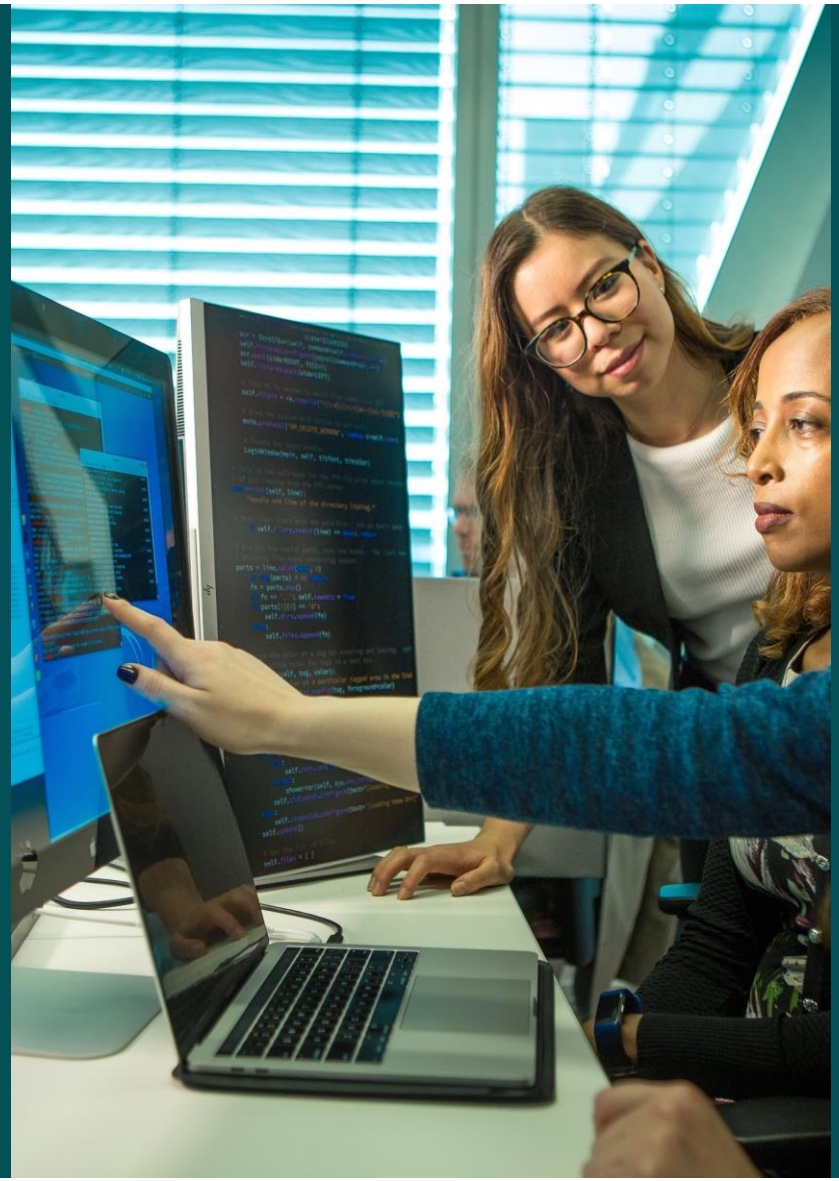
Booz | Allen | Hamilton®



TECH EXCELLENCE DATA SCIENCE PROGRAM CAPSTONE

By: Lauren Bryant and Enneye Makonnen

DECEMBER 2024

AGENDA

MEET THE TEAM

PROJECT GOAL AND OBJECTIVE

METHODS

RESULTS

NEXT STEPS

QUESTIONS

MEET THE TEAM

Lauren Bryant is a Senior Consultant with five years of experience in federal healthcare consulting, healthcare analytics, project management, and quality improvement initiatives. She currently provides portfolio operations and data integration support to the VA Health Connect portfolio. Lauren has led cross-functional teams to implement successful programs and conduct mixed-methods program assessments, driving evidence-based decision-making for the Veterans Health Administration and commercial health strategy. She holds a Master of Public Health degree from Tulane University and a B.A. in Government with a minor in Bioethics from the University of Virginia.

Enneye currently serves as a Healthcare Data Scientist, where she provides analytical support and collaborates with clients to identify gaps and deliver innovative solutions that strengthen the medical device supply chain. She is passionate about improving the quality and safety of global health artificial intelligence (AI) technology and enhancing AI strategy in public service business operations. Enneye's background also includes nine years as a registered nurse and public health professional, enriching her expertise in global health and patient care.

