Studying the most relevant risk factors for heart disease

Francesco Di Flumeri frdf@kth.se

Pablo Laso plaso@kth.se

GOALS

- Identify the most relevant risk factors.
- Building a reliable algorithm that can predict early cases of heart diseases.

RESULTS OVERVIEW

- High model performances were obtained.
- The most important features in the prediction were represented by the size of cardiovascular vessels.

Key ethical & sustainability

KEY ETHICAL

- Obtaining consent for using the dataset.
- Respect patients privacy, by keeping anonymity and data protection.
- Avoiding plagiarism.
- Computer-aided support tool (not intended for replacement).

SUSTAINABILITY ISSUES

- Improving patient medical treatments (3rd goal SGDs).
- Reduce hospitalization expenses for facilities (3rd goal SGDs).
- Trying to maintain and improve the relationship among doctors and patients.
- Reduce fatigue for physicians.

Problems & Research Questions

PROBLEM STATEMENT

- Heart disease is the world's biggest killer.
- Heart disease causes high number of hospitalizations.
- Heart disease causes high financial and lives costs.

RESEARCH QUESTIONS

- Which are the most important features in an ML model for predicting if a patient will experience a heart disease?
- Is a prediction model feasible (trained on these features)?

Related work & hypothesis

RELATED WORK

- Moen et al. worked on the same dataset for conducting a study on predicting cardiovascular diseases.
- Different studies have been conducted for providing doctors with diseases' prediction by starting from raw data.

HYPOTHESIS

- Prediction of heart disease is possible by given certain information about patients
- More generalized results due to the presence of data collections coming from different part of the worlds.
- Hypertension (HT), valvular disease (VHD) and age as major causes of heart failure.

Methodology

RESEARCH METHODOLOGY

- **Analytical method mainly.**
- **Empirical techniques might be used in the analysis of the dataset.**

DATASET

- **Pre-existing data collections coming from the UCI organization.**
  - Combination of four datasets, from different organizations.
  - 76 attributes per each subject.

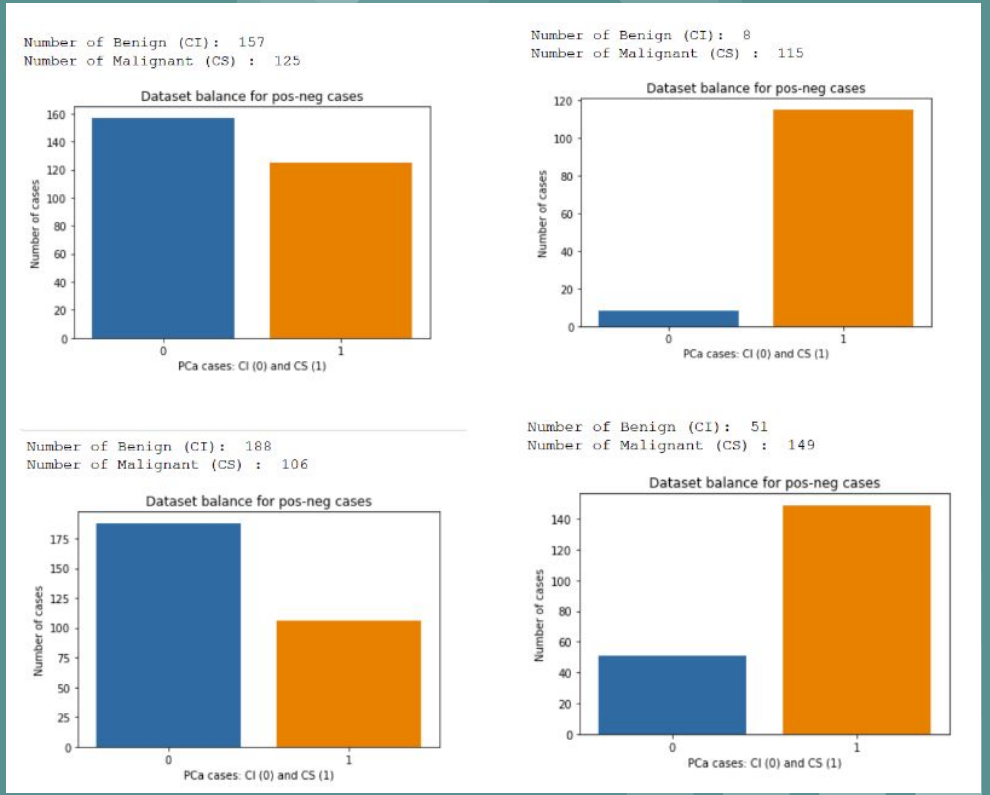
DATA CLEANING

- **Duplicate removal, imputation, integrity constraints violations, and outliers.**

