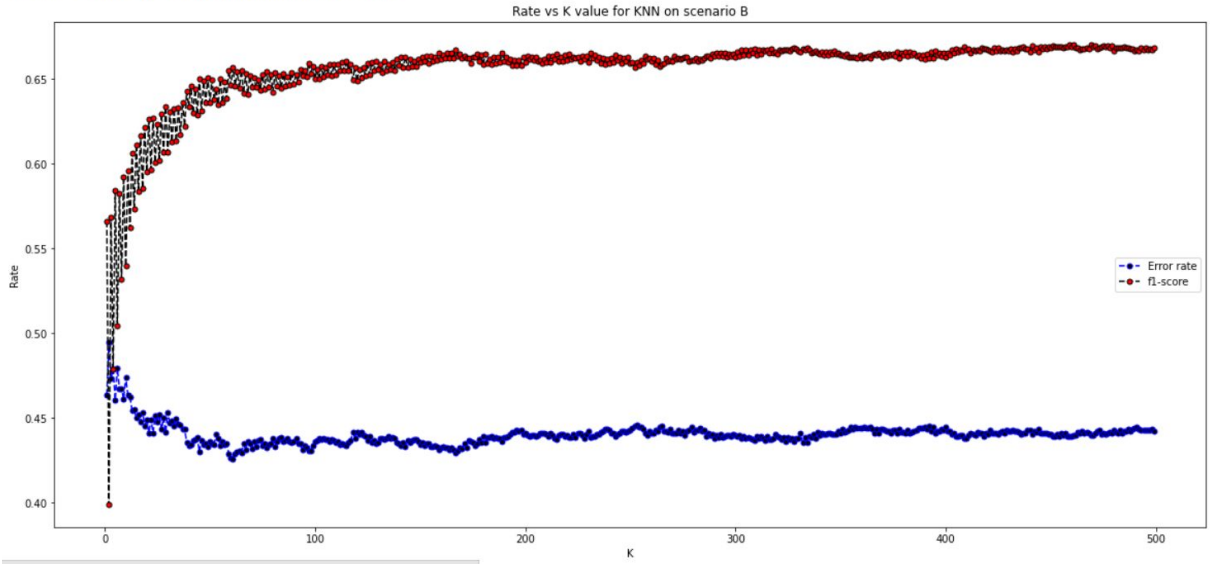
Choice of best value of K

Minimal error is 0.4305019305019305 at k = 467
Maximum f1 score is 0.6694861492916051 at k = 445



5 features and 4 components

Minimal error is 0.4258135686707115 at k = 60
Maximum f1 score is 0.6702523789822094 at k = 460

