University of Prince Edward Island

Assignment 2: Paper on Applied Machine Learning

Luke Schipper
CS4120: Machine Learning
Dr. Antonio Bolufe-Rohler
19 March 2021

Abstract

This paper applies classical classification algorithms to a dataset of 299 heart failure patients. Resulting models predict whether such patients are at risk of death from heart failure given eleven distinct features. Models are selected by scoring 4-fold cross-validation; in one phase, F1 is used as the scoring metric, and in the other, accuracy is used. The classifiers used in this paper are logistic regression, logistic regression with polynomial features, and decision trees. In order to improve the performance of these models on an imbalanced dataset, parameter-tuning and feature reduction is performed. The final performance of these models is then compared to the research of Chicco & Jurman (2020).

The best-performing model from cross-validation is logistic regression with balanced weights, trained on ejection fraction, serum creatinine, and age. Decision trees also performed well, though on average, they did slightly worse than logistic regression. Polynomial logistic regression performed considerably worse than both. The maximum accuracy achieved was 80%, and the maximum F1 was 0.75. These are slightly higher than Chicco & Jurman's (2020) random forest classifier, but these results can be considered optimistic. However, it is clear the results are in range of Chicco & Jurman's.

Feature reduction identified ejection fraction and serum creatinine as the most important features for predicting death and survival, confirming Chicco & Jurman's (2020) conclusion. In addition, age was identified as another important feature; in general, training with ejection fraction, serum creatinine, and age improved model performance. This paper concludes that doctors should consider age in addition to ejection fraction and serum creatinine when assessing death risk in their heart failure patients.

Introduction

According to the Government of Canada, heart disease is the second most common cause of death for Canadians, with nearly twelve adults dying every hour ("Heart Disease in Canada", 2017). Heart failure, caused by the heart's poor performance in circulating blood, along with other cardiovascular diseases, manages to kill around 17 million people each year (Chicco & Jurman, 2020). It is clear there is an urgent need to determine whether an individual is at risk of perishing from heart disease. Doctors must be able to estimate risk of death given the information they have about their patients.

Research published in 2020 (Chicco & Jurman) developed a machine learning model for predicting death outcomes of heart failure patients and determining what patient information is most relevant to doctors. The research applied various classification techniques on a dataset of 299 heart failure patients, including random forests, regression, and neural networks. The dataset, collected from a hospital in Pakistan in 2015, contains twelve features (see Appendix A, *Table 1*), ranging from clinical measurements like sodium creatinine to categorial features like sex. The target value, "death event", indicates whether the patient died from their heart condition within the time period of the study. Chicco & Jurman performed feature selection on this dataset, narrowing down to ejection fraction and serum creatinine; using these two features, they developed a model with 74% accuracy. Chicco & Jurman conclude that it is mostly those two features that doctors should focus on when assessing risk of death in heart failure patients.

In contrast, the goal of this paper is to confirm Chicco & Jurman's conclusion and potentially improve on the results of their model. Differing learning and feature selection techniques are used to determine if such results are commonly replicable. Included in these techniques are logistic regression, decision trees, and recursive feature selection. Furthermore, feature scaling and dimensionality reduction are used to improve algorithm tuning and performance.

Background

The 299 patients in the dataset had previous history with heart failure, classifiable by New York Heart Association standards, so they were already at risk of heart failure complications (Chicco & Jurman, 2020). There are no "healthy" individuals represented in the data: a model would have to differentiate between already extreme clinical measurements to classify deaths and survivals. As seen in Appendix A (see *Table 2, 3*), the data is imbalanced with only 33% of records indicating a death occurred ("death event" = 1). Clearly, heart failure alone cannot be used as a sole predictor for death outcomes: outliers/variance in clinical measurements might also be used to identify risk of death in already unhealthy patients. The dataset indicates that most heart failure patients are men (see *Graph 1, 2*); furthermore, smokers or patients with diabetes are heavily represented. These categorical features are equally common in patients who died as they are patients who survived, however. The only categorical features that might help differentiate deaths from survivals are anemia and high blood pressure. As seen in *Graph 2*, anemia and high blood pressure are slightly more common in deaths than survivals.

