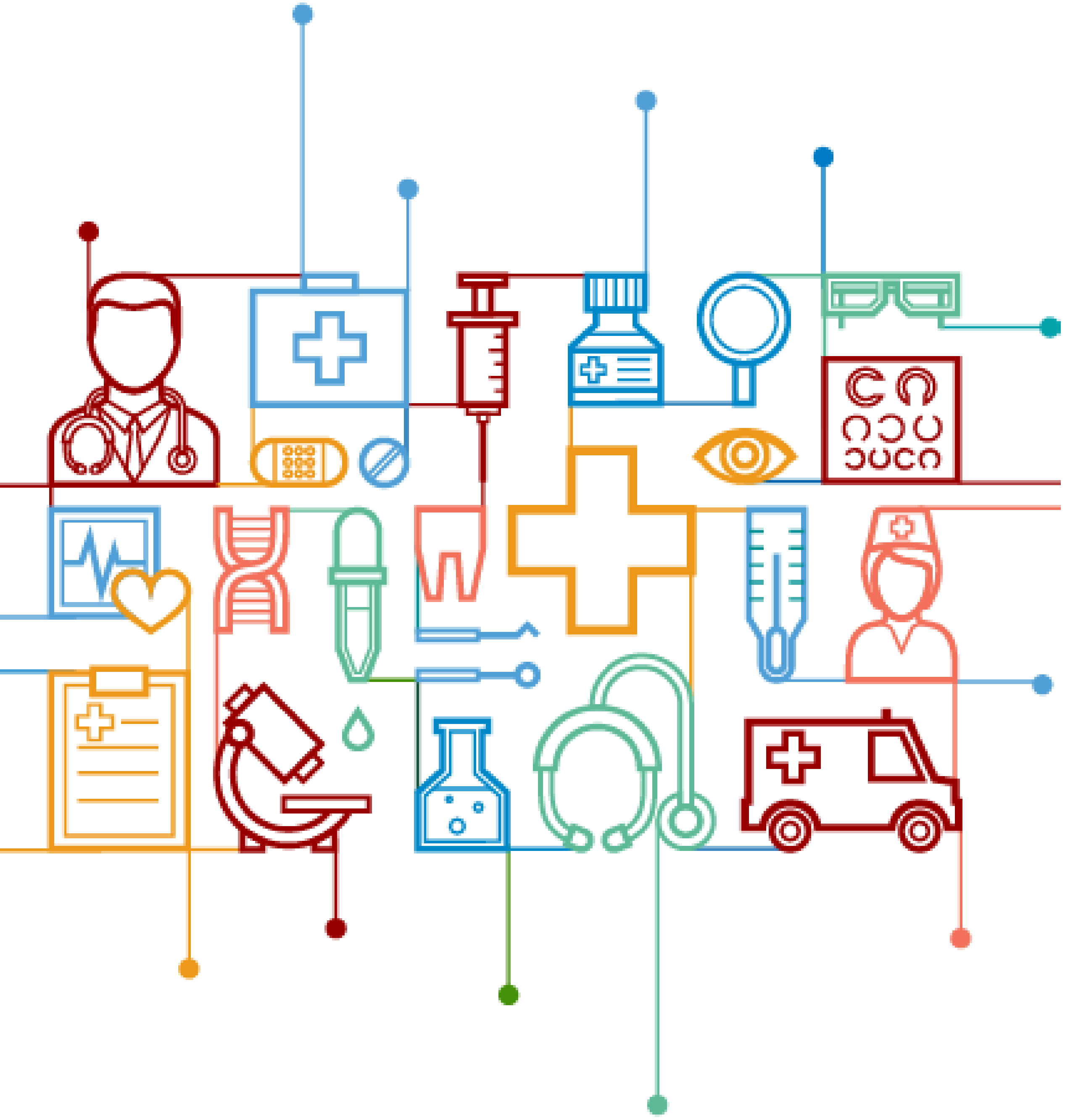
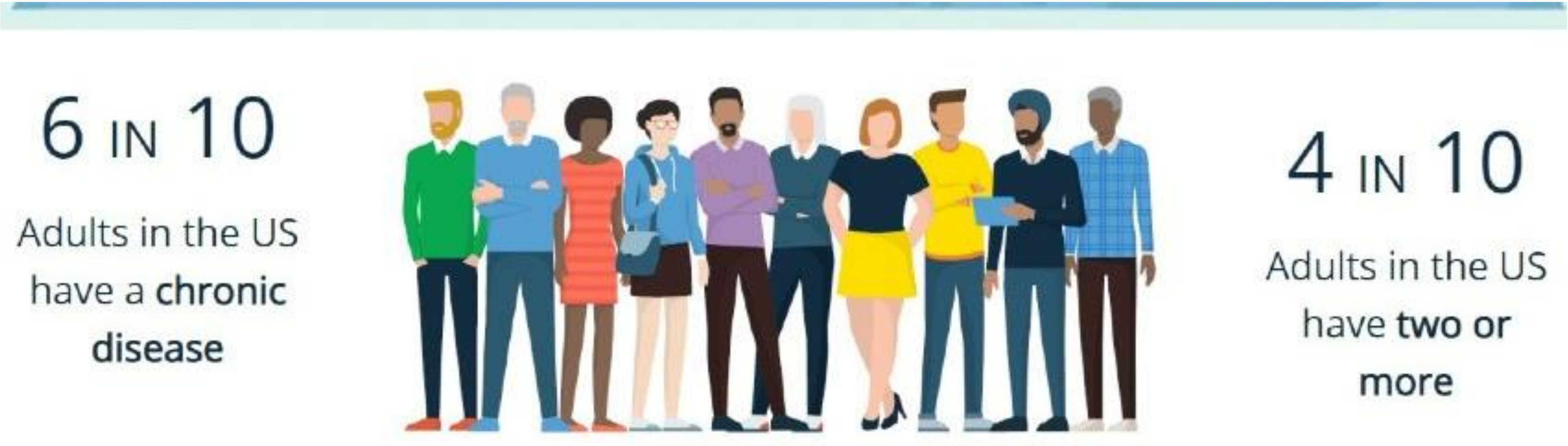


Uncovering Patterns and Predicting Chronic Disease Risks in the U.S

Van Le - Stanley He - James Li


6 IN 10
Adults in the US have a chronic disease
4 IN 10
Adults in the US have two or more

The Burden of Chronic Diseases

❑ Chronic diseases are the leading causes of death and disability in the U.S.

❑ Importance to predict and manage disease outbreaks.

❑ These indicators are essential for surveillance, prioritization, and evaluation of public health interventions.

Data Mining in Health Informatics

❑ Uncovering health patterns in large datasets.

❑ Predictive analytics power to inform healthcare strategies and interventions.

❑ Health risk assessment and disease prevention.

Project Objectives

Classification of risk levels at the state level that are essential for the evaluation of public health interventions.

Pattern of chronic disease across different states.

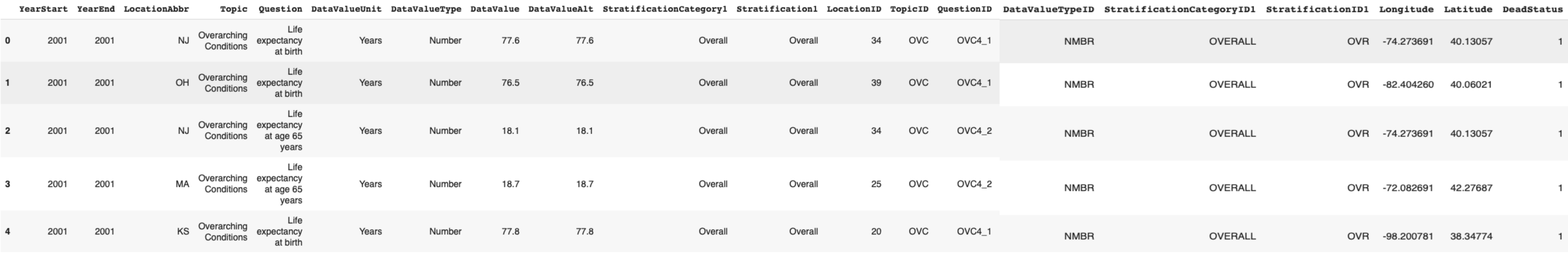
Predictive models of mortality risk that important for early detection and future strategies.

