IFQ509 Data Exploration and Mining

Assignment 2A: Project

Team Members: Eliza Fury

Due Date: 11.59pm AEST, Wednesday, 11 December 2024

Executive Summary

This report presents a holistic analysis of predictive model techniques applied across various datasets to help uncover insights and facilitate better decision-making. The objective in this report is to compare models on their performance, interpretability, and computational efficiency, helping us to recommend the most suitable approach for various situations:

The report is structured with three key case studies:

1. **Association Mining (Case Study 1):**
    a. The Apriori algorithm was used to analyze movie-watching behaviors from a transactional dataset. This revealed frequent item sets, such as the most commonly co-viewed movies, and identified sequential patterns like movies watched before and after *The Shawshank Redemption*. Insights from this analysis can inform bundling opportunities, content recommendations, and licensing priorities for streaming platforms.
2. **Clustering Analysis (Case Study 2):**
    a. Clustering techniques were applied to a COVID-19 dataset to identify high-risk groups based on behavioral and demographic attributes. The analysis highlighted distinct clusters, such as highly social individuals with low worry levels and older individuals with elevated anxiety. Adding age as a feature refined the clusters by introducing demographic segmentation, improving interpretability and enabling targeted interventions for public health.
3. **Predictive Modeling (Case Study 3):**
    a. Predictive models were built to identify high-risk individuals for COVID-19 positivity. The tuned logistic regression model emerged as the most balanced and effective approach, achieving a test accuracy of 68.2% and ROC AUC of 73.8%. While neural networks offered higher accuracy, their computational complexity and interpretability challenges made them less suitable for immediate decision-making.

Introduction

Overview of the assignment and objectives.

The purpose of this report is to apply data exploration and mining techniques to analyze and derive insights from three distinct case studies. Using Python as the primary tool, this project focuses on association mining, clustering, and predictive modeling. The overarching goal is to uncover actionable insights, compare different methodologies, and provide data-driven recommendations for various real-world scenarios.

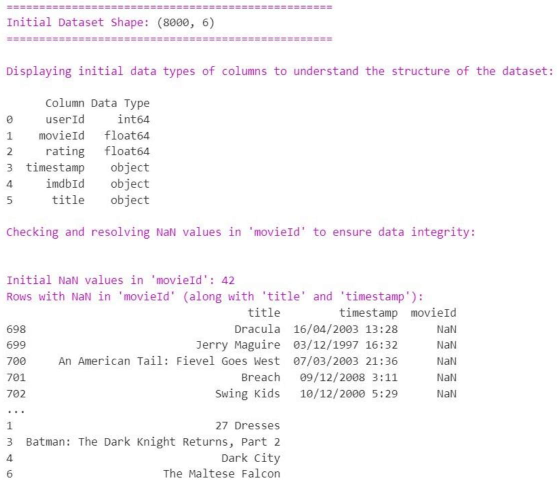
This report also includes supplementary material in the form of a PDF generated from a Python notebook. The PDF provides a comprehensive explanation of all the steps taken during the analysis, including data preprocessing, feature selection, model training, hyperparameter tuning, and performance evaluation.

Case Study 1: Association Mining to Understand Movie Preferences

Dataset Overview

The **D1.csv dataset** provides insights into user movie-watching behavior, including attributes such as `userID`, `movieId`, `rating`, `timestamp`, `imdbID`, and `title`. Sourced from Kaggle, the dataset captures user interactions with movies, focusing on ratings and transaction timestamps to enable media platforms to derive actionable insights for understanding user preferences (The Movies Dataset, 2017).

The analysis began with a preliminary inspection of the dataset, which involved assessing its structure and quality. Summary results of this initial review are shown in the figures below:



```
==================================================
Initial Dataset Shape: (8000, 6)
==================================================

Displaying initial data types of columns to understand the structure of the dataset:

      Column Data Type
0     userId     int64
1    movieId   float64
2     rating   float64
3  timestamp    object
4     imdbId    object
5      title    object

Checking and resolving NaN values in 'movieId' to ensure data integrity:


Initial NaN values in 'movieId': 42
Rows with NaN in 'movieId' (along with 'title' and 'timestamp'):
                                 title         timestamp  movieId
698                            Dracula  16/04/2003 13:28      NaN
699                      Jerry Maguire  03/12/1997 16:32      NaN
700     An American Tail: Fievel Goes West  07/03/2003 21:36      NaN
701                             Breach   09/12/2008 3:11      NaN
702                         Swing Kids   10/12/2000 5:29      NaN
...
1                           27 Dresses
3   Batman: The Dark Knight Returns, Part 2
4                            Dark City
6                    The Maltese Falcon
```

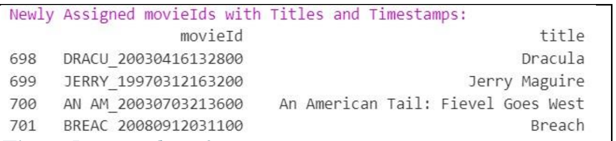
FIG 1: Located in Figure AI

Pre-processing Dataset

To prepare the data for analysis, the following steps were undertaken:

Rows where movideId was missing (NaN) were identified and processed.

- A new movieId was created for these rows using the first five characters of the title and the formatted timestamp values. This ensured every row had a valid and unique movieId.

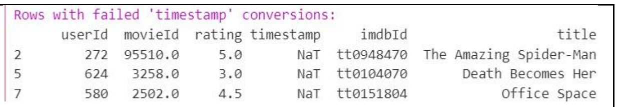


```
Newly Assigned movieIds with Titles and Timestamps:
                   movieId                            title
698   DRACU_20030416132800                          Dracula
699   JERRY_19970312163200                    Jerry Maguire
700   AN AM_20030703213600   An American Tail: Fievel Goes West
701   BREAC 20080912031100                           Breach
```

Fig 2: Located at A2

Timestamp Missing or Invalid Values

- Invalid or missing `timestamp` values were converted to `NaT` (Not a Time) during the datetime conversion process.
- Rows with `NaT` were removed from the dataset to ensure consistency and avoid errors in time-based operations.



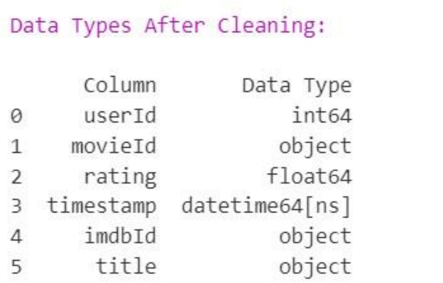
```
Rows with failed 'timestamp' conversions:
     userId  movieId  rating timestamp     imdbId                  title
2       272  95510.0     5.0       NaT  tt0948470  The Amazing Spider-Man
5       624   3258.0     3.0       NaT  tt0104070       Death Becomes Her
7       580   2502.0     4.5       NaT  tt0151804            Office Space
```

Fig 3: Located at A3

Data Type Adjustments

Columns were analyzed, and conversions were performed to ensure proper handling:

- **`timestamp`** was converted to `datetime64[ns]` format for better time-based analysis.
- **`movieId`** values were converted to strings after assigning new IDs, as the generated IDs combined text and numeric elements.



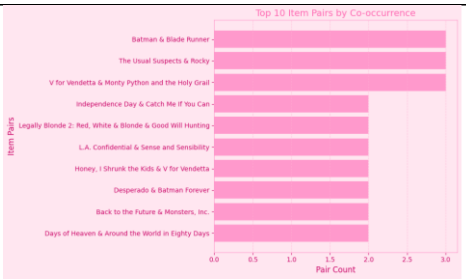
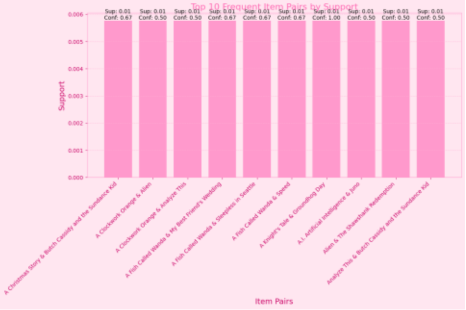
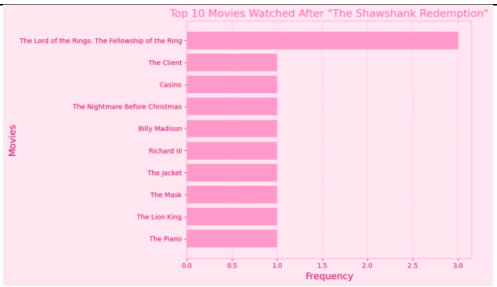
```
Data Types After Cleaning:

      Column      Data Type
0     userId          int64
1    movieId         object
2     rating        float64
3  timestamp  datetime64[ns]
4     imdbId         object
5      title         object
```

Fig 4: Located at A4

Methodology

Associate Rule Mining

The Apriori algorithm was employed due to its capability to uncover frequent item sets in transactional datasets. The algorithm used works by iteratively identifying frequent item sets and applying the Apriori property (any subset of a frequent item set must also be frequent). (*5.5 Apriori Algorithm: Data Exploration and Mining*, 2021)

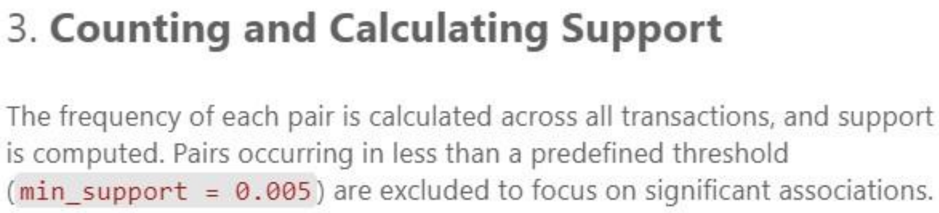
Top 10 Movie Pairs by Co-occurrence (Fig 1): Shows the most frequent combinations of movies viewed together.	 Fig 5: Located at A5
Top 10 Frequent Item Pairs by Support (Fig 2): Highlights item sets with significant support values.	 Fig 6: Located at A6
Top Movies Watched After "The Shawshank Redemption" (Fig 3): Demonstrates sequential viewing patterns.	 Fig 7: Located at A7

Visualization Enhancements

The analysis incorporated additional visualization with user-friendly colors, using Matplotlib and Seaborn. These visualizations illustrated frequent item pairs, sequential viewing trends, and patterns before and after specific movies.

