Heart Disease Classification Using Neural Networks

Annegret Henninger
Anastasiia Bazhanova
Alena Calma

Ryerson University

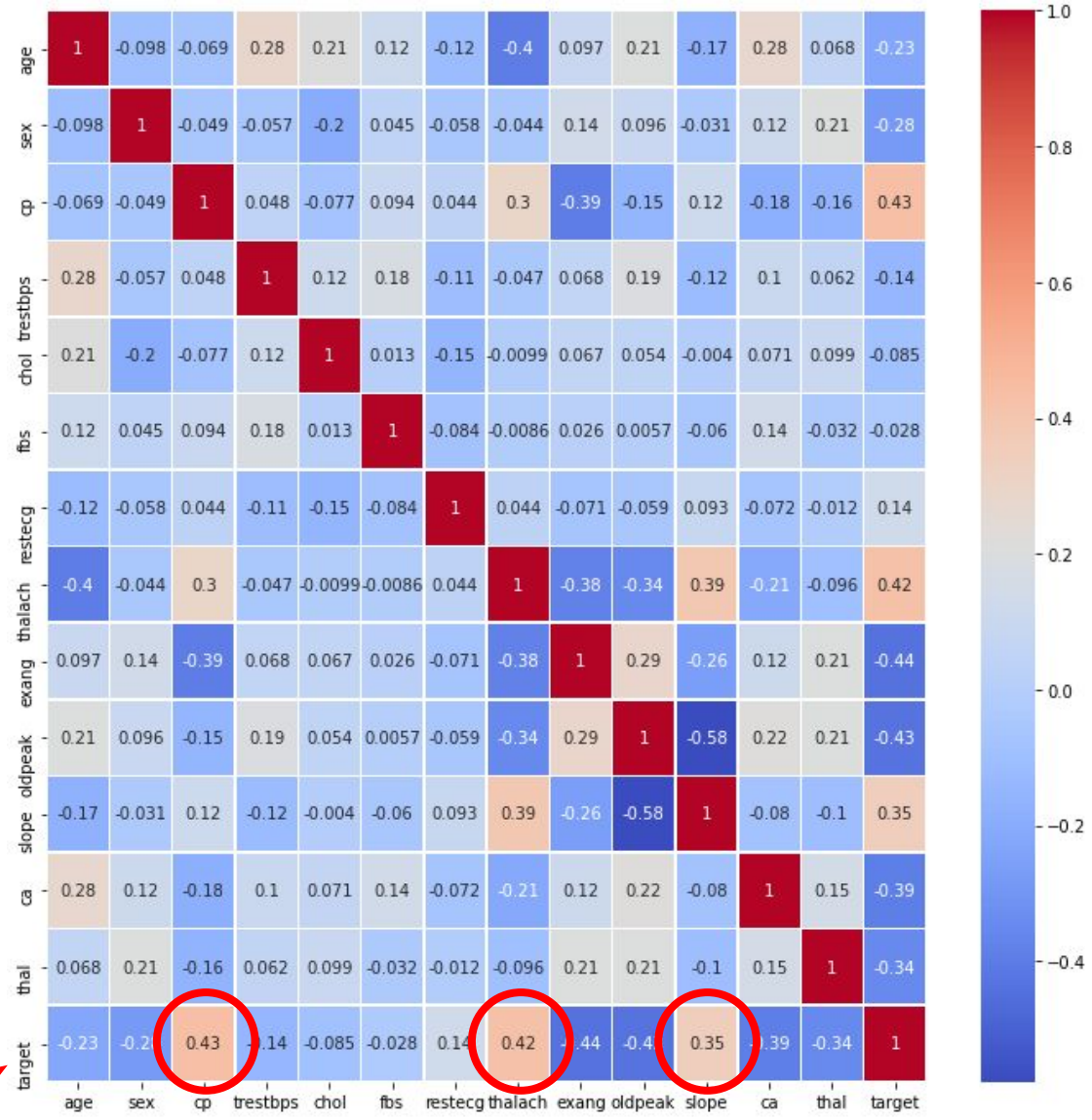
Data Exploration

Strongest Connection to CVD:

1. Chest Pain;
2. Max. Achieved Heart Rate;
3. Slope of the Peak ST Segment.

Factors Leading to ***higher rates*** of Heart Disease:

- Females of a certain age;
- Three types of less common chest pains;
- ST-T Wave Abnormality;
- Absence of Exercise Induced Angina;
- Downsloping ST Segment;
- Absence of Major Vessels;
- Cholesterol Levels above 200 mg/dl;
- A Max. Achieved Heart Rate of above 150.