OVERVIEW

- **Business Question:**
  - What health and/or demographic factors (e.g., age, underlying conditions, gender) contribute to the increased risk of ICU admission for hospitalized COVID-19 patients?

- **Objective:**
  - Apply at least two classification models to predict the likelihood of ICU admission in hospitalized COVID-19 patients, utilizing demographic data, health status, and relevant risk factors.

- **Machine Learning Techniques:**
  - **Decision Trees** - Comprises assigning a class label to a set of unclassified datasets
  - **Logistic Regression** - Deals with situations in which the outcome for a dependent variable can have two possible values (e.g., yes/no, passed/not passed)

SUMMARY OF DATASET

- **Covid-19 Dataset:**
  - Covid-19 patient dataset provided by Mexican government
  - 1,048,576 unique, anonymized patients
  - Most features (exceptions: age, classification final, date died) were Boolean with 1 meaning “yes”, and 2 meaning “no” (98 and 99 represented missing data)
  - 1-3 in classification final meant the patient was diagnosed with covid-19. 4+ meant the test was negative or inconclusive.
- **21 Unique Features:**
  - Patient Demographic (e.g., age, sex)
  - Risk Factors (e.g., tobacco, obesity)
  - Health Condition (e.g., Pneumonia, COPD, Hypertension, etc.)
  - Hospital Status (e.g., ICU, Patient type, Medical Unit, date died, etc.)

METHODS

- **Business Question**
  - What health and/or demographic factors contribute to higher risk for admission to **ICU** in **hospitalized Covid-19** patients?
- **Target Variable:** ICU
- **Predictive Features:**
  - All except date died, medical unit, USMER (Viral Respiratory Disease Monitoring Units), and intubation
    - Irrelevant or highly correlated with ICU admission
- **Data Preparation**
  - Removed Classification_final 4 or greater (negative/unclear Covid diagnosis)
  - Removed Patient Type = 1 (returned home)
  - Transformed Pregnant = 98 (all males) to Pregnant = 2 (not pregnant)
  - Removed rows with missing data (98 or 99)
  - 108,160 rows and 17 columns remained

```
data['SEX'].value_counts() #1 = female, 2 = male
```

```
SEX
1    525064
2    523511
Name: count, dtype: int64
```

```
male_97preg = len(data[(data['SEX']==2) & (data['PREGNANT'] ==97)])
print(male_97preg)
```

```
523511
```

```
data['PREGNANT'] = data['PREGNANT'].replace(97,2)
data['PREGNANT'].value_counts()
```

```
PREGNANT
2     1036690
1        8131
98       3754
Name: count, dtype: int64
```

METHODS

- **Data Preparation continued**
  - Random Under Sampling:
    - Randomly sampled the majority class to match the number of instances in the minority class, improving model performance without overfitting
  - Randomized the dataset:
    - If the data is not randomly sorted, then this may lead to a poor model
  - Split the data into a training and a testing set:
    - Used 90% for training, 10% for testing

```
ICU
2    0.905649
1    0.094351
Name: count, dtype: float64
```

```
from imblearn.under_sampling import RandomUnderSampler
rus = RandomUnderSampler(random_state=42)
x_resampled, y_resampled = rus.fit_resample(x, y)
```

```
y_resampled.value_counts()/y_resampled.count()
```

```
ICU
1    0.5
2    0.5
Name: count, dtype: float64
```

The data is now balanced

METHODS

Logistic Regression – Additional Steps:

- **Scaled Age**:
  - Apply scaling to the "age" feature to ensure it has a mean of 0 and a standard deviation of 1 (or scaled to a range between 0 and 1).

- **Checked for Multicollinearity:**
  - Used variance inflation factor (VIF) to identify correlated variables and removed or combined them to ensure model stability.