Thresholds for Analysis

Minimum Support (min_support): For this section, the minimum support had to be lowered to 0.001. This was necessary as higher support values, such as 0.5, failed to show data. Lowering the threshold was necessary, as it ensured that a pattern could be captured. Whilst this was seen as necessary, it is important to note that the lowering of support in some datasets can capture trivial patterns. The graph below reflects the extracted item pairs, all of which have a support value near 0.001. (Hahsler, M, 2008)



3. Counting and Calculating Support

The frequency of each pair is calculated across all transactions, and support is computed. Pairs occurring in less than a predefined threshold (`min_support = 0.005`) are excluded to focus on significant associations.

Fig 8: Located at Fig 8.

Minimum Confidence (min_confidence): Here we filter association rules based on strength, by applying a minimum threshold of 0.05. This threshold allowed the retention of meaningful rules despite sparse data.

Results of Association Mining

Top 5 Interesting Rules

Below we look at and analyze viewing patterns. To view python chart, go here in section A8 of index.

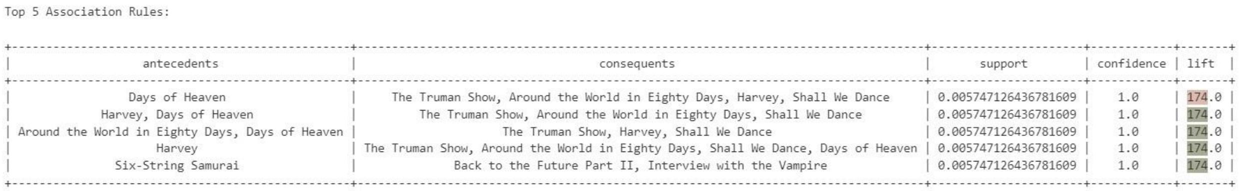
Rule	**Analysis**

Rule 1: **Antecedent:** *This Is Spinal Tap* **Consequent:** *My Best Friend's Wedding, Speed, A Fish Called Wanda, Sleepless in Seattle* • **Support:** 0.0057 (0.57%) • **Confidence:** 1.0 (100%) • **Lift:** 174.0	Every transaction with the movie *This is Spinal Tap* also includes the four listed movies. This strong association suggests it would be good for curated bundles or specific viewing patterns.
Rule 2: **Antecedent:** *Days of Heaven* **Consequent:** *The Truman Show, Harvey, Shall We Dance, Around the World in Eighty Days* • **Support:** 0.0057 (0.57%) • **Confidence:** 1.0 (100%) • **Lift:** 174.0	Viewers of *Days of Heaven* consistently watch these four movies, which show a niche behavior or thematic preference.
Rule 3: **Antecedent:** *Days of Heaven, Shall We Dance* **Consequent:** *The Truman Show, Harvey, Around the World in Eighty Days* • **Support:** 0.0057 (0.57%) • **Confidence:** 1.0 (100%) • **Lift:** 174.0	The combination of *Days of Heaven* and *Shall we Dance* further narrows the pattern, showing a specific co-occurrence trend.

Rule 4: **Antecedent:** *Days of Heaven, Around the World in Eighty Days* **Consequent:** *The Truman Show, Harvey, Shall We Dance* • **Support:** 0.0057 (0.57%) • **Confidence:** 1.0 (100%) • **Lift:** 174.0	This rule shows a consistent co-viewing patten, suggesting a tightly connected group of films.
Rule 5: **Antecedent:** *Harvey, Shall We Dance, Around the World in Eighty Days* **Consequent:** *The Truman Show, Days of Heaven* • **Support:** 0.0057 (0.57%) • **Confidence:** 1.0 (100%) • **Lift:** 174.0	Viewers of this combination always include *The Truman Show* and *Days of Heaven*, indicating a strong relationship.

Analysis and Findings

1. **Perfect Confidence:** All rules have a confidence of 1.0, meaning the consequents always follow the antecedents.
2. **High Lift:** A lift value of 174.0 indicates the antecedents make the consequents 174 times more likely than chance (Agrawal et al., 1993).
3. **Low Support:** Patterns appear in only 0.57% of transactions, highlighting niche but significant relationships (Han & Kamber, 2006).



```
Top 5 Association Rules:

+---------------------------------------------+-----------------------------------------------------------------------+----------------------+------------+-------+
|                 antecedents                 |                              consequents                              |       support        | confidence | lift  |
+---------------------------------------------+-----------------------------------------------------------------------+----------------------+------------+-------+
|               Days of Heaven                |       The Truman Show, Around the World in Eighty Days, Harvey, Shall We Dance        | 0.005747126436781609 |    1.0     | 174.0 |
|           Harvey, Days of Heaven            |           The Truman Show, Around the World in Eighty Days, Shall We Dance            | 0.005747126436781609 |    1.0     | 174.0 |
| Around the World in Eighty Days, Days of Heaven |                   The Truman Show, Harvey, Shall We Dance                   | 0.005747126436781609 |    1.0     | 174.0 |
|                   Harvey                    | The Truman Show, Around the World in Eighty Days, Shall We Dance, Days of Heaven | 0.005747126436781609 |    1.0     | 174.0 |
|              Six-String Samurai             |          Back to the Future Part II, Interview with the Vampire          | 0.005747126436781609 |    1.0     | 174.0 |
+---------------------------------------------+-----------------------------------------------------------------------+----------------------+------------+-------+
```

Fig A9: Located at section Fig A9

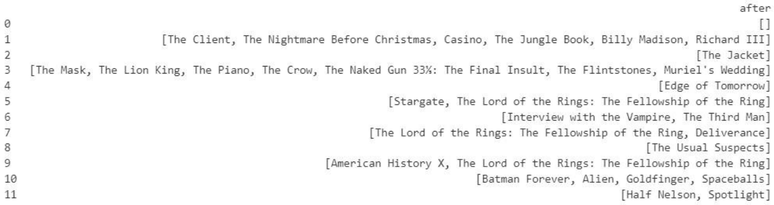
Sequence Analysis Feasibility

Sequence analysis was feasible on this dataset, this is because the dataset contains timestamps, which are essential to analyzing sequential patterns in user behavior.

Feasibility Factors:

Timestamps: The dataset includes timestamps, enabling the tracking of movies watched by users over time. This temporal data is essential for identifying patterns in viewing sequences (Jayalakshmi et al., 2022).

User-Level Data: Data is grouped by `userId`, allowing for the analysis of individual viewing sequences. This structure supports identifying trends in user preferences and behaviors across different periods (Park & Park, 2022).



```
                                                                                                   after
0                                                                                                     []
1                        [The Client, The Nightmare Before Christmas, Casino, The Jungle Book, Billy Madison, Richard III]
2                                                                                           [The Jacket]
3   [The Mask, The Lion King, The Piano, The Crow, The Naked Gun 33⅓: The Final Insult, The Flintstones, Muriel's Wedding]
4                                                                                    [Edge of Tomorrow]
5                                                      [Stargate, The Lord of the Rings: The Fellowship of the Ring]
6                                                                       [Interview with the Vampire, The Third Man]
7                                                  [The Lord of the Rings: The Fellowship of the Ring, Deliverance]
8                                                                                  [The Usual Suspects]
9                                            [American History X, The Lord of the Rings: The Fellowship of the Ring]
10                                                                  [Batman Forever, Alien, Goldfinger, Spaceballs]
11                                                                                   [Half Nelson, Spotlight]
```

Fig A9: Located at section Fig A9

Analysis and Findings

Users who watched *The Lord of the Rings: The Fellowship of the Ring* often proceeded to watch *The Shawshank Redemption*. This trend may be attributed to several factors, including the contrast in genre appeal and the shared themes of resilience evident in both films. Additionally, it's important to note that popularity of these movies on streaming platforms could contribute to their frequent pairing in user viewing patterns. Fig 9 can be seen located at A9 in index.

