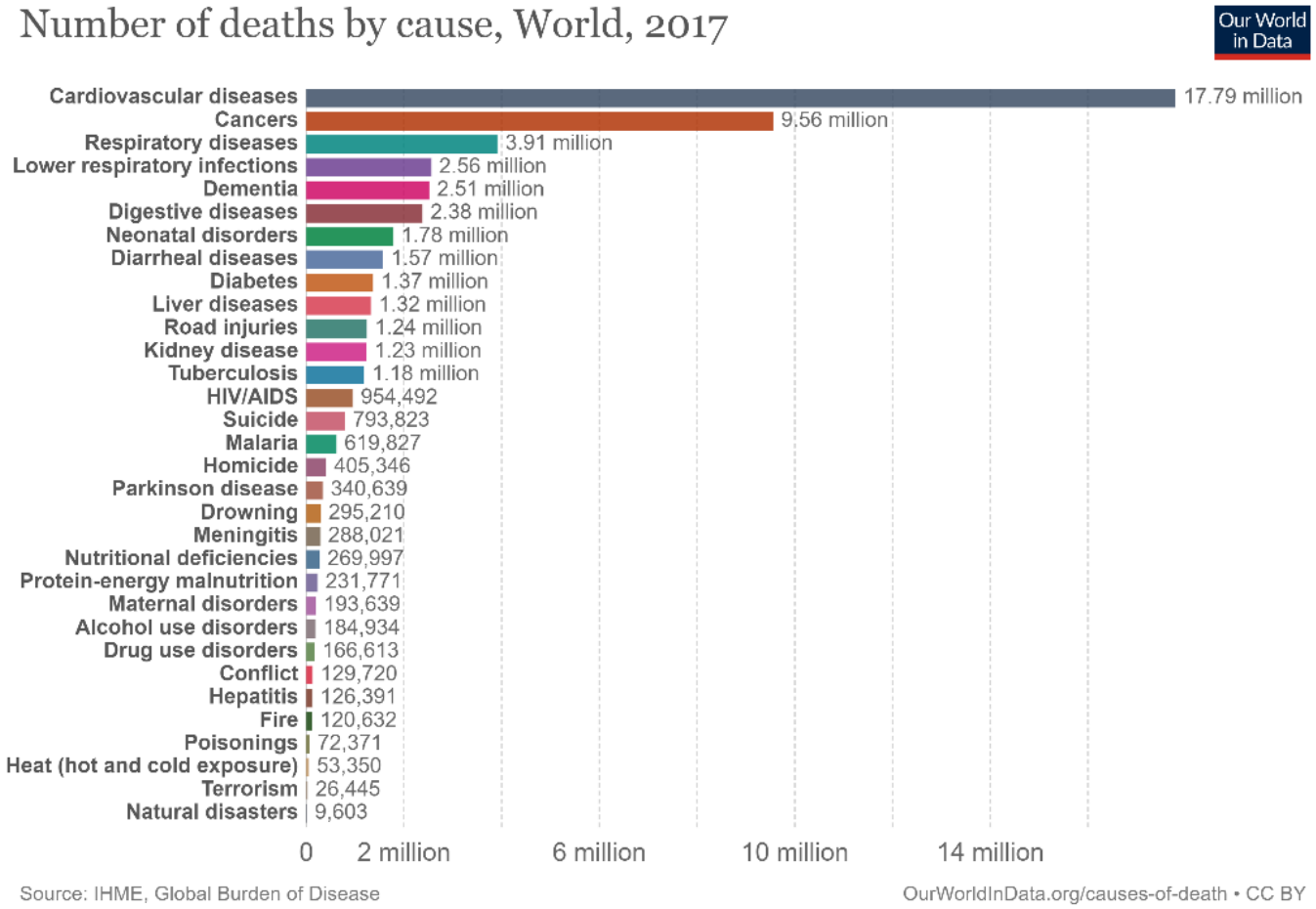
PREDICTING DEATH EVENTS FROM HEART FAILURES



1. BUSINESS PROBLEM

- HEART DISEASE MAJOR CAUSE OF DEATH.
- TECHNOLOGICAL IMPROVEMENTS AVAILABLE SUCH AS ML.
- ML IMPROVES EFFICIENCY AS IT PROVIDES BETTER MANAGEMENT OF RESOURCES.

