# Acute Liver Failure Patient Risk Factors

## Problem Statement and Background

Acute Liver Failure is a rare disease in which damage to the liver creates a loss of function. While there are many known diseases that cause ALF, treatment at the current time is limited and often involves transplant. Identifying patient demographic risk factors, may allow for earlier screening and prevention to improve patient outcome.

The aim of this model is to use this data set to find feature data most correlated with ALF. Modeling this will allow us not only to have a better understanding of risks related to ALF, but will also help predict a patient's predisposition to ALF when key indicators are considered.

## Data Wrangling

A Kaggle dataset was used in this study, which contained 8785 rows and 30 columns. A positive ALF status was marked with a 1, while the absence of ALF was notated by 0. It was found that 2785 instances in the dataset had no ALF status, these were removed to avoid error in later analysis. Missing data of continuous variables were filled with the mean of that variable.

Due to many features being correlated heavily with other, such as obesity and body mass index, it was decided to plot a heatmap of features with a max correlation of 0.8. No features were above this threshold with our target feature to begin with, so loss of function with the data model was not a concern.

## Exploratory Data Analysis

From the confusion matrix, features were selected for further analysis. Some features that had low correlation and no known clinical relevance were removed to simplify the model. The features no longer considered at this point in the analysis were: Height, Minimum Blood Pressure, Good Blood Pressure, Income, Education, Physical Activity, and Unmarried status.

Continuous features were visualized via boxplot against the target feature. Categorical features were visualized against ALF using a bar chart.

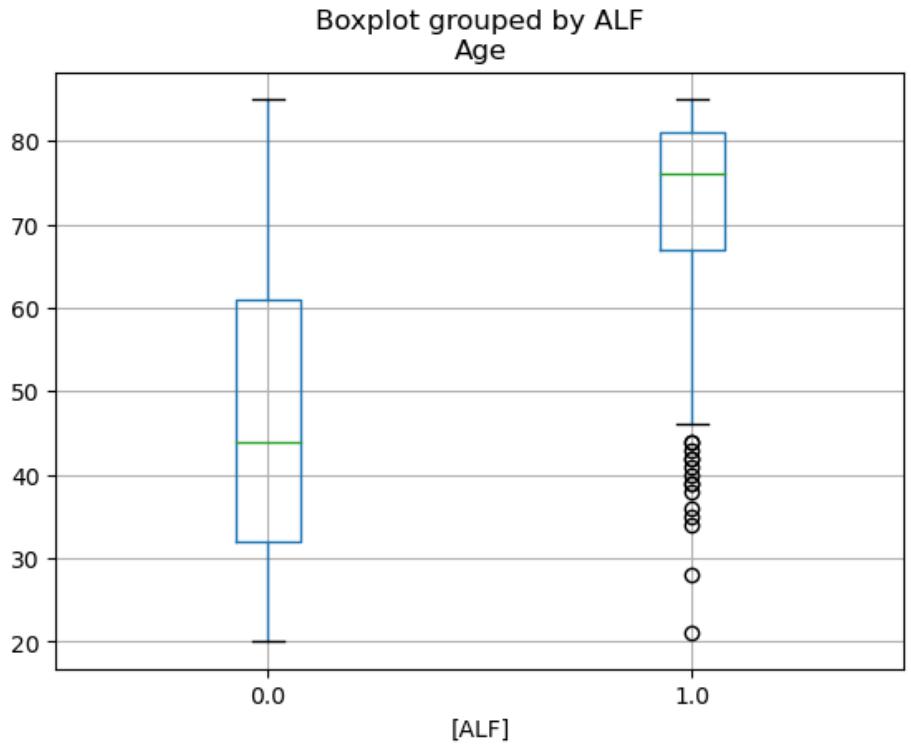


*Figure 1 shows a box plot of one of the selected continuous features(Age) vs the target feature ALF*

