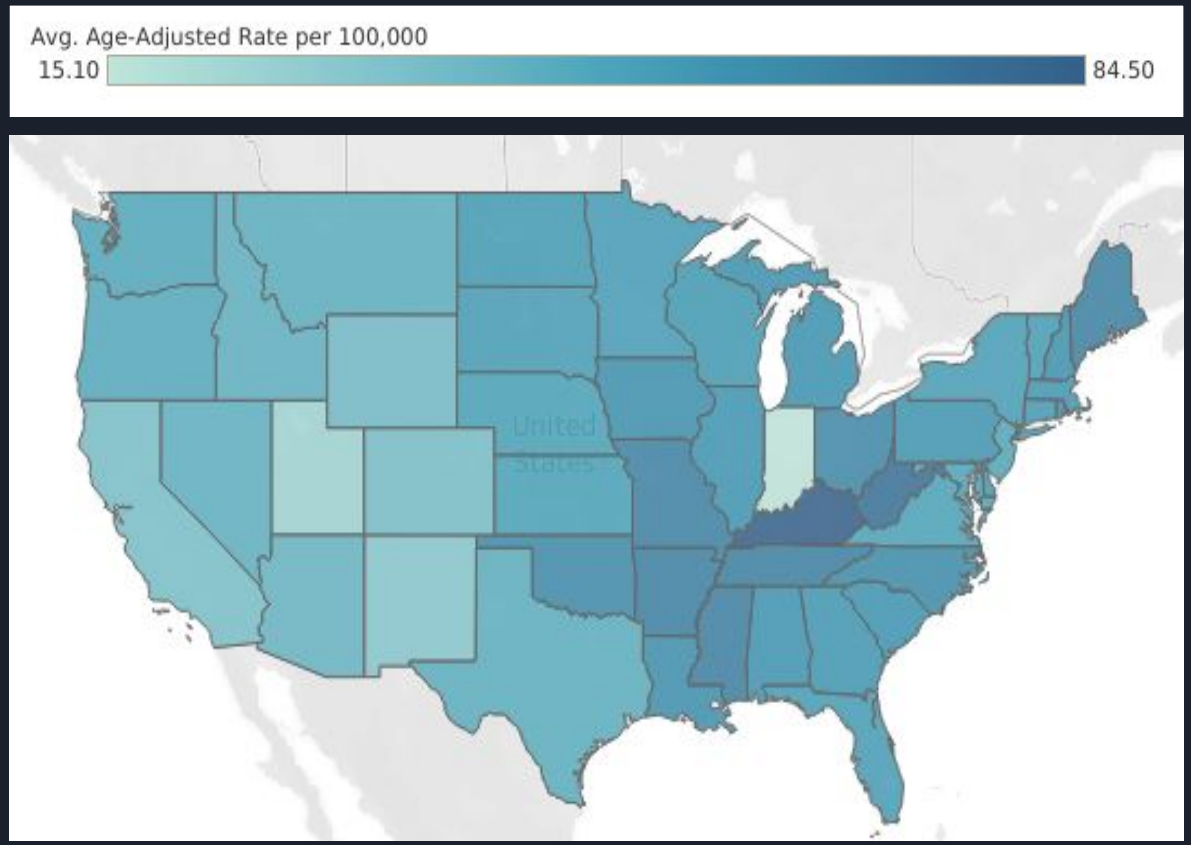
Exploring Lung Cancer Using Machine Learning Models:

Early Detection Methods for Non-Small Cell Lung Cancer

Nefertiti Muhammad, Adrian De La Cruz, Joe Coffaro, Alexander Walden



SOUTHEASTERN U.S.
Highest Rates: Kentucky (84.5 per 100,000) and West Virginia (76.1 per 100,000)