When observing the continuous features of the dataset, mostly composed of clinical blood measurements, differences in ejection fraction are immediately noticeable in those who died (see *Table 2, 3*). The mean ejection fraction is 7% lower in those who died, indicating such patients had even poorer heart performance than usual. The standard deviation of ejection fraction is slightly higher, indicating more variance and/or outliers. It is possible extremely low values of ejection fraction are strong predictors of death, given that the measurement is a summarizer of heart performance, with poor performance generally leading to heart failure. A similar pattern can be observed in serum creatinine and creatinine phosphokinase, with a higher mean and standard deviation present in patients who died. These three features are likely to provide the most information to machine learning models. In *Graph 3,* a pair plot is used to graph ejection fraction and serum creatinine against each other. A high-density cluster for survival outcomes is immediately apparent. One need only find a model that can best approximate the boundaries of clusters such as these.

This paper uses a generalized approach to train and score the presented models. The dataset has been split into a training set (80% of the data) and a test set (20%). During hyperparameter tuning, 4-fold cross-validation is performed so that approximately 20% of the data is allocated in a cross-validation set on each run. Upon completion of 4-fold cross-validation, the mean score of each run is taken. The model with the highest mean score is selected as the best-performing model. Before training occurs, the dataset is standardized with z-score. Standardization is favored over mean normalization in a dataset like this, since z-score tends to retain information about features despite the presence of outliers (Liu, 2020). Outliers are likely present in the dataset, but they may be indicative of a death outcome, so information may be lost by removing them. Indeed, standardization seems to improve model performance across the board with this dataset.

Most models are trained using eleven features, but in some cases feature reduction is used to improve results. “Time” is ignored since it is a feature specific to the original study. The value of “time” is highly dependent on both the length of the study and the date on which the recording was taken. It is possible that including “time” would slightly improve model results, as was observed early on in model training. However, this feature is not replicable beyond the context of the study and thus prevents generalizability.

F1 and accuracy are used to score cross-validation, the highest mean dictating the best model. Since the dataset is imbalanced, F1 is a preferable metric to promote true positives (predicting “death event” = 1). Accuracy rewards true positives and true negatives, but an ideal model in this context would have high recall for “death event” = 1. False positives are preferred over false negatives in a life-or-death situation. Since F1 does not reward true negatives, models scored with F1 are more aggressive in predicting death outcomes rather than survival outcomes. Accuracy is used as a second scoring metric to observe if the use of F1 makes a noticeable difference in model performance. Since Chicco & Jurman (2020) use accuracy as the primary metric in their research, this paper is able to compare model results.

Results

Training began first with standard logistic regression. The classifier performed admirably without tuning. However, F1 slightly improved when using the "liblinear" solver. Scikit-learn's documentation (2020) recommends the liblinear solver for smaller datasets. Since the heart failure dataset has 299 records, one might expect liblinear to perform suitably in most situations. While the choice of solver did not impact final results significantly, liblinear was chosen as the primary solver for all logistic models presented in this paper.

Rather, tuning class weights made the most impact on logistic regression, as was the case with every other model presented in this paper. Scikit-learn allows the programmer to provide custom weights to each target class; a higher weight assigned to a class causes greater penalization for incorrect outputs during training. As described in scikit-learn's documentation (2020), setting this parameter to "balanced" automatically tunes the class weights to be inversely proportional to the class frequency. Since the heart failure data is imbalanced, a "balanced" logistic regressor assigns greater weight to "death event" = 1. The model more aggressively predicts death outcomes rather than survival, even at the expense of precision. This change in precision is justified, however, since false positives are less harmful than false negatives. In terms of final model results, setting the class weights to "balanced" improved F1 significantly and accuracy slightly. The results for all models are summarized in Appendix B.