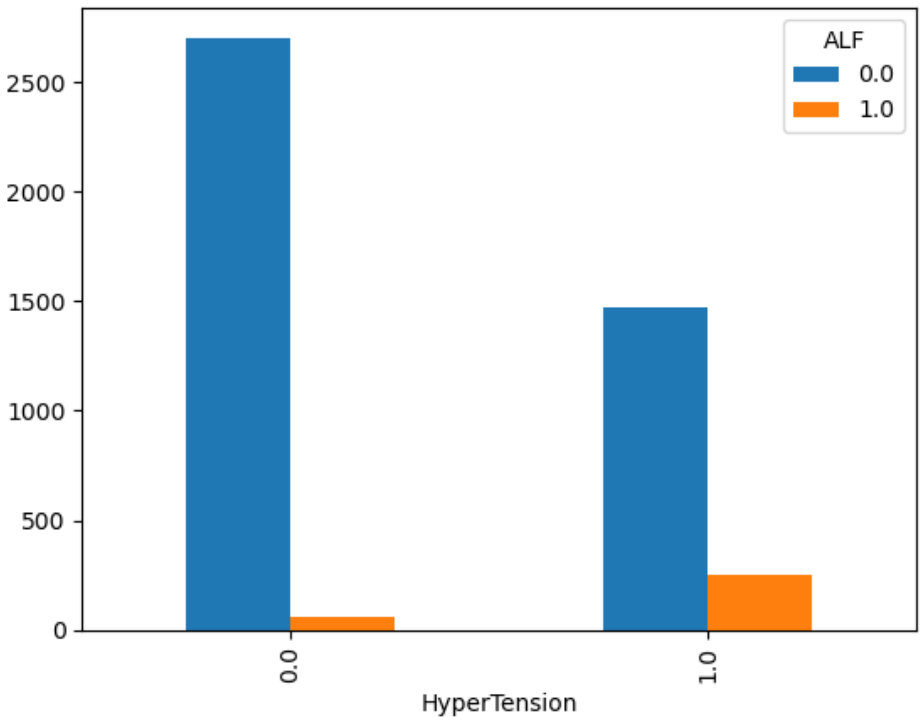


*Figure 2 shows a box plot of one of the selected categorical features (Hypertension) vs the target feature ALF*

Statistical significance between features and the target variable were also ran via chi-squared and t-test for continuous and categorical features, respectively.

| Categorical Feature | Chi-Sq P-Value |
|---|---|
| Obesity | 0.507297 |
| PVD | 7.4614e-33 |
| Dyslipidemia | 0.626036 |
| PoorVision | 2.69356e-09 |
| Alcohol Consumption | 0.000236681 |
| HyperTension | 4.64872e-56 |
| Family HyperTension | 0.0026312 |
| Diabetes | 6.52425e-24 |
| Family Diabetes | 0.255857 |
| Hepatitis | 5.71934e-39 |
| Family Hepatitis | 0.253568 |

*Figure 3 shows the chi-squared value of the categorical features and ALF*

| Distinct Feature | T-Test P-Value |
|---|---|
| Age | 8.37202e-141 |
| Weight | 0.549932 |
| BMI | 0.662988 |
| Waist | 2.26304e-05 |
| Maximum Blood Pressure | 3.2786e-47 |
| Bad Cholesterol | 0.0145368 |
| Total Cholesterol | 0.263 |

*Figure 4 shows the T-test values if the distinct features and ALF*

## Modelling

To model the data, 5 different supervised machine learning models were applied to the data. To preprocess the data, dummy variables were created on the categorical features selected: PVD, Poor Vision, Alcohol Consumption, Hypertension, Family Hypertension, Diabetes, Hepatitis. X was defined as the following features: Age, Waist, Obesity, Maximum Blood Pressure, Bad Cholesterol, PVD, Poor Vision, Alcohol Consumption, Hypertension, Family Hypertension, Diabetes, Hepatitis. ALF, the target feature, was assigned to the y variable. Sklearn's StandardScaler was used to fit and transform the X variable. Sklearn's train_test_split was then used to create a test-train split with a 20% test size.

The models selected for supervised classification were Decision Tree, Gradient Boost, K nearest neighbors, Gaussian, and Random Forest. The model with the best precision score for presence of ALF was also gradient boosting.

**Decision Tree**

Due to the low data of our target feature, ‘balanced’ was assigned to class_weight in the Decision Tree Classifier. Max depths were iterated by plotting accuracy against the training data and the cross-validation accuracy at tree depths ranging from 1 to 10. Training set accuracy was optimal at a max depth of 2, without sacrificing accuracy on mean cross validation.