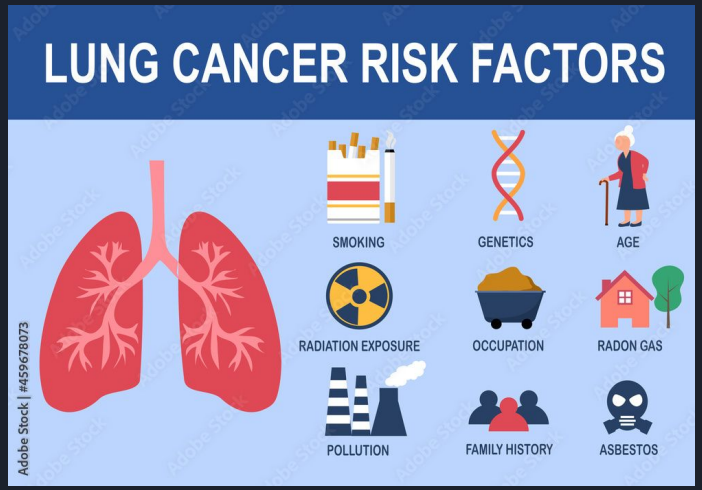
WESTERN U.S.
Lowest Rates: Utah (24.9 per 100,000) and New Mexico (32.7 per 100,000)

Geographic differences may reflect differences in risk factors such as smoking prevalence, access to healthcare, and pollution

Lung Cancer Incidence by State (2017-2021)

Lung Cancer Statistics and Risk Factors

- Leading Cause of Cancer Deaths in the U.S. (~25%)
- Second after breast cancer in prevalence in U.S
- Two types: Non-Small Cell Lung Cancer (NSCLC) (80-85%) and Small Cell Lung Cancer (SCLC) (10-15%)
- Risk Factors: smoking (~85% of cases), secondhand smoke, pollution, and occupational hazards
- Often diagnosed at advanced stages
- Has a low 5-year survival rate (~21%)



Hypothesis: Smoking history along with other diagnostic factors can provide a ML model the data to predict survival.

Datasets & Methodology

CODING LANGUAGES/LIBRARIES

- Pandas
- Matplotlib
- Tensorflow
- Scikit Learn
- Tableau

MACHINE LEARNING APPROACH

- Logistic Regression
- Neural Network
- Random Forest Classifier

DATA SETS

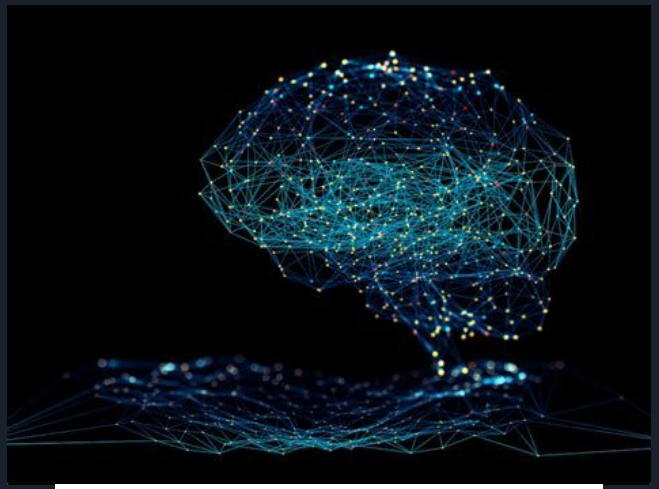
https://www.countyhealthrankings.org/health-data/health-factors/health-behaviors/tobacco-use/adult-smoking?year=2024

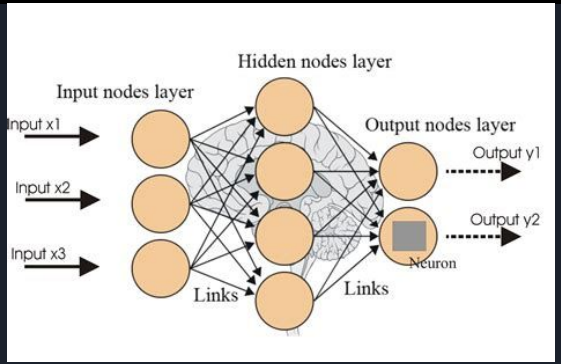
NSCLC TCGA Broad 2016 Dataset

https://gis.cancer.gov/canceratlas/tableview/?d=1&a=1&r=1&s=33

Neural Network Model

- Cleaned data; coded string and object values as numeric, selected our preferred target and features
- Used Keras tensorflow to create a deep learning model
- Target variable: survival coded as ‘0’: Living and ‘1’: Deceased
- Scikitlearn library: training/splitting and scaling
- Compiled model with loss, activation, and accuracy metrics





NN Model (cont.)