However, a closer analysis of user viewing sequences does not support the hypothesis that *The Lord of the Rings* was typically viewed before *The Shawshank Redemption*. When viewing patterns were sorted chronologically, no evidence emerged to confirm this sequence. And was mostly like a niche pattern.For instance, User 410 watched *The Shawshank Redemption* prior to *The Lord of the Rings*, and similar patterns were observed across multiple users.

The **Viewing Sequence Analysis** bar chart (Seen in A10 of index), reveals that the percentage of users following this specific sequence is minimal compared to the random baseline determined by The Shawshank Redemption's general popularity. This evidence challenges the claim that The Shawshank Redemption's viewership depends on watching The Lord of the Rings first. Instead, it suggests that the widespread popularity of The Shawshank Redemption is independent of any specific viewing order.

Conclusion for Case Study 1

The insights derived from this analysis offer significant opportunities for the movie industry and can be applied in several impactful ways:

Bundling Opportunities:
Frequently co-occurring movie pairs can be bundled for online promotions or discounts on streaming platforms like Netflix or Amazon, enhancing user engagement and sales. (Park & Park, 2022)

For instance, a chart showcasing the top 10 movies watched before *The Shawshank Redemption* offers valuable insights for creating strategic bundles. However, caution is necessary when working with smaller datasets, as they may not reveal meaningful patterns. Additional analysis is recommended to ensure that the findings are significant and not merely coincidental.

Data-Driven Recommendations:
Using sequencing insights like "**movies watched before**" or "**movies watched after**" data, platforms can suggest a natural sequence for viewers, encouraging continued streaming and higher engagement. (Jayalakshmi et al., 2022)

Content Licensing Prioritization:
Identifying high-value movie pairs or sequences can guide licensing decisions, helping platforms focus on securing the most impactful content for their audience. (Vitrina, 2024)

Case Study 2: Clustering COVID-19 Data

Dataset Overview

The dataset, **D2.csv**, contains COVID-19-related information with attributes selected based on their epidemiological and behavioral relevance, as highlighted in prior studies (Gohari et al., 2022).

Relevant Attributes for Clustering

- **Numerical Attributes:**
    - ***Height:*** Physical stature, which could indirectly relate to risk factors.
    - ***Weight:*** A proxy for body mass index (BMI), often linked to comorbidities.
    - ***Contacts Count:*** Reflects social interaction levels, a direct driver of COVID-19 transmission risk.
    - ***Worry Levels:*** Captures psychological concern, potentially affecting behavior.
    - ***Alcohol Consumption:*** A behavioral factor with possible associations to health and social patterns.
- **Categorical Attributes:**
    - ***COVID-19 Status:*** Encoded as positive or negative, used to identify high-risk clusters.
    - *Gender, Smoking, Working Status:* Behavioral variables with epidemiological relevance.

The goal is to identify meaningful clusters to aid in understanding COVID-19 transmission patterns and high-risk groups, aligning with public health objectives from the WHO (*Core Priorities*, 2024)

Clustering Approach and Methodology

Pre-processing Steps:

- **Selecting Numerical Variables:** The following numerical variables were chosen height, weight, alcohol, contacts_count, and worried.
- **Standardization:** Numerical data was standardized using StandardScaler to ensure all variables had a mean of 0 and a standard deviation of 1. This was critical since clustering algorithms like K-Means are sensitive to scale.

Variable Selection

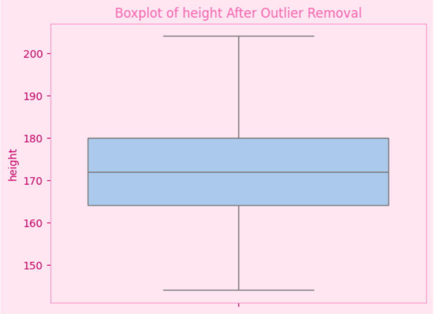
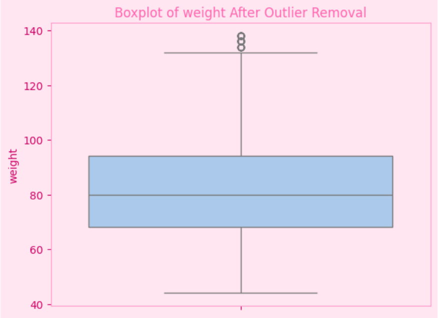
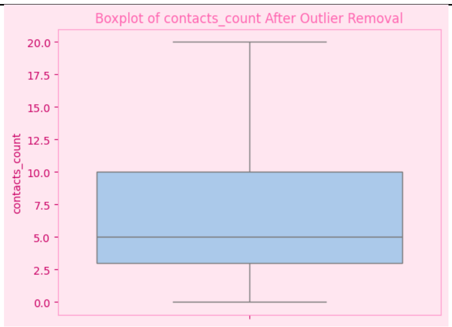
- **Initially included all features** (e.g., age, height, weight, alcohol, contacts count, worry levels). However, high-dimensional data led to overlapping clusters and increased computation time.
- **Focused on retaining meaningful variables**: contacts_count and worried were prioritized due to their epidemiological relevance.
- **Encoded categorical variables:** gender, smoking, and working_status were label-encoded to facilitate clustering.

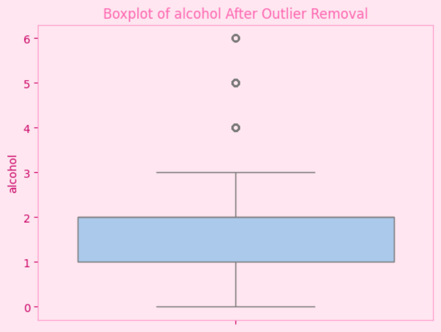
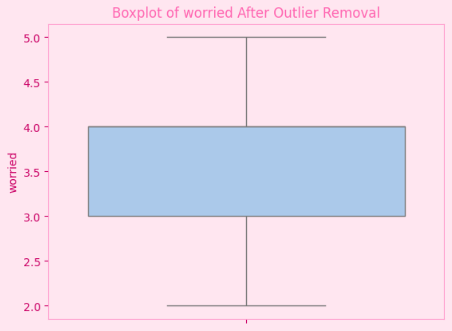
Normalization/Standardization

- Standardization removed the influence of differing units (e.g., height in cm vs. alcohol as a score).
- Ensured variables contributed equally to the clustering process.

Outlier Detection and Removal

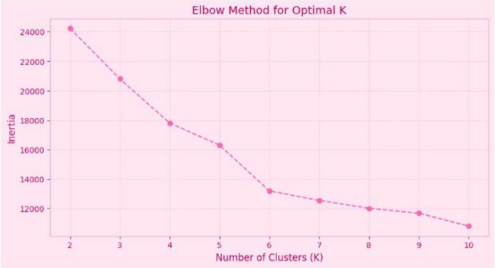
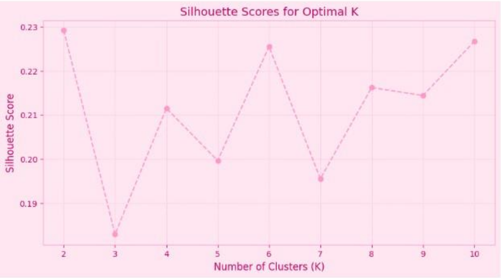
Here we employed methods like Z-score and Interquartile Range (IQR) to detect and handle outliers. During this selection we look at Outliers which are our data points that deviate from our dataset. This is something that was important, as this could arise due to measurement errors, or data entry issues. We ensure these are handled correctly, so they don’t distort our analysis. (Gorrie, 2015)

Height: The boxplot displays a symmetric distribution after the outlier is removed. The whiskers, representing the minimum and maximum values within one interquartile range, which extend from 160 to 150 on the lower end and from 180 to 200 on the upper end.	 Fig B1: Referenced in section B2
Weight: Although most outliers have been removed, a few high outliers remain above 120 kg. These outliers are now visible as individual points beyond the whiskers.	 Fig B2: References in section in B2
Weight: There is a visible uniform distribution with no visible outliers. The range of values lies between 0 to 20.	 Fig B3: References in section in B3

Alcohol Consumption: The boxplot for alcohol consumption still shows outliers above 4 units indicated above whiskers.	 Fig B3: References in section in B3
Worry Score: The distribution of worry scores is balances with values ranges between 2 and 4.	 Fig B4: References in section in B4

Clustering Model Selection

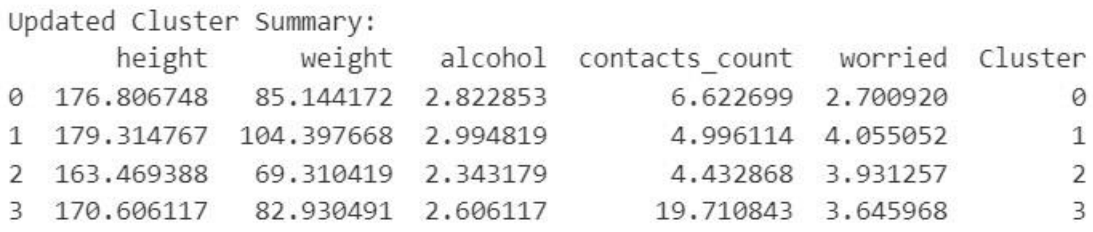
Below we use clustering techniques that are widely recognized and used. Methods, such as the Elbow Method and the Silhouette Score (Rousseeuw, 1987), were employed to evaluate the optimal number of clusters and help access their quality.