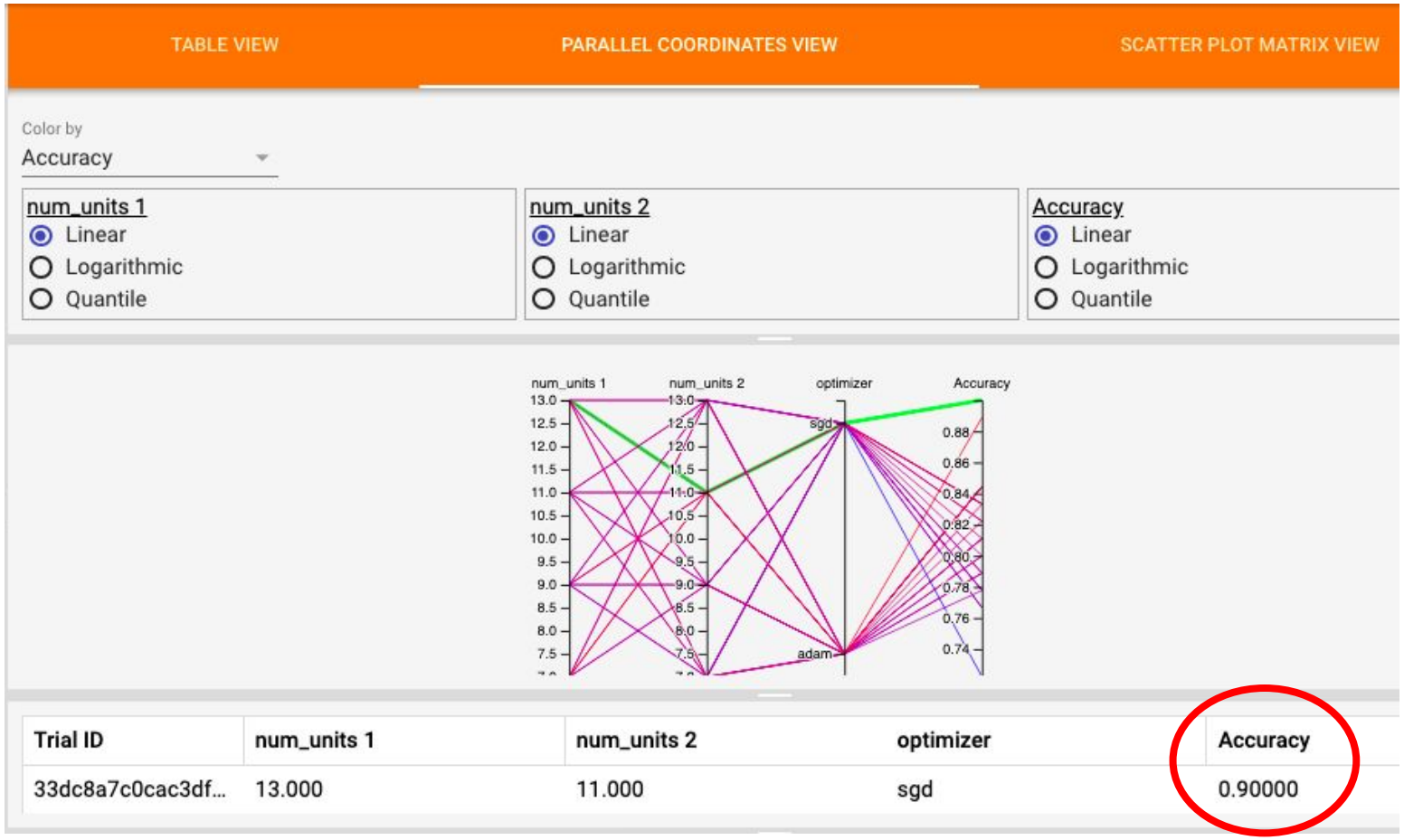
target

Learning Method #1: ANN using Tensorboard to visualize results

After running 13 distinctives test cases, these features have shown higher accuracy in the train set:

- 2 layers (diagram)
- more nodes in each layer
- more epochs (i.e. 100)
- a mix of activation functions, i.e. relu + tanh (diagram)
- adam optimizer statistically overperformed sgd, however, the highest accuracy was achieved with sgd (diagram)

Downside: Overfitted! Test set accuracy is **76.74%**.