Dataset Overview

❑ `U.S. Chronic Disease Indicators (CDI)` was developed by consensus that allows states and large metropolitan areas to uniformly define, collect, and report chronic disease data that are important to public health practice. (catalog.data.gov)

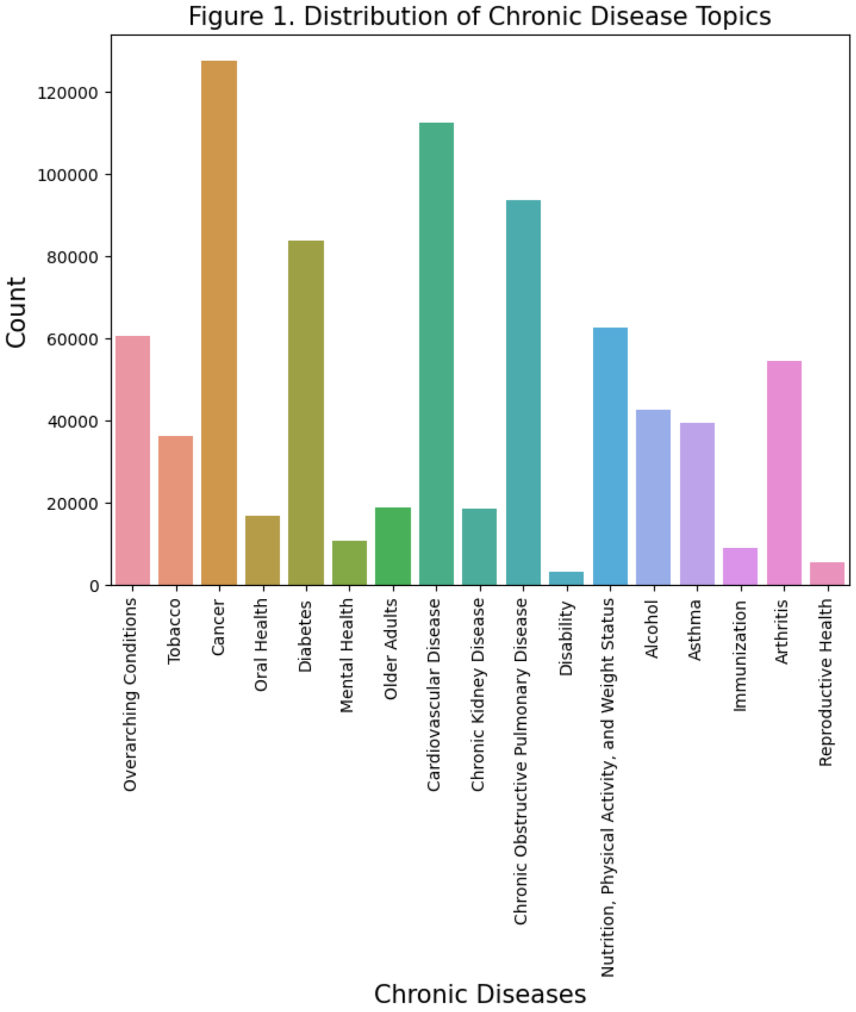
❑ The Chronic Condition Indicator was created to facilitate health services research on diagnoses using administrative data.

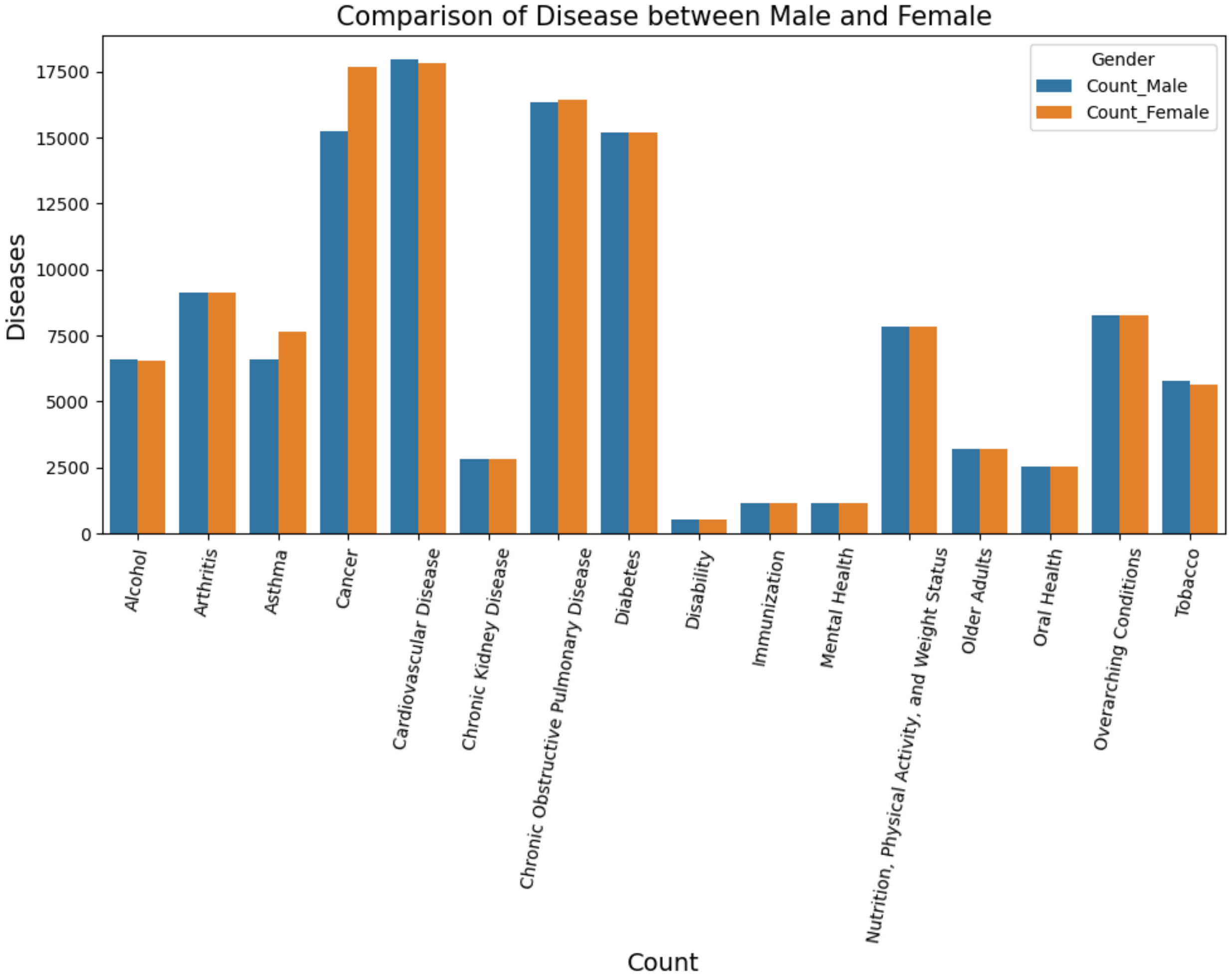
❑ There are more than 900K data rows in our dataset, with more than 24 indicators over 50 states that include 17 topic groups.