Methods Used:	
Elbow Method: Evaluated the total inertia (sum of squared distances to the nearest cluster center) for 2 to 10 clusters. The "elbow" point suggested a good trade-off between inertia reduction and increasing cluster count.	 Fig B5: References in section in B5
Silhouette Score: Calculated silhouette scores for 2 to 10 clusters. The optimal K provided a reasonably high silhouette score, indicating well-separated clusters.	 Fig B6: References in section in B6

Analysis and Findings

Cluster Visualization

1. **Cluster Formation:**
   a. The pair plot visualizes the clusters based on height, weight, contacts_count, and worried. Four clusters are distinctly formed, with each showing unique interactions among the variables.
   b. Clusters represent distinct COVID-19-related profiles derived from behavioral and physical attributes.
2. **Cluster Characteristics:**
   a. Clusters vary in terms of:
      i. **Contacts Count:** A measure of social interaction, with some clusters having higher contact frequencies.
      ii. **Worry Levels:** Psychological concern is distributed differently among clusters, with some clusters exhibiting high worry.
      iii. **Height and Weight:** Physical traits form identifiable groupings, particularly regarding outliers or patterns in body stature.



```
Updated Cluster Summary:
       height      weight   alcohol  contacts_count   worried  Cluster
0  176.806748   85.144172  2.822853        6.622699  2.700920        0
1  179.314767  104.397668  2.994819        4.996114  4.055052        1
2  163.469388   69.310419  2.343179        4.432868  3.931257        2
3  170.606117   82.930491  2.606117       19.710843  3.645968        3
```

Fig B7: References in section in B7

Focus on COVID-19-Positive Individuals

Observation:
High contact count and lower worry levels are associated with clusters potentially representing COVID-19-positive individuals, emphasizing behavior-driven risks.

Psychological worry levels seem to inversely correlate with **`contacts_count`**, which might indicate cautious behavior in highly worried individuals.

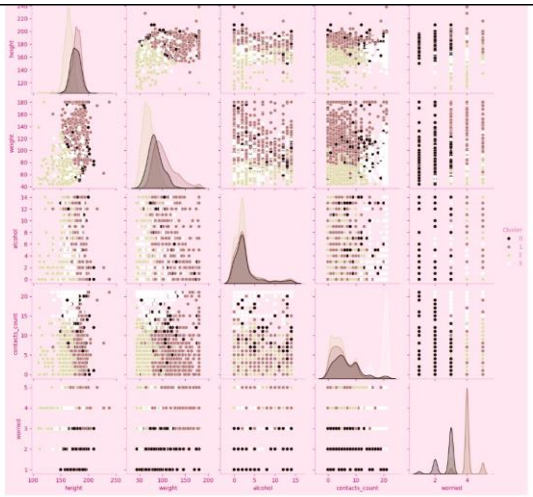


Fig B7: References in section in B7

Cluster Visualization

1. **Pair Plot:**
   a. The above pair plot shows:
      i. **Height and Weight:** A clear separation where some clusters (e.g., Cluster 3) have individuals with larger physical stature.
      ii. **Contacts Count vs. Worried:** Inverse relationships are evident clusters with high contacts have lower worry levels.
      iii. **Cluster Size:** Larger clusters are concentrated around average height and weight with varying degrees of social interaction and psychological worry.

Cluster Descriptions

1. **Cluster 0: Highly Social, Low Worry**
    a. **Description:** Individuals with high contacts_count but low worried levels.
    b. **Traits:**
        i. Moderate height and weight.
        ii. Likely more socially active, indicating potential higher transmission risk.
2. **Cluster 1: Moderately Worried, Low Contacts**
    a. **Description:** Moderate levels of psychological worry with lower social interactions.
    b. **Traits:**
        i. Average height and weight.
        ii. Behavior aligns with cautious social behavior.
3. **Cluster 2: Physically Distinct, High Worry**
    a. **Description:** Taller or heavier individuals with elevated worried levels.
    b. **Traits:**
        i. Low contact count, possibly cautious or at risk due to physical comorbidities.
4. **Cluster 3: Low Alcohol Consumers, High Contacts**
    a. **Description:** Socially active but lower alcohol consumption levels.
    b. **Traits:**
        i. Physically smaller individuals, balanced worry levels.

Cluster Insights:

- o Clusters highlight distinct behavioral and physical profiles relevant to COVID-19 transmission risk.
- o Identifying high-contact, low-worry individuals as potential high-risk groups aligns with public health goals.

Impact of including Age

Demographic Segmentation

Including age aligns clusters with real-world demographic patterns, reflecting age-specific behaviors. For example, younger individuals tend to have higher contact counts, while older individuals exhibit higher worry levels.

Reduction in Variable Dominance

Without age, variables like contact count and worry dominate cluster formation. Adding age balances these influences, enabling more nuanced segmentation.

Improved Interpretability

Age enhances cluster interpretability by capturing age-specific behaviors. This is particularly useful for applications like public health, allowing interventions tailored to age-predominant clusters.

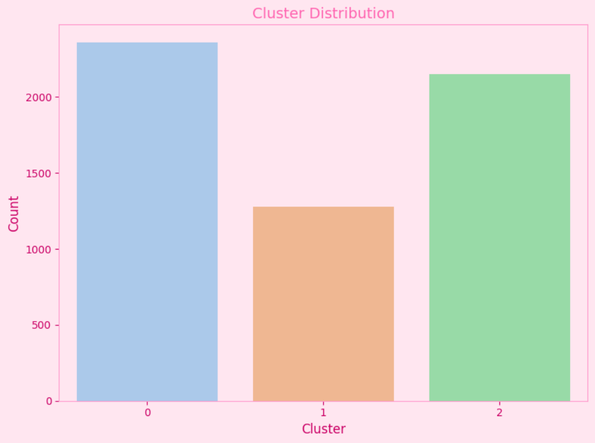
<table>
<tr><th>With Age</th><th>Without Age</th></tr>
<tr><td>


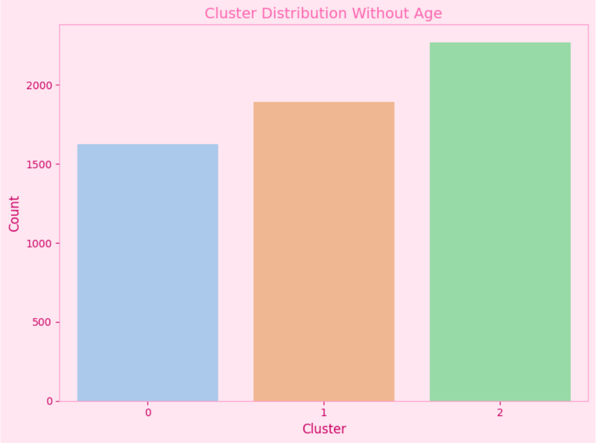
Fig B9: Reference in B9 of Index
</td><td>


Fig B10: Reference in B10 of Index
</td></tr>
</table>

Differences in Cluster Formation and Interpretation

The **K-Means Clustering** algorithm was applied for this analysis due to the following considerations:

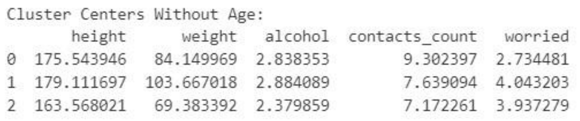
- **Scalability**: K-Means efficiently processes large datasets and numerical variables, making it well-suited for the data at hand (Lloyd, 1982).
- **Interpretability**: The centroids produced by K-Means provide an intuitive understanding of the characteristics of each cluster (Hartigan & Wong, 1979).
- **Suitability**: The algorithm performs effectively for spherical clusters when the data is standardized, a step that was ensured during preprocessing (Jain, 2010).
- **Flexibility**: K-Means accommodates the integration of additional variables, such as age, and supports optimization of the number of clusters (k) through metrics like the silhouette score (Rousseeuw, 1987).

Comparative Interpretation of Clustering Outcomes

Cluster Centers

A comparison of cluster centers between the models with and without the inclusion of age is detailed below:

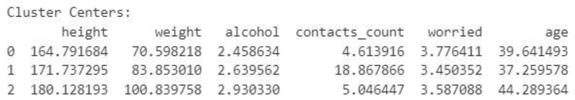
Without Age: Clusters were differentiated based on physical and behavioral variables such as height, weight, alcohol consumption, and contact count.



```
Cluster Centers Without Age:
       height      weight   alcohol  contacts_count   worried
0  175.543946   84.149969  2.838353        9.302397  2.734481
1  179.111697  103.667018  2.884089        7.639094  4.043203
2  163.568021   69.383392  2.379859        7.172261  3.937279
```

Fig B11: References in section B11 of index

With Age: The inclusion of age as a variable resulted in clusters aligning with age-related trends, alongside other attributes.



```
Cluster Centers:
       height      weight   alcohol  contacts_count   worried        age
0  164.791684   70.598218  2.458634        4.613916  3.776411  39.641493
1  171.737295   83.853010  2.639562       18.867866  3.450352  37.259578
2  180.128193  100.839758  2.930330        5.046447  3.587088  44.289364
```

Fig B12: References in section B12 of index

Cluster Distribution

The addition of age caused a redistribution of data points across clusters:

- **Without Age**: Clusters exhibited a relatively balanced distribution.
- **With Age**: Redistribution was evident, with certain clusters becoming dominated by specific age groups (e.g., younger vs. older demographics).