2. DATA ACQUISITION AND CLEANSING

KAGGLE.COM USED AS DIRECT SOURCE

-Data was indirectly collected from UCI (University of California, Irvine) repository

-Dataset was used before for other Machine Learning models

-No missing values were found

-Total datapoints: 300 patients

-Different features as: Gender, Smoker or not, Blood pressure, levels of serum sodium…etc

VARIABLES

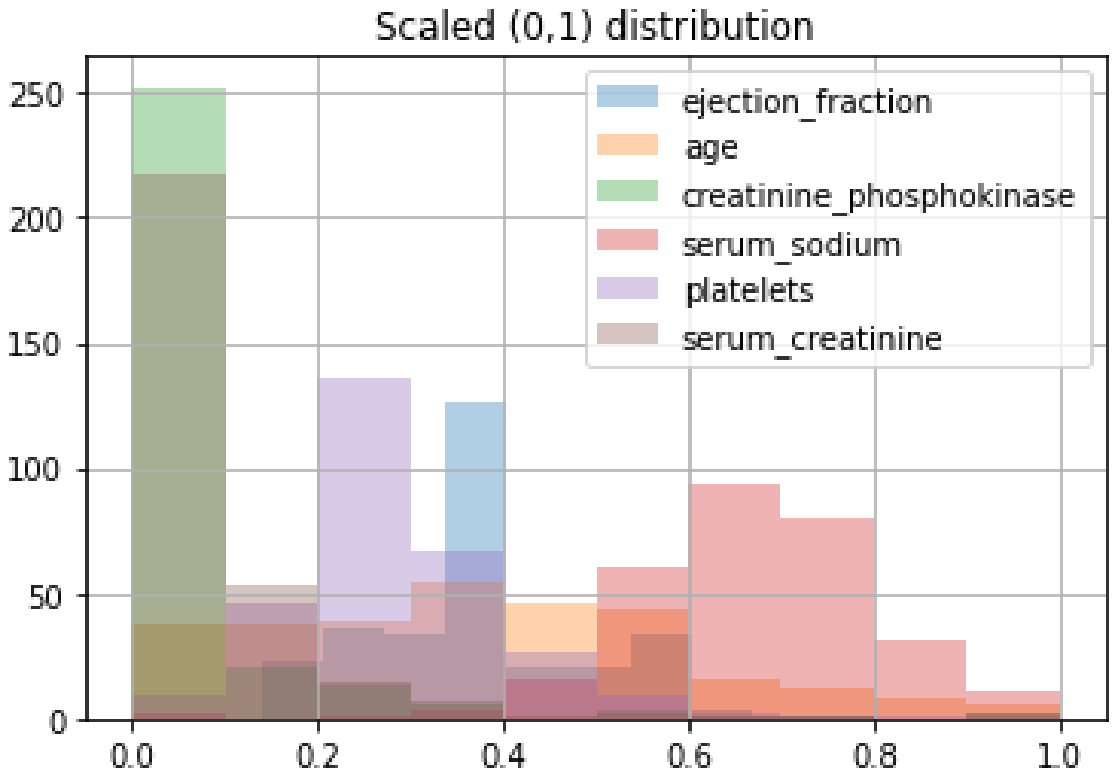
FEATURES AND TARGET VARIABLE

Features	Target variable
Sex/gender	Death event (Dummy variable)
Person is a smoker (dummy/boolean)	
High blood pressure or Hypertension (dummy/boolean)	
Creatinine phosphokinase (CPK) levels	
Ejection fraction (percentage of blood leaving the heart at each contraction)	
Serum creatinine levels	
Platelets in the blood	
Serum sodium levels	
Diabetes (dummy/boolean)	
Anaemia (dummy(boolean)	
Age	