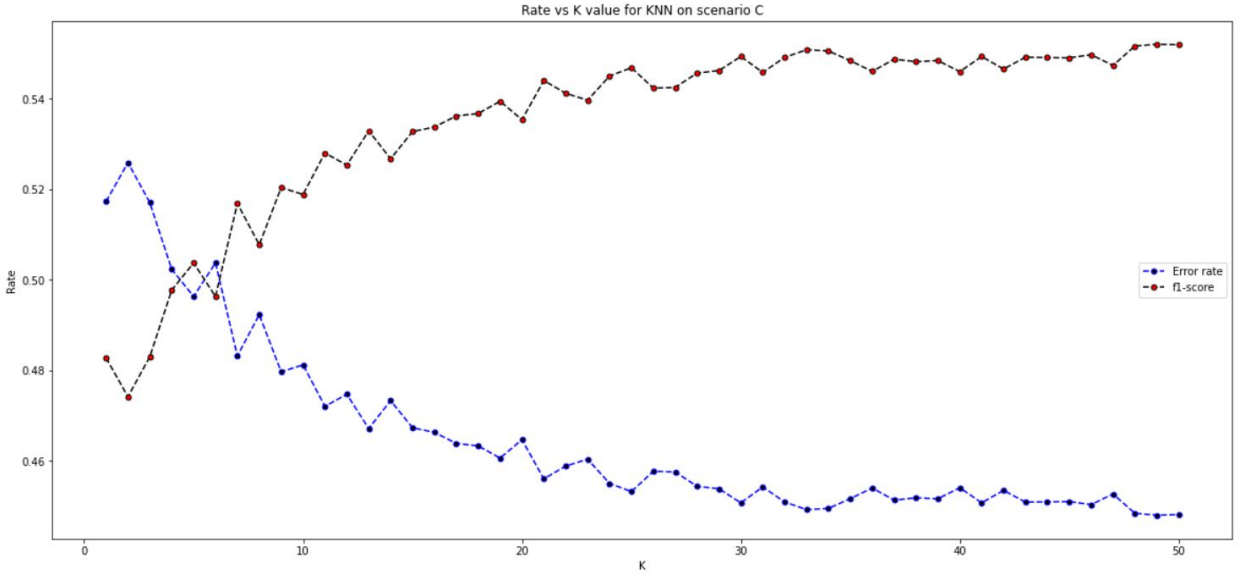


10 features and 8 components

KNN classifier

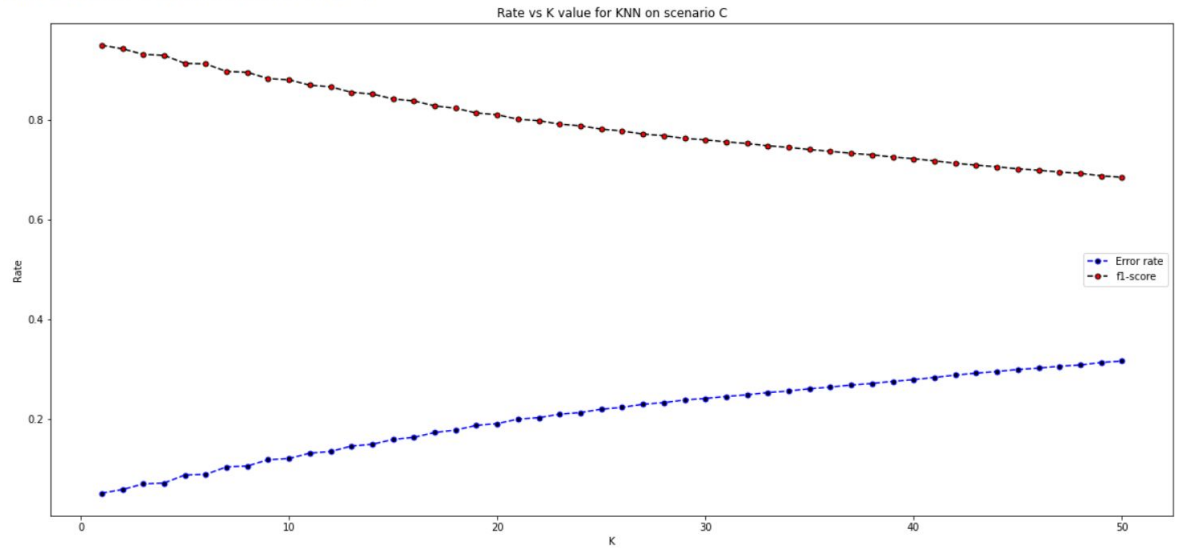
Choice of best value of K

Minimal error is 0.4478886886145209 at k = 49
Maximum f1 score is 0.5521113113854791 at k = 49



5 features and 4 components

Minimal error is 0.05066862349764539 at k = 1
Maximum f1 score is 0.9493313765023547 at k = 1



10 features and 8 components

Results - Scenario A

Table 1: Metrics for scenario A with 5 features and 4 components (k = 47)

