Predicting Inpatient or Outpatient Treatment

Patients in a hospital can receive care in different settings and by numerous methods. While patients can receive treatment at home, the two main settings for treatment are inpatient hospitals or an outpatient clinic. An inpatient visit requires the patient to stay overnight at the hospital to receive treatment and observation whereas an outpatient visit allows the patient to go home the same day. Another difference between the two visits is that outpatient visits occur in a clinic setting away from the hospital whereas an inpatient visit requires admission into the hospital. However, outpatient care can occur occasionally at a hospital if the patient is not admitted into the hospital. It is important to determine whether the patient's next treatment should be inpatient or outpatient because the type of care that can be provided differs, and it is also an indication of the acuity level of the patient. The patient's provider determines where the next treatment should occur. This model could aid in that process for the provider and provide a recommendation. Providers have limited time during the day and care for multiple patients at a time, so this model would reduce decision-fatigue. For these reasons, the model would be pitched as a solution to reduce decision-fatigue and free up more time for the provider. This would lead to better care within the hospital. Ultimately, the provider has the final say. There are outside factors that the model would not consider, for example the patient's ability to travel or access to an outpatient clinic. The scheduling for an outpatient clinic can also take weeks, which may be a factor in the decision.

The data used in this analysis is from a patient treatment classification dataset found on Kaggle. It has results from various laboratory test performed on the blood such as hematocrit, hemoglobin, erythrocyte, leucocyte, thrombocyte, MCH, MCHC, and MCV. It also contains the patient's age, gender, and whether the patient's next treatment was performed in an inpatient or outpatient setting.

When building the model, only patients of age 18 or older were kept. Since adult medicine and pediatric medicine differ in practice, the focus was solely on adults. Additionally, the gender and inpatient/outpatient variables were transformed to binary indicators for the model.

From the exploratory data analysis phase, a difference in histograms between inpatient and outpatient populations was observed for a few laboratory tests.

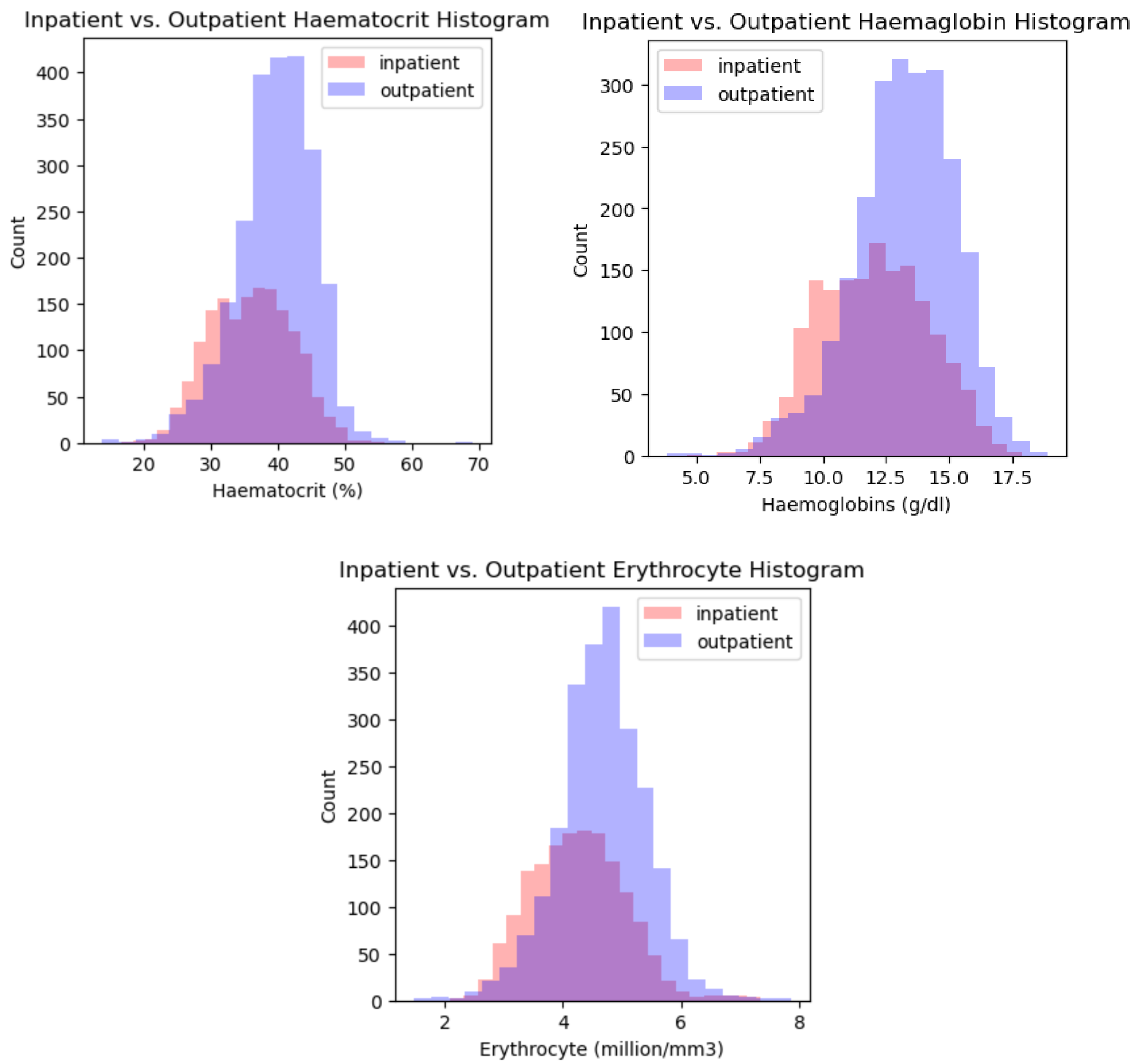

Figure 1: Distribution differences between inpatient and outpatient treatments for Hematocrit, Hemaglobin, and Erythrocyte labrotory tests

