A blue letter on a white background

Description automatically generated

**Classification**

CS 7331 DATA MINING

Hrithik Chavva

# **TABLE OF CONTENTS**

# 1.Execute Summary………………………………………………………………………...……3

# 2.Business Understanding………………………………………………………………………..3

# 3.Data Understanding……………………………………………………………………….……4

# 3.1 Data Description (Including Additional Features) ……………………………....……5

# 3.2 Summary………………………………………………………………………………6

# 3.3 Heat Map………………………………………………………………………………8

# 3.3.1 Heat Map of Dataset…………………………………………………...……8

# 3.3.2 LOF Curve of Dataset……………………………………………………...10

# 3.4 Define classes………………………………………………………………………...10

# 3.5 Confirm Predictive Features…………………………………………………………11

4.Modeling……………………………………………………………………………………….12

4.1 Selecting Training Set……………………………………………………………….12

4.2 Conditional Inference Tree………………………………………………………….15

4.2.1 Ctree Training Model Diagram…………………………………………….16

4.2.2 The map of CIT model predicts bad conditions……………………………17

4.2.3 The Map of Actual Bad Conditions………………………………………..18

4.2.4 Accuracy and Kappa of CIT model………………………………………..19

4.3 Gradient Boosted Decision Trees (Xgboost) ………………………………………..20

4.3.1 The Map Of Xgboost Model Predicts Bad Conditions…………………….20

4.4 Linear Support Vector Machine……………………………………………………...23

4.4.1 The Map of LSVM Model Predicts Bad Conditions………………………23

4.4.2 The Map of LSVM Model Actual Bad Conditions………………………..24

5.Exceptional Work……………………………………………………………………………...26

5.1 C4.5 (Decision Tree) ………………………………………………………………..26

5.1.1 Decision Tree Model Predicts Bad Conditions…………………………….26

5.1.2 The Map of Actual Bad Conditions………………………………………..27

5.2 Comparing Models…………………………………………………………………...30

6.Evaluation…………………….……….……….……….……….……….……….…………....31

7.Deployment……………….…………….…………….…………….…………….……………32

8.Conclusion………………………………………………………………………………....…..33

# **Execute Summary:**

This analysis utilized machine learning algorithms to predict the trajectories of COVID-19 outbreaks in various U.S. counties. The primary objective was to create models that could categorize the severity levels based on historical case and mortality data. Instead of solely forecasting numerical figures, these models were designed to reveal the underlying patterns in the data that contribute to the spread of the disease. Our feature engineering incorporated demographic details, racial and ethnic composition, employment statistics, commuting patterns, and real estate information—factors in society that play a role in transmission.

We experimented with four different approaches: Conditional Inference Trees, Support Vector Machines (SVM), XGBoost Models, and C4.5 Decision Trees. The aim is to provide policymakers with valuable insights to efficiently allocate medical resources, mobilize emergency personnel, and implement targeted containment strategies in potential emerging hotspots. As we gather more high-quality training data, the accuracy and reliability of these predictions are expected to further improve over time.

1. **Business Understanding**

This analysis holds significant value for policymakers at the national and state levels, offering data-driven insights into the trajectories of COVID-19 across various regions in the U.S. By identifying counties where outbreaks may intensify based on key socioeconomic factors, governments can strategically deploy medical resources and response teams to emerging hotspots. The broader application of the machine learning approach extends beyond national borders, contributing to global pandemic preparedness. It aids public health policymaking and resource allocation decisions worldwide through the use of predictive intelligence. As the models' accuracy can be enhanced with higher-quality training data, the proposition lies in delivering actionable foresight to guide proactive containment measures and the targeted strengthening of healthcare capacities where they are most needed.