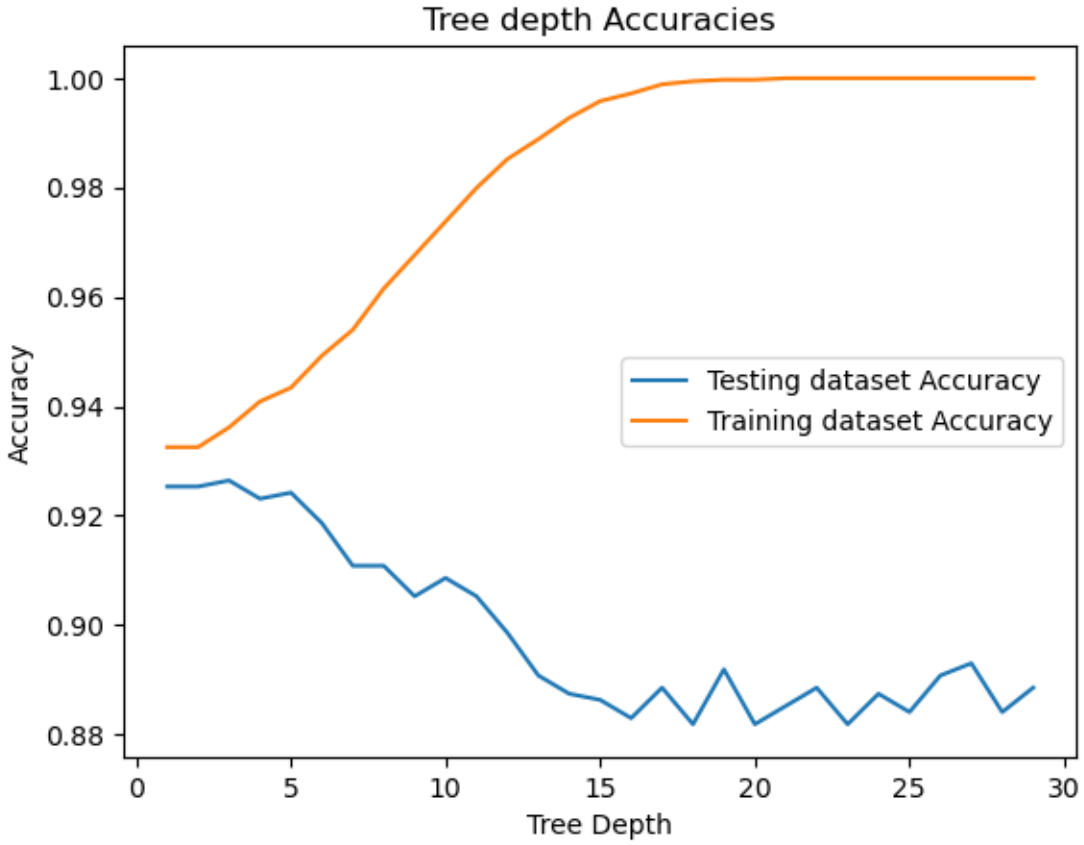


*Figure 5 shows tree depths plotted for train accuracy and mean cross validation accuracy*