Logistic regression was then retrained with polynomial features. Ideally, this would result in less bias and improved results, but no improvements were observed. Attempting to raise the polynomial's degree even slightly with eleven features significantly increased training time. To prevent this, polynomial expansion was applied after linear discriminant analysis (LDA). With the data projected into one dimension, training time and the number of polynomial features were reduced.

Starting at the second degree, up to the eleventh degree was tested. Results are summarized in *Graphs 4, 5*. When scoring cross-validation with F1, increasing the degree only fluctuated training and cross-validation (CV) score. When scoring with accuracy, increasing the degree improved CV score but not training score. Since an improvement would usually be expected in training score, it is likely this is not a generalizable result. Without more data, it is difficult to determine whether overfitting is occurring, since all degree values seem to generalize equally well. It is apparent, however, that at the eight degree and beyond, the hypothesis function boundaries become incompatible with the dataset, and both training and CV score decrease drastically. The eighth degree can be considered a cut-off point where the model becomes biased. However, even the best results of polynomial logistic regression place it behind standard logistic regression in every metric (see *Table 4, 5*). It is evident increasing the degree does not improve the model.

The last model trained was a decision tree classifier. The hyperparameters that made the most impact on performance were "min_samples_leaf" and "min_samples_split". According to scikit-learn's documentation (2020), "min_samples_leaf" forces a minimum number of samples at each leaf node, while "min_samples_split" forces a minimum number of samples to continue

splitting the tree. Clearly, both parameters can be used to reduce overfitting by preventing the tree from growing too large. According to Mithrakumar (2019), the optimal values for these parameters tend to fall between one through twenty and two through forty, respectively. To keep training time reasonable, a decision tree model was trained for every combination of values in these ranges. The best performing model from that tuning had "min_samples_leaf" set to nine and "min_samples_split" set to 31. F1 and accuracy were quite low; the tree seemed to slightly overfit the data, so it was tuned again with a maximum depth. With depth set to four, the model achieved its best results (see *Table 4, 5*).

Besides being quite small, it was clear the dataset contained some noise that prevented better performance. While a model that considers many features might be more generalizable, it is possible that better performance might be attained with feature reduction. Lastly, each model was tuned again; however, feature reduction based on model-assigned weights was performed recursively. The best feature subset was then selected, based entirely on CV score. Ideally, removing the least important features would improve performance.

For both logistic regression and decision trees, seven features were removed when scoring cross-validation with F1. The four features remaining—ejection fraction, serum creatinine, age, and sex—were used to train the models. Logistic regression improved in precision for "death event" = 0 and recall for "death event" = 1, but otherwise performed slightly worse. The decision tree model improved overall. Reducing to ejection fraction and serum creatinine was expected since they are referenced as the most important features by Chicco & Jurman (2020). Age and sex, however, differ from the findings of Chicco & Jurman. Age is not as surprising, since the mean age of deaths is about seven years higher than the mean age of survivals. Sex is more surprising, since the data analysis does not show a major disparity in sex between deaths and survivals (see *Table 2, 3*). When scoring cross-validation with accuracy, feature reduction did not include sex; training logistic regression and decision trees with just ejection fraction, serum creatinine, and age produced the best performing models in this paper, with highest results in every metric (see *Table 5, Diagram 2*).

Comparing the results of the models, one can see that even some of the worst performing models manage to outperform Chicco & Jurman's (2020). Chicco & Jurman's best performing model is a random forest classifier, which achieved an F1 score of 0.547 and an accuracy of 74% (see *Table 6*). This paper's best performing model achieved an F1 score of 0.75 and an accuracy of 80% (see *Table 5*). While it is possible this model is better than a random forest, there is more to the story. Chicco & Jurman implemented a more sophisticated cross-validation, where each validation and test set were guaranteed to contain 33% death outcomes. This allowed more accurate representation of the data distribution and made the model more generalizable. Furthermore, their results were computed from the mean of a hundred train-test runs, with random seeding for each run. Therefore, it is likely the results presented in this paper are optimistic.