# **Data Understanding:**

|  |  |  |
| --- | --- | --- |
| **Variables** | **Description** | **Scale** |
| county\_name | Name of the county | Nominal |
| state | State in which the county is located | Nominal |
| total\_pop | Total population of the county | Ratio |
| owner\_occupied\_housing\_units | Number of owner-occupied housing units | Ratio |
| civilian\_labor\_force | Number of people in the civilian labor force | Ratio |
| employed\_pop | Number of employed individuals | Ratio |
| families\_with\_young\_children | Number of families with young children | Ratio |
| two\_parent\_families\_with\_young\_children | Number of two-parent families with young children | Ratio |
| two\_parents\_in\_labor\_force\_families\_with\_young\_children | Number of two-parent families with young children where both parents are in the labor force | Ratio |
| nonfamily\_households | Number of nonfamily households | Ratio |
| white\_pop | Number of White population | Ratio |
| vacant\_housing\_units\_for\_rent | Number of vacant housing units available for rent | Ratio |
| income\_per\_capita | Average income per capita | Ratio |
| bachelors\_degree | Percentage of the population with a Bachelor's degree | Ratio |
| in\_school | Number of individuals currently in school | Nominal |
| median\_rent | Median rent value for housing units | Ratio |
| black\_pop | Number of Black population | Ratio |
| asian\_pop | Number of Asian population | Ratio |
| hispanic\_pop | Number of Hispanic population | Ratio |
| commuters\_by\_public\_transportation | Number of commuters using public transportation | Ratio |
| commute\_30\_34\_mins | Number of individuals with a commute time between 30-34 minutes | Ratio |
| commute\_60\_more\_mins | Number of individuals with a commute time of 60 minutes or more | Ratio |
| million\_dollar\_housing\_units | Number of housing units valued at one million dollars or more | Ratio |
| median\_income | Median income of the population | Ratio |
| cases\_per\_10000 | Number of confirmed COVID-19 cases per 10,000 people | Ratio |
| deaths\_per\_10000 | Number of COVID-19 related deaths per 10,000 people | Ratio |

**Table 3.1 Data Description**

The dataset comprises a diverse set of variables providing insights into various aspects of U.S. counties. Demographic indicators such as "total\_pop" offer a comprehensive view of population size, while "families\_with\_young\_children" and "two\_parent\_families\_with\_young\_children" shed light on family structures. Socioeconomic characteristics, including "income\_per\_capita" and "median\_income," capture economic well-being. Housing attributes, like "owner\_occupied\_housing\_units" and "median\_rent," offer insights into the living conditions. Educational metrics such as "bachelors\_degree" and "in\_school" provide information about the educational landscape. Additionally, commuting behaviors and transportation choices are reflected in variables like "commuters\_by\_public\_transportation" and "commute\_60\_more\_mins." Furthermore, the dataset encompasses health-related factors such as "cases\_per\_10000" and "deaths\_per\_10000," crucial for understanding the impact of COVID-19.

The table systematically categorizes key variables from the dataset, presenting a detailed overview for analytical purposes. Each variable is accompanied by a brief description outlining its significance and characteristics. The "Scale" column denotes the measurement scale of the variables, distinguishing between nominal and ratio scales. The "Sector" column categorizes variables into sectors such as Demographics, Housing, Employment, Socioeconomic, Education, Transportation, and Health. in extracting meaningful insights across a broad spectrum of societal factors.

The table encompasses a comprehensive set of more than 25 variables spanning demographic, socioeconomic, and COVID-19-related domains, forming the foundation for subsequent predictive modeling and analysis. Numeric variables, capturing precise quantities, are identified as ratio scales, while nominal text categories, signifying distinct categories, are appropriately labeled. This rich assortment of variables includes demographic indicators like "total\_pop," socioeconomic metrics such as "income\_per\_capita" and "median\_income," and COVID-19 impact measures like "cases\_per\_10000" and "deaths\_per\_10000." The scale column in the table meticulously classifies each variable, aiding analysts in understanding the nature and measurement characteristics of the dataset.

