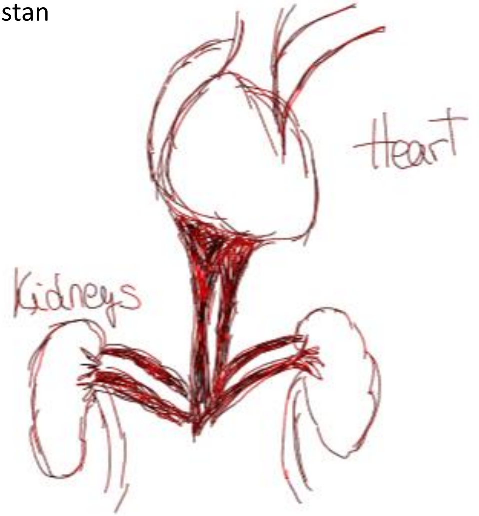
Revisited Machine Learning Approach to predict Mortality from Cardiovascular Disease

Lara Malinov (malinov@stanford.edu)

Data

The data consists of 299 medical records of patients with left ventricular systolic dysfunction collected by the Faisalabad Institute of Cardiology and the Allied Hospital In Faisalabad in Punjab, Pakistan



Simplified Schema of the Cardiovascular System

- Sex, Age, Smoker, Diabetes, High Blood Pressure
- Anemia: characterized by low levels of red blood cells which carry oxygen
- Creatinine phosphokinase (CPK): indicates muscle tissue damage
- Ejection fraction: percentage of blood expulsed at every heart contraction
- Platelets: responsible for stopping bleeding
- Serum Sodium: high levels indicate kidney's failure to evacuate it from the blood

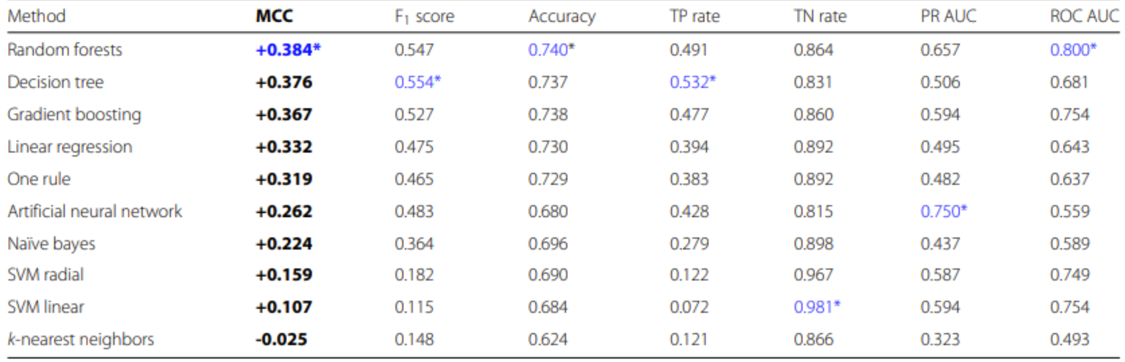
Results and Key Findings

The research was motivated by the erroneous comparison of models [1] without the proper data preprocessing. This especially applies to unscaled data used with distance-based models such as the k-Nearest Neighbors algorithm or Support Vector Machines. The results show that, when the data is scaled and balanced, the majority of the models have improved performance on the test predictions across different evaluation metrics (see top Table 1), even though significantly less observations were used (100 vs. 239). Showing that sometimes even using less data can improve performance under the right preprocessing choices. In addition, AdaBoost proved to outperform the models in [1] and the lowest scorers in Table 2 are among the best performing models in Table 1, namely Naïve Bayes and Support Vector Machines with radial basis function.

Table 1: Test prediction performance with scaled and balanced training set

	MCC	F1 score	Accuracy	TP rate	TN rate	PR AUC	ROC AUC
AdaBoost	0,445	0,585	0,779	0,517	0,892	0,633	0,742
Random Forest	0,439	0,582	0,769	0,500	0,896	0,633	0,743
Naïve Bayes	0,428	0,552	0,804	0,585	0,861	0,609	0,705
SVM radial	0,406	0,559	0,754	0,477	0,888	0,613	0,726
Logistic Regression	0,381	0,542	0,729	0,444	0,890	0,605	0,717
Gradient Boosting	0,327	0,504	0,714	0,420	0,869	0,568	0,684
SVM linear	0,320	0,500	0,709	0,414	0,868	0,565	0,681
Multi-Layer Perceptron	0,250	0,454	0,673	0,370	0,849	0,526	0,643
k-Nearest Neighbors	0,185	0,396	0,678	0,350	0,820	0,466	0,601

Table 2: Result of Chicco and Jurman (2020) [1]



Method	**MCC**	F_1 score	Accuracy	TP rate	TN rate	PR AUC	ROC AUC
Random forests	**+0.384***	0.547	0.740*	0.491	0.864	0.657	0.800*
Decision tree	**+0.376**	0.554*	0.737	0.532*	0.831	0.506	0.681
Gradient boosting	**+0.367**	0.527	0.738	0.477	0.860	0.594	0.754
Linear regression	**+0.332**	0.475	0.730	0.394	0.892	0.495	0.643
One rule	**+0.319**	0.465	0.729	0.383	0.892	0.482	0.637
Artificial neural network	**+0.262**	0.483	0.680	0.428	0.815	0.750*	0.559
Naïve bayes	**+0.224**	0.364	0.696	0.279	0.898	0.437	0.589
SVM radial	**+0.159**	0.182	0.690	0.122	0.967	0.587	0.749
SVM linear	**+0.107**	0.115	0.684	0.072	0.981*	0.594	0.754
k-nearest neighbors	**-0.025**	0.148	0.624	0.121	0.866	0.323	0.493

Impact of ML on medical diagnosis: case study

Age: 59
Sex: Male

Diabetes
Smoker
No Anemia
No High Blood Pressure

CPK: 66 mcg/L
*Ejection Fraction: 20 %
*Platelets: 70,000 kiloplatelets/mL
*Serum Creatinine 2.4 mg/dL
Serum Sodium: 134 mEq/L

Follow up appointment in 135 days

This is the case of man which AdaBoost rightfully predicted would die. He has some abnormal values(*) but only had a follow up appointment 4 months after his values were taken. This is an example of how ML techniques can be used to complement diagnostics and highlight cases that need more attention and medical care.

[1] 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, 20(1), 1-16.