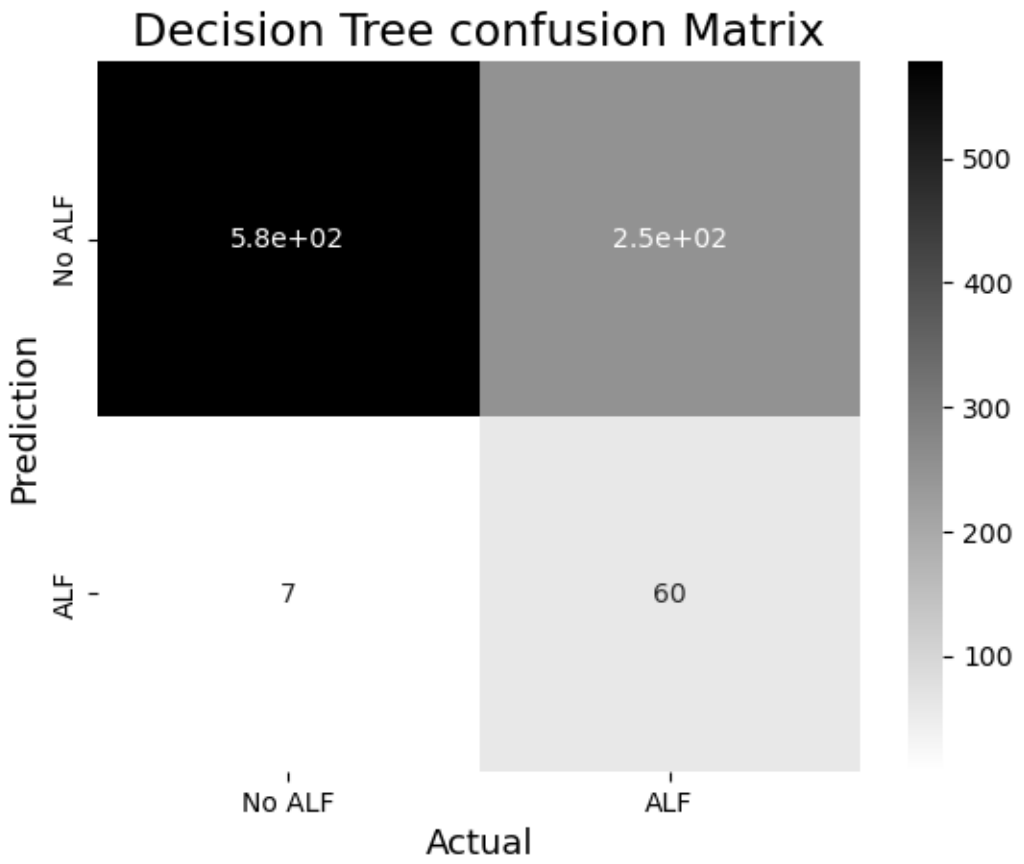


*Figure 6 shows the confusion matrix for the Decision Tree Model*

**Gradient Boost**

Gradient boosting, a boosting method that reduces bias by using the combination of several built estimators was used to model the data. This specific ensemble method was chosen due to the size of the dataset. Gradient boosting can be used as a classifier and is flexible in the many parameters that can be provided. Max depth was also specified on this model to minimize the effects of the dataset size, by implementing a slower learning rate. Max depth was also applied by using the same max depth iterator that was used in the decision tree model but substituting for gradient boost.