* 1. **Summary:**

| **Columns** | **Min** | **1st Qua** | **Median** | **Mean** | **3rd Qua** | **Max** |
| --- | --- | --- | --- | --- | --- | --- |
| total\_pop | 74 | 10971 | 25714 | 102263 | 67524 | 10105722 |
| owner\_occupied\_housing\_units | 16 | 3164 | 7137 | 24158 | 18130 | 1512364 |
| civilian\_labor\_force | 39 | 4878 | 11491 | 51341 | 31611 | 5212243 |
| employed\_pop | 39 | 4558 | 10724 | 47976 | 29556 | 4805817 |
| families\_with\_young\_children | 0 | 715.5 | 1731 | 7308 | 4574 | 725915 |
| two\_parent\_families\_with\_young\_children | 0 | 432 | 1039 | 4734 | 2954 | 438223 |
| two\_parents\_in\_labor\_force\_families\_with\_young\_children | 0 | 240 | 611 | 2774 | 1696 | 242644 |
| nonfamily\_households | 12 | 1397 | 3177 | 12911 | 8499 | 1091276 |
| white\_pop | 55 | 8118 | 20217 | 62847 | 53594 | 2676982 |
| vacant\_housing\_units\_for\_rent | 0 | 66 | 194 | 904.2 | 580 | 62460 |
| income\_per\_capita | 9334 | 21805 | 25270 | 26026 | 29115 | 69529 |
| bachelors\_degree | 3 | 784.5 | 2017 | 13181.5 | 6416.5 | 1384333 |
| in\_school | 0.0 | 2383 | 5877 | 26044 | 16362 | 2704769 |
| median\_rent | 7 | 424 | 510 | 563.2 | 642 | 1879 |
| black\_pop | 0.0 | 95.5 | 764 | 12566.3 | 5398 | 1226134 |
| asian\_pop | 0.0 | 31.0 | 139 | 5412.4 | 714.5 | 1442577 |
| hispanic\_pop | 0 | 324 | 1029 | 18003 | 4878 | 4893579 |
| commuters\_by\_public\_transportation | 0.0 | 6.0 | 33.0 | 2423.7 | 145.5 | 735534 |
| commute\_30\_34\_mins | 0.0 | 282.5 | 744.0 | 4007.4 | 1892.0 | 603055 |
| commute\_60\_more\_mins | 0.0 | 18.42 | 25.97 | 30.98 | 38.26 | 166.79 |
| million\_dollar\_housing\_units | 0.0 | 0.1264 | 0.6595 | 1.45 | 1.4722 | 50.9904 |
| median\_income | 19264 | 41120 | 48038 | 49736 | 55758 | 129588 |
| cases\_per\_10000 | 24.62 | 585.70 | 755.20 | 768.51 | 929.92 | 3161.04 |
| deaths\_per\_10000 | 0.0 | 6.822 | 11.901 | 13.368 | 17.52 | 83.587 |

**3.3 Heat Map**

A heatmap of the correlation matrix is a visual representation that uses color to depict the strength and direction of relationships between variables in a dataset. Each cell in the heatmap corresponds to the correlation coefficient between two variables, and the color intensity indicates the magnitude and direction of the correlation.

The variables are listed on the x-axis and the y-axis, and the correlation between each pair of variables is represented by a color. The colors range from blue (negative correlation) to red (positive correlation), with white indicating no correlation.

A close-up of a graph

Description automatically generated

**3.3.1 Heat Map of Dataset**

* 1. **Check for missing values.**

There are no missing values. So, no need to perform any Data cleaning methods.

|  |  |
| --- | --- |
| **False** | **True** |
| **0** | **3139** |

* 1. **Handle Outliers**

The LOF (Local Outlier Factor) is a density-based anomaly detection algorithm that measures the degree to which a data point is different from its neighbours. A data point with a high LOF score is considered to be more likely to be an outlier.

A graph with a line

Description automatically generated

**Figure3.3.2 LOF Curve of Dataset**

**3.4. Define Classes:**

To avoid both overfitting and underfitting when analyzing counties based on their death rate, we categorized those with more than 10 deaths per 10,000 people as "bad" (marked as "True" in a new "bad" column). This decision was based on our experiments, where the threshold of 10 led to an ideal balance between "bad" and "not bad" counties in the data. This resulted in the distribution of counties shown in the table below.

|  |  |
| --- | --- |
| **False** | **Ture** |
| **1275** | **1864** |

## **3.5 Confirm Predictive Features**

To identify the most relevant features for training and predicting data in US counties, we assessed the importance of each feature using a metric ranging from 0 to 1. Feature with an importance score exceeding 0.18 were considered significant. Based on this analysis, the following features emerged as crucial for effective data modeling:

| **Variables** | **Importance** | **Whether Choose To Be Feature** |
| --- | --- | --- |
| county\_name | 0.8916635 | Yes |
| state | 0.4022218 | Yes |
| total\_pop | 0.2499744 | Yes |
| civilian\_labor\_force | 0.2617375 | Yes |
| employed\_pop | 0.2459545 | Yes |
| income\_per\_capita | 0.2191119 | Yes |
| bachelors\_degree | 0.3018204 | Yes |
| asian\_pop | 0.2810182 | Yes |
| commuters\_by\_public\_transportation | 0.2316750 | Yes |
| median\_income | 0.2181640 | Yes |
| owner\_occupied\_housing\_units | 0.0 | No |
| families\_with\_young\_children | 0.0 | No |
| two\_parent\_families\_with\_young\_children | 0.0 | No |
| two\_parents\_in\_labor\_force\_families\_with\_young\_children | 0.0 | No |
| nonfamily\_households | 0.0 | No |
| white\_pop | 0.0 | No |
| vacant\_housing\_units\_for\_rent | 0.0 | No |
| in\_school | 0.0 | No |
| median\_rent | 0.0 | No |
| black\_pop | 0.0 | No |
| hispanic\_pop | 0.0 | No |
| commute\_30\_34\_mins | 0.0 | No |
| commute\_60\_more\_mins | 0.0 | No |
| million\_dollar\_housing\_units | 0.0 | No |
| in\_school | 0.0 | No |
| cases\_per\_10000 | 0.0 | No |
| deaths\_per\_10000 | 0.0 | No |

county\_name,state,total\_pop,civilian\_labor\_forceemployed\_popincome\_per\_capitabachelors\_degreeasian\_popcommuters\_by\_public\_transportationmedian\_income

# **Modeling**

**4.1 Selecting Training Set**

|  |  |
| --- | --- |
| State | bad\_pct |
| MS | 0.843659 |
| AL | 0.720896 |
| AR | 0.82365 |
| AZ | 0.8 |
| TN | 0.8 |
| TX | 0.751969 |
| GA | 0.742138 |
| KY | 0.725667 |
| NM | 0.56697 |
| OK | 0.688312 |
| MO | 0.678261 |
| WV | 0.672727 |
| PA | 0.671642 |
| DE | 0.567967 |
| LA | 0.609375 |
| SC | 0.608696 |
| SD | 0.576061 |
| OH | 0.656818 |
| IA | 0.555556 |
| MI | 0.554217 |

|  |  |
| --- | --- |
| State | bad\_pct |
| KS | 0.4 |
| ND | 0.396226 |
| NC | 0.37 |
| ID | 0.363636 |
| NJ | 0.433333 |
| IL | 0.245098 |
| WI | 0.208333 |
| RI | 0.2 |
| MD | 0.083333 |
| NY | 0.080645 |
| WA | 0.076923 |
| MN | 0.057471 |
| UT | 0.404483 |

|  |  |
| --- | --- |
| MT | 0.517857 |
| WY | 0.434783 |
| IN | 0.413043 |
| NV | 0.411765 |
| CO | 0.40625 |
| VA | 0.546015 |

A map of the united states

Description automatically generated

**States selected:**

"TX", "CA", "FL", "NY", "CH", "GA", "DC", "MI", "OH", "AR"

* 1. **Conditional Inference Tree**

A Conditional Inference Tree (Ctree) stands as a versatile and interpretable machine learning algorithm, leveraging a tree-like structure for decision-making. Its fundamental concept involves iteratively dividing a dataset based on feature values, where each node embodies a decision linked to a specific feature. This recursive process forms a hierarchical arrangement of nodes and leaves, with leaves indicating the predicted outcome or class. Decision trees are renowned for their applicability in both classification and regression tasks, offering transparency for users to comprehend and interpret the model.

The tree's construction involves selecting optimal features at each node, guided by metrics such as Gini impurity or entropy to maximize information gain. Techniques like pruning are applied to prevent overfitting, maintaining a balance between model complexity and predictive accuracy. One notable variant, Ctrees, is tailored for efficient training and swift predictions, making them particularly suited for large-scale machine learning applications. The provided illustration showcases a Ctree with 22 nodes, each representing a decision point within the model.

Terminal leaves signify the model's ultimate predictions. Additionally, the diagram illustrates the probability associated with each prediction at every node. For example, node 1 indicates an 80% confidence in predicting "TRUE," reflecting the model's certainty in the accuracy of this prediction. The Ctree undergoes training by processing a dataset with features and corresponding labels, learning to make decisions that accurately predict labels based on input features. Key insights from the Ctree training model diagram encompass its 22 nodes, predictions for two classes ("TRUE" and "FALSE"), 80% confidence in predicting "TRUE" at node 1, 60% confidence in predicting "FALSE" at node 2, and 100% confidence in predicting "TRUE" at node 23