Methodology

DATA BALANCE

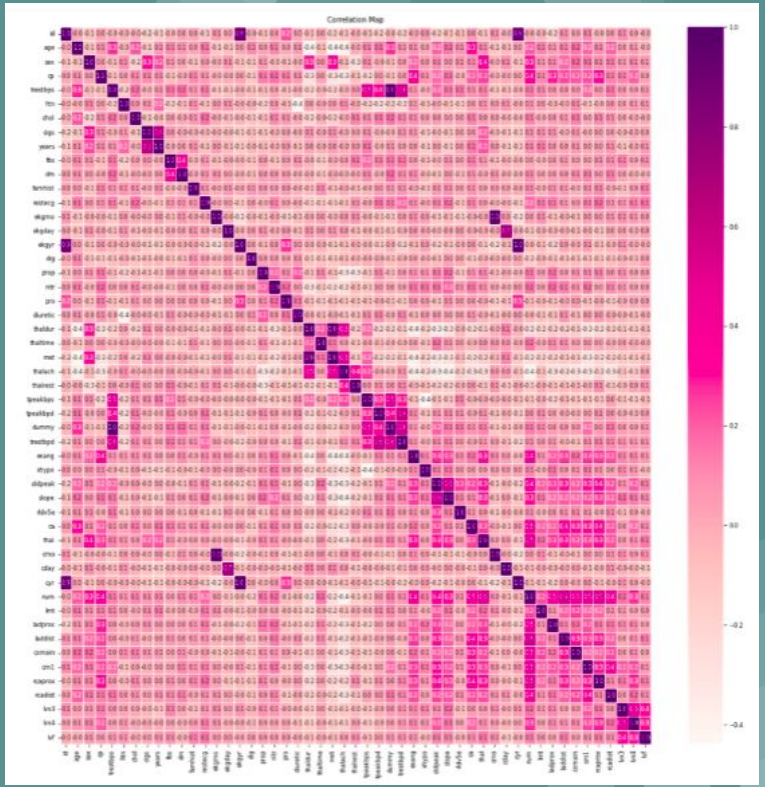
- We represented the data by using a histogram which shows the number of positive and negative instances.
- Some datasets are not equally balanced.



Methodology

EXPLORATORY DATA ANALYSIS

- Understanding the dataset issues.
- Get information about data types, shape of the dataset and descriptive metrics.
- Extract information about the relevancy of some features over others.
- Identify outliers, missing values or human error.
- Understand the relationship, or lack of, between variables.
- Maximize the insight into the dataset and minimize the potential error that may occur during the next steps in the analysis process.