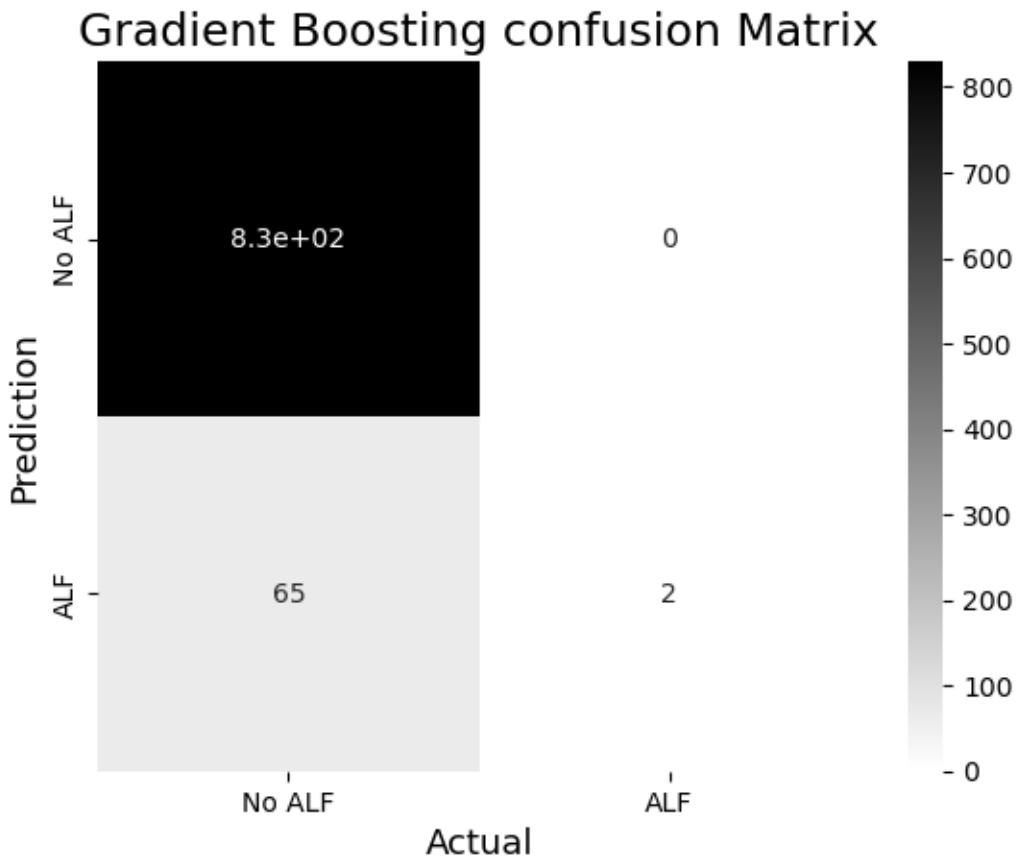


*Figure 7 shows the confusion matrix for the Gradient boosting Model*

**KNN**

Nearest neighbors' classification calculates the distance between data points in order to classify them. Setting the value of "k" when using k neighbors affects the model by imposing less noise/less distinct boundaries (high K value) or more noise/more distinct values. So, setting the K value affects predictiveness of model, and it is important to keep in mind if specificity or sensitivity is valued more. In the case of this project, specificity for our ALF target feature is more valued, as the aim is to value detecting ALF over being accurate overall.