Learning Method #2: Feature Values

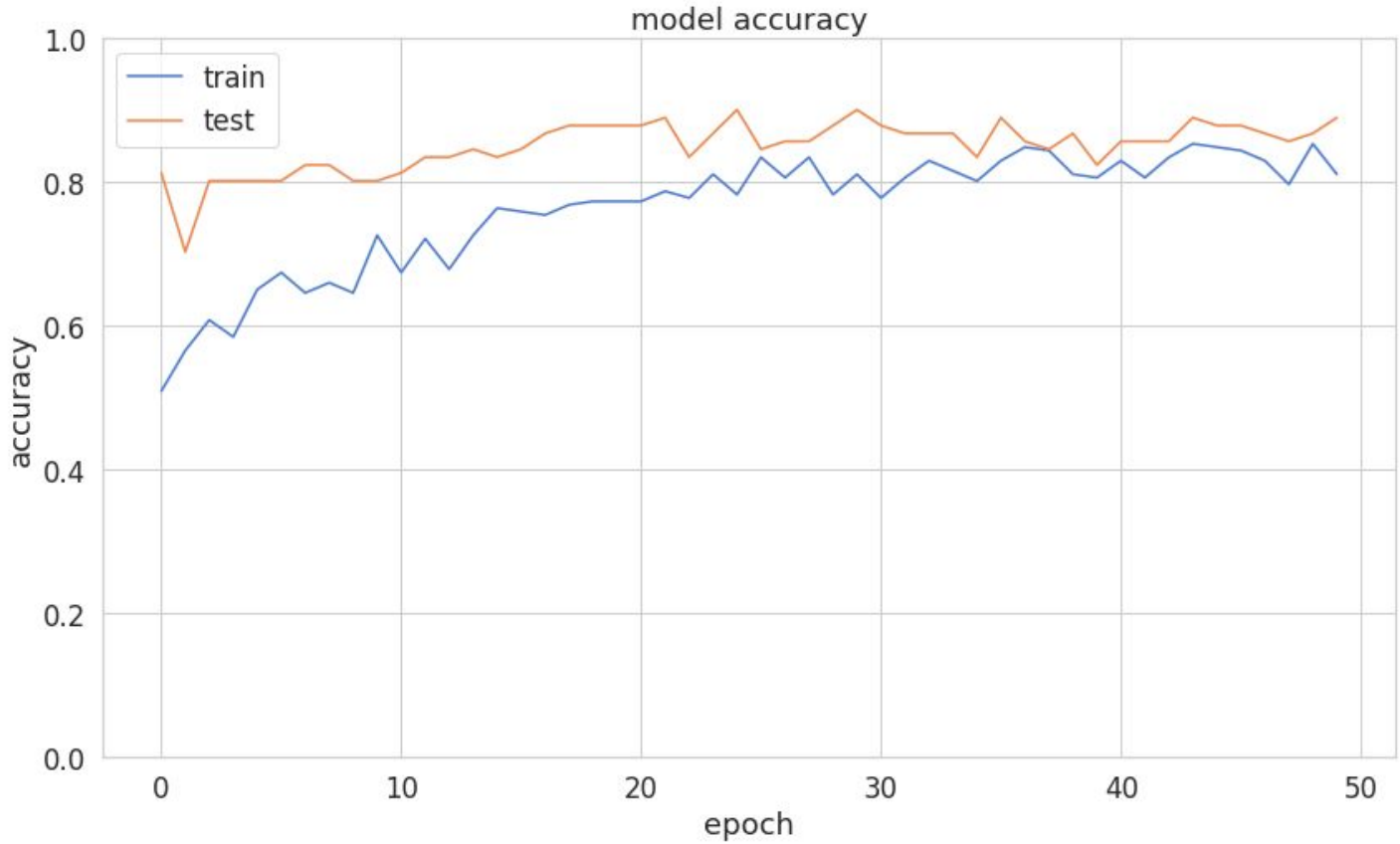
We also used feature columns which allows us to make the features easier for the model to compute, and to create cross columns from correlating features.