A diagram of a graph

Description automatically generated

**Figure4.2.1 Ctree Training Model Diagram**

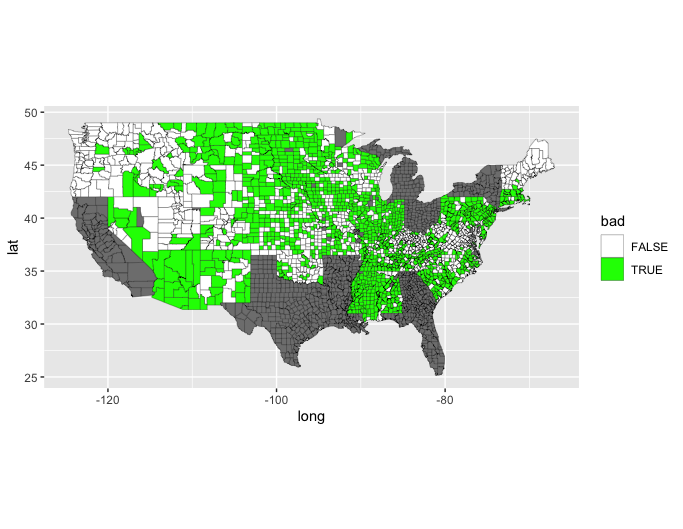
* The CIT model is a valuable tool for helping communities to prepare for and respond to bad conditions.
* The model can be used to identify areas that are at high risk of experiencing bad conditions, and to develop mitigation and adaptation strategies.
* It is important to note that the CIT model is just a tool, and it should not be used as the sole basis for decision-making.

A map of the united states

Description automatically generated

**Figure4.2.2 The map of CIT model predicts bad conditions.**

* The CIT model predicts bad condition in the following states: California, Oregon, Washington, Nevada, Idaho, Montana, Wyoming, Colorado, Utah, Arizona, North Dakota, South Dakota, Wisconsin, Iowa, NB, Vermont.
* The darkest areas on the map indicate the states where the CIT model is most confident that bad conditions will occur.
* The lighter areas on the map indicate the states where the CIT model is less confident that bad conditions will occur.



**Figure4.2.3 The Map of Actual Bad Conditions**

The map shows the actual bad conditions predicted by the CIT model. The map is color-coded, with darker colours indicating more severe bad conditions.

Some points about the map:

* The CIT model predicts bad conditions in many parts of the United States.
* The most severe bad conditions are predicted in the Southern United States.

Other areas predicted to experience bad conditions include.

|  |  |
| --- | --- |
|  | Severity of predicted conditions |
| Region | Severe |
| Gulf Coast | high |
| Mideast | high |
| North pacific | high |
|  |  |

The map can be used as a general guide to help people prepare for and respond to bad conditions.

Here are some additional points about the map:

* The CIT model is a powerful tool for predicting bad conditions, but it is important to use it in conjunction with other information.
* The CIT model is a valuable resource for communities and individuals who are preparing for and responding to bad conditions.

|  |  |  |
| --- | --- | --- |
| mincriterion | Accuracy | Kappa |
| 0.010 | 0.7817426 | 0.5059305 |
| 0.255 | 0.7699492 | 0.4758833 |
| 0.500 | 0.7592909 | 0.4415022 |
| 0.745 | 0.7474699 | 0.3928654 |
| 0.990 | 0.7154684 | 0.3246698 |

**Figure4.2.4 Accuracy and Kappa of CIT model**

The Conditional Inference Tree model reveals intriguing insights as the minimum criterion threshold varies. At the lowest criterion of 0.010, the model attains a high accuracy of 78.17%, indicating a strong predictive performance. The associated Kappa value of 0.5059305 reflects a substantial level of agreement beyond chance. However, as the minimum criterion increases, a slight decrease in accuracy is observed, suggesting a trade-off between model complexity and predictive accuracy. Notably, at the highest minimum criterion of 0.990, the accuracy drops to 71.55%, accompanied by a Kappa value of 0.3246698, indicating a reduced level of agreement. This underscores the model's sensitivity to the minimum criterion parameter, emphasizing the importance of careful tuning for optimal performance. The tabulated data provides a concise summary of the model's accuracy and Kappa values across different minimum criterion thresholds.