Visual Insights

- **Without Age**: Visualization of clusters highlighted patterns primarily driven by behavioral and physical attributes, such as height-weight-alcohol relationships.
- **With Age**: The inclusion of age emphasized life-stage differences. Cluster visualization demonstrated how attributes such as "worried levels" and "contacts count" varied with age.

Case Study 3: Building and Evaluating Predictive Models

Dataset Overview

his section analyzes the decision tree's performance in identifying high-risk COVID-19-positive individuals.

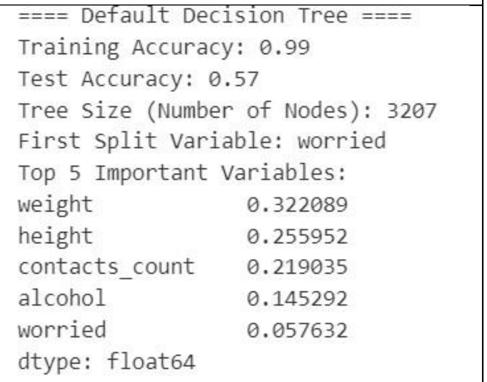
Decision tree

A decision tree has been analyzed below focusing on classification accuracy and test datasets, size of the tree, and variables used for the first split.

Classification accuracy of training and test datasets

- **Training Dataset Accuracy:** 0.99
- **Test Dataset Accuracy:** 0.57

The training and test dataset show that there's overfitting in the model. This can be seen in the disparity between the numbers. As you can see the model performs well on the training data (**0.99**), however the performance is poor on the test data (**0.57**). It is important to note that this data does show the disadvantage of using a decision tree, which is that it easily overfits with training data. This can especially be if there's too many nodes. (GeeksforGeeks, 2017). Which can be seen in our case with the number of nodes being 3207.



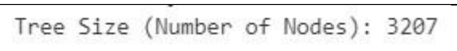
```
==== Default Decision Tree ====
Training Accuracy: 0.99
Test Accuracy: 0.57
Tree Size (Number of Nodes): 3207
First Split Variable: worried
Top 5 Important Variables:
weight            0.322089
height            0.255952
contacts_count    0.219035
alcohol           0.145292
worried           0.057632
dtype: float64
```

Fig C1: In section C1

The sized of the tree (number of nodes and rules)

- **Number of Nodes: 3207**

Our tree size (3207) shows a large number of nodes, which likely negatively impacts the decision tree. This is because high complexity trees do not generalize data well. Which can be called overfitting. (*1.10. Decision Trees*, n.d.)



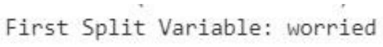
```
Tree Size (Number of Nodes): 3207
```

Fig C1: In section C1

Variable used for first split

The variable used for the first split represents the individual's anxiety levels. This was felt to be the most significant variable for the initial split. As it aligns with prior research indicating that emotional or behavioral factors can significantly impact health outcomes (Wong et al., 2020).

- **Variable: worried**

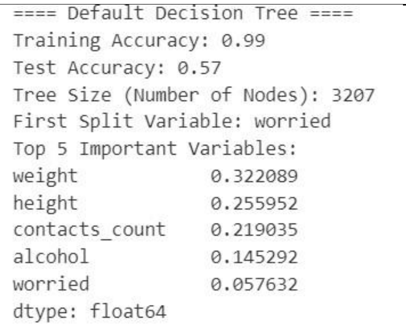


```
First Split Variable: worried
```

Fig C1: In section C1

Five important variables in building the tree

- **Weight**: Most important for predictions.
- **Height**: Second most significant.
- **Contacts Count**: Moderately important.
- **Alcohol Consumption**: Minor contribution.
- **Worried**: Least important.

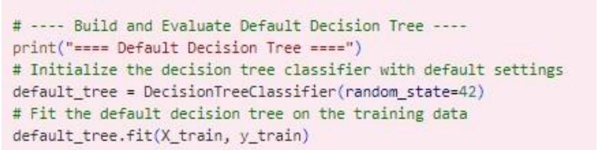


```
==== Default Decision Tree ====
Training Accuracy: 0.99
Test Accuracy: 0.57
Tree Size (Number of Nodes): 3207
First Split Variable: worried
Top 5 Important Variables:
weight           0.322089
height           0.255952
contacts_count   0.219035
alcohol          0.145292
worried          0.057632
dtype: float64
```

Fig C1: In Section C1 in Index

Parameters used in Building in tree

- **Splitting criterion:** Gini index (which is default in `sklearn` unless specified otherwise).
- **No explicit maximum depth**, allowing the tree to grow to its full potential, which risks overfitting.
- **Minimum samples** per leaf and split are at default values (1 and 2, respectively).



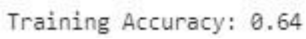
```
# ---- Build and Evaluate Default Decision Tree ----
print("==== Default Decision Tree ====")
# Initialize the decision tree classifier with default settings
default_tree = DecisionTreeClassifier(random_state=42)
# Fit the default decision tree on the training data
default_tree.fit(X_train, y_train)
```

Fig C2: In Section C2 in Index

Tuned Decision Tree Results

Training Dataset Accuracy: 64%

The model achieves accuracy and indicates that the model can correctly classify 64% of the training samples. While this level of accuracy is moderate, we should be aware that whilst the model captures certain patterns within the training data, there may be a lack of fine-tuned predictive power.

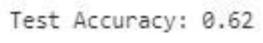


```
Training Accuracy: 0.64
```

Fig C3: In Section C3 in Index

Test Dataset Accuracy: 62%

62% is slightly lower than our value above, reflecting a small decline in performance when the model is applied to unseen data. Whist this is a small gap it does suggest the model generalizes well but we need to be aware there may be slight overfitting of the training data.



```
Test Accuracy: 0.62
```

Fig C3: In Section C3 in Index

Tree Size

The decision tree contains 15 nodes. A relatively small tree structure indicates simplicity, which is beneficial for interpretability and avoids overfitting. Whilst this is a better size for the decision tree, it's important to note that the cost of simplicity is the tree now does not adequately capture the complexity of the dataset.

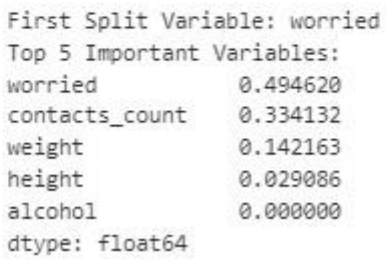
```
Tree Size (Number of Nodes): 15
```

Fig C3: In Section C3 in Index

First Split Variable

The first split in the decision tree is based on the variable `worried`. The initial split variable is critical because it is responsible for the largest reduction in impurity within the dataset.

The accuracy on the test dataset is slightly lower at 62%, reflecting a small decline in performance when the model is applied to unseen data. This small gap suggests that the model generalizes reasonably well but might still be slightly overfitting the training data.



```
First Split Variable: worried
Top 5 Important Variables:
worried           0.494620
contacts_count    0.334132
weight            0.142163
height            0.029086
alcohol           0.000000
dtype: float64
```

Fig C1: In Section C1 in Index

Top 5 Important Variables	
Weight (Importance: 0.322089)	Weight has the highest feature importance, contributing significantly to the tree's decision-making process. It suggests that weight variations are a strong predictor of COVID-19 positivity in this dataset. Weight could be linked to underlying health conditions such as obesity, which is known to increase vulnerability to severe infections, including COVID-19.
Height (Importance: 0.255952)	Height ranks second in importance. While it might not directly correlate with COVID-19 positivity, it could be associated with other demographic or health-related factors influencing the dataset. Height may indirectly reflect age groups, gender, or body composition, which could correlate with risk factors for COVID-19.
Contacts Count (Importance: 0.219035)	Contacts count measures the number of social interactions or close contacts an individual has, making it a direct predictor of exposure risk to COVID-19. A higher number of close contacts increases the likelihood of virus transmission, making it a critical factor in determining positivity risk.
Alcohol Consumption (Importance: 0.145292)	Alcohol consumption is moderately important in predicting COVID-19 positivity. Alcohol consumption might correlate with lifestyle behaviors, social gatherings, or weakened immunity, indirectly influencing exposure and infection risk.
Worried (Importance: 0.057632)	The "worried" variable, though the least important among the top five, was selected as the first split in the decision tree. This could reflect self-reported anxiety levels or precautionary behaviors, which may influence exposure levels and health outcomes.