Cross feature 1: CP and Slope

Cross feature 1: age and Thal

Best accuracy: 89.01%

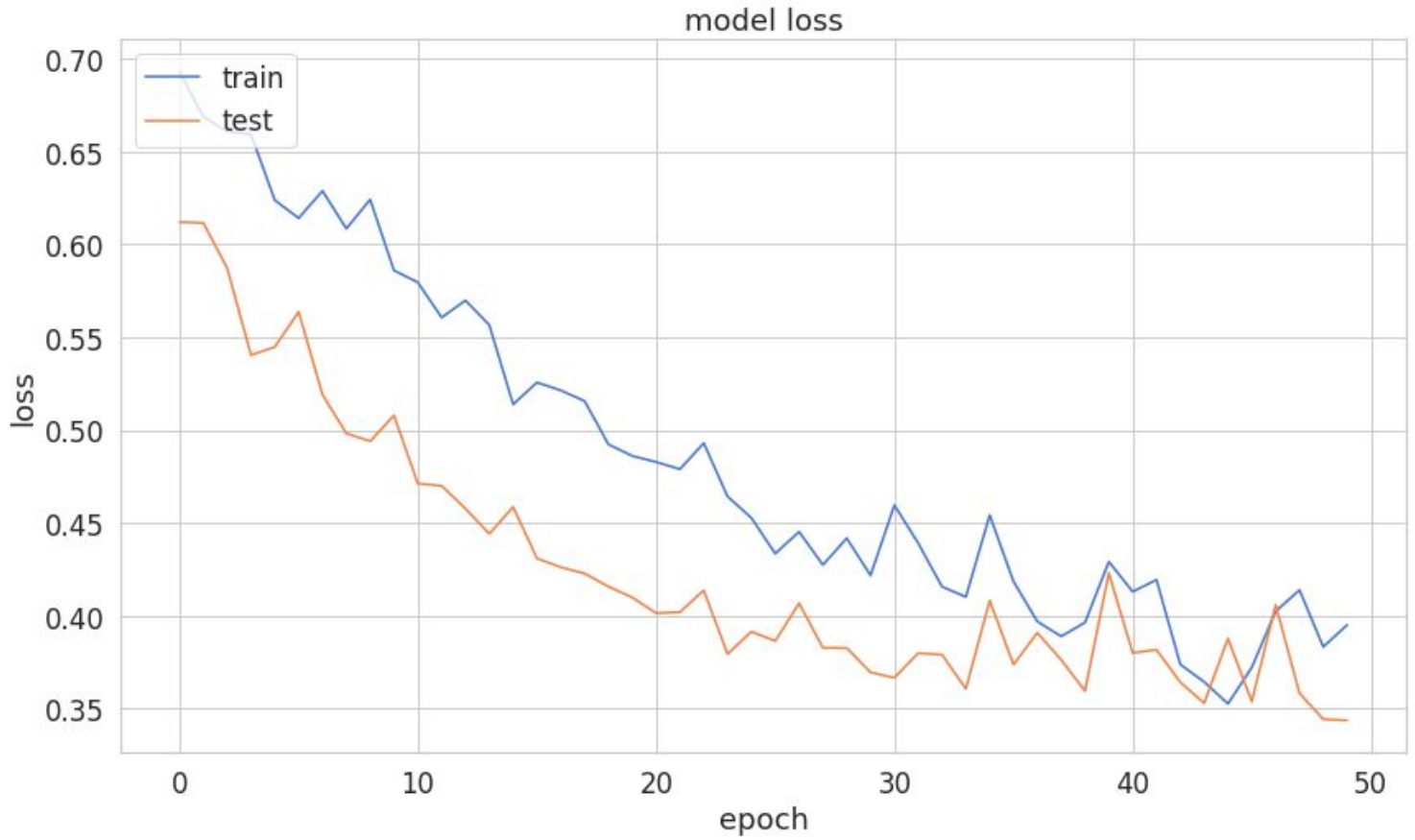
model { batch:32, hidden layer 1: Tahn 13n, hidden layer 2: ReLU 11 n, sigmoid output, cross-entropy loss function.



Learning Method #2: Feature Values

loss: 0.3426

Our top model is also the best because of the low loss function at 0.3426. Reducing loss by adjusting weights is the main function of the training process.