SCALING

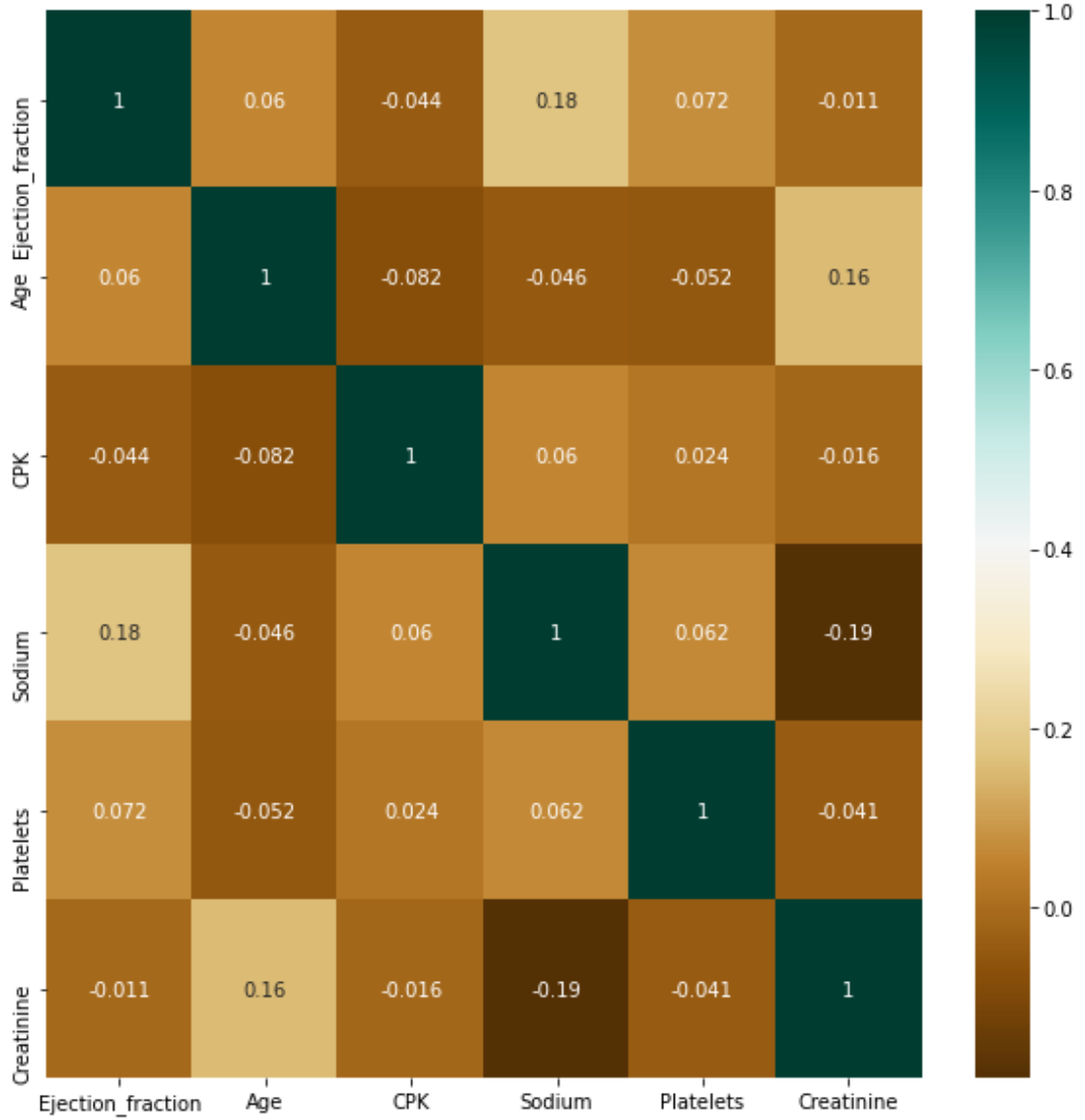
APPLICATION OF SCALING METHOD IN ORDER TO APPLY CLASSIFICATION ALGORITHMS

$$X' = \frac{X - X_{MIN}}{X_{MAX} - X_{MIN}}$$



CORRELATION HEATMAP

NO MULTICOLLINEARITY FOUND BETWEEN CONTINUOUS VARIABLES



3. EXPLORATORY DATA ANALYSIS

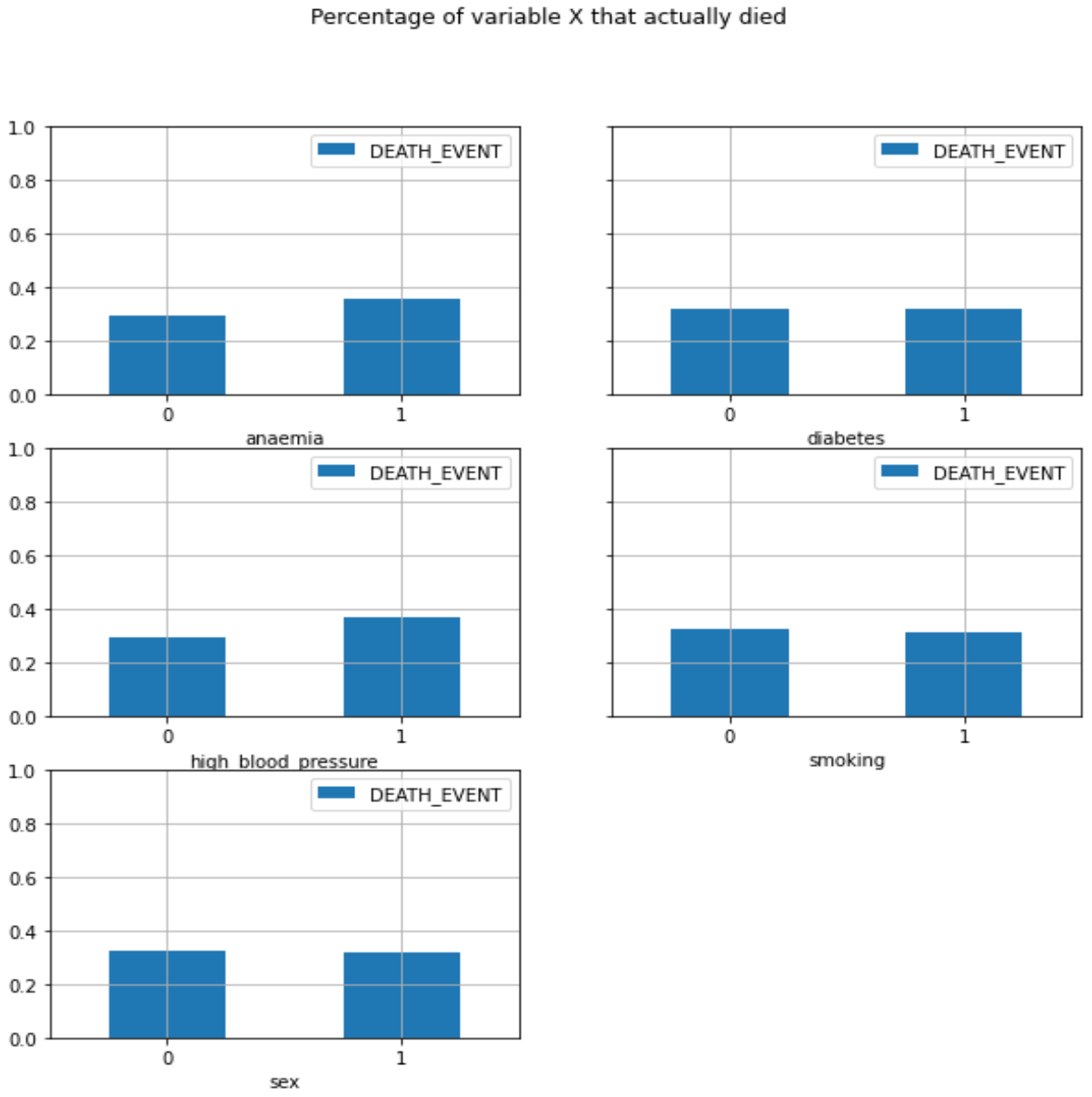
-Relationship between categorical variables and target variable.