Conclusion

As can be seen in *Table 5* and *Table 6*, the models presented in this paper are comparable to that of Chicco & Jurman (2020): final reported metrics are slightly higher than Chicco & Jurman's random forest classifier. In future research, more train-test cycles should be performed with random seeding for the cross-validation and test sets. Taking the mean score of these cycles, it is likely the metrics will converge closer to the random forest classifier. Furthermore, the results of feature reduction showed ejection fraction and serum creatinine as the most important features in predicting death from heart failure, consistent with the findings of Chicco & Jurman.

In general, logistic regression is the best performer, regardless of cross-validation scoring technique (F1 or accuracy). Decision tree classifiers consistently performed slightly worse than logistic regression in most metrics. Polynomial features failed to improve logistic regression, making results worse. *Graph 6* plots the decision boundary generated by the best performing model, logistic regression with balanced weights and reduced features, when trained only with ejection fraction and serum creatinine. While the data is noisy and sporadic, the line clearly distinguishes between larges clusters of death and survival.

While adding more weight to death outcomes helped improve model performance, scoring cross-validation with F1 did not find the best performing model. Cross-validation accuracy, along with reduction to three features, found the best model instead. During feature reduction, an extra feature, sex, was removed when scoring with accuracy, which likely caused the increase in model performance. A new feature, age, was ranked alongside ejection fraction and serum creatinine as most important, indicating that death from heart failure is associated with advanced age. In general, doctors should focus on ejection fraction, serum creatinine, and age when assessing death risk in their heart failure patients.

Reference List

Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., & Raza, M. A. (2017, July 20). [Heart failure clinical records data set]. Retrieved from https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records

Chicco, D., Jurman, G (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*. Retrieved from https://doi.org/10.1186/s12911-020-1023-5

Heart Disease in Canada. (2017, February 9). Retrieved from https://www.canada.ca/en/public-health/services/publications/diseases-conditions/heart-disease-canada.html

Liu, C. (2020, April). Data transformation: Standardization vs normalization. Retrieved from https://www.kdnuggets.com/2020/04/data-transformation-standardization-normalization.html

Mithrakumar, M. (2019, November 11). How to tune a decision tree? Retrieved from https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680

Scikit-learn documentation. (2020). Retrieved from https://scikit-learn.org/stable/modules/classes.html

Appendix A: Data Analysis

Feature Name	Description	Unit	Range
Age	Of patient	Years	40-95
Anemia	Decrease of red blood cells	Boolean	0, 1
High blood pressure	N/Y	Boolean	0, 1
Creatinine phosphokinase	Level in blood	mcg/L	23-7861
Diabetes	N/Y	Boolean	0, 1
Ejection fraction	Percent of blood leaving heart on each contraction	Percent	14-80
Sex	F/M	Boolean	0, 1
Platelets	Level in blood	kiloplatelets/mL	25.01-850.00
Serum creatinine	Level in blood	mg/dL	0.50-9.40
Serum sodium	Level in blood	mEq/L	14-148
Smoking	N/Y (does patient smoke)	Boolean	0, 1
Time	From this recording being taken to end of study or death event (follow-up period)	Days	4-285
DEATH EVENT	Survived/died in follow-up period	Boolean	0, 1

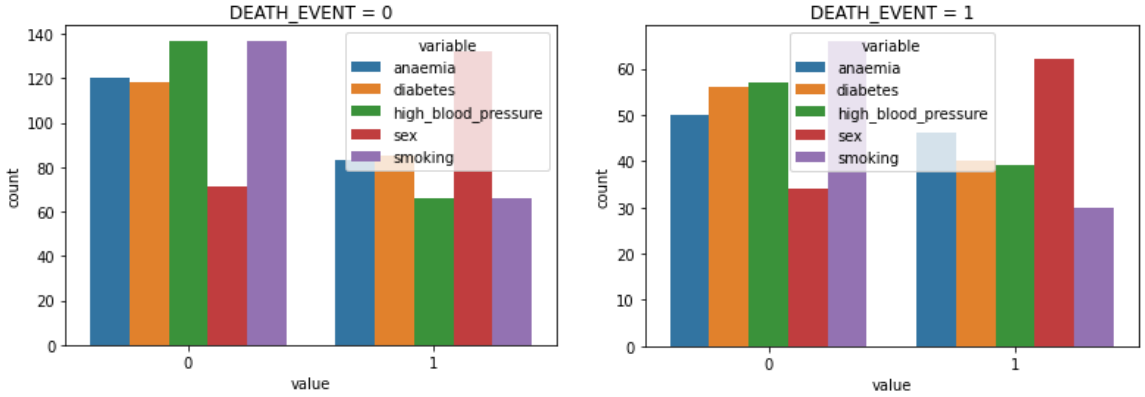
Table 1. Summary of features (Chicco & Jurman, 2020).