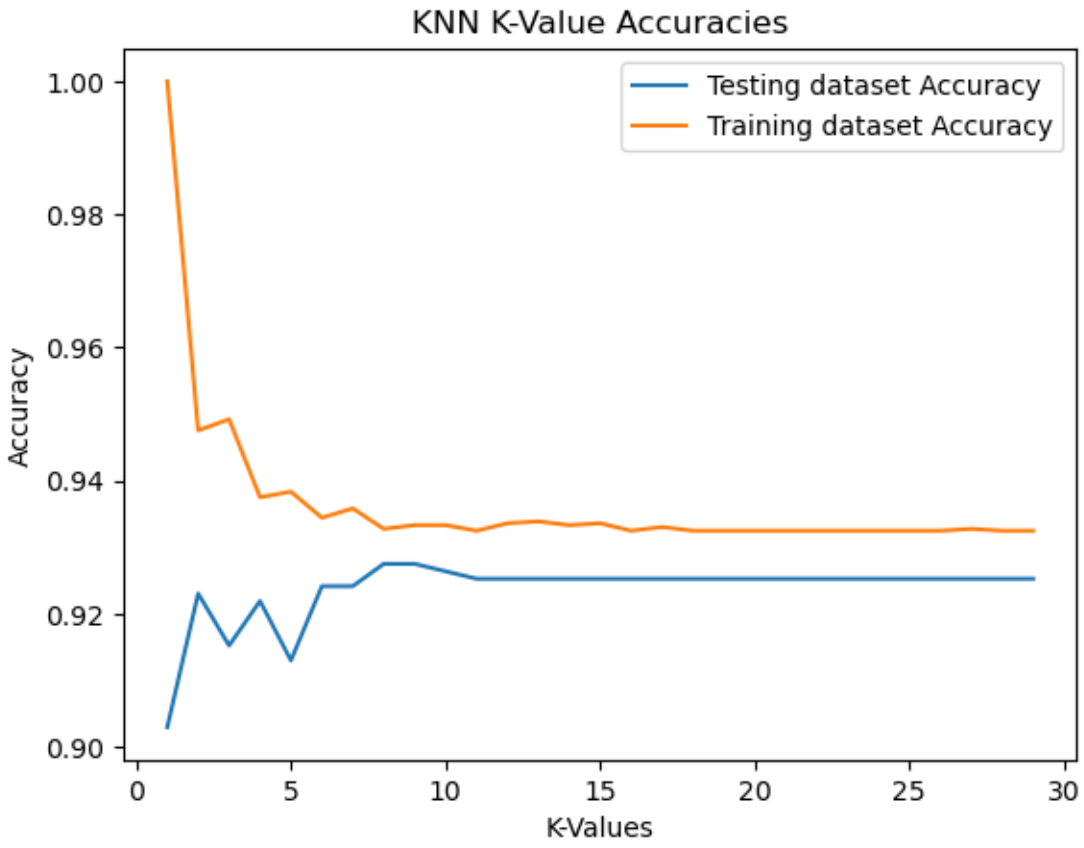


*Figure 8 shows K-Values and their accuracy on the training and test set*

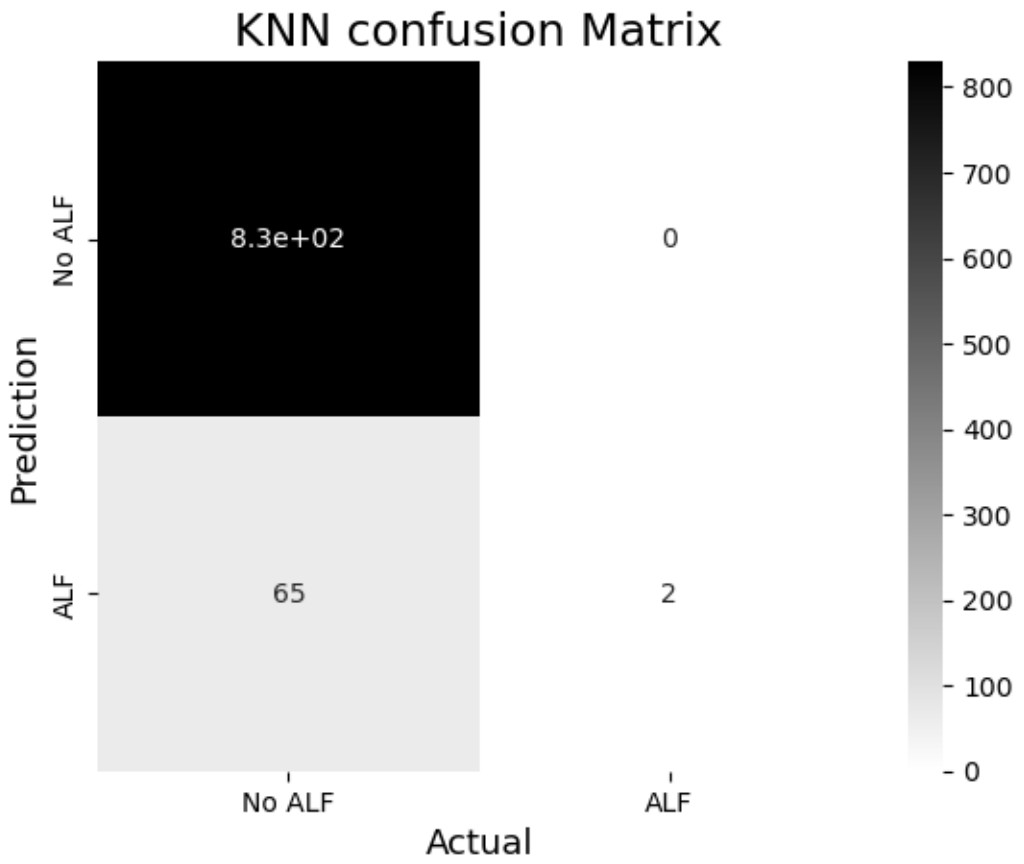


*Figure 9  shows the confusion matrix for the KNN model*

**Gaussian**

Many of the continuous features selected for in the dataset follow a normal distribution curve. Gaussian naïve bayes also follows a normal distribution and can be used for supervised learning, which is why it was chosen as a model to test in this study.