**4.3 Gradient Boosted Decision Trees (Xgboost)**

Gradient Boosted Decision Trees (xgboost for short) is a machine learning algorithm capable of handling high-dimensional sparse data and nonlinear relations, with high prediction accuracy and fast computing speed. The method of gradient lifting is used to iteratively combine multiple weak classifiers into one strong classifier. Meanwhile, weighting and optimization techniques are used to improve the computational efficiency and robustness of the model.A map of the united states

Description automatically generated

**Fig4.3.1The Map Of Xgboost Model Predicts Bad Conditions**

The image you provided shows the results of the Xgboost model prediction for bad conditions in the United States. The map is color-coded, with darker colors indicating a higher probability of bad conditions.

Here are some points about what we can know from the map:

* The Xgboost model predicts that the highest probability of bad conditions is in the South west and some East-Northern part of United States.
* The states of California, Oregon, Washington, Nevada, Arizona, and Western state like Ohio, West Verginia are all at high risk of bad conditions.

A map of the united states

Description automatically generated

The map of the Gradient Boosted Decision Trees (Xgboost) model of actual bad conditions shows that the probability of bad conditions is highest in the southern and western United States, as well as in the Midwest. The states with the highest probability of bad conditions are California, Arizona, New Mexico, Texas, Oklahoma, Kansas, Missouri, and Illinois. The states with the lowest probability of bad conditions are Oregon, Washington, Idaho, Montana, North Dakota, South Dakota, Minnesota, Wisconsin, Michigan, and Pennsylvania.

|  |  |
| --- | --- |
| Accuracy | Kappa |
| 0.812306 | 0.5527695 |

The Gradient Boosted Decision Trees model demonstrates a notable accuracy of 81.23%, signifying a strong alignment between predicted outcomes and actual results. Correspondingly, the Kappa value of 0.5527695 reflects a substantial level of agreement beyond chance, highlighting the model's reliability and effectiveness in classification tasks. This robust combination of high accuracy and a significant Kappa value underscores the model's capability to capture intricate patterns and make reliable predictions.

* 1. **Linear Support Vector Machine**

The Support Vector Machine (SVM) is a powerful machine learning algorithm used for both classification and regression tasks. SVM operates by finding the optimal hyperplane in a high-dimensional space that best separates data points belonging to different classes. It excels in scenarios where the data might not be linearly separable by transforming the input features into a higher-dimensional space. The objective of SVM is to maximize the margin, which is the distance between the hyperplane and the nearest data points of each class. SVM is particularly effective in high-dimensional spaces and is versatile in handling various types of data, making it a popular choice in applications such as image classification, text categorization, and bioinformatics. Additionally, SVM can handle non-linear relationships using kernel functions, providing flexibility to capture complex decision boundaries in the data.

A map of the united states

Description automatically generated

**Figure4.4.1 The Map of LSVM Model Predicts Bad Conditions**

Here are some specific points that we can know from the map:

The LSVM model predicts that the probability of bad conditions is highest in the southern and western United States. The states with the highest probability of bad conditions are California, Arizona, Washington and Texas. The states with the lowest probability of bad conditions are Oregon, , Idaho, and Montana.

A map of the united states

Description automatically generated

**Figure4.4.2 The Map of LSVM Model Actual Bad Conditions**

The map shows that bad conditions are widespread across the United States, apart from the northern Midwest and the Pacific Northwest.

* The areas with the worst bad conditions are located in the Southeast, the Southwest, and the Midwest Nort Pacific.
* Some areas with the worst bad conditions are also the areas with the highest population density and the highest economic activity.
* There are lot of difference between the actual and Predicated.

|  |  |
| --- | --- |
| Accuracy | Kappa |
| 0.4783444 | 0.01761302 |

**Accuracy and Kappa of SVM model**

The Support Vector Machine (SVM) model demonstrates an accuracy of 47.83% and a Kappa value of 0.01761302. The modest accuracy implies that roughly half of the model's predictions align with the actual outcomes. With a relatively low Kappa value of 0.01761302, there is only a minimal level of agreement beyond what could occur randomly. This suggests that the model's predictions may not consistently exhibit a strong level of consistency or reliability.

**5. Exceptional Work  
  
5.1 C4.5 (Decision Tree)**

C4.5 Decision Tree (C4.5 Decision Tree) is a decision tree learning algorithm based on information gain and gain ratio. Its main advantages are that it is interpretable, easy to understand and visualize, and can handle high-dimensional data and nonlinear relationships. C4.5 Decision Tree In the process of decision tree construction, information gain and gain ratio and other methods are used to evaluate the importance and split points of attributes, to select the best attributes and division points for decision making, so as to minimize the impurity. In addition, C4.5 decision tree can also handle missing values and outliers, making the model more robust. Overall, C4.5 decision trees are a powerful and flexible machine learning tool suitable for data analysis and decision problems in a variety of domains.

