Predicting Heart Disease in Adult Patients with a Classification Model

Sarah Lueling

Assignment Task

Interest

- Heart disease is the leading cause of death

Goal

- Predict possible heart disease and help with early detection and management

Classification (Binary)

- Target: heart disease or no heart disease

Data

1 Appearance of data

age	sex	chest_pain_type	resting_blood_pressure	cholestoral	fasting_blood_sugar	rest_ecg	Max_heart_rate	exercise_induced_angina	oldpeak	slope	vessels_colored_by_flourosopy	thalassemia	target
52	Male	Typical angina	125	212	Lower than 120 mg/ml	ST-T wave abnormality	168	No	1.0	Downsloping	Two	Reversable Defect	0
53	Male	Typical angina	140	203	Greater than 120 mg/ml	Normal	155	Yes	3.1	Upsloping	Zero	Reversable Defect	0
70	Male	Typical angina	145	174	Lower than 120 mg/ml	ST-T wave abnormality	125	Yes	2.6	Upsloping	Zero	Reversable Defect	0
61	Male	Typical angina	148	203	Lower than 120 mg/ml	ST-T wave abnormality	161	No	0.0	Downsloping	One	Reversable Defect	0

2 Summary of data

Data contains 1,025 observations

- 80% in the training set
- 20% in the test set

Each classified as having heart disease (1) or not (0)

- 526 have it (1)
- 499 don't have it (0)

13 features recorded for each observation

- 8 numerical
- 5 categorical

3 Pre-processing steps taken

- Label Encoding
- Change attribute names

Classification Algorithms

- DECISION TREE CLASSIFIER
- RANDOM FOREST CLASSIFIER
- LOGISTICAL REGRESSION CLASSIFIER

Logistical Regression Classifier

Reason

- Used to predict binary outcome

Explanation

- Independent variables are analysed to determine the binary outcome with the results falling into one of two categories
- Goal: find a best-fitting relationship between the dependent and independent variables

Accuracy score: 0.79

Tuning

- Tuning: solver = liblinear, random state, l2

Decision Tree Classifier

Reason

- Good for binary target data

Explanation

- It hierarchically splits data into subsets which are then split again into the smaller subsets until they become “pure”.

Accuracy Score: 0.971

Tuning

- Parameter values : criterion='entropy', random_state = 42

Random Forest Classifier

Reason

- Random decision forests correct for decision trees' habit of overfitting to their training set

Explanation

- Creates decision tree at each step but only uses a random subset at each stage (grows many trees)
- Tests all possible 'branches' and selects the one with the highest accuracy

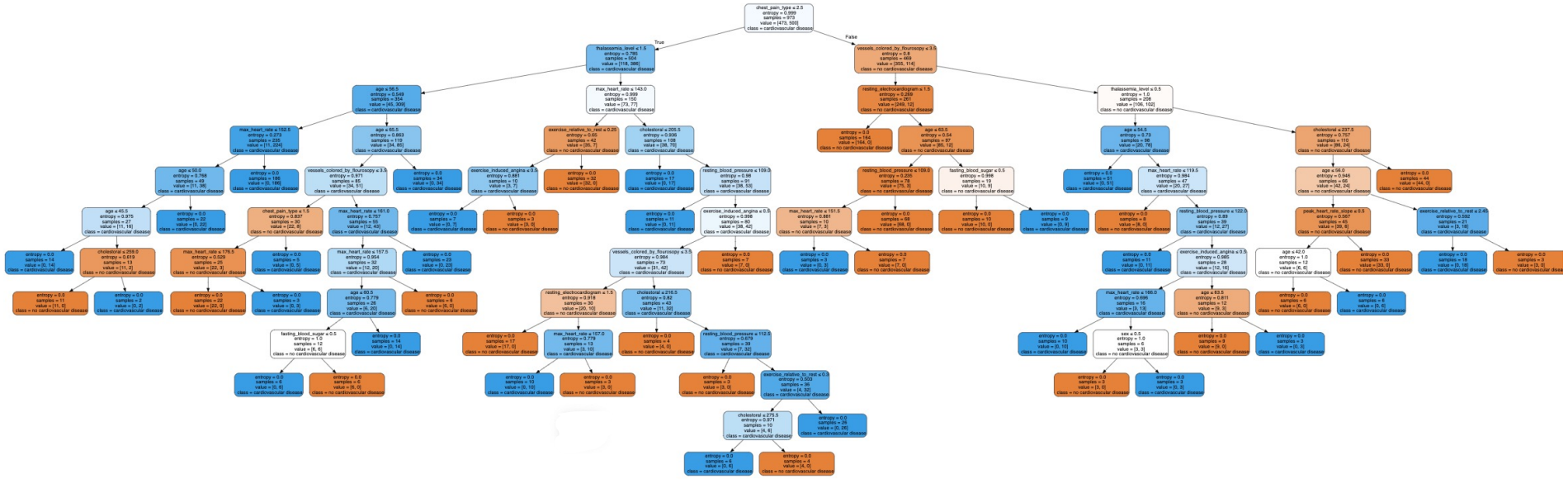
Accuracy Score: 0.985

Tuning

- Tuning: Accuracy only changes when we set random state

Comparison of Classification Models

	Logistical Regression	Decision Tree	Random Forest
Pros	Less prone to over-fitting with regularization techniques	Requires less effort for data preparation Intuitive and easy to explain	Lower risk of overfitting Works well with non linear data
Cons	Not appropriate for non-linear problems Need to pre-process data	Can lead to overfitting of the data For large dataset it's can become too complex to interpret and generalize	Takes longer to train



Result Analysis

- High accuracy of the Random Forest Classifier
- Error Analysis:
    1. Analyse raw data
    2. How was it created?
    3. Accuracy of annotations
    4. Tune parameters
    5. **Improper splitting of training and test data**

Thanks For Listening