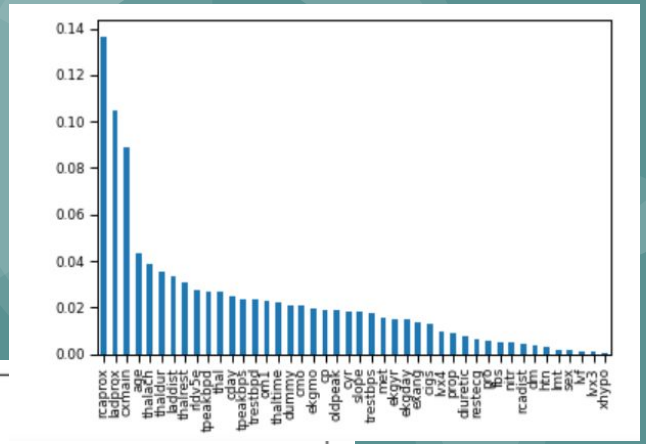
Results

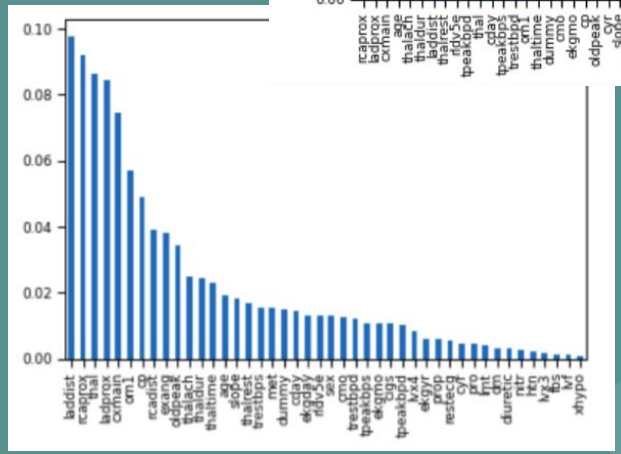
MODEL BUILDING

- Logistic Regression.
- KNN.
- Decision Tree.
- Neural Networks.
- Random Forest.

FEATURE SELECTION

- F-score based.
- Mutual Information.

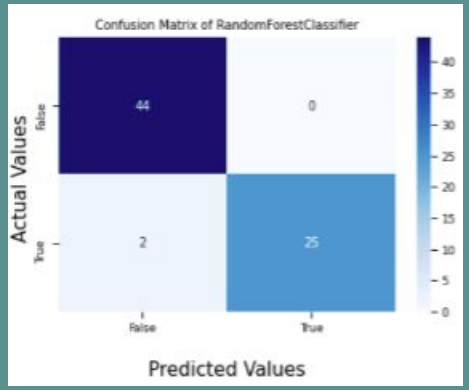




Results

MODEL EVALUATION

- Accuracy.
- Recall.
- Precision.
- F1-score.
- Confusion matrix.



	Train Accuracy	Train Recall	Train Precision	Train F1	Validation Accuracy	Validation Recall	Validation Precision	Validation F1	Test Accuracy	Test Recall	Test Precision	Test F1
Logistic Regression FS	0.947867	0.887755	1.0	0.940541	0.948173	0.887895	1.000000	0.938354	0.985915	0.962963	1.000000	0.981132
Logistic Regression	1.000000	1.000000	1.0	1.000000	0.952624	0.896947	1.000000	0.945623	0.985915	0.962963	1.000000	0.981132
KNN Classifier FS	0.947867	0.887755	1.0	0.940541	0.943522	0.887895	0.987500	0.933592	0.985915	0.962963	1.000000	0.981132
KNN Classifier	0.886256	0.755102	1.0	0.860465	0.834330	0.644211	1.000000	0.781708	0.859155	0.629630	1.000000	0.772727
Decision Tree Classifier FS	0.947867	0.887755	1.0	0.940541	0.943411	0.877368	1.000000	0.931967	0.985915	0.962963	1.000000	0.981132
Decision Tree Classifier	1.000000	1.000000	1.0	1.000000	0.924363	0.918421	0.923103	0.918192	0.901408	0.925926	0.833333	0.877193
Neural Network Classifier FS	0.947867	0.887755	1.0	0.940541	0.948173	0.887895	1.000000	0.938354	0.985915	0.962963	1.000000	0.981132
Neural Network Classifier	1.000000	1.000000	1.0	1.000000	0.891362	0.838421	0.924624	0.878018	0.915493	0.925926	0.862069	0.892857
Random Forest Classifier FS	0.947867	0.887755	1.0	0.940541	0.948173	0.887895	1.000000	0.938354	0.985915	0.962963	1.000000	0.981132
Random Forest Classifier	1.000000	1.000000	1.0	1.000000	0.924695	0.879474	0.954474	0.913159	0.971831	0.925926	1.000000	0.961538

Discussion & Conclusion

DISCUSSION & CONCLUSION

- Most important features in an ML model for predicting heart disease risk are related to medical parameters measuring size of blood vessels (*rcaproxorladdist*).
- Medical, physiological, and psychological information is necessary for having a medical picture about the heart condition of a patient and build a successful ML model.
- Possible to early predict heart disease, but reliability means no data be excluded.
- Heart disease is strongly related to the blood vessels shape (not age, but presence of hypertension (HT) or vascular disease (VHD)).

Limitations and future work

LIMITATIONS

- Authors' restricted medical knowledge.
- Data coming just from two world locations.
- Difficulty of applying the model to the real world because rarely all these data about patients are available.

FUTURE WORK

- Partial dependencies graphs to analyze model behavior in relation to predictors' trends.
- Gather the same information collected in UCI data-set, from other parts of the world.
- Discover different patients subgroups (for personalized treatment).

QUESTIONS