Ryerson University

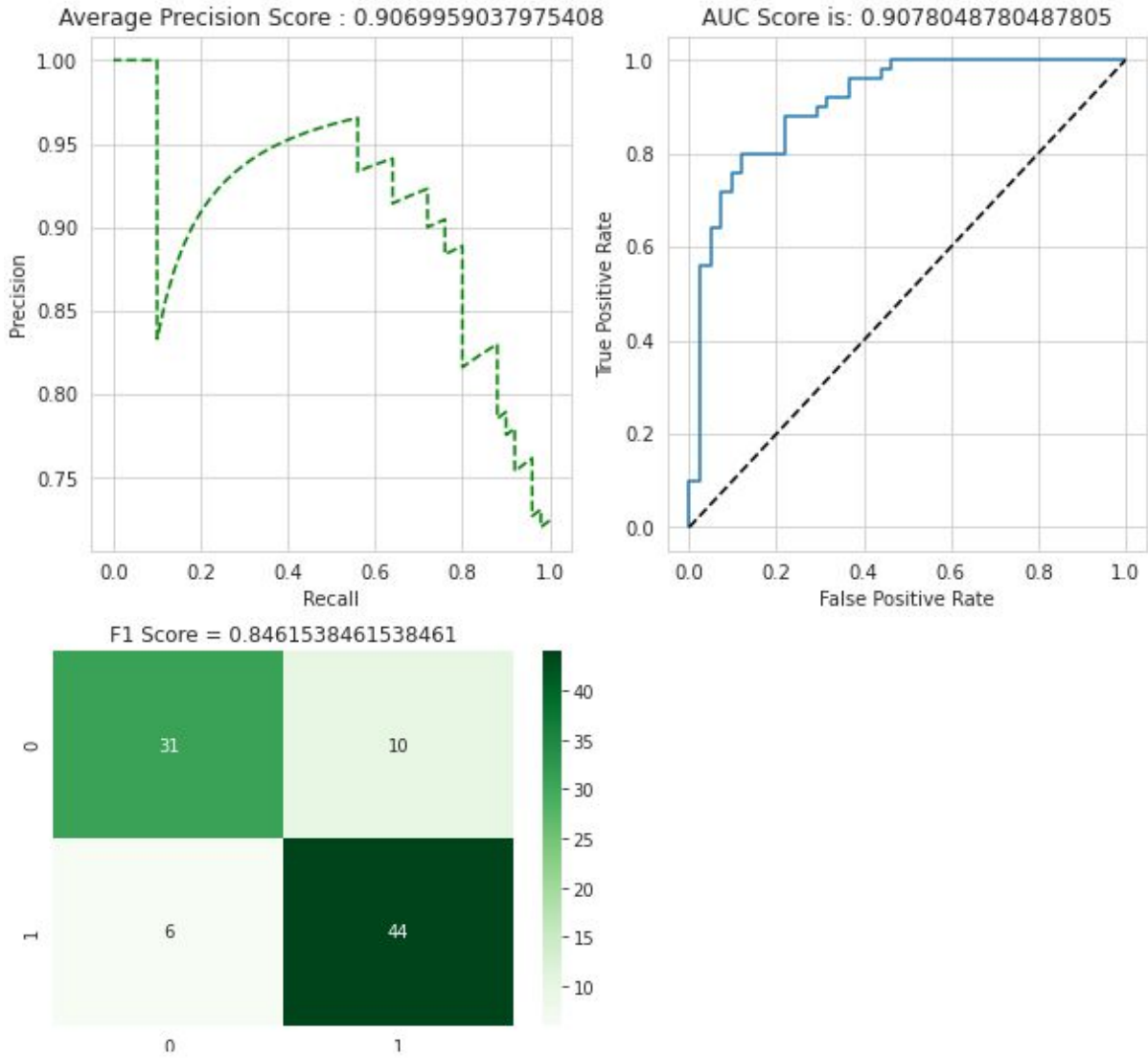
Learning Method #3: Classical Machine Learning Techniques

Although the goal of the project was to use Artificial Neural Network, we also used classical machine learning techniques to see if we can achieve better results considering we only had 303 instances.

We used the following techniques:

1. KNN
2. Logistic Regression
3. SVM
4. Random Forest
5. Decision Tree

We achieved the highest accuracy of 75.88% against the test set with logistic regression.



Evaluation

Best performing parameters:

- Using 2 hidden layers
- Using 10-13 neurons in the hidden layers
- Using ReLU and Tanh functions in the hidden layers
- Top model: used feature columns, 89% accuracy, 0.34 Loss.

Best machine learning technique

- Logistic regression (75.88% accuracy, 0.91 AUC). Particularly suitable for binary classification.

Discussions

For datasets with a larger number of features and varying data, we believe that using feature columns will become increasingly useful. Our top-performing classical machine learning model did not perform as well as ANN. But this could be because the purpose of this paper was to test neural networks, and more time was spent building the ANN models.

Improvements could be achieved by focusing more on classical techniques, exploring more advanced models like multilayer perceptron neural networks, and by determining severity levels as opposed to binary predictions, i.e. low risk, medium risk, high risk.

Conclusion and Future Work

Our best performing learning method was ***Method #2: Feature Values*** which achieved a test set accuracy of 89.01%.

- **Having lower epochs increased the accuracy.** Our best performing model had an epoch of 50. We noticed that our models tended to get overfitted between 75 and 100 epochs, depending on the parameters we set.
- **A higher number of neurons did not automatically mean that it would be more accurate.** We tried large networks with 128, and 256 neurons on 3-6 layers, but the accuracy did not improve.
- **A mix of activation functions leads to a higher accuracy.** In our best performing model, we used several activation functions: Tanh, ReLU and Sigmoid.