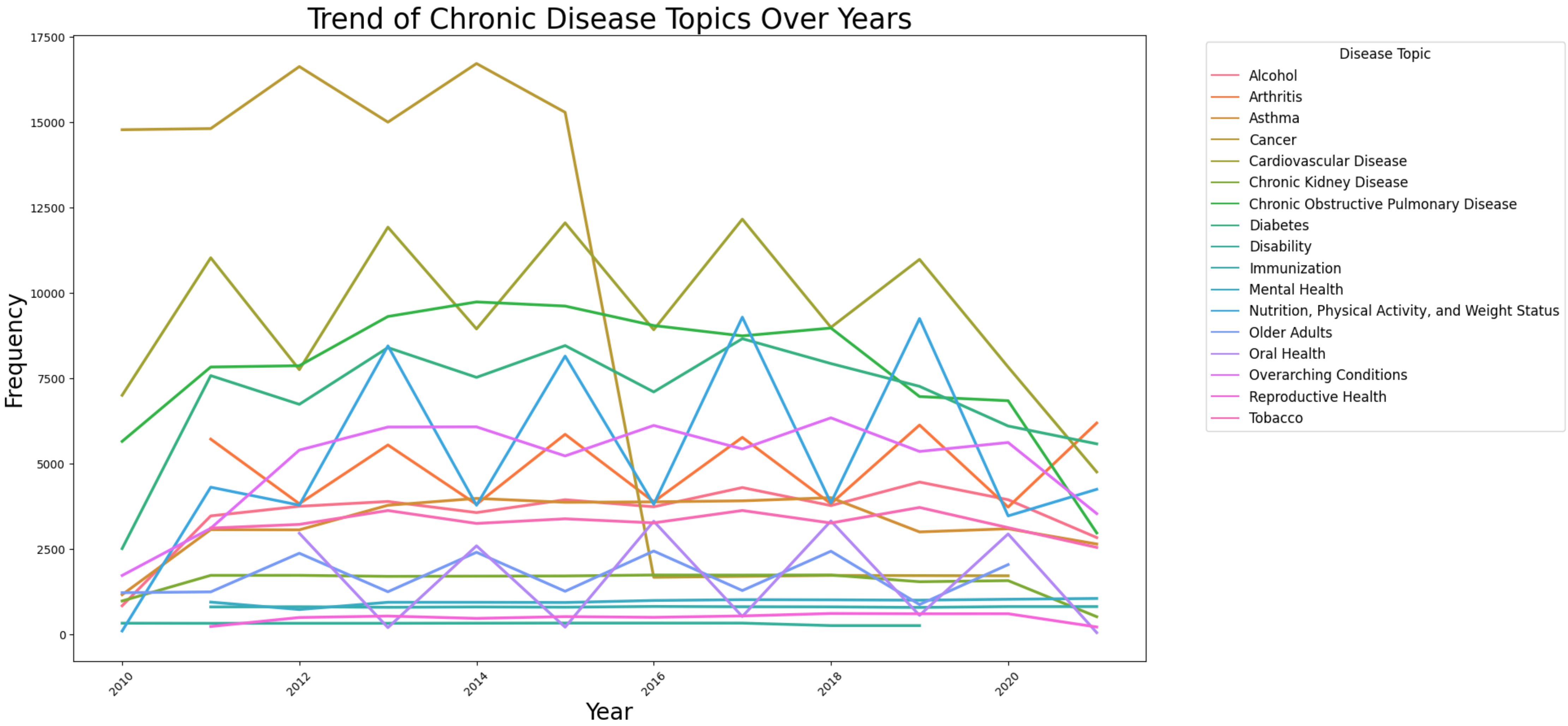
	YearStart	YearEnd	LocationAbbr	Topic	Question	DataValueUnit	DataValueType	DataValue	DataValueAlt	StratificationCategory1	Stratification1	LocationID	TopicID	QuestionID	DataValueTypeID	StratificationCategoryID1	StratificationID1	Longitude	Latitude	DeadStatus
0	2001	2001	NJ	Overarching Conditions	Life expectancy at birth	Years	Number	77.6	77.6	Overall	Overall	34	OVC	OVC4_1	NMBR	OVERALL	OVR	-74.273691	40.13057	1
1	2001	2001	OH	Overarching Conditions	Life expectancy at birth	Years	Number	76.5	76.5	Overall	Overall	39	OVC	OVC4_1	NMBR	OVERALL	OVR	-82.404260	40.06021	1
2	2001	2001	NJ	Overarching Conditions	Life expectancy at age 65 years	Years	Number	18.1	18.1	Overall	Overall	34	OVC	OVC4_2	NMBR	OVERALL	OVR	-74.273691	40.13057	1
3	2001	2001	MA	Overarching Conditions	Life expectancy at age 65 years	Years	Number	18.7	18.7	Overall	Overall	25	OVC	OVC4_2	NMBR	OVERALL	OVR	-72.082691	42.27687	1
4	2001	2001	KS	Overarching Conditions	Life expectancy at birth	Years	Number	77.8	77.8	Overall	Overall	20	OVC	OVC4_1	NMBR	OVERALL	OVR	-98.200781	38.34774	1

Data Summary & Statistics





Trends of Topic Over Years

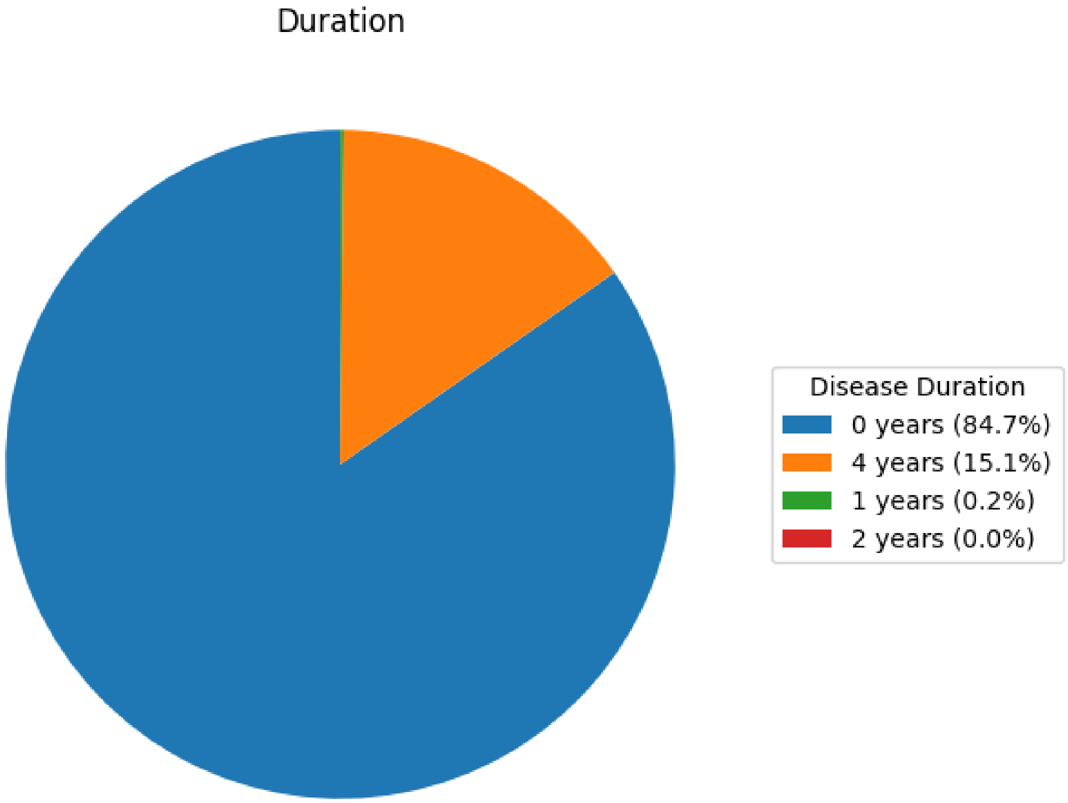


Recovery Time Analysis

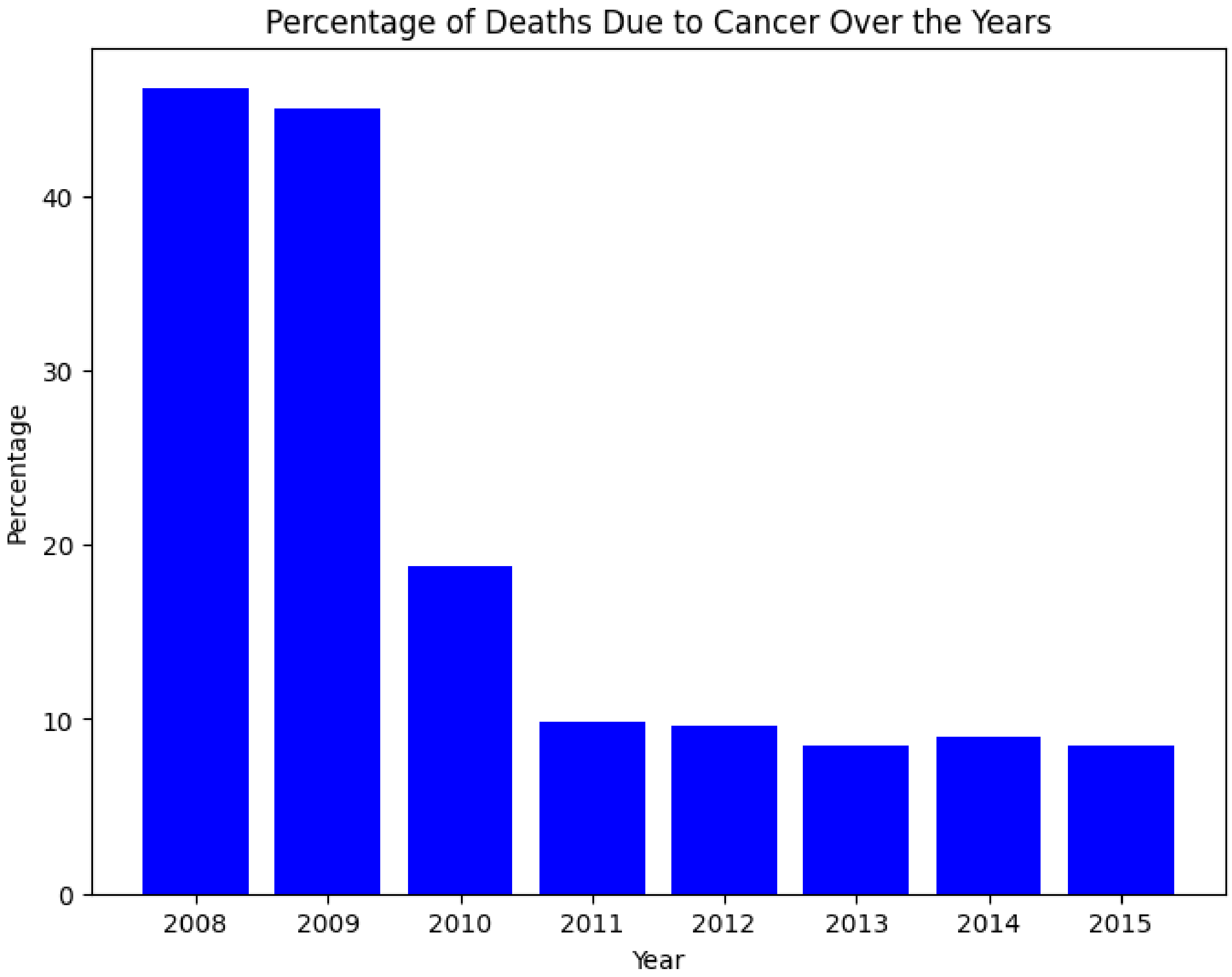
- ❑ The majority recovered swiftly during the initial year.