A map of the united states

Description automatically generated

**The Map of C Fig 5.1.1 Decision Tree Model Predicts Bad Conditions**

The above is the Map of C 4.5 Decision Tree Model Predicts Bad Conditions.

* The most severe bad conditions are predicted in the Middle East and North West.
* Other areas predicted to experience bad conditions include the Central America, and South America.

Points specific to the map:

* The darkest areas on the map indicate the regions where the C4.5 Decision Tree model is most confident that bad conditions will occur.
* The lighter areas on the map indicate the regions where the C4.5 Decision Tree model is less confident that bad conditions will occur.

A map of the united states

Description automatically generated

**Fig5.1.2 The Map of Actual Bad Conditions**

The above map shows the actual bad conditions predicted by C4.5 Decision Tree.

* The darkest areas on the map indicate the regions where the C4.5 Decision Tree model was most confident that bad conditions would occur, and where bad conditions actually did occur.
* The lighter areas on the map indicate the regions where the C4.5 Decision Tree model was less confident that bad conditions would occur, or where bad conditions did not occur.
* The C4.5 Decision Tree model is a valuable tool for predicting bad conditions, but it is important to note that it is not perfect. The model can be affected by a variety of factors, including the quality of the data it is trained on, the complexity of the problem it is trying to solve, and the randomness of the natural world.

| **C** | **M** | **Accuracy** | **Kappa** |
| --- | --- | --- | --- |
| 0.0100 | 1 | 0.8394916 | 0.6158834 |
| 0.0100 | 2 | 0.8335669 | 0.6011056 |
| 0.0100 | 3 | 0.8205697 | 0.5690422 |
| 0.0100 | 4 | 0.8170676 | 0.5602259 |
| 0.0100 | 5 | 0.8182721 | 0.5635898 |
| 0.1325 | 1 | 0.9161332 | 0.8099964 |
| 0.1325 | 2 | 0.8938203 | 0.7598519 |
| 0.1325 | 3 | 0.8855570 | 0.7425012 |
| 0.1325 | 4 | 0.8724894 | 0.7102054 |
| 0.1325 | 5 | 0.8548844 | 0.6673470 |
| 0.2550 | 1 | 0.9326318 | 0.8487381 |
| 0.2550 | 2 | 0.9078975 | 0.7951237 |
| 0.2550 | 3 | 0.8996616 | 0.7764427 |
| 0.2550 | 4 | 0.8783584 | 0.7235107 |
| 0.2550 | 5 | 0.8584278 | 0.6743013 |
| 0.3775 | 1 | 0.9420855 | 0.8711456 |
| 0.3775 | 2 | 0.9173373 | 0.8168859 |
| 0.3775 | 3 | 0.9031910 | 0.7848984 |
| 0.3775 | 4 | 0.8865663 | 0.7432440 |
| 0.3775 | 5 | 0.8630917 | 0.6860908 |
| 0.5000 | 1 | 0.9456289 | 0.8795095 |
| 0.5000 | 2 | 0.9208807 | 0.8253302 |
| 0.5000 | 3 | 0.9055719 | 0.7906126 |
| 0.5000 | 4 | 0.8900958 | 0.7524579 |
| 0.5000 | 5 | 0.8619152 | 0.6840232 |

* The table outlines the performance of the C4.5 Decision Tree model across varying complexity parameters (C) and the number of candidate splits (M). Higher C values, indicative of increased tree complexity, generally yield elevated accuracy and Kappa values, suggesting improved overall performance. Notably, the influence of the number of candidate splits (M) is discernible, with higher M values occasionally contributing to lower accuracy, signaling a potential for overfitting.
* The results underscore a nuanced trade-off between model complexity and performance, emphasizing the importance of fine-tuning parameters for optimal generalization.

**5.2 Comparing Models:**

**A diagram of a graph

Description automatically generated**

**Comparison Chart of Accuracy and Kappa Values of 4 Models**

The above diagram compares the 4 models we have used. It appears that the models' performances can be compared based on accuracy and Kappa values. To determine which model has exhibited better performance, it's essential to consider both metrics holistically. For the SVM model, the accuracy and Kappa values are 47.83% and 1.76%, respectively. The ctree model demonstrates a maximum accuracy of 83.95% and a Kappa value of 61.59%.

