Data mining techniques applied to medicine

Miguel Alcón Doganoc
Universitat Politècnica de Catalunya
Barcelona, Spain
miguel.alcon@est.fib.upc.edu

ABSTRACT

The increasing amount of data collected in whatever field leads to Data Mining and Machine Learning to increase its importance in our lives. Because of that, we want to use this data to impact directly in the most important aspect of people's lives: health. In this work, we focus on the detection of the degree of presence of heart disease in a patient. We try different classifier methods to predict it using the well-known Hearth Disease data set [1] of the UCI Machine Learning Repository.

1 INTRODUCTION

2 DATA SET

Our work is based on the processed Cleveland data offered by the Hearth Disease data set. The Cleveland data is composed of 14 features of the patient and its heart, which are:

- **Age.** In years.
- **Sex.** Binary representation of the patient's gender.
- **Chest pain type.** Integer value between 0 and 3, representing if the pain is a typical angina, atypical angina, non-anginal pain or asymptomatic.
- **Resting blood pressure**. In mm Hg on admission to the hospital.
- **Serum cholesterol.** In mg/dl.
- **Fasting blood sugar.** Binary representation of: the fasting blood sugar of the patient > 120 mg/dl.
- **Resting electrocardiographic results.** Integer values between 0 and 2, representing the severity of the results.
- **Maximum heart rate achieved.** Integer value.
- **Exercise induced angina.** Binary representation of 'yes' or 'no'.
- **Old peak.** Decimal value that represents the ST depression induced by exercise relative to rest.
- **Slope.** Three integer values representing the slope of the peak exercise ST segment.
- **Number of major vessels colored by fluoroscopy** Integer value between 0 and 3.
- **Thalassemia.** Three integer values representing whether it is a fixed or reversible defect, or the observations are normal.
- **Target**. Integer value between 0 and 4 that represents the presence of hearth disease in the patient.

For more information about the features take a look at [1].

3 OBJECTIVES

The main objective of our work is to achieve the best possible accuracy with a restricted number of classifier methods, which are Naive Bayes, Nearest Neighbors, Decision Trees and Super Vector Machines. We select these classifiers due to its relevance and renown.

4 DESIGN OF THE EXPERIMENT

Our experiment suffered several modifications along its lifetime. We started with a straightforward experiment that divides data into train and test, trains different models using each of the classifiers and predicts the target value. With this, we wanted to know if one of the classifiers

5 IMPLEMENTATION

6 CONCLUSION

REFERENCES

[1] Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. Heart disease. https://archive.ics.uci.edu/ml/datasets/heart+Disease, 2019.