Metric	Fisher LDA	Euclidean	Mahalanobis	NaiveBayes	KNN	SVM
Accuracy	74.98%	73.23%	75.00%	73.09%	75.67%	75.17%
Sensitivity	78.93%	69.10%	78.71%	70.69%	83.29%	77.76%
Specificity	71.03%	77.36%	71.28%	75.48%	68.05%	72.59%
Precision	73.15%	75.32%	73.27%	74.25%	72.27%	73.94%
F1 score	0.76	0.72	0.76	0.67	0.77	0.76

Table 2: Metrics for scenario A with 10 features and 8 components KNN

Metric	k=1	k=3	k=8	k=15
Accuracy	94.87%	92.70%	88.45%	82.50
Sensitivity	98.15%	99.29%	98.82%	95.95%
Specificity	91.59%	86.11%	78.09%	69.04%
Precision	92.11%	87.73%	81.85%	75.61%
F1 score	0.95	0.93	0.90	0.85

Table 3: Metrics for scenario A with 10 features and 8 components (k = 1)

Metric	Fisher LDA	Euclidean	Mahalanobis	NaiveBayes	KNN	SVM
Accuracy	74.75%	72.58%	74.77%	71.30%	94.87%	74.72%
Sensitivity	76.47%	66.94%	76.42%	65.38%	98.15%	76.21%
Specificity	73.03%	78.21%	73.12%	77.22%	91.59%	73.23%
Precision	73.93%	75.45%	73.98%	74.16%	92.11%	74.01%
F1 score	0.75	0.71	0.75	0.69	0.95	0.75

Results - Scenario B

Table 4: Metrics for scenario B with 5 features and 4 components (k = 467)

Metric	Fisher LDA	Euclidean	Mahalanobis	NaiveBayes	KNN	SVM
Accuracy	56.45%	56.34%	56.51%	57.09%	56.84%	56.67%
Sensitivity	78.86%	68.37%	66.36%	80.10%	81.76%	80.67%
Specificity	30.81%	42.58%	45.24%	30.75%	28.33%	29.21%
Precision	56.60%	57.67%	58.10%	56.96%	56.62%	56.60%
F1 score	0.66	0.63	0.62	0.67	0.67	0.67

Table 5: Metrics for scenario B with 10 features and 8 components(k=460)

Metric	Fisher LDA	Euclidean	Mahalanobis	NaiveBayes	KNN	SVM
Accuracy	56.87%	56.62%	57.17%	56.51%	56.07%	56.84%
Sensitivity	75.76%	58.81%	54.58%	77.26%	83.15%	75.76%
Specificity	35.25%	54.11%	54.58%	32.76%	27.07%	35.19%
Precision	57.24%	59.46%	59.96%	56.80%	55.95%	57.22%
F1 score	0.65	0.59	0.60	0.65	0.67	0.65

Results - Scenario C

Table 6: Metrics for scenario C with 5 features and 4 components (k = 49)

Metric	NaiveBayes	KNN	SVM
Accuracy	52.10%	55.21%	54.01%
Sensitivity weighted	52.10%	55.21%	54.01%
Precision weighted	50.03%	55.36%	52.48%
F1 score weighted	0.49	0.55	0.46

Table 7: Metrics for scenario C with 10 features and 8 components (k = 1)

Metric	NaiveBayes	KNN	SVM
Accuracy	52.49%	94.93%	54.54%
Sensitivity weighted	52.49%	94.93%	54.54%
Precision weighted	51.03%	95.07%	52.74%
F1 score weighted	0.51	0.95	0.50

Discussions

- Scenario A and B: less features improved performance (except for KNN in scenario B)
- Scenario C: more features improved performance
- NaiveBayes is not adequate: features don’t assume any particular distribution

Conclusions

- Scenario A - Good distinction between positive and negative cases, with F1 score above 0.71 for each classifier. Best option: 10 best features and 8 components for PCA, using KNN classifier
- Scenario B - Poor distinction between the two heart conditions on the sample, the algorithm classifies more than half of Myocardial Infarction cases as Coronary Heart Diseases, in all predictors.
- Scenario C - Worst scenario, the only classifier able to separate classes well was KNN

Thank you for your attention!