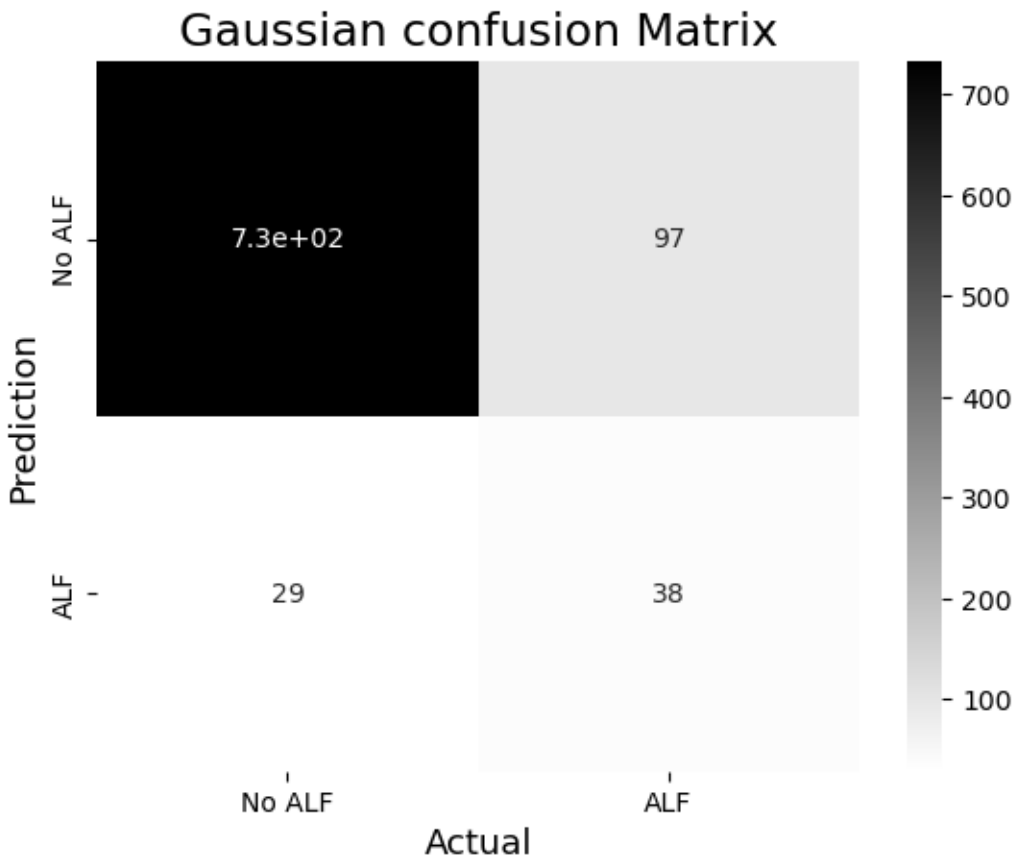


*Figure 10  shows the confusion matrix for the Gaussian Naives Bias Model*

**Random Forest**

The last supervised classification model ran on the dataset, was Random Forest. Random Forest is a decision tree classifier that utilizes averaging to prevent against overfitting of the data. Like other models tried in this project, Random Forest classifier has many hyperparameters that can be used to help tune the model. Bootstrapping, max depth, and class weight were specified for hyperparameter tuning.

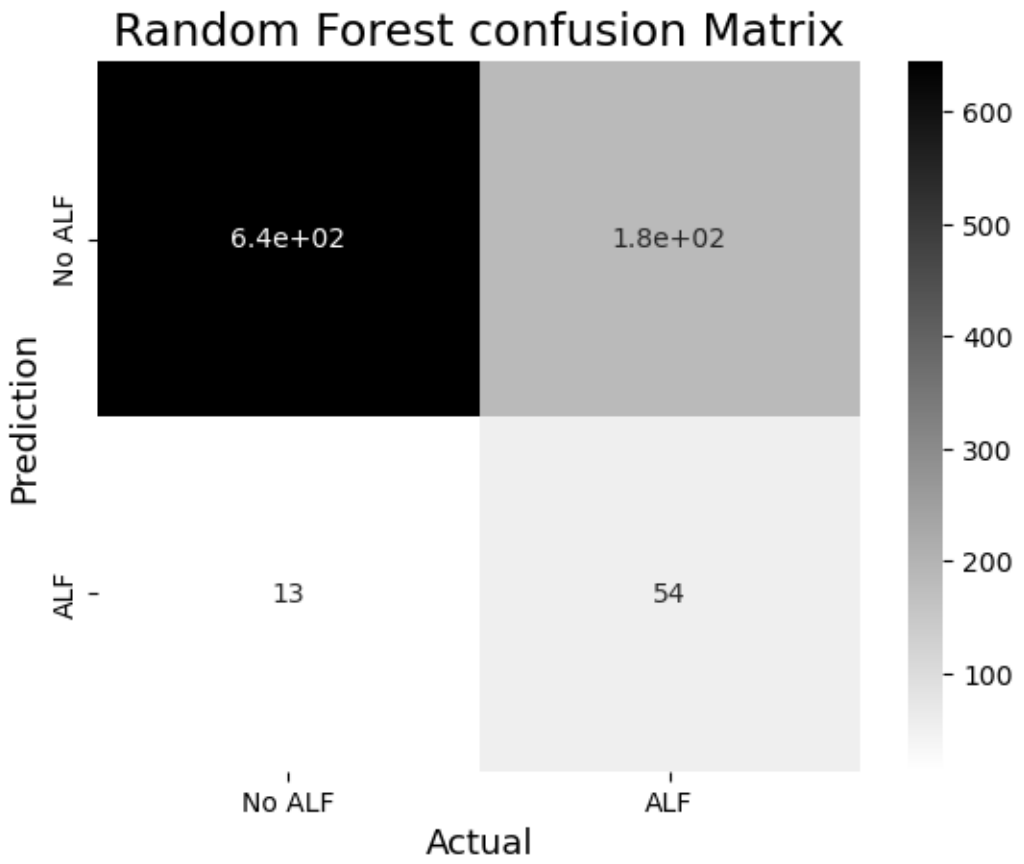


*Figure 11 shows the confusion matrix for the Random Forest Model*

All models were compared via accuracy and a classification report of their confusion matrix. Gradient boosting was the most accurate model ran with the dataset.