- ❑ A significant portion required 4 years for recovery.

- ❑ A notable subset recovered after 1 or 2 years.



Cancer Death Rate Throughout The Years



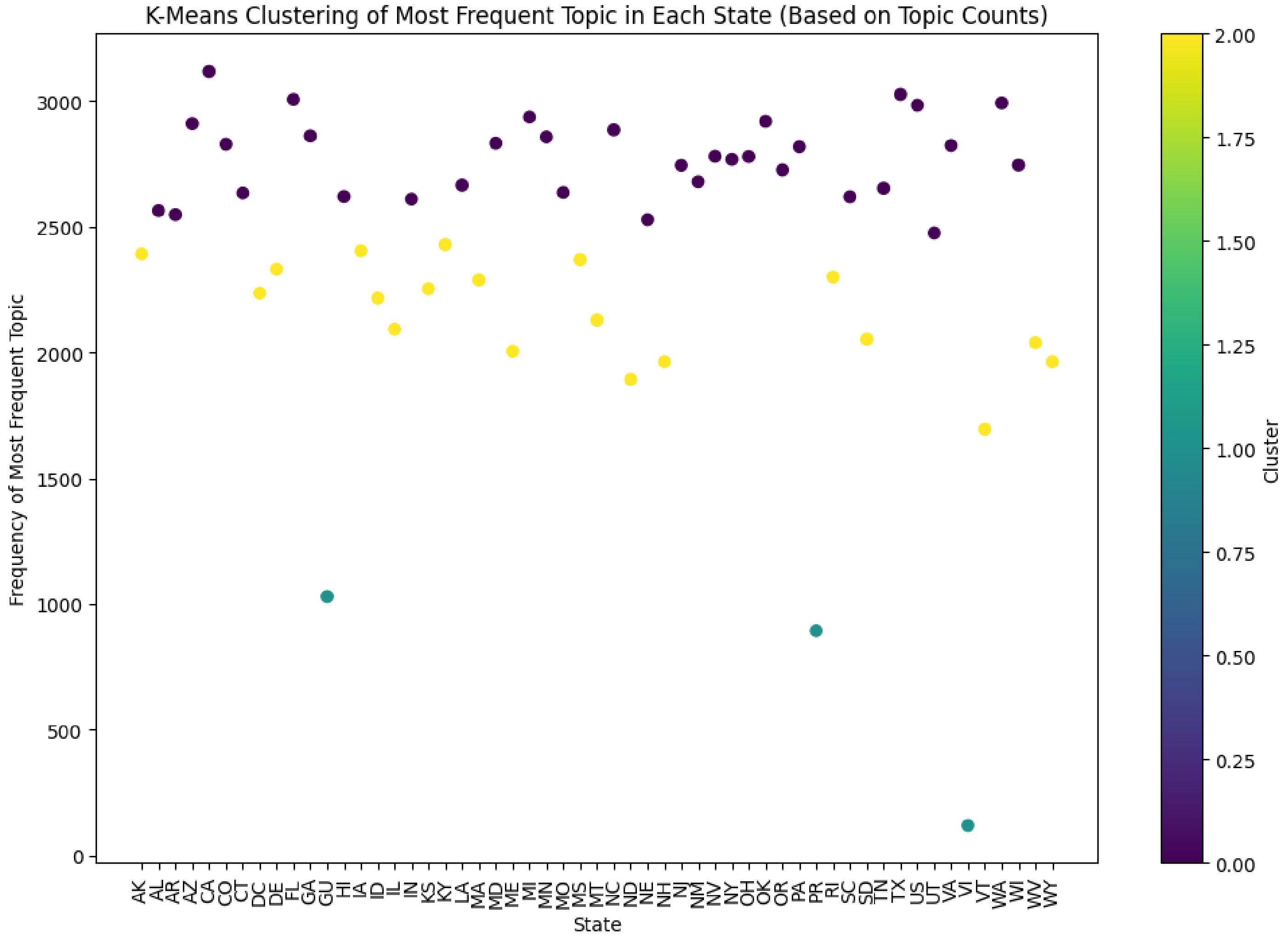
K-means Clustering Analysis

❑ Purpose: Determine patterns in disease prevalence across states.

❑ Input data consists of the health issues or the chronic diseases that are the most frequent in each state in the US.

❑ The K-means cluster analysis was chosen because it is relatively easy to understand and implement.

❑ Silhouette score: 0.63622 which is relatively high.