	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium	age
count	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000
mean	540.054187	40.266010	266657.489901	1.184877	137.216749	58.761906
std	753.799572	10.859963	97531.202283	0.654083	3.982923	10.637890
min	30.000000	17.000000	25100.000000	0.500000	113.000000	40.000000
25%	109.000000	35.000000	219500.000000	0.900000	135.500000	50.000000
50%	245.000000	38.000000	263000.000000	1.000000	137.000000	60.000000
75%	582.000000	45.000000	302000.000000	1.200000	140.000000	65.000000
max	5209.000000	80.000000	850000.000000	6.100000	148.000000	90.000000

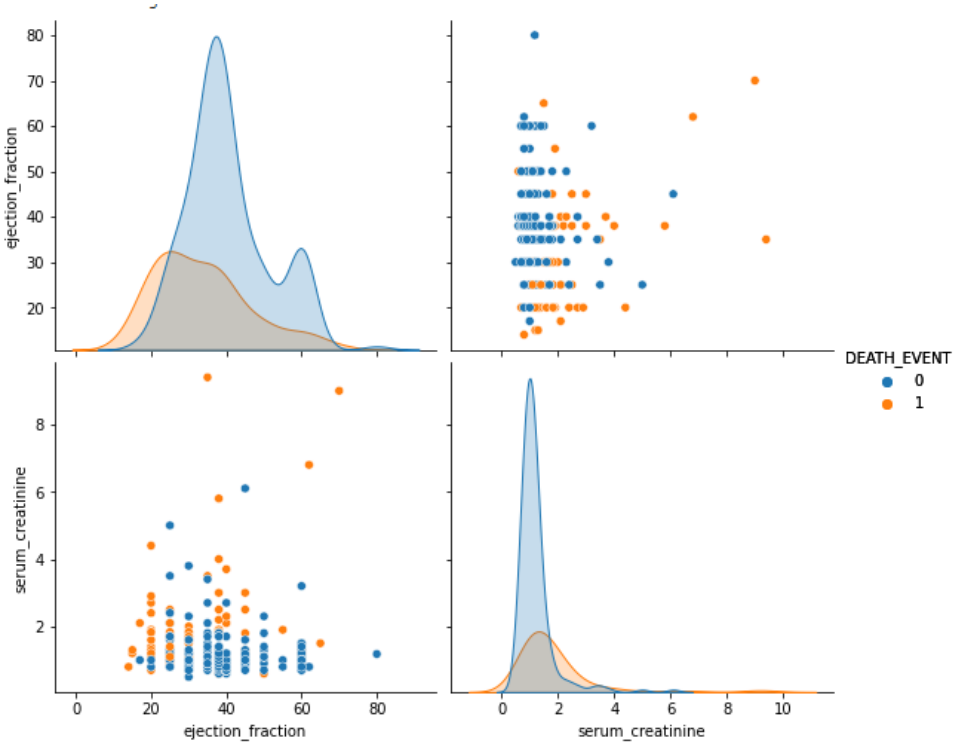
Table 2. Continuous features of survival outcomes.