| Model | Accuracy scores |
|---|---|
| Decision Tree | 0.71126 |
| Gradient Boosting | 0.928651 |
| K Nearest Neighbors | 0.927536 |
| Gaussian | 0.859532 |
| Random Forest | 0.925307 |

*Figure 12 shows the accuracy of the individual models tested*

The **Decision Tree classification report** is:

| | **precision** | **recall** | **f1-score** | **support** |
|---|---|---|---|---|
| 0.0 | 0.99 | 0.70 | 0.82 | 830 |
| 1.0 | 0.19 | 0.90 | 0.32 | 67 |
| accuracy | | | 0.71 | 897 |

| | | | | |
|---|---|---|---|---|
| macro avg | 0.59 | 0.80 | 0.57 | 897 |
| weighted avg | 0.93 | 0.71 | 0.78 | 897 |

The **Gradiant Boost classification report** is:

| precision | recall | f1-score | support |
|---|---|---|---|
| 0.0 | 0.94 | 0.99 | 0.96 | 830 |
| 1.0 | 0.57 | 0.19 | 0.29 | 67 |
| accuracy | | | 0.93 | 897 |
| macro avg | 0.75 | 0.59 | 0.63 | 897 |
| weighted avg | 0.91 | 0.93 | 0.91 | 897 |

The **KNN classification report is:**

| **precision** | **recall** | **f1-score** | **support** |
|---|---|---|---|
| 0.0 | 0.93 | 1.00 | 0.96 | 830 |
| 1.0 | 1.00 | 0.03 | 0.06 | 67 |
| accuracy | | | 0.93 | 897 |
| macro avg | 0.96 | 0.51 | 0.51 | 897 |
| weighted avg | 0.93 | 0.93 | 0.89 | 897 |

The **Gaussian classification** report is:

| **precision** | **recall** | **f1-score** | **support** |
|---|---|---|---|
| 0.0 | 0.96 | 0.88 | 0.92 | 830 |
| 1.0 | 0.28 | 0.57 | 0.38 | 67 |
| accuracy | | | 0.86 | 897 |
| macro avg | 0.62 | 0.73 | 0.65 | 897 |
| weighted avg | 0.91 | 0.86 | 0.88 | 897 |

The **Random Forest classifiaction** report is:

| **precision** | **recall** | **f1-score** | **support** |
|---|---|---|---|
| 0.0 | 0.93 | 0.99 | 0.96 | 830 |
| 1.0 | 0.50 | 0.07 | 0.13 | 67 |
| accuracy | | | 0.93 | 897 |
| macro avg | 0.72 | 0.53 | 0.55 | 897 |
| weighted avg | 0.90 | 0.93 | 0.90 | 897 |

## Conclusion and Future Research

Datasets with target features being a minority in the dataset express difficulty in classification. While the model may have a high accuracy, precision in these cases outweighs accuracy in importance. The goal is to create a model that can predict for acute liver failure based on patient features. The risk of a false positive is minimal compared to the detriment of a false negative.

Due to the nature of this dataset, no model was impressive in terms of predicting ALF through the Given features. KNN and Gradient Boosting had the highest precision for ALF, but this is likely due to the lack of ALF targets in those models. The Gaussian model had more ALF Targets, but only a precision of 0.28 for ALF. The Decision Tree and Random Forest models allowed for balancing of class weights and max depth specification. Even with the allowance of ALF targets in the range of the non-ALF quantity, these models failed to give a precise ALF score.

After reviewing these results, it is concluded that the dataset is not comprehensive enough to accurately predict risk factors of ALF. More positive ALF data is needed to improve model precision and performance. Future models and methods should focus on balancing within datasets and possibly adding features.

The application of machine learning in healthcare is becoming more frequent. Teams with a wide volume of patient data and testing should consider machine learning to assist in early detection in order to improve patient outcomes. With this rapidly growing technology, it is important to assess whether a data set and model are actually presenting real-world findings, or if overfitting and tuning are allowing for a connection to be made when there is not one.