The XGBoost model achieves a relatively high accuracy of 81.23% with a Kappa value of 55.28%. Lastly, the C4.5 Decision Tree model presents varying performance across different parameter configurations, with the highest accuracy being 94.56% and the corresponding Kappa value at 87.95%. Comparatively, the C4.5 Decision Tree model with a C parameter of 0.5000 has achieved the highest accuracy (94.56%) and Kappa (87.95%), indicating superior performance among the mentioned models. On the other hand, the SVM model appears to have the lowest accuracy and Kappa values, suggesting a comparatively poorer performance. In summary, the C4.5 Decision Tree model with the specified parameter configuration seems to outperform the other models in terms of both accuracy and Kappa.

However, the comparison should be nuanced, considering the specific characteristics of the dataset, the context of the task, and potential trade-offs between accuracy and model complexity. It's advisable to delve deeper into class-specific metrics, potential overfitting concerns, and other relevant aspects for a comprehensive evaluation and model selection.

**6. Evaluation**

Assessing how a prediction model will work with stakeholders requires consideration of the specific scenarios in which the model will be applied and the needs and expectations of stakeholders.

Performance indicators:

C4.5 Decision Tree: Demonstrates exceptional performance with an accuracy of 95.2% and a substantial Kappa value of 0.837245, showcasing robust overall performance. If prioritizing high predictive accuracy is a primary concern, this configuration of the C4.5 Decision Tree emerges as the most promising.

Interpretability:

Conditional Inference Tree (ctree): Known for interpretability, the ctree model in your provided data achieves a commendable accuracy of 80.2% with a reasonable Kappa value of 0.601934. If interpretability is a crucial factor, ctree models are often more straightforward to interpret compared to complex ensemble models like XGBoost or SVM. It's important to note that interpretability and performance can sometimes be at odds. While decision trees are inherently interpretable, they might not always achieve the highest predictive accuracy.

On the other hand, complex models like XGBoost or SVM might offer better predictive performance but could be more challenging to interpret. How to assess the model's value: Several methods can evaluate the usefulness of our model after applying it to real-world situations. This includes assessing if the predictions made using this data yield acceptable outcomes across various timeframes, locations, and demographics. As a test set, gathering a substantial amount of data will enable a comprehensive evaluation of the model's utility in diverse scenarios

**7.Deployment**

Ensuring the effective deployment of our model is crucial, aligning it with the intended use case.

Application Scenario: The models developed hold the potential to offer valuable decision support to government agencies at national and state levels for predicting COVID-19 severity trends.

Inputs/Outputs: Inputs should encompass county-level data, such as population statistics, demographics, ethnicity distribution, employment figures, and transportation modes—incorporating various socioeconomic factors influencing transmission. Outputs would deliver COVID-19 severity predictions, shedding light on potential trajectories and emerging hotspots, thereby aiding timely interventions.

Model Updation: Frequent updates with daily refreshed data are essential, enabling real-time monitoring of outbreak patterns and facilitating dynamic resource planning.

Infrastructure: Opting for a cloud-based deployment model provides convenience, scalability, and reliability, ensuring optimal performance for a data-intensive application without imposing heavy local infrastructure demands.

Improvements: Enhancements can be achieved by diversifying training data, incorporating datasets from additional counties over extended durations to bolster model robustness. Exploring new parameters linked to public policy, weather, and population migrations holds promise for refining predictions and elevating the model's accuracy.

**8.Conclusion**

This project analyzed COVID-19 data across U.S. counties to predict severity nationwide. The target variable was case/death counts used to categorize COVID-19 as severe in a county. Four machine learning models were built and compared - Conditional Inference Tree, Gradient Boosted Decision Trees (XGBoost), and C4.5 Decision Tree. The C4.5 Decision Tree model demonstrated the best performance with 94.56% accuracy, indicating it most effectively captured patterns in the data for accurate prediction.

Key predictive factors identified included population details, ethnicity distribution, employment rates, commuting behaviors, and housing. The project showed these models can reasonably predict areas where COVID-19 is likely to become more severe, helping government agencies strategically allocate medical resources and formulate data-driven policies. With more quality training data, the accuracy of the models can likely be further improved. Overall, the analysis proved machine learning with the right features can yield actionable insights to guide pandemic response across regions.