-Relationship between continuous variables and target variable.

-Proposed Machine Learning models.

CATEGORICAL VARIABLES AND TARGET VARIABLE

MEAN OF DEATH EVENTS CONDITIONING BY{0,1}



CATEGORICAL VARIABLES AND TARGET VARIABLE

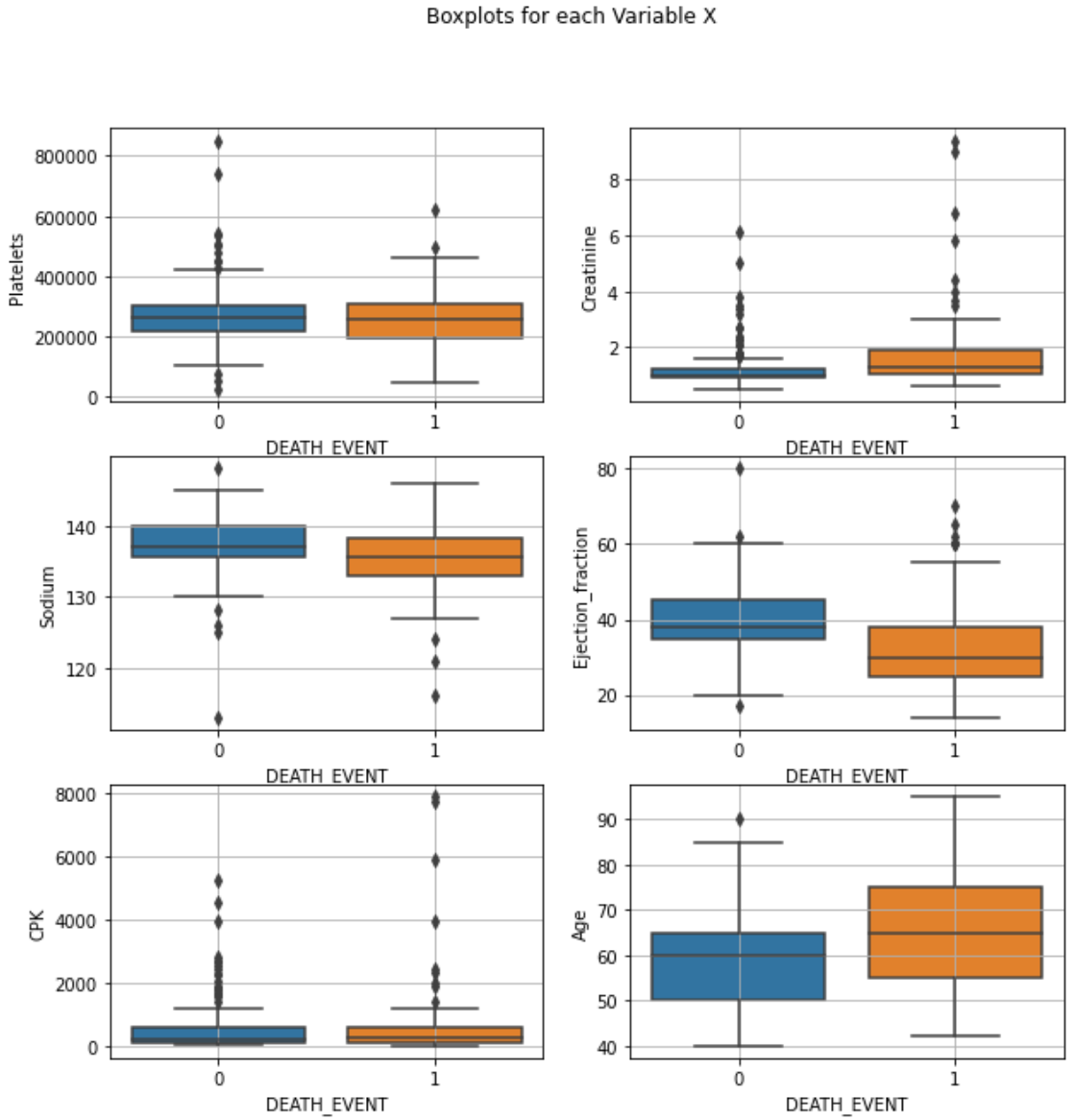
ANOVA TEST (INFERENCE TEST) FOR CHECKING SIGNIFICANT DIFFERENCES BETWEEN THE MEANS.