- Used Keras tuner methods to find optimized combination of parameters
- Structure: layer 1 with 9 neurons, two with 5, three with 3, four with 9 and output layer with 1
- Activation function: tanh
- Fit model with 20 epochs
- Best model accuracy: 73%, short of the threshold we sought

```python
# Create a method that creates a new Sequential model with hyperparameter options
def create_model(hp):
    nn_model = tf.keras.models.Sequential()

    # Allow kerastuner to decide which activation function to use in hidden layers
    activation = hp.Choice('activation',['relu','tanh','sigmoid'])

    # Allow kerastuner to decide number of neurons in first layer
    nn_model.add(tf.keras.layers.Dense(units=hp.Int('first_units',
        min_value=1,
        max_value=10,
        step=2), activation=activation, input_dim=len(X_train_scaled[0])))

    # Allow kerastuner to decide number of hidden layers and neurons in hidden layers
    for i in range(hp.Int('num_layers', 1, 4)):
        nn_model.add(tf.keras.layers.Dense(units=hp.Int(f'units_{i}',
            min_value=1,
            max_value=10,
            step=2),
            activation=activation))

    nn_model.add(tf.keras.layers.Dense(units=1, activation="sigmoid"))

    # Compile the model
    nn_model.compile(loss="binary_crossentropy", optimizer='adam', metrics=["accuracy"])

    return nn_model
```

✓ 0.0s

```python
# Import the kerastuner library
import keras_tuner as kt
# Define method to create tuner instance
def run_tuner(epochs):

    tuner = kt.Hyperband(
        create_model,
        objective="val_accuracy",
        max_epochs=30,
        hyperband_iterations=2)
    tuner.search(X_train_scaled,y_train,epochs=30,validation_data=(X_test_scaled,y_test))
    return tuner
```

✓ 0.0s

Random Forest Classifier

- Random forest classification is one alternative ML model, might better suit features
- Created an instance of a random forest classifier
  - Fit a model according to the training data
  - Made predictions and evaluated accuracy
- Yielded 76% accuracy, predicting more reliably than NN model

```
[129]
# Create a random forest classifier
model = RandomForestClassifier(n_estimators=300, random_state=78)
# Model fitting
model = model.fit(X_train_scaled, y_train)
```

```
[131]
rf_predictions = model.predict(X_test_scaled)
```

```
[66]
# Calculating the accuracy score
acc_score = accuracy_score(y_test, rf_predictions)
```

```
[133]
print("Classification Report\n")
print(f"Accuracy score: {acc_score}")
print(classification_report(y_test,rf_predictions))
```

```
...  Classification Report

Accuracy score: 0.7027972027972028
              precision    recall  f1-score   support

      0:LIVING       0.77      0.94      0.85       175
    1:DECEASED       0.70      0.32      0.44        71

    accuracy                           0.76       246
   macro avg       0.74      0.63      0.65       246
weighted avg       0.75      0.76      0.73       246
```

Regression Model

- Dataset cleaned by removing any null data and replacing binary string factors with boolean values (i.e. 0 or 1)
- Used LogisticRegression with
  - solver as ‘lbfgs’
  - max iterations at 1,200
- Results:

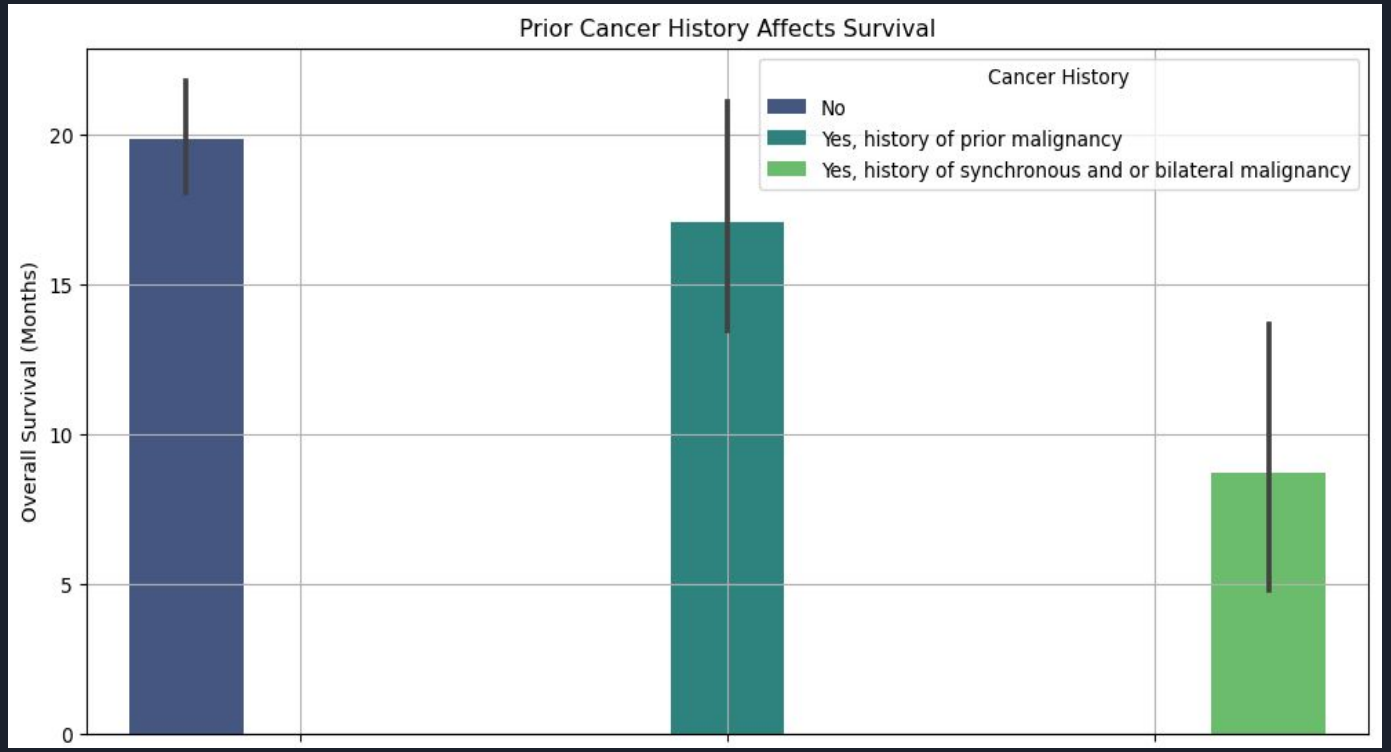
```
[[121   0]
 [ 61   0]]
              precision    recall  f1-score   support

   0:LIVING       0.66      1.00      0.80       121
  1:DECEASED      0.00      0.00      0.00        61

    accuracy                           0.66       182
   macro avg       0.33      0.50      0.40       182
weighted avg       0.44      0.66      0.53       182
```

- Opportunities to improve precision
  - Only using the factors of
    - Smoking History
    - Prior Cancer Diagnosis Occurrence
  - Larger, more comprehensive dataset
- Final precision for surviving prediction increased to 0.74 by using two factors above