Risk Level Assessment

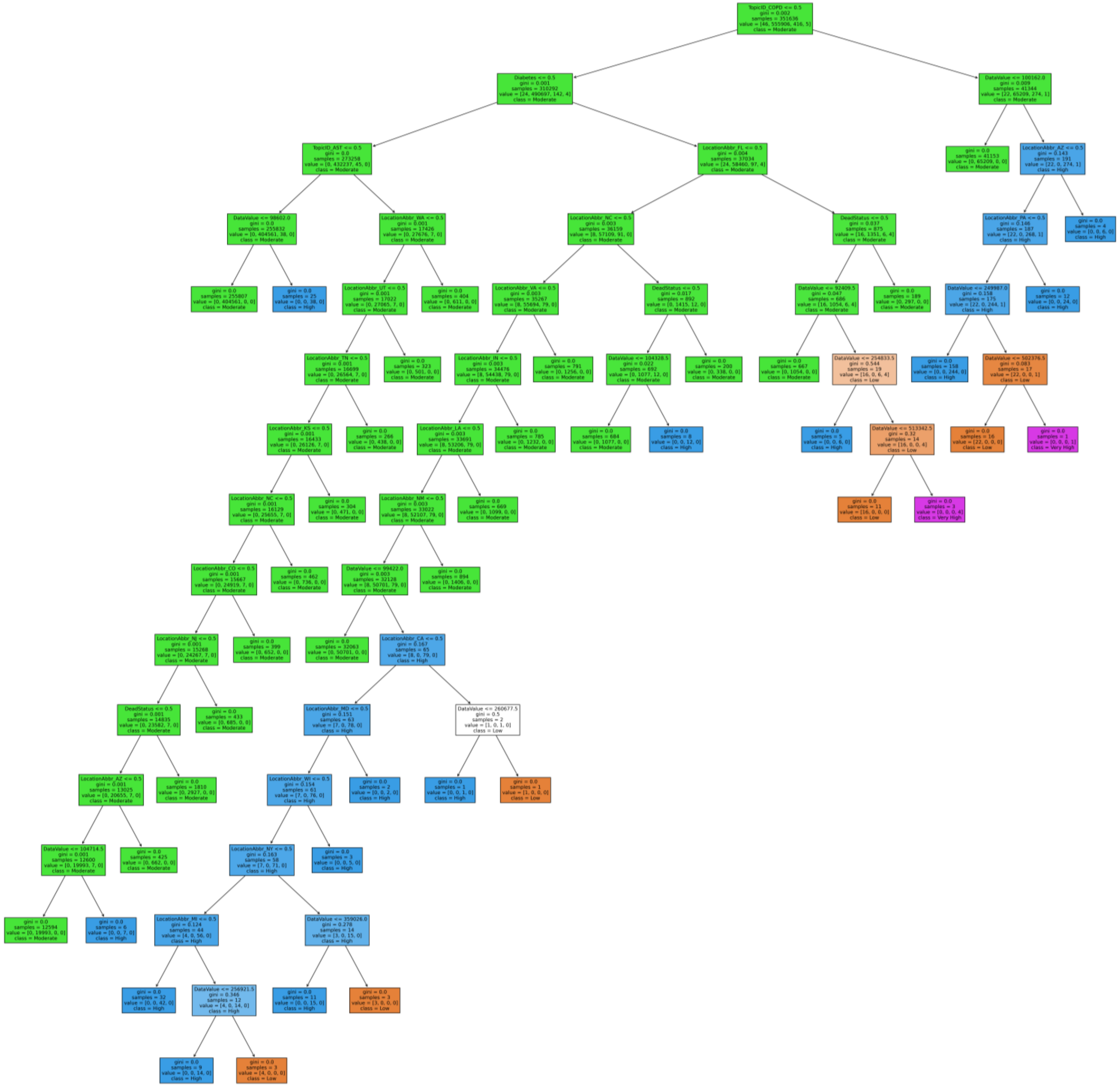
❑ Random Forest Classifier

❑ Attributes: `LocationAbbr`, `YearStart`, `DataValue`, `DeadStatus`, `TopicID`

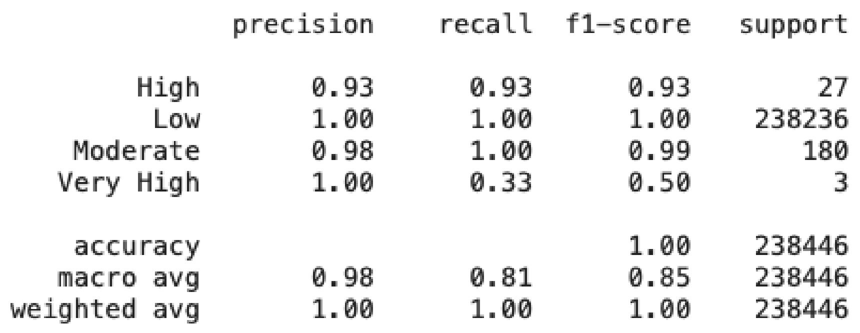
❑ Target: `RiskLevel` that is defined based on `DataValue` and `DataValueAlt`.

Risk Level	DataValue
Low	[-1, 500000)
Medium	[500000 – 1000000)
High	[1000000 – 2000000)
Very High	>= 2000000

Structure of Random Forest


TopicID_COPD <= 0.5 gini = 0.002 samples = 351636 value = [46, 555906, 416, 5] class = Moderate
Diabetes <= 0.5 gini = 0.001 samples = 310292 value = [24, 490697, 142, 4] class = Moderate
DataValue <= 100162.0 gini = 0.009 samples = 41344 value = [22, 65209, 274, 1] class = Moderate
TopicID_AST <= 0.5 gini = 0.0 samples = 273258 value = [0, 432237, 45, 0] class = Moderate
LocationAbbr_FL <= 0.5 gini = 0.004 samples = 37034 value = [24, 58460, 97, 4] class = Moderate
gini = 0.0 samples = 41153 value = [0, 65209, 0, 0] class = Moderate
LocationAbbr_AZ <= 0.5 gini = 0.143 samples = 191 value = [22, 0, 274, 1] class = High
DataValue <= 98602.0 gini = 0.0 samples = 255832 value = [0, 404561, 38, 0] class = Moderate
LocationAbbr_WA <= 0.5 gini = 0.001 samples = 17426 value = [0, 27676, 7, 0] class = Moderate
LocationAbbr_NC <= 0.5 gini = 0.003 samples = 36159 value = [8, 57109, 91, 0] class = Moderate
DeadStatus <= 0.5 gini = 0.037 samples = 875 value = [16, 1351, 6, 4] class = Moderate
LocationAbbr_PA <= 0.5 gini = 0.146 samples = 187 value = [22, 0, 268, 1] class = High
gini = 0.0 samples = 4 value = [0, 0, 6, 0] class = High
gini = 0.0 samples = 255807 value = [0, 404561, 0, 0] class = Moderate
gini = 0.0 samples = 25 value = [0, 0, 38, 0] class = High
LocationAbbr_UT <= 0.5 gini = 0.001 samples = 17022 value = [0, 27065, 7, 0] class = Moderate
gini = 0.0 samples = 404 value = [0, 611, 0, 0] class = Moderate
LocationAbbr_VA <= 0.5 gini = 0.003 samples = 35267 value = [8, 55694, 79, 0] class = Moderate
DeadStatus <= 0.5 gini = 0.017 samples = 892 value = [0, 1415, 12, 0] class = Moderate
DataValue <= 92409.5 gini = 0.047 samples = 686 value = [16, 1054, 6, 4] class = Moderate
gini = 0.0 samples = 189 value = [0, 297, 0, 0] class = Moderate
DataValue <= 249987.0 gini = 0.158 samples = 175 value = [22, 0, 244, 1] class = High
gini = 0.0 samples = 12 value = [0, 0, 24, 0] class = High
LocationAbbr_TN <= 0.5 gini = 0.001 samples = 16699 value = [0, 26564, 7, 0] class = Moderate
gini = 0.0 samples = 323 value = [0, 501, 0, 0] class = Moderate
LocationAbbr_IN <= 0.5 gini = 0.003 samples = 34476 value = [8, 54438, 79, 0] class = Moderate
gini = 0.0 samples = 791 value = [0, 1256, 0, 0] class = Moderate
DataValue <= 104328.5 gini = 0.022 samples = 692 value = [0, 1077, 12, 0] class = Moderate
gini = 0.0 samples = 200 value = [0, 338, 0, 0] class = Moderate
gini = 0.0 samples = 667 value = [0, 1054, 0, 0] class = Moderate
DataValue <= 254833.5 gini = 0.544 samples = 19 value = [16, 0, 6, 4] class = Low
gini = 0.0 samples = 158 value = [0, 0, 244, 0] class = High
DataValue <= 502376.5 gini = 0.083 samples = 17 value = [22, 0, 0, 1] class = Low
LocationAbbr_KS <= 0.5 gini = 0.001 samples = 16433 value = [0, 26126, 7, 0] class = Moderate
gini = 0.0 samples = 266 value = [0, 438, 0, 0] class = Moderate
LocationAbbr_LA <= 0.5 gini = 0.003 samples = 33691 value = [8, 53206, 79, 0] class = Moderate
gini = 0.0 samples = 785 value = [0, 1232, 0, 0] class = Moderate
gini = 0.0 samples = 684 value = [0, 1077, 0, 0] class = Moderate
gini = 0.0 samples = 8 value = [0, 0, 12, 0] class = High
gini = 0.0 samples = 5 value = [0, 0, 6, 0] class = High
DataValue <= 513342.5 gini = 0.32 samples = 14 value = [16, 0, 0, 4] class = Low
gini = 0.0 samples = 16 value = [22, 0, 0, 0] class = Low
gini = 0.0 samples = 1 value = [0, 0, 0, 1] class = Very High
LocationAbbr_NC <= 0.5 gini = 0.001 samples = 16129 value = [0, 25655, 7, 0] class = Moderate
gini = 0.0 samples = 304 value = [0, 471, 0, 0] class = Moderate
LocationAbbr_NM <= 0.5 gini = 0.003 samples = 33022 value = [8, 52107, 79, 0] class = Moderate
gini = 0.0 samples = 669 value = [0, 1099, 0, 0] class = Moderate
gini = 0.0 samples = 11 value = [16, 0, 0, 0] class = Low
gini = 0.0 samples = 3 value = [0, 0, 0, 4] class = Very High
LocationAbbr_CO <= 0.5 gini = 0.001 samples = 15667 value = [0, 24919, 7, 0] class = Moderate
gini = 0.0 samples = 462 value = [0, 736, 0, 0] class = Moderate
DataValue <= 99422.0 gini = 0.003 samples = 32128 value = [8, 50701, 79, 0] class = Moderate
gini = 0.0 samples = 894 value = [0, 1406, 0, 0] class = Moderate
LocationAbbr_NJ <= 0.5 gini = 0.001 samples = 15268 value = [0, 24267, 7, 0] class = Moderate
gini = 0.0 samples = 399 value = [0, 652, 0, 0] class = Moderate
gini = 0.0 samples = 32063 value = [0, 50701, 0, 0] class = Moderate
LocationAbbr_CA <= 0.5 gini = 0.167 samples = 65 value = [8, 0, 79, 0] class = High
DeadStatus <= 0.5 gini = 0.001 samples = 14835 value = [0, 23582, 7, 0] class = Moderate
gini = 0.0 samples = 433 value = [0, 685, 0, 0] class = Moderate
LocationAbbr_MD <= 0.5 gini = 0.151 samples = 63 value = [7, 0, 78, 0] class = High
DataValue <= 260677.5 gini = 0.5 samples = 2 value = [1, 0, 1, 0] class = Low
LocationAbbr_AZ <= 0.5 gini = 0.001 samples = 13025 value = [0, 20655, 7, 0] class = Moderate
gini = 0.0 samples = 1810 value = [0, 2927, 0, 0] class = Moderate
LocationAbbr_WI <= 0.5 gini = 0.154 samples = 61 value = [7, 0, 76, 0] class = High
gini = 0.0 samples = 2 value = [0, 0, 2, 0] class = High
gini = 0.0 samples = 1 value = [0, 0, 1, 0] class = High
gini = 0.0 samples = 1 value = [1, 0, 0, 0] class = Low
DataValue <= 104714.5 gini = 0.001 samples = 12600 value = [0, 19993, 7, 0] class = Moderate
gini = 0.0 samples = 425 value = [0, 662, 0, 0] class = Moderate
LocationAbbr_NY <= 0.5 gini = 0.163 samples = 58 value = [7, 0, 71, 0] class = High
gini = 0.0 samples = 3 value = [0, 0, 5, 0] class = High
gini = 0.0 samples = 12594 value = [0, 19993, 0, 0] class = Moderate
gini = 0.0 samples = 6 value = [0, 0, 7, 0] class = High
LocationAbbr_MI <= 0.5 gini = 0.124 samples = 44 value = [4, 0, 56, 0] class = High
DataValue <= 359026.0 gini = 0.278 samples = 14 value = [3, 0, 15, 0] class = High
gini = 0.0 samples = 32 value = [0, 0, 42, 0] class = High
DataValue <= 256921.5 gini = 0.346 samples = 12 value = [4, 0, 14, 0] class = High
gini = 0.0 samples = 11 value = [0, 0, 15, 0] class = High
gini = 0.0 samples = 3 value = [3, 0, 0, 0] class = Low
gini = 0.0 samples = 9 value = [0, 0, 14, 0] class = High
gini = 0.0 samples = 3 value = [4, 0, 0, 0] class = Low