(SEE REPORT IN ORDER TO KNOW WHY THIS RESULTS)

Variable	P-Value
Anaemia	0.25
Diabetes	0.97
Blood pressure (Hypertension)	0.17
Smoking	0.82
Sex	0.94

CONTINUOUS VARIABLES AND TARGET VARIABLE

BOXPLOTS FOR EACH ‘DEATH EVENT’ VALUE



KEY IDEAS

KEY IDEAS AND PROPOSED MODELS

-No categorical variable seem to be relevant for explaining 'Death event'. Blood pressure could be used as it's the variable with least p-value (0.17)

-Continuous variable are core variables in order to explain differences: Creatinine levels, blood fraction ejection, age and blood pressure.

-Proposed models: Logistic Regression, KNN, SVM.

4.MODELING AND RESULTS

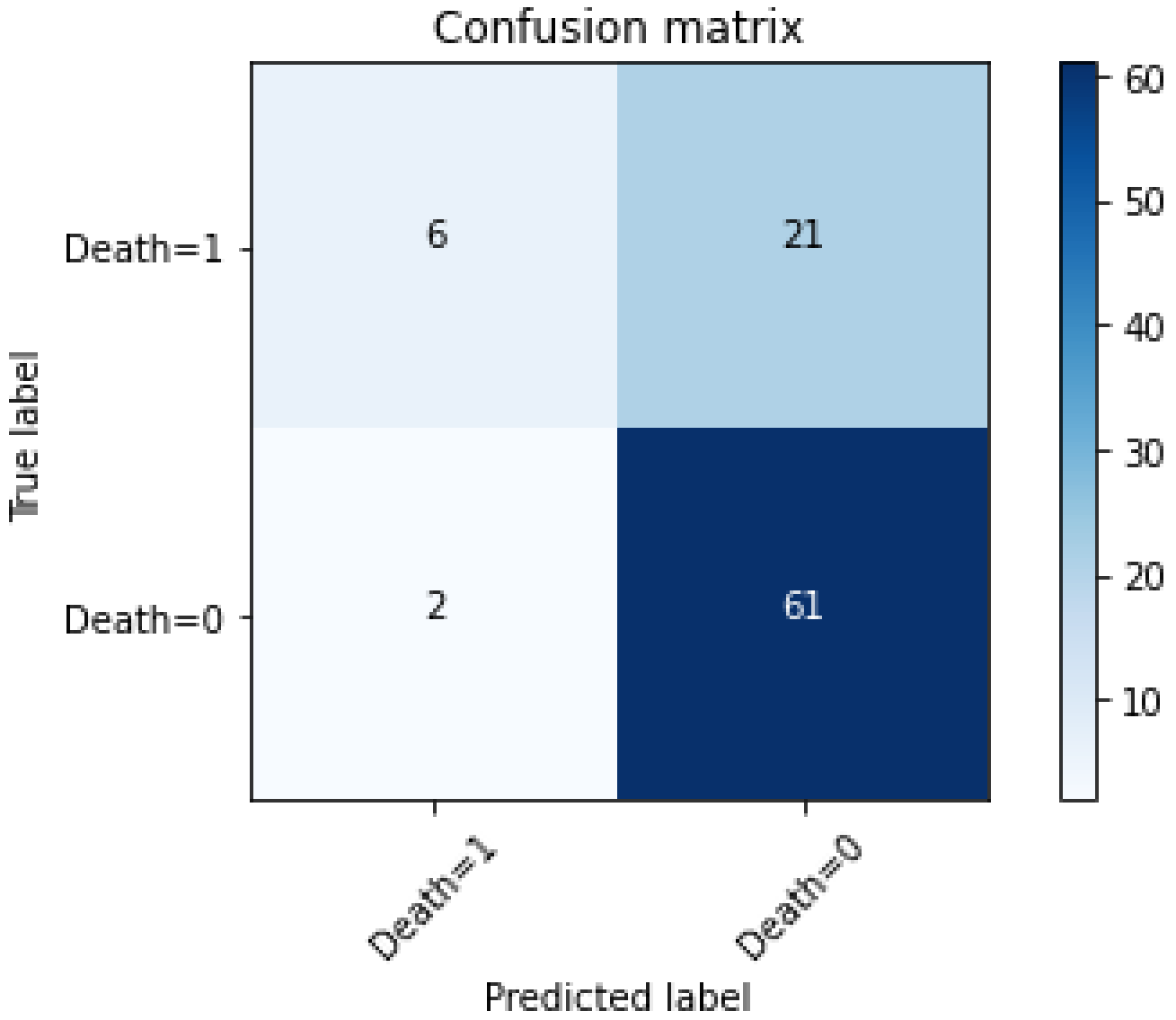
LOGISTIC REGRESSION

KNN ALGORITHM

SUPPORT VECTOR MACHINE

LOGISTIC REGRESSION

CONFUSION MATRIX



Variable	Precision	Recall	F1-score	Support
0	0.74	0.97	0.84	63
1	0.75	0.22	0.43	27
Accuracy			0.74	90
Macro avg	0.75	0.60	0.59	90
Weighted avg	0.75	0.74	0.69	90

MAIN METRICS

GOOD PREDICTION WITH LABEL=0, BAD PREDICTION WITH LABEL=1

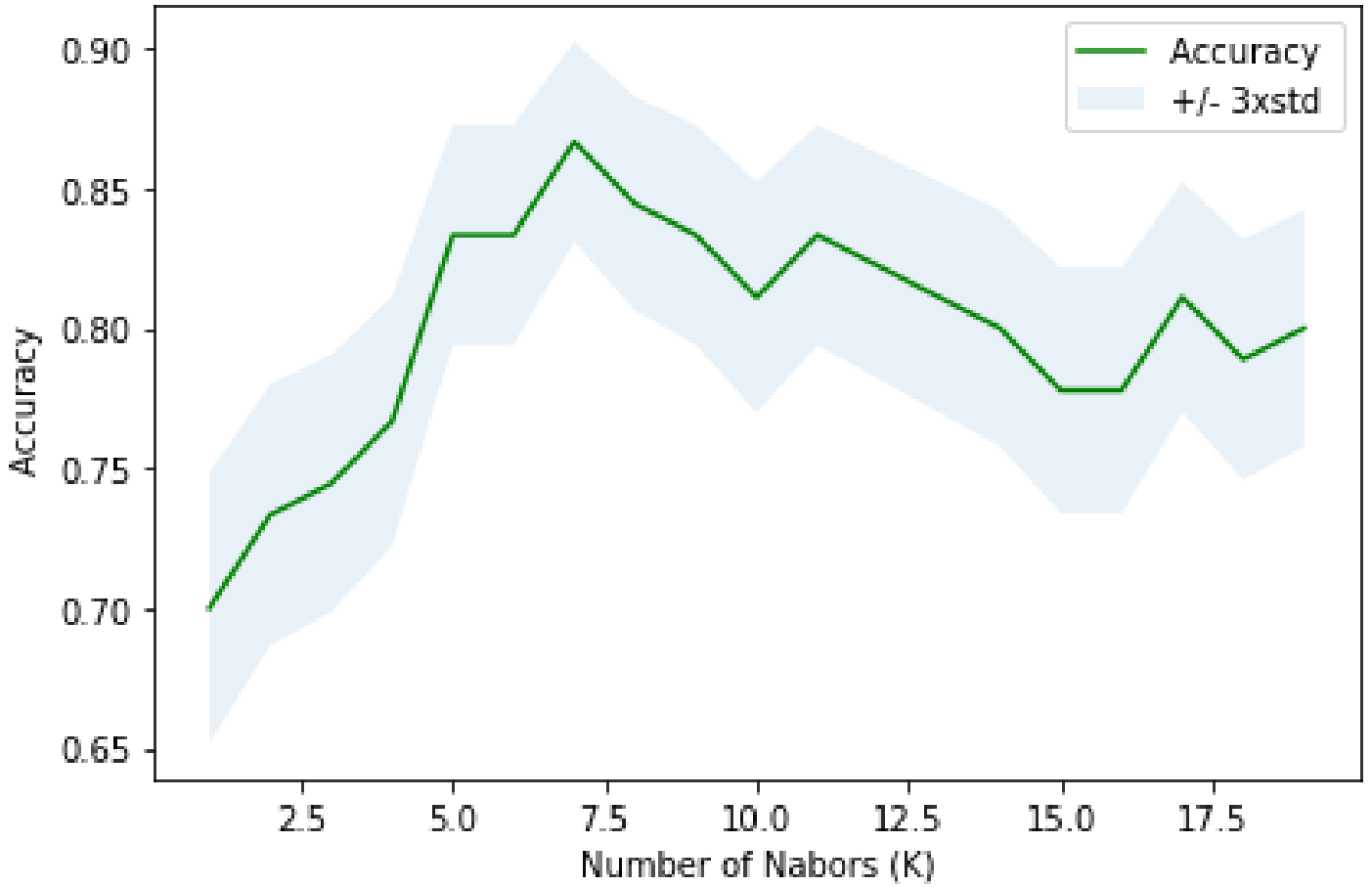
LogLoss model above	Jaccard-index model above	Mean of CrossValidation (cv=10)
0.53	0.74	0.72