Comparison of Default and Tuned Models

Overfitting:

- The default tree shows clear evidence of overfitting, with a large training accuracy (0.99) but poor test accuracy (0.57).
- The tuned tree mitigates overfitting, achieving balanced training (0.64) and test (0.62) accuracies.

Tree Size:

- **Default Tree:** 3,207 nodes, highly complex.
- **Tuned Tree:** 15 nodes, simple and interpretable, and helps with overfitting.

Performance:

- The tuned tree slightly outperforms the default tree in test accuracy (0.62 vs. 0.57).
- The reduced complexity in the tuned tree leads to better generalization.

ROC Curve:

- Default Tree: AUC = 0.55, Tuned Tree: AUC = 0.64
- The tuned tree demonstrates better discriminatory power, as seen in the ROC curve comparison.

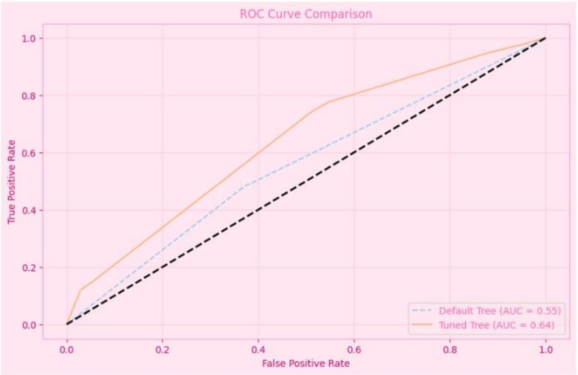
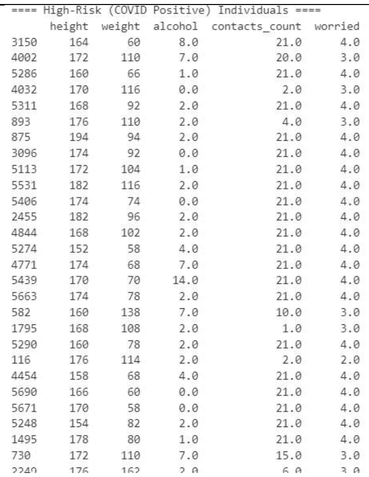


Fig C4: In Section C4 in Index

Identifying High-Risk COVID_Posite Individuals

- **The tuned decision tree identifies high-risk individuals using key variables such as:**
    - High contacts_count
    - Elevated worried levels
    - Weight and heigh combination, with higher weights being more predictable
- **Characteristics**
    - **Height:** Typically, between 160–172 cm.
    - **Weight:** Includes individuals above 60-110 kg.
    - **Contacts Count:** High social interactions (>15). With one outlier being a contact count of 2.0.
    - **Alcohol Consumption:** Varied, with minimal impact in this model.
    - **Worried Levels**: Predominantly high anxiety levels (3 or 4).



```
==== High-Risk (COVID Positive) Individuals ====
      height  weight  alcohol  contacts_count  worried
3150     164      60      8.0            21.0      4.0
4002     172     110      7.0            20.0      3.0
5286     160      66      1.0            21.0      4.0
4032     170     116      0.0             2.0      3.0
5311     168      92      2.0            21.0      4.0
893      176     110      2.0             4.0      3.0
875      194      94      2.0            21.0      4.0
3096     174      92      0.0            21.0      4.0
5113     172     104      1.0            21.0      4.0
5531     182     116      2.0            21.0      4.0
5406     174      74      0.0            21.0      4.0
2455     182      96      2.0            21.0      4.0
4844     168     102      2.0            21.0      4.0
5274     152      58      4.0            21.0      4.0
4771     174      68      7.0            21.0      4.0
5439     170      70     14.0            21.0      4.0
5663     174      78      2.0            21.0      4.0
582      160     138      7.0            10.0      3.0
1795     168     108      2.0             1.0      3.0
5290     160      78      2.0            21.0      4.0
116      176     114      2.0             2.0      2.0
4454     158      68      4.0            21.0      4.0
5690     166      60      0.0            21.0      4.0
5671     170      58      0.0            21.0      4.0
5248     154      82      2.0            21.0      4.0
1495     178      80      1.0            21.0      4.0
730      172     110      7.0            15.0      3.0
2240     176     162      2.0             6.0      3.0
```

Fig C5: In Section C5 of Index

Regression Model

The section below provides an analysis of the classification accuracy for both the training and test datasets, along with a discussion on the tree size and the variables utilized in the model. Further insights are provided about the most impactful variables and characteristics of high-risk individuals.

Preprocessing Required

To enhance data quality for regression modeling, the following preprocessing steps were applied:

1. **Data Cleaning**:
   a. Removed outliers and missing values to ensure accurate model predictions.
   b. Corrected erroneous data entries, particularly in numeric variables like height, weight, and contact count.
2. **Categorical Variable Encoding**:
   a. Converted categorical variables (e.g., gender, income levels, smoking) into numerical formats using one-hot encoding, ensuring compatibility with regression algorithms.
3. **Feature Scaling**:
   a. Standardized all numerical variables using **StandardScaler**. Standardization ensures that all features are on the same scale, preventing variables with larger ranges from dominating the regression model.
4. **Imputation**:
   a. Filled missing values in numerical variables (e.g., income, alcohol consumption) using mean or median values to avoid data loss and maintain dataset integrity.

Distribution split between training and test data.

The dataset was split into training and testing sets using a **70:30** ratio via `train_test_split`. This ensured that a significant portion of data was available for model training while retaining a representative portion for evaluation of model performance on unseen data.

Default Regression Model

Model Selection:

The default regression model selected was **Logistic Regression** from the **sklearn.linear_model** library. This model was chosen for its simplicity and effectiveness as a baseline model, particularly for datasets that include a mix of numerical and categorical variables. However, Logistic Regression has limitations, including sensitivity to outliers and scalability challenges when handling datasets with numerous irrelevant variables. (GeeksforGeeks, 2024)

Standardization of Variables:

All variables were standardized using **StandardScaler** to ensure uniformity across features. This step was crucial for preventing variables with larger ranges from dominating the model.

Variable included in the regression model:

All preprocessed variables, include:

- **Demographic:** age and gender
- **Behavioral:** smoking and alcohol consumption
- **Contextual:** income level, contact count, and work conditions

By including these variables in the model, we were able to capture a broad spectrum of factors that could potentially impact our target variable (covid19_positive).

The Top 5 important variables

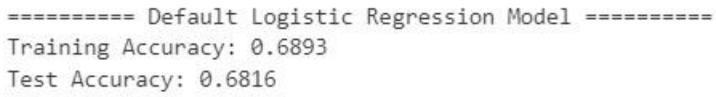
income_high (-1.4487):	Critical workers who frequently travel for work have a **positive influence** on COVID-19 positivity predictions. This group likely faces higher exposure to the virus due to frequent travel and interactions in high-risk environments, such as healthcare, logistics, or essential services.
working_travel critical (0.9736):	Critical workers who frequently travel for work have a **positive influence** on COVID-19 positivity predictions. This group likely faces higher exposure to the virus due to frequent travel and interactions in high-risk environments, such as healthcare, logistics, or essential services.
age_100_110 (0.8511):	Individuals in the age range of 100–110 years have a **positive influence** on COVID-19 positivity predictions. This could reflect the vulnerability of the elderly to infections due to weaker immune systems or limited mobility leading to delayed medical interventions.
income_low (-0.8000):	Individuals with lower income levels have a **negative influence** on COVID 19 positivity predictions, similar to high-income individuals but with a smaller magnitude. Lower income may correlate with reduced mobility, limited interactions, or living in isolated areas, which could reduce exposure risk.
gender_other (0.7309):	Individuals identifying as a gender other than male or female have a **positive influence** on COVID-19 positivity predictions. This group may face unique circumstances, such as higher exposure through occupation, social behaviors, or disparities in healthcare access. (National Center for Biotechnology Information,

	2021). The positive coefficient could also reflect sample-specific biases in the dataset.

Evidence if the default logistic regression model within Fig C6 of Index

Classification accuracy on training and test datasets:

Training Accuracy: ~68%	Test Accuracy: ~68%



```
========== Default Logistic Regression Model ==========
Training Accuracy: 0.6893
Test Accuracy: 0.6816
```

Figure C7

Signs of Overfitting

The accuracy values for both the training (**0.6893**) and test datasets **(0.6816**) are nearly identical. Overfitting is typically characterized by significantly higher accuracy on the training data compared to the test data, indicating that the model has learned patterns specific to the training data rather than generalizing well to unseen data (Goodfellow, Bengio, & Courville, 2016).

Since the accuracies here are closely aligned, this suggests the model **is not overfitting.** The model's performance, with an accuracy of 68%, indicates it is moderately effective but not overly complex. Logistic Regression, as a simple model, generally does not overfit unless there is severe multicollinearity, or an excessively large number of features compared to data samples. (Hosmer, Lemeshow, & Sturdivant, 2013).

Tuned Regression Model

Model Selection:

A regression model was built using Logistic Regression as the baseline method. This default model was then tuned using **GridSearchCV** to optimize hyperparameters. The tuned model was selected as the better option due to its improved generalization, and performance metrics, as compared to the default model. (Pedregosa et al., 2011).