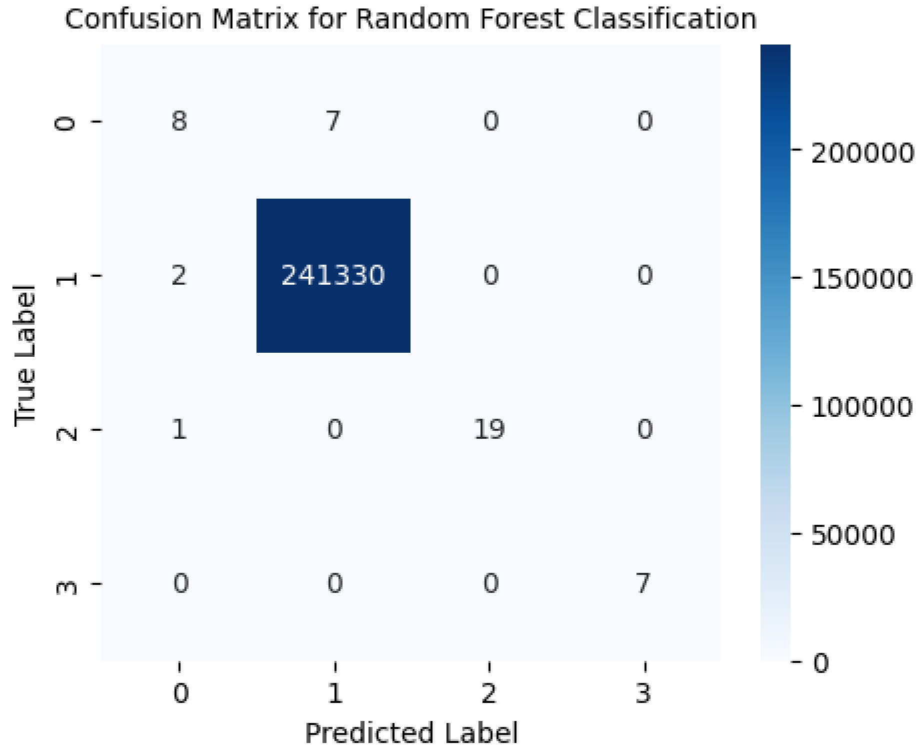
RandomForest Classifier Evaluation



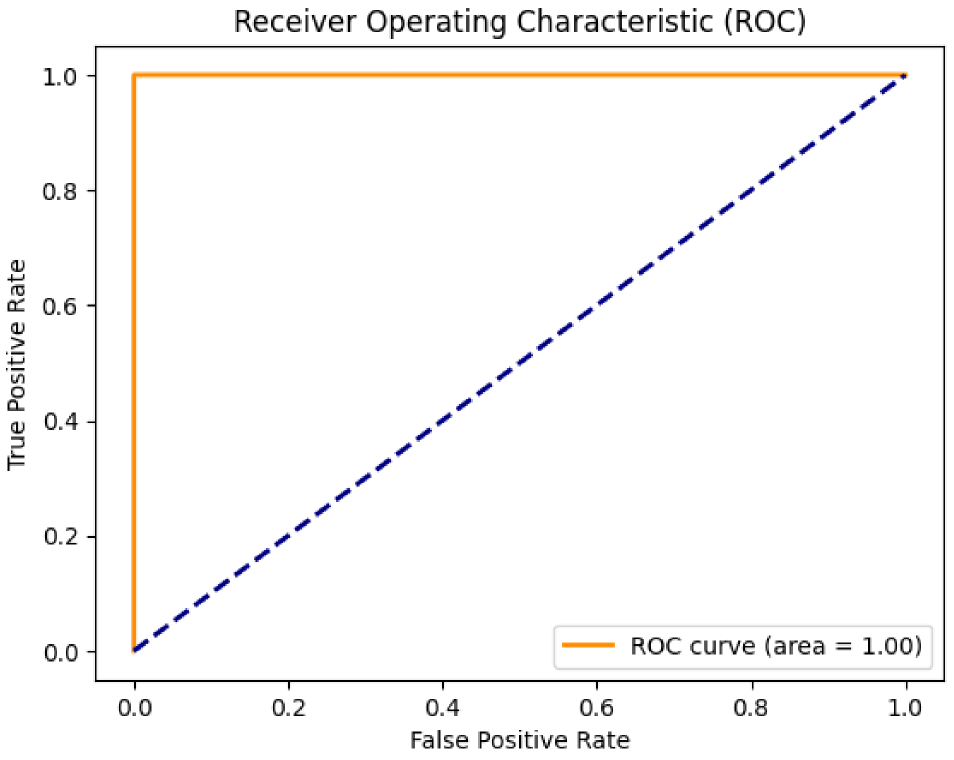
```
              precision    recall  f1-score   support

        High       0.93      0.93      0.93        27
         Low       1.00      1.00      1.00    238236
    Moderate       0.98      1.00      0.99       180
   Very High       1.00      0.33      0.50         3

    accuracy                           1.00    238446
   macro avg       0.98      0.81      0.85    238446
weighted avg       1.00      1.00      1.00    238446
```



Confusion Matrix for Random Forest Classification



Receiver Operating Characteristic (ROC)