From these distributions, the peaks of the histograms can be observed at lower levels for the inpatient populations compared to the outpatient populations. This is an indication that there are statisical differences between the two populations. It should be noted that 40.75% of the observations resulted in inpatient treatment and the remaining 59.25% resulted in outpatient treatment. This accounts for the outpatient histograms having higher peaks in their histograms. There are also moderate correlations between laboratory test results that should be noted.

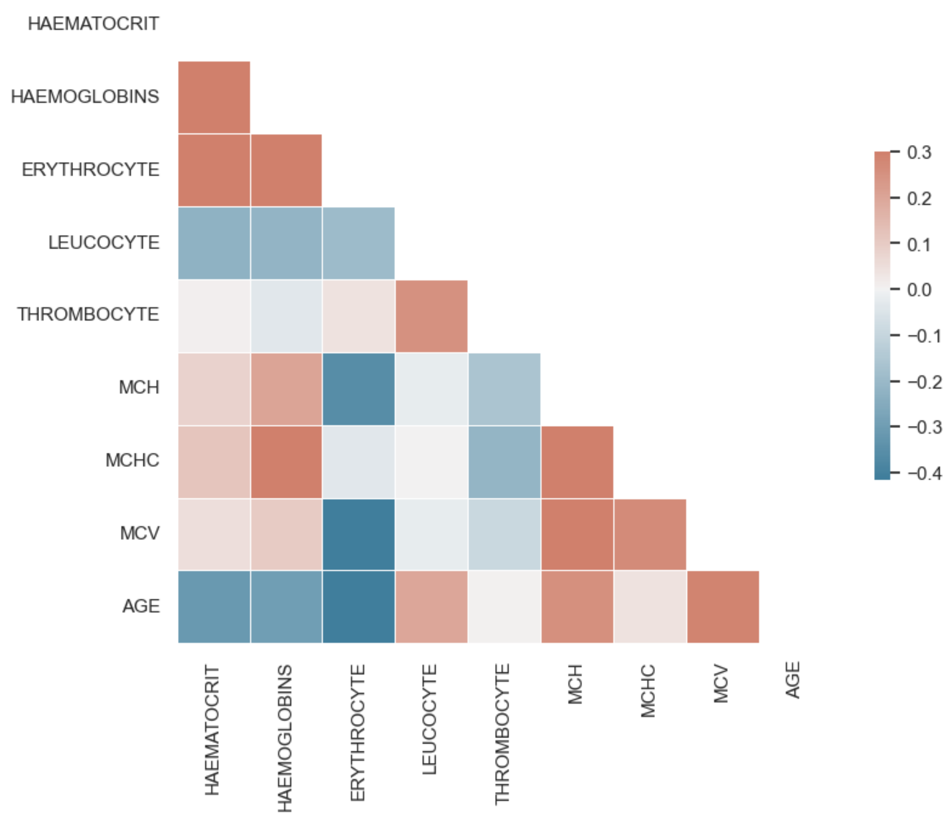

Figure 2: Heat map indicating correlations between laboratory test results

There are negative, moderate correlations with the Erythrocyte test and age, MCV and MCH. There are also positive, moderate correlations between many more tests such as Hematocrit with Hemaglobin and Erythrocyte, MCH with MCV and MCHC, and Hemaglobin with MCHC.

Four different classification models were fit to the data after the exploratory data analysis and preparations steps: K-Nearest Neighbors, Decision Tree, Random Forest, and Logistic Regression. With each model, the accuracy, precision, recall, and F1 scores were calculated to measure the effectiveness of the model. The model with the best accuracy was the Random Forest Classifier at 75.22&%. As mentioned earlier, 40.75% of the observations resulted in inpatient treatment and the remaining 59.25% resulted in outpatient treatment. Due to this imbalance, accuracy may not the best statistic to determine the best model. If the model only chose outpatient treatment, then the accuracy would still be almost 60% accurate. Instead, the F1 score is a better statistic to detemine which model fits best. The F1 score is a combination of precision and recall, and the best F1 score was also the Random Forest Classifier at 66.89%. Therefore, this was the best model at predicting the patient's next treatment setting.

This Random Forest Classifier's readiness for deployment would depend on it's intended use. If the provider would like it to competely make the decision for them, then it would not be ready. With an accuracy around 75%, that means that potentially 25% percent of their patients could be misclassified as outpatient when they should be inpatient. The patient would receive less care than they needed and could have consequences towards the patient's health and the hospital's rating. If the model is used simply as a tool to confirm what the provider was already thinking since ultimately they make the decision, the model would be ready. If the model does not match their intuition, the can override it and make the switch.

One recommendation that the model would benefit from is diagnosis or reason for the initial visit. Specific lab tests might be concerning based on the patient's diagnosis and warrant an inpatient treatment, whereas the same lab tests could warrant no further concerns due to a different diagnosis. Lab tests are also used along with imaging, therefore the imaging data would also be beneficial for the same reasons.