CS58 Final Report

Fareya Ikram

Julian Castoro

We chose to implement the Random Forest(Fareya’s Implementation) and Neural Network for each of the datasets below, we chose these two specifically because of their popularity in industry. If a specific dataset performs exceptionally poorly, we will try another algorithm.

Each algorithm’s performance is evaluated using stratified cross-validation with k=10.



**The Hand-Written Digits Recognition Dataset**

*The goal is to analyze 8x8 pictures of hand-written digits and classify them as belonging to one of 10 possible classes.*

With the above goal in mind, our group will report all performance metrics however have a greater weight of importance placed on the accuracy for this dataset. We are assuming this is not being used in any critical to life scenario and we are simply concerned with how often we are correct, for both classes.

**The Titanic Dataset**

*The goal, here, is to predict whether a given person was likely to survive this tragedy. The dataset contains data corresponding to 887 real passengers of the Titanic.*

With the above goal in mind, our group will report all performance metrics however have a greater weight of importance placed on the accuracy for this dataset. While lives were tragically lost in this event, none are on the line with this prediction. We are simply concerned with how often we can classify accurately.

Chart, scatter chart

Description automatically generated

The highest accuracy occurred at n\_trees = 50 so we explored around this number to find the optimal n\_trees, all charts shown below.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated



Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated



We found that the highest accuracy was at n\_trees = 52

**The Loan Eligibility Prediction Dataset**

*The goal is to automatically (and accurately) predict whether a given person should qualify for a loan.*

With the above goal in mind, our group will report all performance metrics however have a greater weight of importance placed on the precision for this dataset. A process like loan eligibility should not be left entirely to the decision of an algorithm however we imagine in this case it is being used as a decision making aid for a human banker. Precision is used as the most important metric as a higher precision would lead to a higher chance someone who should get a loan is actually getting one and we are avoiding False Positives. A False Positive in this case would lead to the greatest loss for both the bank and the individual applying for the loan.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

We find the highest precision at n\_trees = 20 so we reran the tests around that point to further zero in on the optimal n\_trees. For the sake of space I will only include the chart for precision

Chart, scatter chart

Description automatically generated



**N\_trees = 14 led to the highest precision of 0.85.**

For the sake of completeness we show the performance if no attributes had been removed.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

**The Oxford Parkinson’s Disease Detection Dataset**

*The goal is to predict whether a particular person is healthy or whether it is a patient with Parkinson’s.*

With the above goal in mind, our group will report all performance metrics however here we really just want to minimize the number of False Negatives as this could lead to someone not receiving the treatment they need. We will look at the Recall specifically for this.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated



Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated



Below we show the average confusion matrix for each n\_tree across each fold. This is to ensure the n\_trees we pick also leads to the lowest false positive rate.

Table

Description automatically generated with medium confidence