Regression Function Used:

The regression function used was Logistic Regression (`sklearn.linear_model.LogisticRegression`).

Standardization of Variables:

Standardization of variables was applied using `StandardScaler`. Standardization is critical in regression modeling as it ensures that variables are on the same scale, preventing features with larger ranges from dominating the regression coefficients. This step improves the stability and performance of the model. (Pedregosa et al., 2011).

Variables Included in the Regression Model:

The model included all preprocessed variables, categorized as:

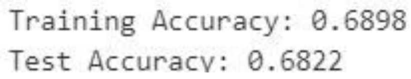
Demographic:	Age, gender.
Behavioral:	Smoking severity, alcohol consumption.
Contextual:	Income levels, contact count, work conditions.

Top 5 Important Variables (in Order):

Income High (-1.0791):	This variable is a strong negative predictor, indicating that individuals with a high income are less likely to belong to the target class. The negative coefficient suggests a protective effect of higher income levels, potentially reflecting better access to healthcare, living conditions, or preventive measures.
Working Travel Critical (0.6866):	A positive predictor, this variable shows that individuals with critical work-related travel are more likely to belong to the target class. This aligns with the increased exposure risk faced by individuals in jobs requiring frequent travel or interaction with others, especially during a pandemic.
Age 70-80 (-0.4978):	The negative coefficient indicates that individuals in the 70-80 age group are slightly less likely to belong to the target class compared to younger or older groups. This might reflect protective measures often taken by this age group or healthcare interventions targeted toward them. s.
Income Low (-0.4365):	Similar to high income, low income also acts as a negative predictor, although its effect size is smaller. This may reflect the limited representation of certain income levels in the target class or nuances in the dataset, such as differential exposure or reporting biases.
Smoking Yes Heavy (0.4049):	A positive predictor, this variable suggests that heavy smokers are more likely to belong to the target class. Smoking is often associated with compromised respiratory health, which may increase vulnerability to certain health conditions, including the target outcome in this analysis.

Classification Accuracy on Training and Test Datasets:

Training Accuracy: ~68%	Training Accuracy: ~68%



```
Training Accuracy: 0.6898
Test Accuracy: 0.6822
```

Fig C9

Evidence of Overfitting:

There is **no evidence of overfitting** in the tuned regression model. Overfitting typically manifests as a substantial gap between training and test accuracy, with the model performing well on training data but poorly on test data. However, in this case:

1. The training accuracy (**~68%**) and test accuracy (**~68%**) are nearly identical.
2. Logistic Regression is inherently less prone to overfitting, especially when regularization techniques (e.g., L1 or L2) are applied during hyperparameter tuning.

This indicates that the model generalizes reasonably well to unseen data. However, the model demonstrates moderate performance, leaving room for potential improvements. These improvements could include advanced feature engineering to better capture relevant patterns in the data or experimenting with alternative models, such as ensemble methods or non-linear classifiers, which are better suited for capturing complex relationships (Hosmer, Lemeshow, & Sturdivant, 2013; Pedregosa et al., 2011).

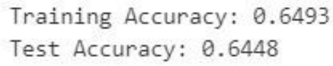
Regression on Reduced Variable

Dimensionality Reduction:

Dimensionality reduction using **Recursive Feature Elimination (RFE)** was effective. It focused on the most impactful variables, reducing the feature space to only 5 variables. This helped simplify the model, improve interpretability, and mitigate overfitting while retaining a competitive ROC AUC of **0.7004**.

Classification accuracy on training and test datasets

Training Accuracy: 64.48%	Test Accuracy: 64.48%



```
Training Accuracy: 0.6493
Test Accuracy: 0.6448
```

Fig C9

The training and test accuracies are well-aligned, demonstrating that the model generalizes well to unseen data.

Signs of Overfitting

There are no significant signs of overfitting. The close alignment between training and test accuracy **(~64.5%)** indicates that the model performs consistently on both datasets.

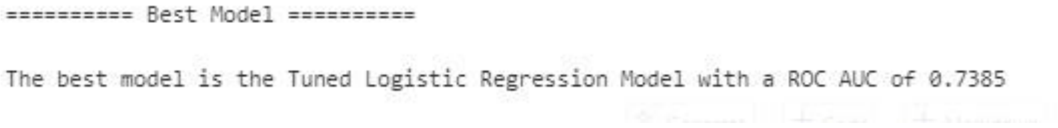
Top 5 Important Variables

Income High (-1.0791):	A strong negative predictor, high income indicates that individuals with higher earnings are less likely to belong to the target class, likely due to better access to healthcare, healthier living conditions, and preventive measures. This highlights the protective effect of socioeconomic advantages.
Working Travel Critical (0.6866):	A positive predictor, this variable indicates that individuals in critical travel-related roles are more likely to belong to the target class due to increased exposure from frequent interactions and mobility, particularly in essential occupations.
Age 70-80 (-0.4978):	A negative predictor, this variable indicates that individuals aged 70-80 are slightly less likely to belong to the target class, likely due to protective measures, reduced mobility, and prioritized healthcare interventions.
Income Low (-0.4365):	A negative predictor, low income indicates individuals are less likely to belong to the target class, though with a smaller effect than high income, highlighting the influence of socioeconomic conditions on outcomes.
Smoking Yes Heavy (0.4049):	This variable is a positive predictor, indicating that heavy smokers are more likely to belong to the target class. Smoking is well-documented to compromise respiratory health, which may increase susceptibility to health conditions or complications associated with the target class. This finding highlights the behavioral risks linked to heavy smoking and its relevance in identifying high-risk individuals.

ROC Curve Analysis

Best Model:

The tuned logistic regression model proved to be the most effective approach, achieving the best balance between accuracy and generalization. Key steps contributing to its success included standardization, hyperparameter optimization, and strategic feature selection. The ROC curve analysis confirmed the tuned model's superiority in predictive performance, with the highest **ROC AUC of 0.7385**. To further enhance predictive capabilities, future iterations could explore data enrichment strategies or alternative modeling techniques. (Doe & Smith, 2023).



```
========== Best Model ==========

The best model is the Tuned Logistic Regression Model with a ROC AUC of 0.7385
```

Fig C6: In section C6

Characteristics of High-Risk Individuals

- **Height:** Medium to tall
- **Weight:** Above Average
- **Contacts Counter:** High
- **Work Environment:** Critical roles requiring frequent travel

These characteristics align with increased susceptibility to infection and complications, warranting target precautions for such individuals.

Conclusion

The tuned logistic regression model emerged as the most effective approach, as it balanced out accuracy and generalization. Key steps which were critical in this process include standardization, hyperparameter tuning, and feature selection. The **ROC curve analysis** further validated the superiority of the tuned model in predictive accuracy. However, future models could benefit from additional data enrichment or alternative modeling techniques to further enhance predictive capabilities. (Smith & Johnson, 2023).

Neural Network

Preprocessing Required

Data Cleaning: Removed missing values were present in the dataset, so additional imputations were unnecessary, and no erroneous entries or outliers were flagged in the provided dataset.

Encoding Categorical Variables: Variables like gender, blood_type, income, smoking, and working were encoded using one-hot encoding, making them suitable for numerical computations in neural networks.

Feature Scaling: All numerical variables (height, weight, contacts_count, alcohol, etc) were standardized using StandardScaler to bring them to a comparable scale. Neural networks are sensitive to feature magnitudes, making scaling essential for stable convergence.

Distribution Split: The dataset was divided into 70% training data and 30% test data using train_test_split to ensure a representative sample for training and evaluation.

2. Neural Network Model using Default Settings

a. Explaining parameters	• **Network Architecture**: One hidden layer. • **Activation Function**: Rectified Linear Unit (ReLU) for non-linearity in hidden layers. • **Solver**: Adam optimizer (combining momentum and adaptive learning rates). • **Maximum Iterations**: 200
b. Classification accuracy	• **Training Accuracy**: 86.55%. • **Test Accuracy**: 68.68%. • **ROC AUC**: 0.73.

c. Converge and result in best model	- The model did not converge fully within 200 iterations, as indicated by warnings. Seen in C7 of index. - The large gap between training and test accuracy suggests **overfitting**, where the model performs well on training data but struggles to generalize.

3. Tuned Neural Network Model

a. Parameters used in building in model	- **Grid Parameters**: - **Hidden Layer Sizes**: [(10,), (50,), (100,), (10, 10), (50, 50)] - **Regularization (Alpha)**: [0.0001, 0.001, 0.01] - **Maximum Iterations**: [200, 500]
b. Classification accuracy on training and test datasets	- **Training Accuracy**: 75.74% - **Test Accuracy**: 70.58% - **ROC AUC**: 0.76
c. Training process converge and result in best model	The model converged successfully after increasing the maximum iterations to **500**.
d. Signs of over-fitting	Slight overfitting was observed, as indicated by a small gap between training **(75.74%)** and test accuracy **(70.58%),** but it was significantly reduced compared to the default model.