AVOIDING OVERFITTING

SIMILAR JACCARD-INDEX WITH MEAN OF CROSS VALIDATION MODELS

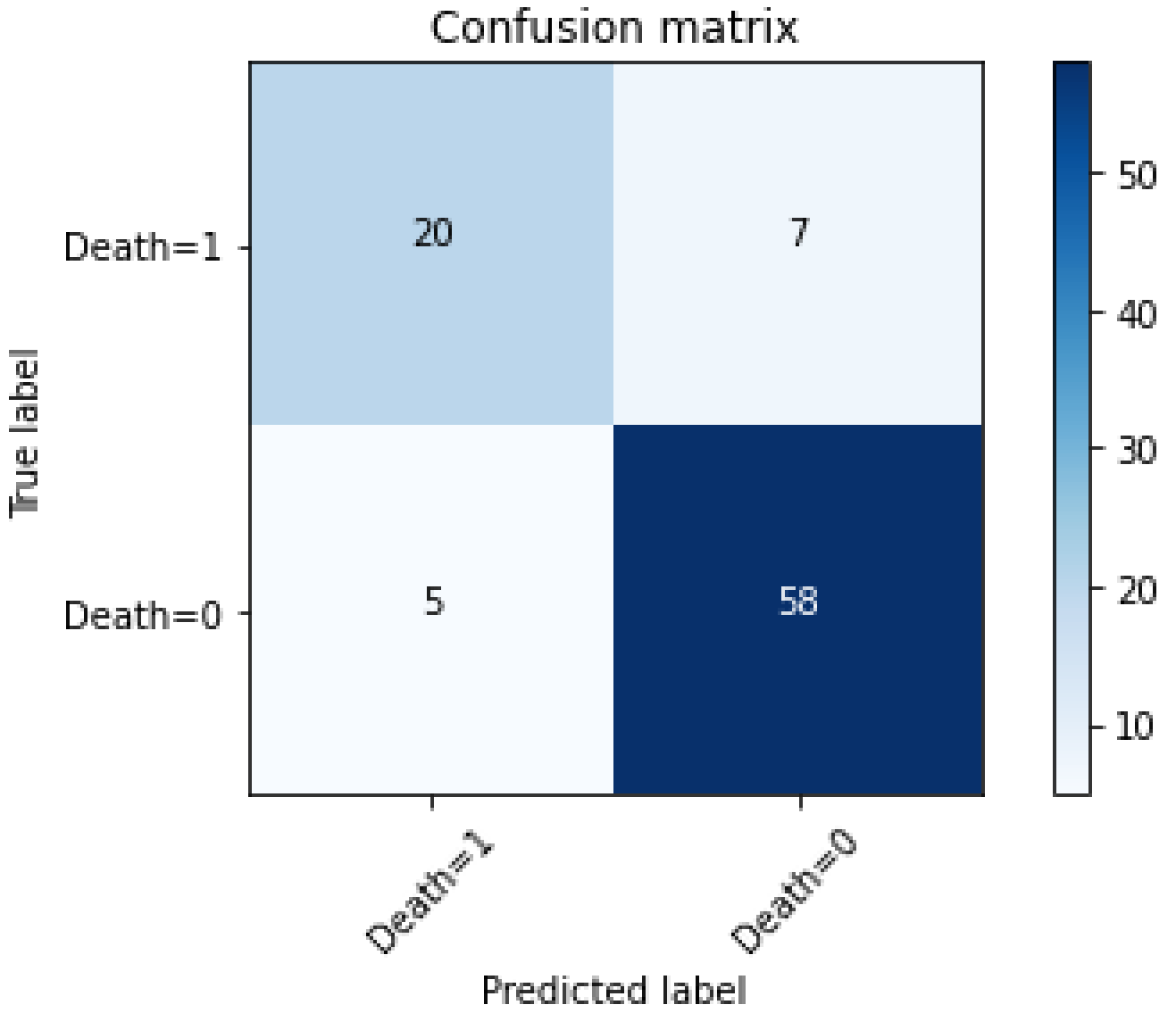
KNN ALGORITHM

ESTIMATING BEST K PARAMETER



KNN ALGORITHM

CONFUSION MATRIX



Variable	Precision	Recall	F1-score	Support
0	0.89	0.92	0.91	63
1	0.8	0.74	0.77	27
Accuracy			0.87	90
Macro avg	0.85	0.83	0.84	90
Weighted avg	0.86	0.87	0.87	90

MAIN METRICS

GOOD METRICS OVERALL.