	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium	age
count	96.000000	96.000000	96.000000	96.000000	96.000000	96.000000
mean	670.197917	33.468750	256381.044792	1.835833	135.375000	65.215281
std	1316.580640	12.525303	98525.682856	1.468562	5.001579	13.214556
min	23.000000	14.000000	47000.000000	0.600000	116.000000	42.000000
25%	128.750000	25.000000	197500.000000	1.075000	133.000000	55.000000
50%	259.000000	30.000000	258500.000000	1.300000	135.500000	65.000000
75%	582.000000	38.000000	311000.000000	1.900000	138.250000	75.000000
max	7861.000000	70.000000	621000.000000	9.400000	146.000000	95.000000

Table 3. Continuous features of death outcomes.

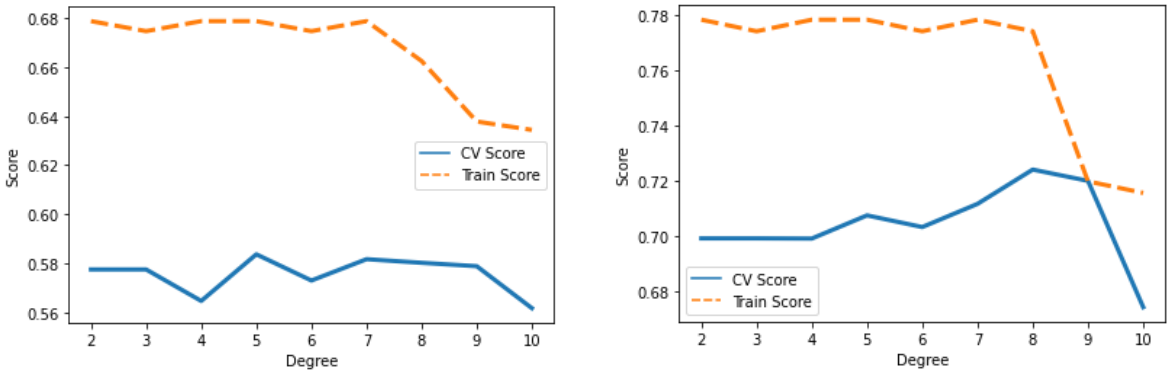


Graphs 1, 2. Categorical features of survival, death.



Graph 3. Pair plot of ejection fraction and serum creatinine, two correlated features.

Appendix B: Parameter Tuning Results



Graphs 4, 5. Degree effect on training and CV score (F1/accuracy respectively) with logistic regression (balanced weights).

	# Features	Balanced Weights	Accuracy	y=1 F1	y=0 F1	y=1 Precision	y=0 Precision	y=1 Recall	y=0 Recall
Logistic regression	11	No	0.72	0.48	0.80	0.80	0.70	0.35	0.95
	11	Yes	0.77	0.68	0.82	0.71	0.79	0.65	0.84
	4	Yes	0.77	0.70	0.81	0.70	0.81	0.70	0.81
Polynomial logistic regression	11	No	0.70	0.47	0.79	0.73	0.69	0.35	0.92
	11	Yes	0.75	0.65	0.81	0.70	0.78	0.61	0.84
Decision tree	11	No	0.63	0.39	0.74	0.54	0.66	0.30	0.84
	11	Yes	0.75	0.67	0.80	0.68	0.79	0.65	0.81
	4	Yes	0.77	0.70	0.81	0.70	0.81	0.70	0.81

Table 4. Model tuning results, scoring cross-validation with F1. When reducing to four features from eleven, features used were ejection fraction, serum creatinine, age, and sex.

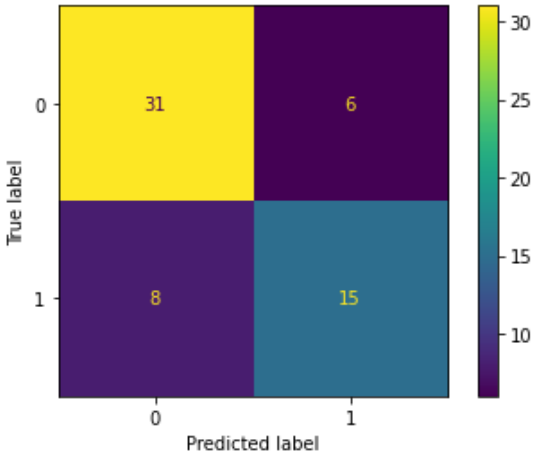


Diagram 1. Confusion matrix for logistic regression (balanced weights, all features), which attained highest F1 score.