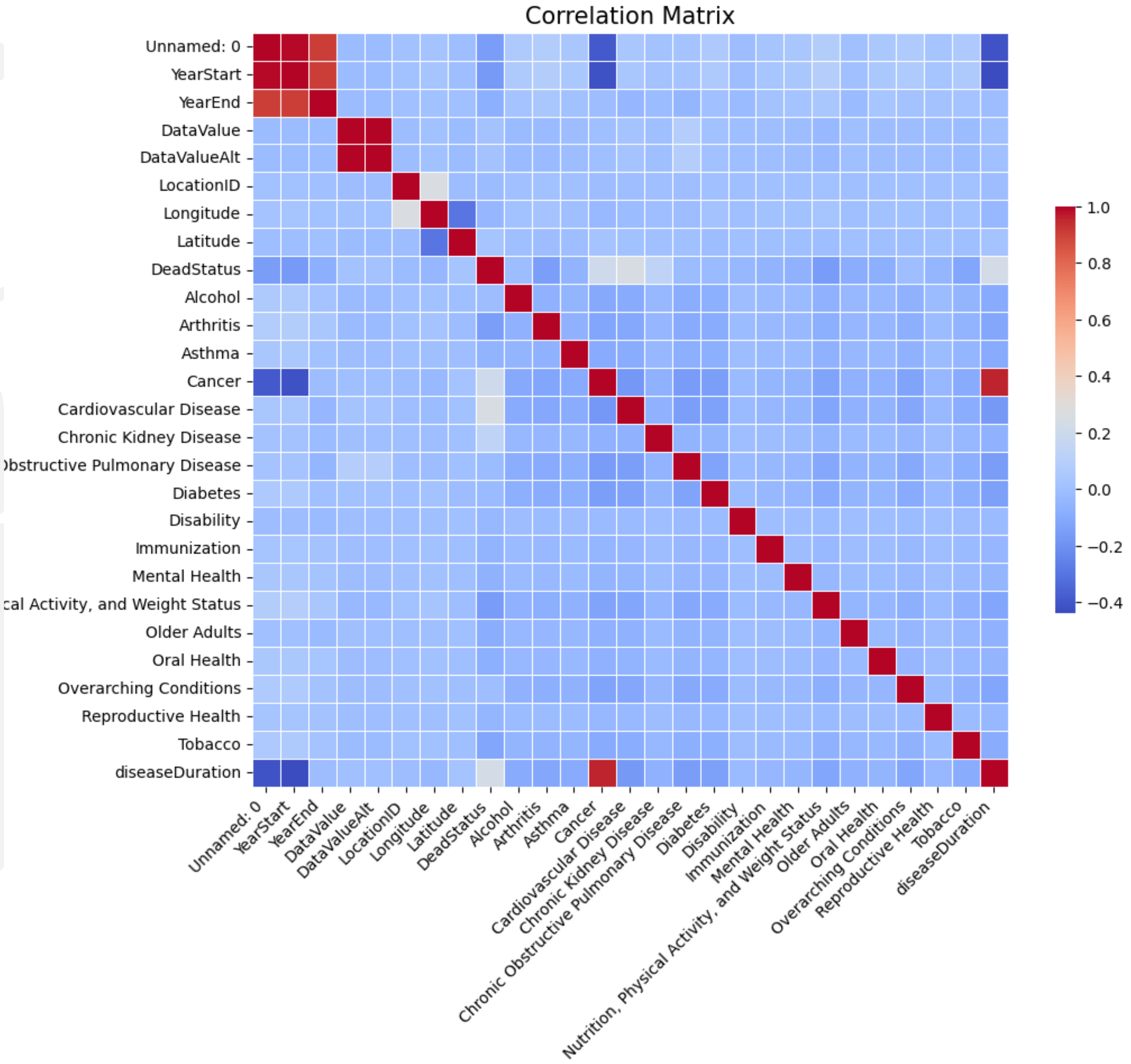
Chronic Disease Mortality Prediction

- ❑ Logistic Regression
  - ▪ Binary classification tasks. (1 = Deceased, 0 = Alive)
- ❑ Naïve Bayes Classification
  - ▪ Able to handle many features
- ❑ Attributes used for model training
  - ▪ \`YearStart\`, \`Arthritis\`, \`Cancer\`, \`Cardiovascular Disease\`, \`Chronic Kidney Disease\`, \`Nutrition, Physical Activity, and Weight Status\`, \`Tobacco\`, \`diseaseDuration\`

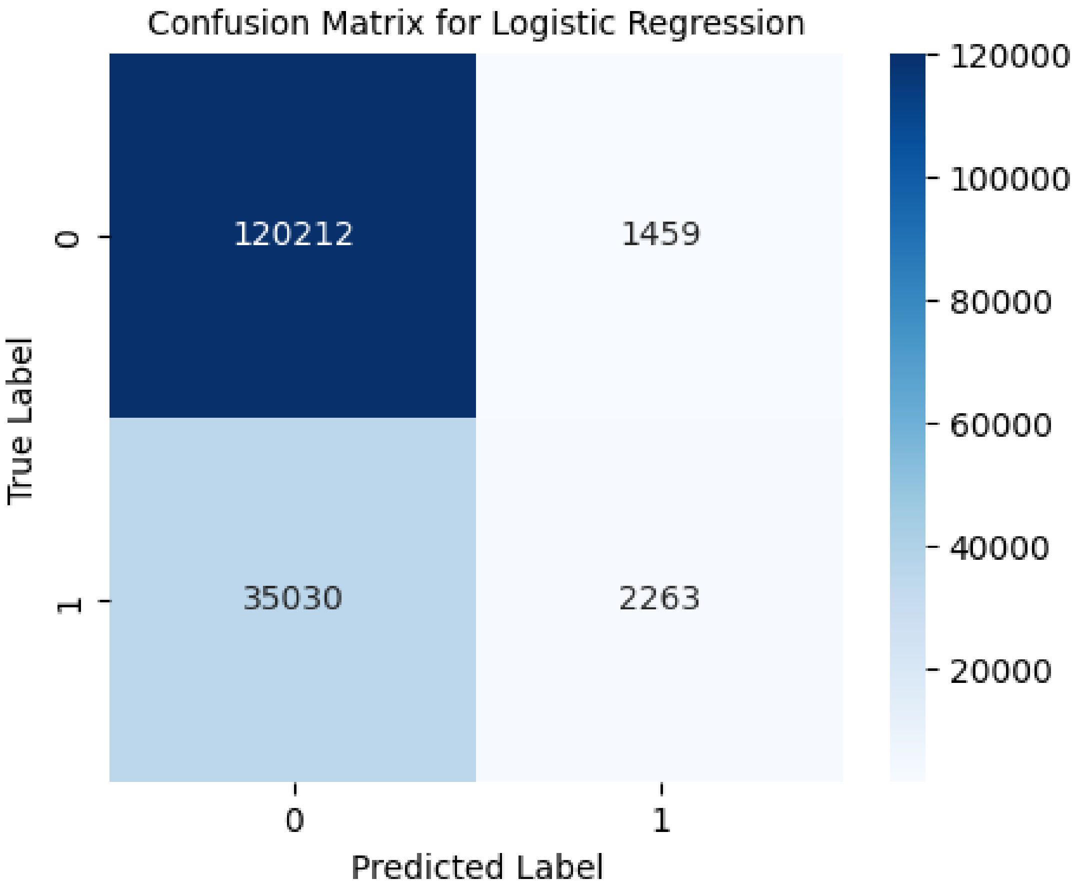
Correlation Matrix

❑ Selected features based on correlation threshold (> 0.1)

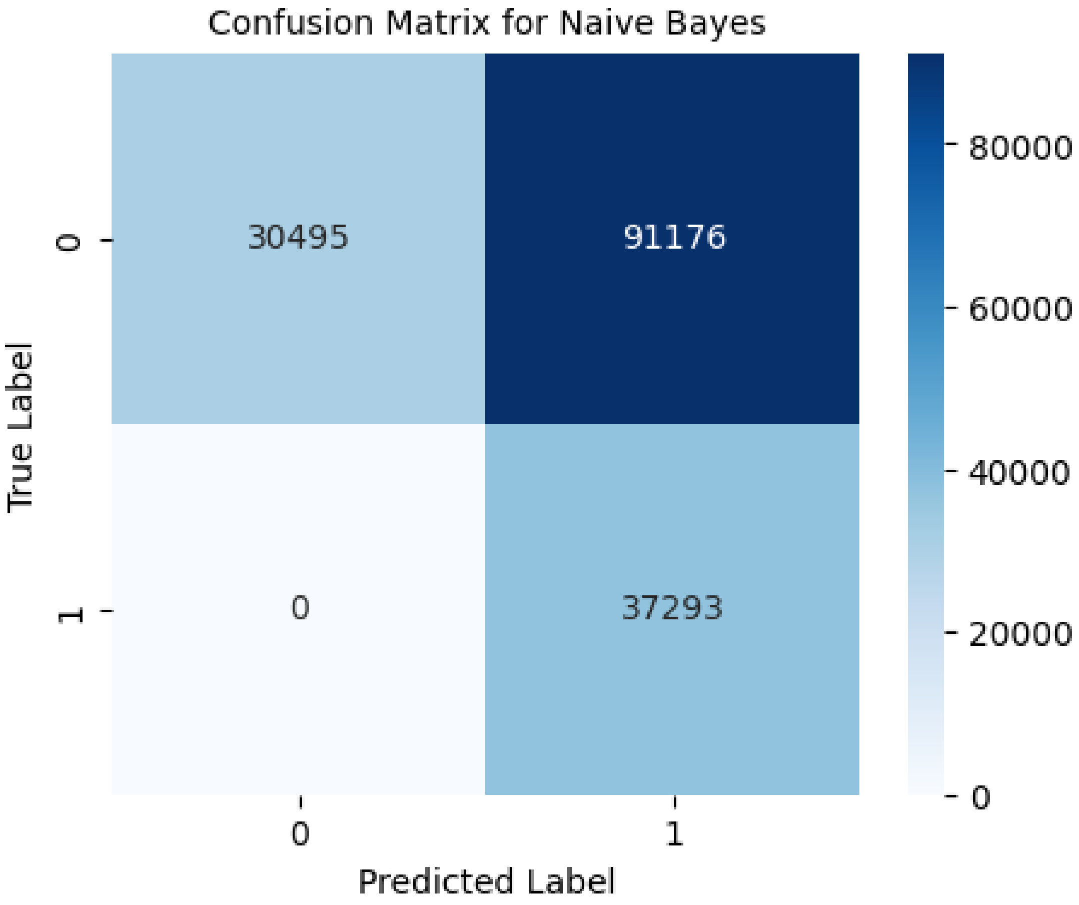
'YearStart', 'DeadStatus', 'Arthritis', 'Cancer', 'Cardiovascular Disease', 'Chronic Kidney Disease', 'Nutrition, Physical Activity, and Weight Status', 'Tobacco', 'diseaseDuration'