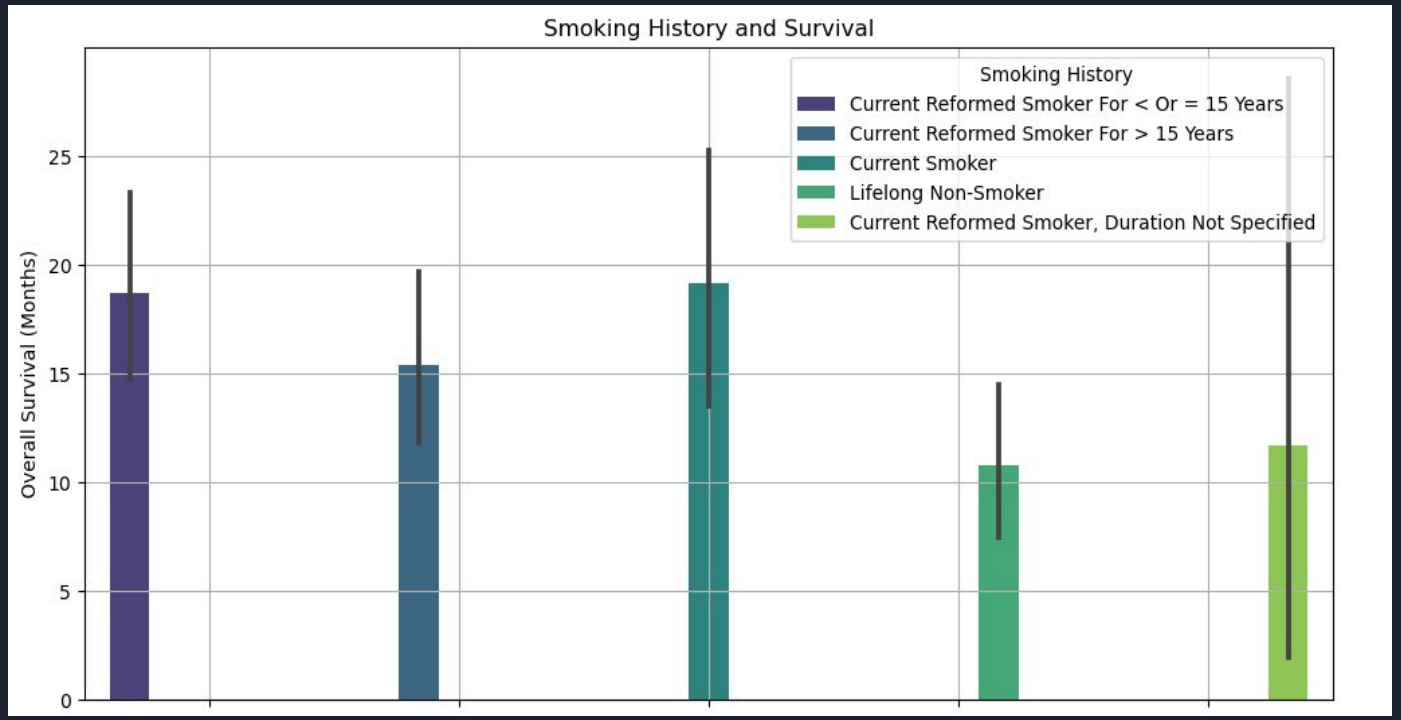
Factors Influencing Survival Prediction



ANOVA p-value (0.1010)

There are no significant differences between the groups. However, patients with cancer in both lungs trend downwards in survival