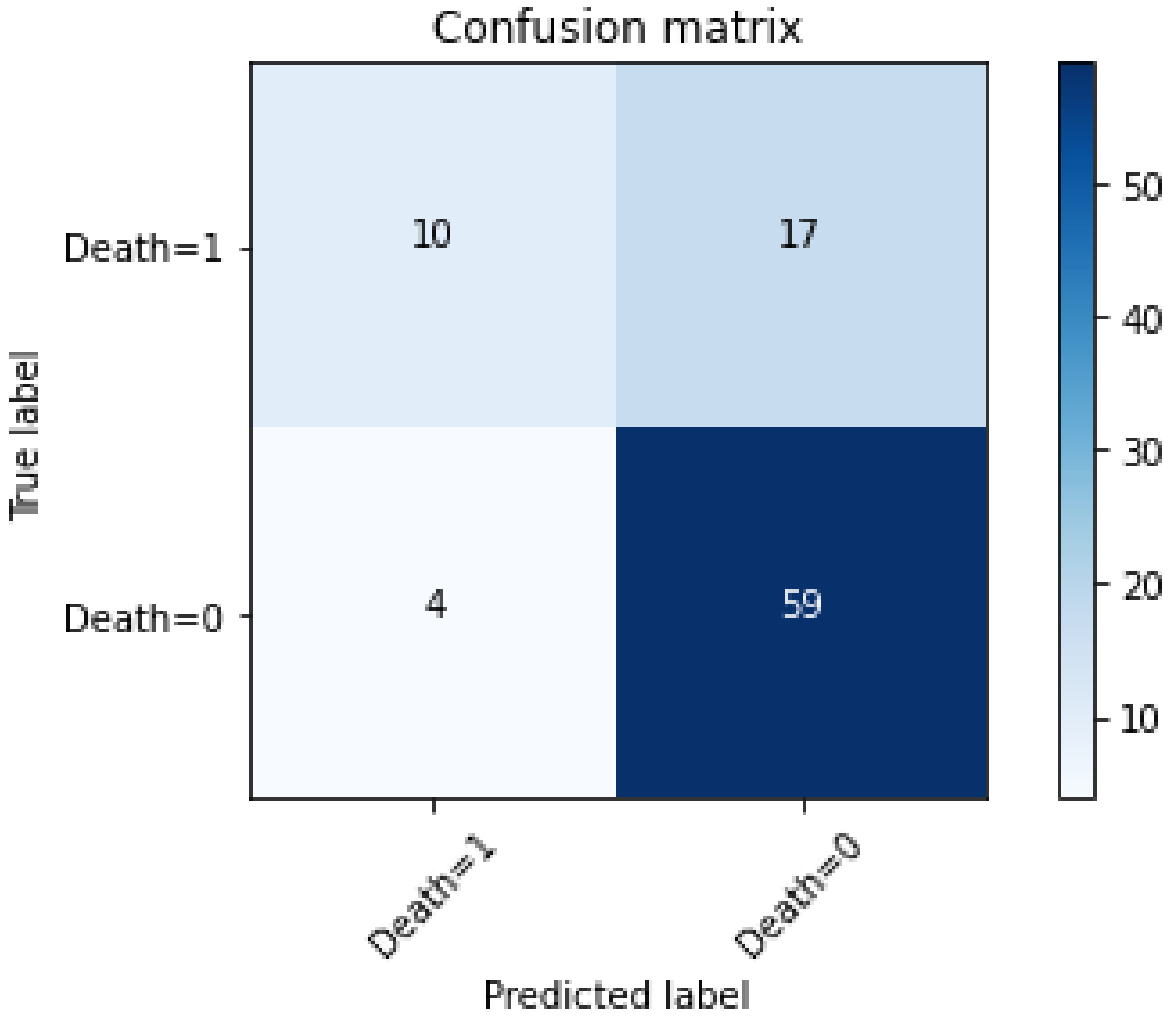
Jaccard-index model above	Mean of CrossValidation (cv=10)
0.87	0.77

AVOIDING OVERFITTING

MODEL DOESN'T SEEM TO BE OVERFITTED.

SUPPORT VECTOR MACHINE

CONFUSION MATRIX



Variable	Precision	Recall	F1-score	Support
0	0.78	0.94	0.83	63
1	0.71	0.37	0.49	27
Accuracy			0.77	90
Macro avg	0.75	0.65	0.67	90
Weighted avg	0.76	0.77	0.74	90

MAIN METRICS

SAME AS WITH LOGISTIC REGRESSION MODEL BUT SLIGHTLY BETTER PERFORMANCE.

Jaccard-index model above	Mean of CrossValidation (cv=10)
0.77	0.73

AVOIDING OVERFITTING

MODEL DOESN'T SEEM TO BE OVERFITTED.

5.DISCUSSION

ABOUT MODELS AND HANDICAPS

Ranking models
1. KNN
2. SUPPORT VECTOR MACHINE
3. LOGISTIC REGRESSION

The best model which is perfectly equilibrated is the KNN model. SVM model and LR show some bad prediction accuracy with label=1

6.CONCLUSION

NEXT STUDIES RELATED TO THIS SUBJECT

-Better Database (larger).

-More technical features.

-Avoid noisy features (smoking can cause heart attack, but doesn't seem to be a key factor in order to predict if the patient dies or not).