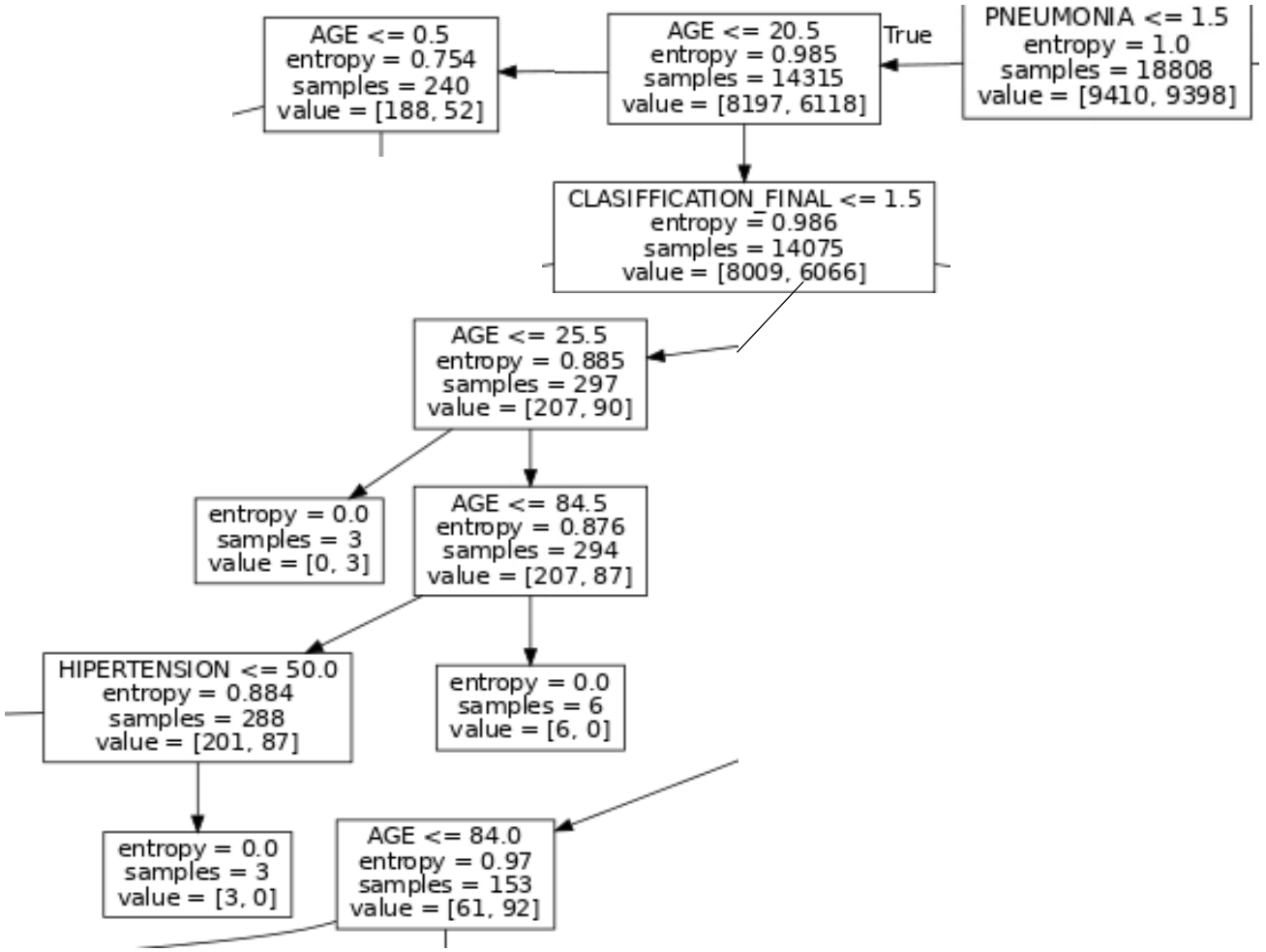
Decision Tree – Additional Steps:

- **Determined Feature Importance:**
  - Ran feature importance analysis to pinpoint the most relevant features, which allows for simplifying the model, improving interpretability, and potentially enhancing accuracy.