Logistic Regression vs. Naïve Bayes



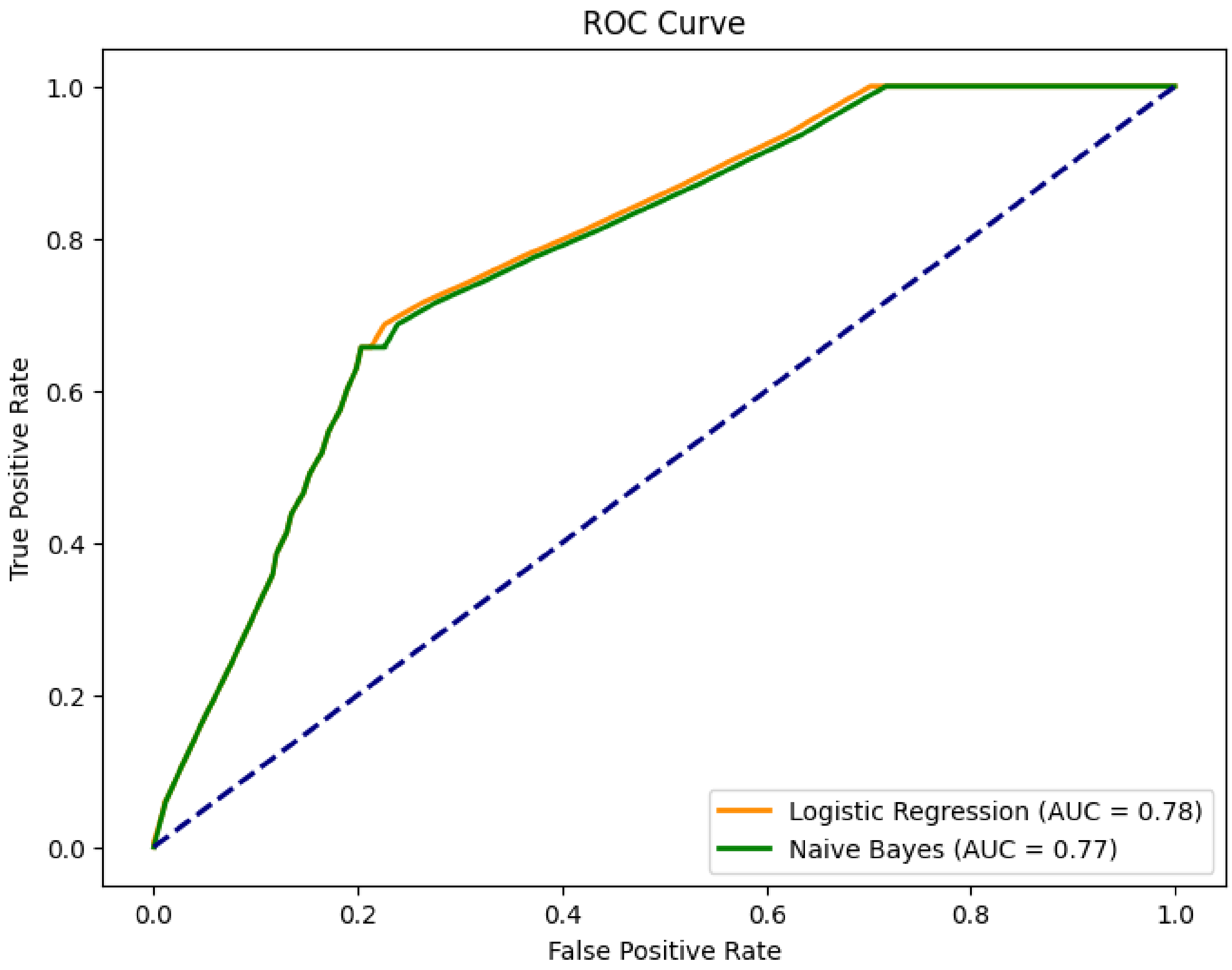
Confusion Matrix for Logistic Regression



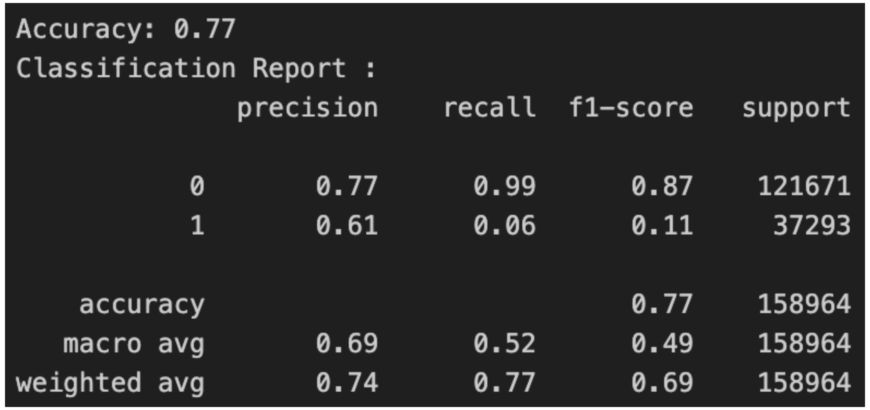
Confusion Matrix for Naive Bayes

Predicted 1, Actual 1 = TP   Predicted 0, Actual 1 = FN
Predicted 1, Actual 0 = FP   Predicted 0, Actual 0 = TN

Logistic Regression vs. Naïve Bayes Result



Logistic Regression Classification Report

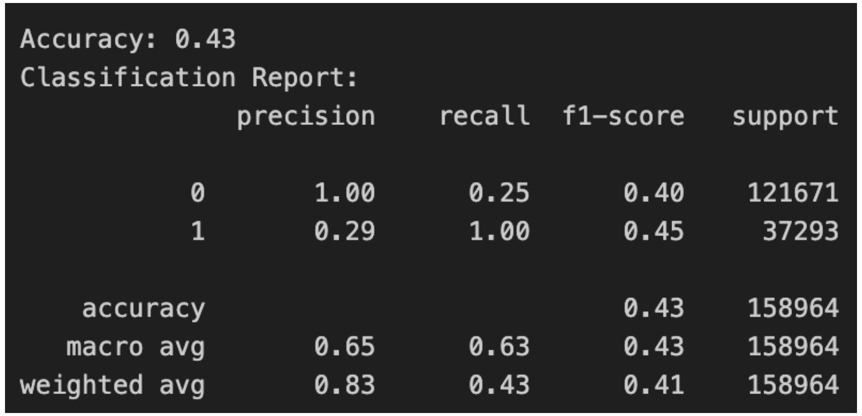


```
Accuracy: 0.77
Classification Report :
              precision    recall  f1-score   support

           0       0.77      0.99      0.87    121671
           1       0.61      0.06      0.11     37293

    accuracy                           0.77    158964
   macro avg       0.69      0.52      0.49    158964
weighted avg       0.74      0.77      0.69    158964
```

Naïve Bayes Classification Report



```
Accuracy: 0.43
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.25      0.40    121671
           1       0.29      1.00      0.45     37293

    accuracy                           0.43    158964
   macro avg       0.65      0.63      0.43    158964
weighted avg       0.83      0.43      0.41    158964
```

Challenges

❑ Unbalanced dataset

❑ Data types are diverse: hard to compile into model training, so we need intensive preprocessing data.

❑ Data was not available for all `Topic` for all years.

why
? what
when
when
who
?
where
what
?
QUESTIONS
how
who
?
what
why
how
how
? when
where

Questions