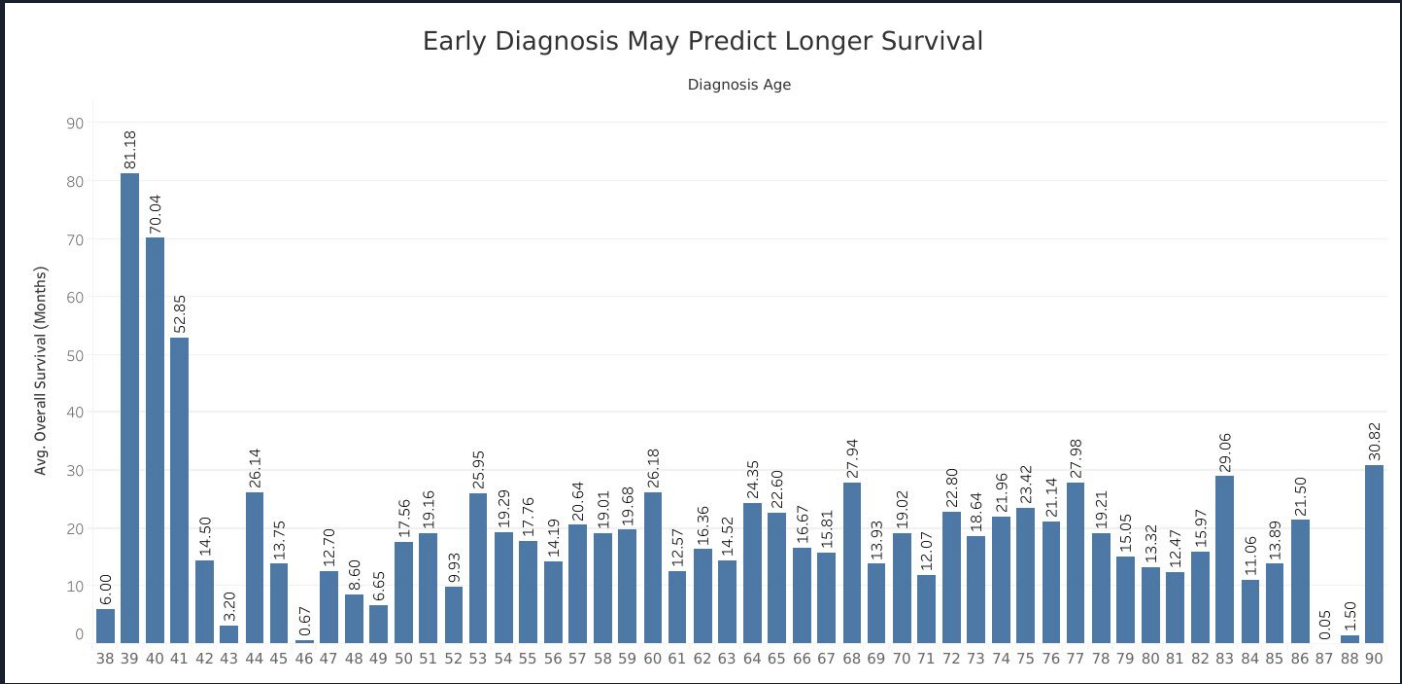
Factors Influencing Survival Prediction



ANOVA results (p-value: 0.20769)
There are no significant differences between the smoking history groups in relation to survival

This points to limitations of the dataset and the complexity of predicting survival outcomes

Factors Influencing Survival Prediction



ANOVA results (p-value: 0.4633) don’t show a strong relationship between age of diagnosis to overall survival, but patients with higher survival rates tend to be younger

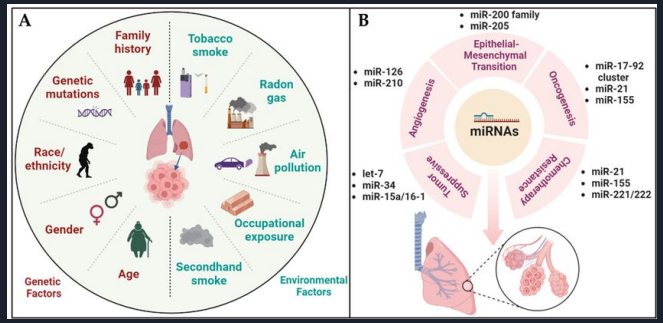
ML Process Problems/Concerns

DATA LIMITATIONS

- Size of dataset for fitting by machine learning models
- Null values might obfuscate trends
- Failure to capture more socially complex factors
- Might be harder to standardize dataset