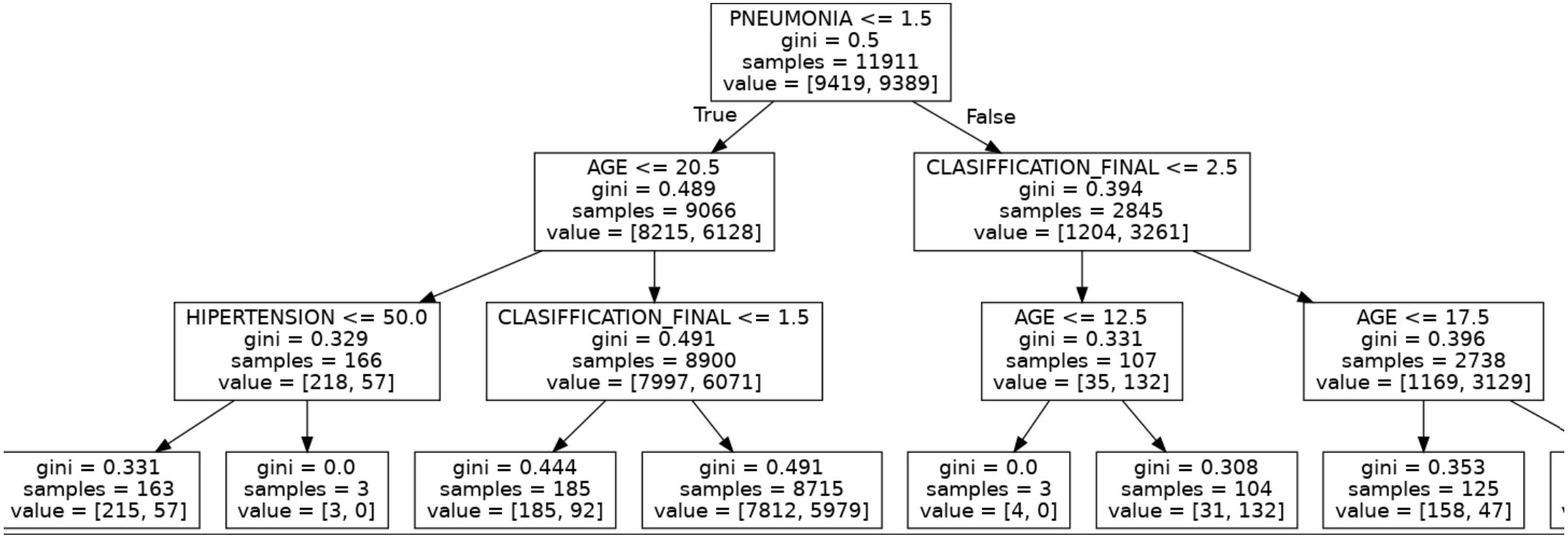
	Important
PNEUMONIA	0.439997
AGE	0.425737
HIPERTENSION	0.060169
DIABETES	0.048537
CLASIFFICATION_FINAL	0.025560
PATIENT_TYPE	0.000000

RESULTS: DECISION TREES

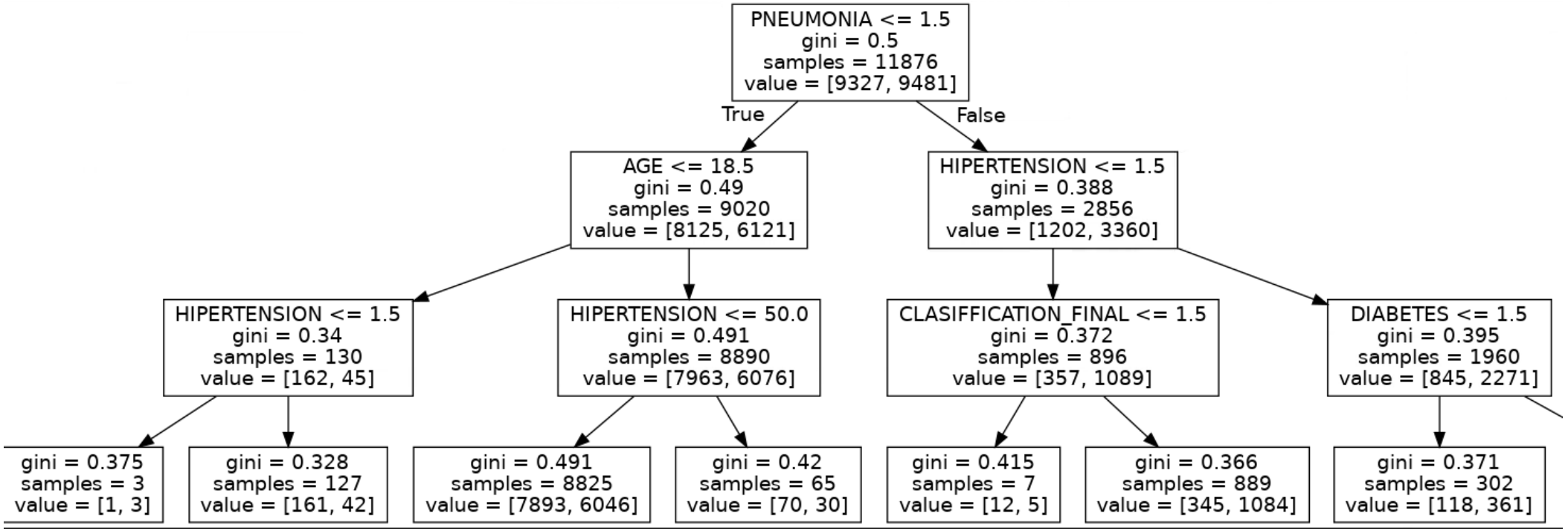
- Pneumonia and age were the most significant factors contributing to the risk of ICU admission in hospitalized COVID-19 patients.
- High entropy values (0.80 - 0.98) indicating lower purity in the target value of the child nodes, affecting the decision tree's ability to make accurate splits.
- Accuracy may not be as important.
- Aggregated multiple decision trees into a Random Forest model to prevent overfitting, enhance generalization, and better capture non-linear relationships.
- Used Gini for better performance metrics.