	# Features	Balanced Weights	Accuracy	y=1 F1	y=0 F1	y=1 Precision	y=0 Precision	y=1 Recall	y=0 Recall
Logistic regression	11	No	0.72	0.48	0.80	0.80	0.70	0.35	0.95
	11	Yes	0.77	0.68	0.82	0.71	0.79	0.65	0.84
	3	Yes	0.80	0.75	0.83	0.72	0.86	0.78	0.81
Polynomial logistic regression	11	No	0.70	0.47	0.79	0.73	0.69	0.35	0.92
	11	Yes	0.75	0.65	0.81	0.70	0.78	0.61	0.84
Decision tree	11	No	0.63	0.27	0.76	0.57	0.64	0.17	0.92
	11	Yes	0.75	0.67	0.80	0.68	0.79	0.65	0.81
	3	Yes	0.80	0.75	0.83	0.72	0.86	0.78	0.81

Table 5. Model tuning results, scoring cross-validation with accuracy. When reducing to three features from eleven, features used were ejection fraction, serum creatinine, and age.

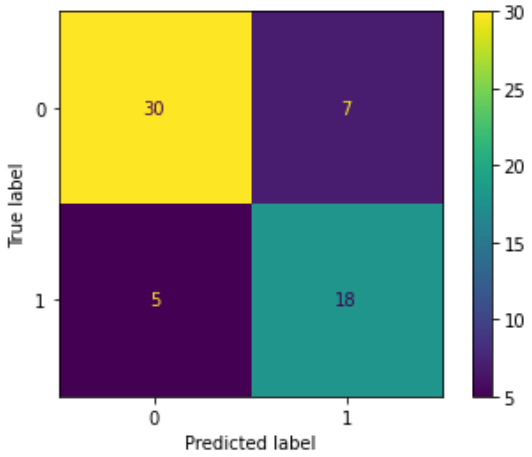
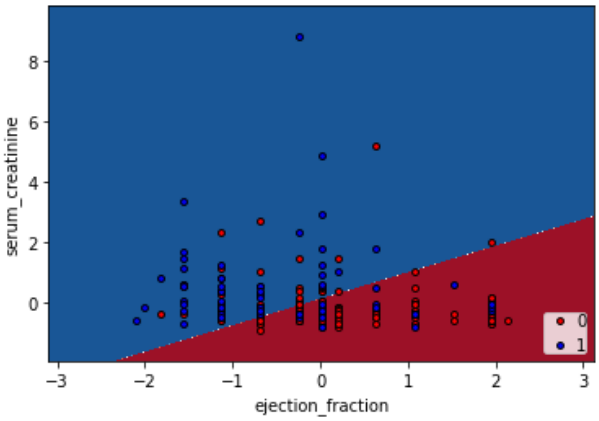


Diagram 2. Confusion matrix for logistic regression/decision tree (balanced weights, reduced features), which both attained highest accuracy score.



Graph 6. Decision boundary drawn by logistic regression/decision tree when using only serum creatinine and ejection fraction as input features.

Method	MCC	F_1 score	Accuracy
Random forests	blue+**0.384***	0.547	blue0.740*
Decision tree	**+0.376**	blue0.554*	0.737
Gradient boosting	**+0.367**	0.527	0.738
Linear regression	**+0.332**	0.475	0.730
One rule	**+0.319**	0.465	0.729
Artificial neural network	**+0.262**	0.483	0.680
Naïve bayes	**+0.224**	0.364	0.696
SVM radial	**+0.159**	0.182	0.690
SVM linear	**+0.107**	0.115	0.684
k-nearest neighbors	**-0.025**	0.148	0.624

Table 6. Results from Chicco & Jurman (2020), based on the mean of a hundred executions.