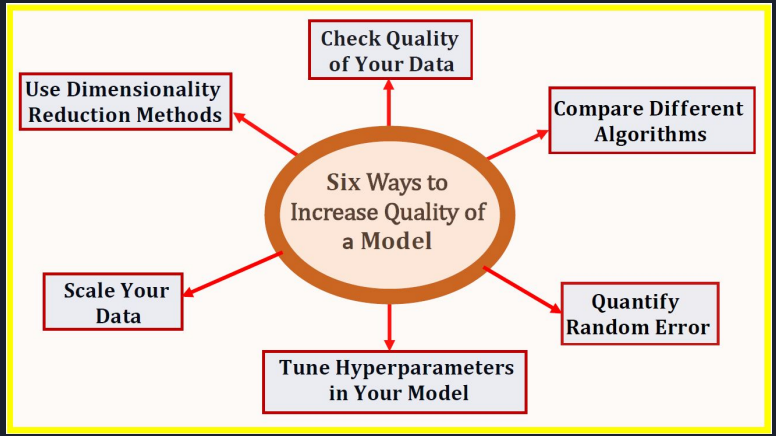
NAMELY: COMPLEXITY OF LUNG CANCER

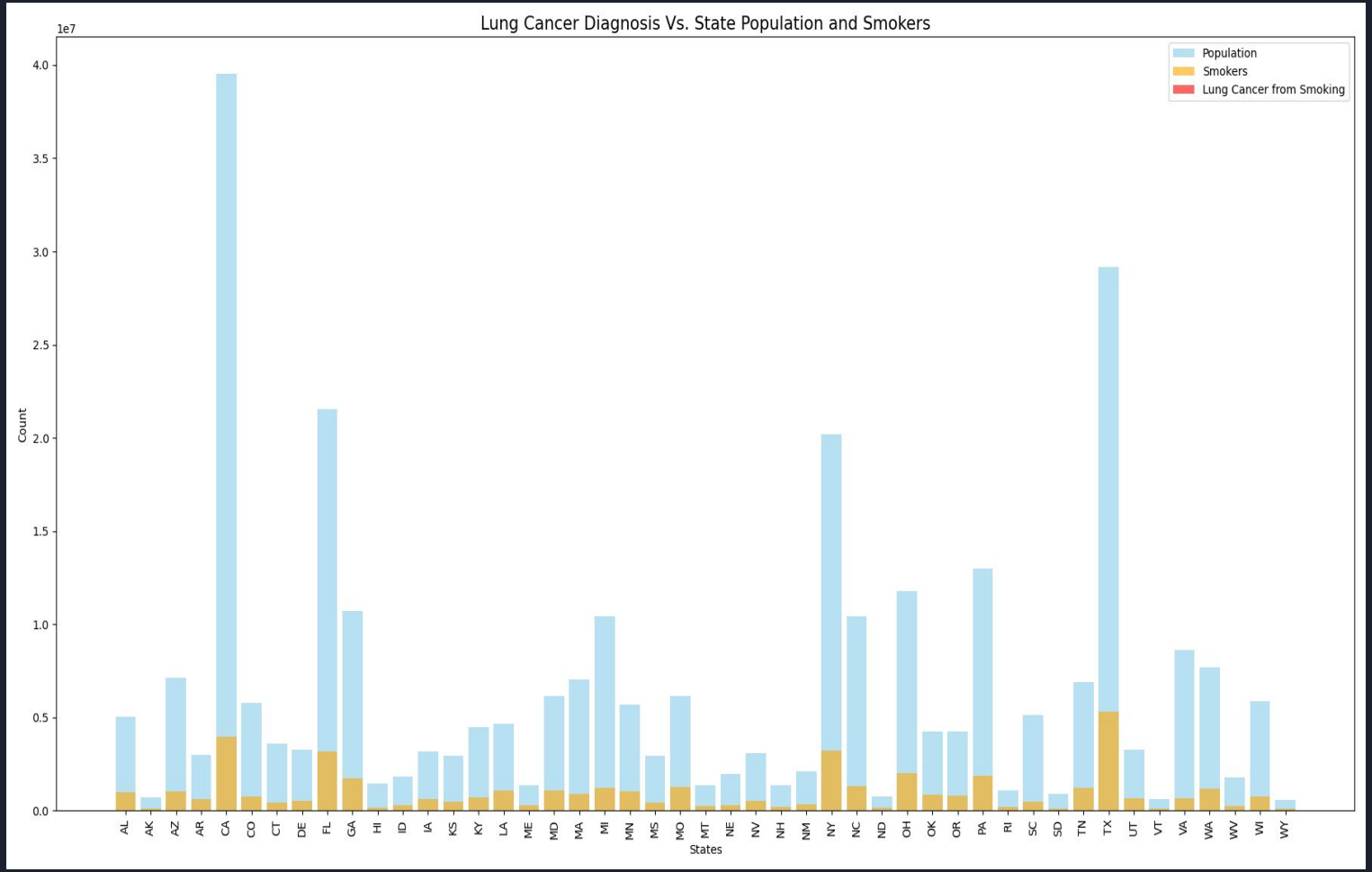
- Environmental factors
- Non-smoker frequency
- Genetics



How to Increase Efficacy

- Size and quality of dataset important for creating robust models
- Better tuning is possible with these elements of data
- Comparing other models to one another
- Monitoring and evaluating metrics
- Further domain (healthcare) knowledge for better insights on relevant features




Lung Cancer Diagnosis Vs. State Population and Smokers
1e7
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Count
Population
Smokers
Lung Cancer from Smoking
AL AK AZ AR CA CO CT DE FL GA HI ID IA KS KY LA ME MD MA MI MN MS MO MT NE NV NH NM NY NC ND OH OK OR PA RI SC SD TN TX UT VT VA WA WV WI WY
States

Limitations, Solutions, Further Studies

Limitations:

Why is the graph showing no data for Lung Cancer Diagnosis from smoking?

- Population by state is so large and the number of smokers compared to the population does not give an accurate representation of lung cancer diagnosis

Solutions and Further Studies:

- Break down the data even further, look into population per city or region to get a closer look and more accurate result of lung cancer diagnosis as a result to smoking

- This data looks at lung cancer diagnosis specifically from the cause of smoking, smoking is not the only way one person can get lung cancer.

- Another reason why the graph was unable to capture Lung Cancer Diagnosis is because the number of cases is so low compared to total population

Conclusion

- Larger dataset would better fit prediction models
- Investigating this question further, perhaps with other forms of cancer
- Multiple factors to lung cancer incidence not covered
  - Environmental pollution
  - Secondhand smoke
  - Chemical exposure
  - Workplace hazards
  - Family history