RESULTS: DECISION TREES



RESULTS: DECISION TREES



RESULTS: LOGISTIC REGRESSION

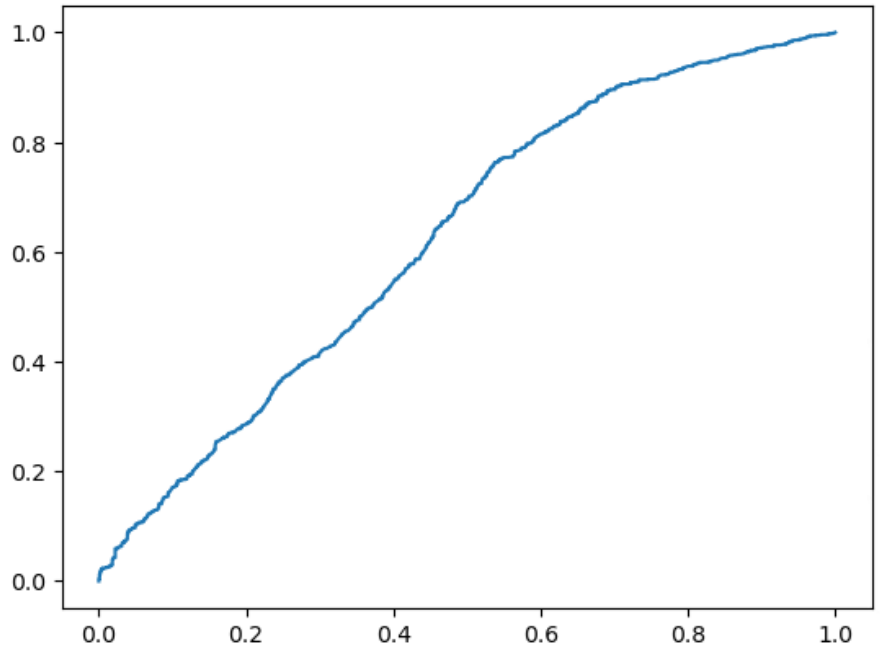
- **Area Under the ROC Curve (AUC) = 0.6**
  - True Positive Rate/False Positive Rate
  - Slightly better than random guessing
- **Precision = 0.84**
  - Positive predictive value:
    - True positive / All positives
  - "Of all the instances predicted as positive, how many are actually positive?“
  - Important when the cost of false positives is high
- **Recall/Sensitivity = 0.56**
  - True Positive / (True Positive + False Negative)
  - "Of all the actual positive instances, how many were correctly identified as positive?“
  - Important when the cost of false negatives is high
- **Mean Square Error (MSE) = 0.40**
- Tried decreasing number of features, which caused the model to perform worse

```
roc_auc_score(y_test, logisticRegr.predict(x_test))
```

```
0.6041294139113232
```

```
fpr,tpr,thresholds = roc_curve(y_test,logisticRegr.predict_proba(x_test)[:,0],pos_label=1)
plt.figure()
plt.plot(fpr,tpr)
```

```
[<matplotlib.lines.Line2D at 0x74e4648e43d0>]
```



DISCUSSION AND NEXT STEPS

- Explore other potential factors that may contribute to risk of ICU admission (e.g., community-acquired infections, environmental exposure, travel)
- Presence of pneumonia and age < 20 as significant contributors to ICU admission suggest this group may differ from the general population

- **Ideas for Additional Research:**
  - Specific age groups (Infants, Children, Adults, Elderly)
  - Removing/changing features used
  - Adding new features

QUESTIONS?