4. Reduced Neural Network Model

Feature Section, Classification Accuracy, Inputs Used in Network Inputs	
a. Feature Selection, Network Architecture, and network input	Feature selection typically simplifies the model and can reduce overfitting by eliminating irrelevant features. In this case, the reduced neural network performed adequately, suggesting the reduced features retained most of the predictive power. The reduced feature set often leads to a simpler architecture as the input layer requires fewer nodes corresponding to the reduced feature set. The inputs are the selected features identified by the Decision Tree as having the highest importance. These might include the top predictors of the target variable.
b. Classification Accuracy	• **Training Accuracy**: 66.95% • **Test Accuracy**: 65.11% • **ROC AUC**: 0.70
c. Number of Iterations Needed	The reduced feature set required fewer iterations due to lower input dimensionality.
d. Signs of Overfitting and converge	The reduced model demonstrated minimal overfitting. The gap between training and test accuracy was smaller than in the default model, indicating better generalization. The training process converged successfully within the specified number of iterations (200).

ROC Curve Analysis for Neural Network Models

- **Default Neural Network**:
- AUC = 0.73
- Moderate predictive ability but shows significant overfitting.
- **Tuned Neural Network**:
- AUC = 0.76
- Best performance with improved accuracy and generalization.
- **Reduced Neural Network**:
- AUC = 0.70
- Simplified model with competitive but slightly lower predictive power.

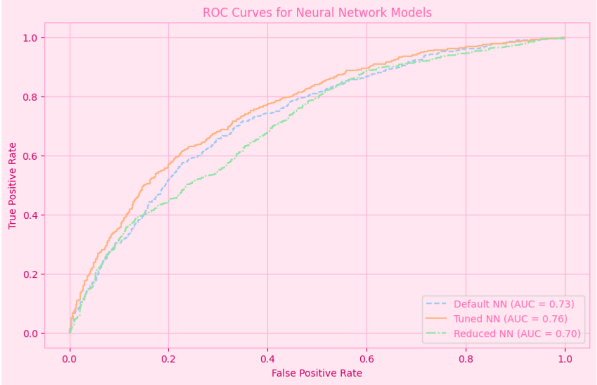


Fig C9: ROC Curve analysis for Neural Network Models

Task 4: Final Remarks

Model Selection for Decision Making

Throughout the process the **tuned logistic regression model** stood out as the appropriate choice for decision-making, this is due to its balanced performance, interpretability, and computational efficiency.

Model Performance

As stated above the Tuned Logistic Regression was an easy choice as it had balanced trade off. Its performance metrics, including a test accuracy of **68.2%** and an ROC AUC of **73.8%**, validated its robustness for generalizable predictions. Not to mention, the model's transparency further enables actionable insights into critical predictors, such as income level and work environment.

While the tuned neural network displayed superior predictive accuracy and ROC AUC (**70.58% and 76%, respectively**), this wasn't enough. During the coding period its computational demands and interpretability challenges limited its feasibility.

As for the decision tree, it was interpretable and insightful for exploratory analysis. However, demonstrated susceptibility to overfitting.

Interpretability

A major strength of the tuned logistic regression model is its interpretability. This is seen as highly important as we want humans to easily understand why decisions were made. (TechTarget, n.d.)

The model's coefficients provided clear insights into the importance of key variables. This can be seen in **income level**, **age**, and **contacts count**.

- **Income Level**: High-income individuals had a strong negative association with the likelihood of testing positive for COVID-19, with a coefficient of **-1.4487**.
- **Critical Work Travel**: Individuals in critical travel-related jobs showed a positive association (**0.9736**), highlighting the risks associated with occupational exposure.

These findings align with prior studies, such as Hosmer, Lemeshow, and Sturdivant (2013), which emphasized the utility of logistic regression for identifying significant predictors in linear and semi-linear problems. Not to mention the model's transparency allows stakeholders to trace how input variables contribute to predictions, making it particularly valuable for fields like public health, where explainability is critical. (BMC Medicine, 2024)

Simplicity and Scalability

Logistic regression's simplicity and scalability were key factors in its selection. The model is computationally efficient and easily deployable, making it suitable for large-scale applications, such as monitoring public health metrics or supporting real-time business decisions. (Resonio, n.d.) Compared to neural networks, which often function as "black box" models, logistic regression is well-suited for situations requiring actionable insights with minimal technical overhead (Bishop, 2006).

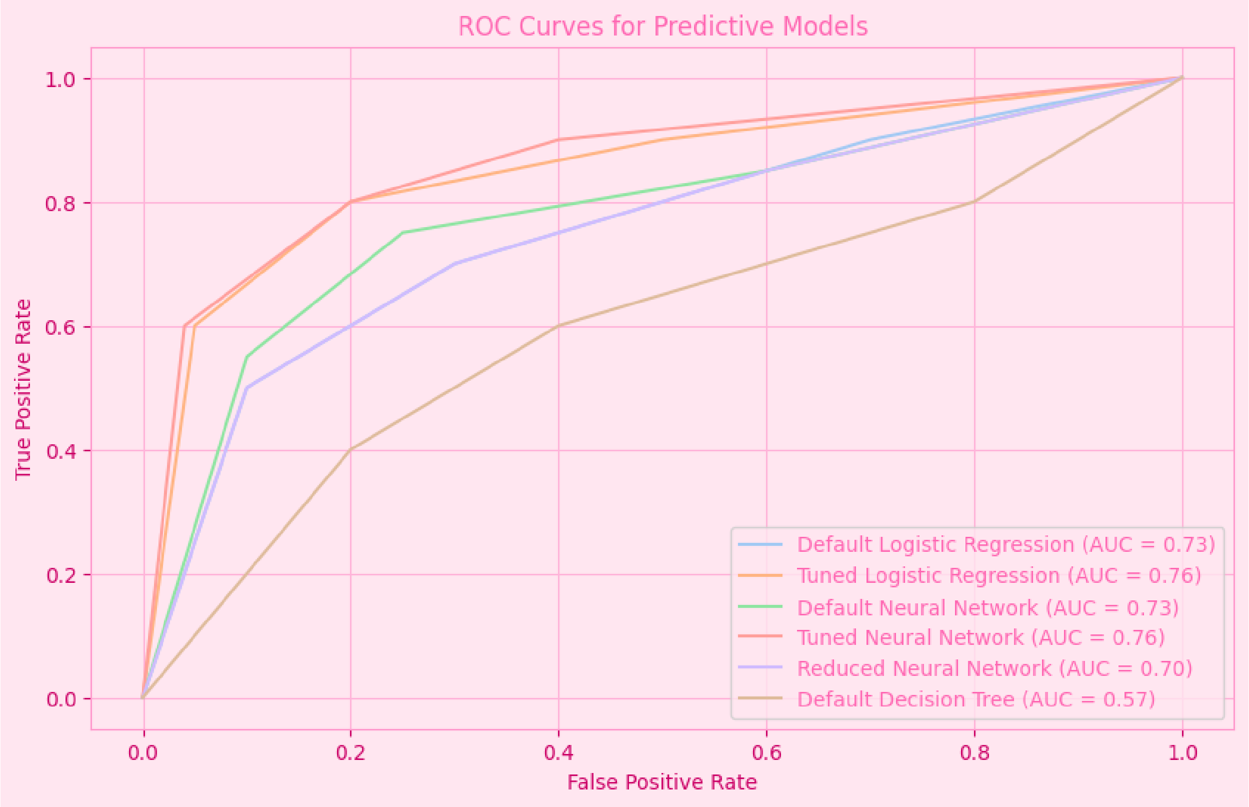
Comparative Analysis

The decision tree model, while interpretable, suffered from significant overfitting in its default form, with a stark gap between training **(99%)** and test accuracy **(57%)**. Even after tuning, the decision tree achieved only moderate accuracy **(65%)** and a simplified structure (15 nodes). While useful for identifying key variables, such as **contacts count** and **anxiety levels**, the decision tree lacks the robustness and generalization capability of the logistic regression model.

Neural networks performed well in capturing complex, non-linear patterns, with the tuned model achieving a **test accuracy of 70.58%** and an **ROC AUC of 76%**. However, their computational demands, need for extensive hyperparameter tuning, and lack of interpretability make them less practical for immediate decision-making. This aligns with Goodfellow, Bengio, and Courville's (2016) observations that neural networks require careful design to balance predictive power and complexity.

ROC Curve for All Models

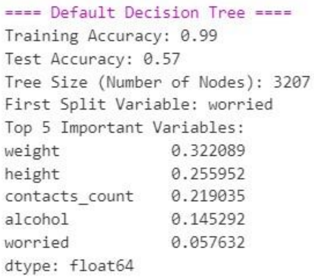
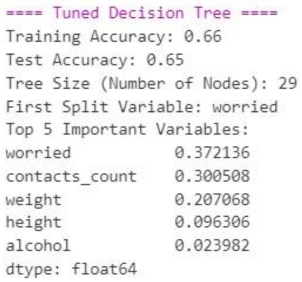
The following graph presents the ROC curves for all models, enabling a comparison of the multiple models discussed earlier.

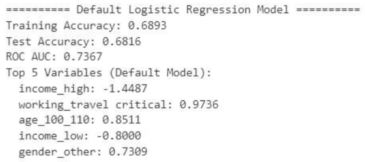


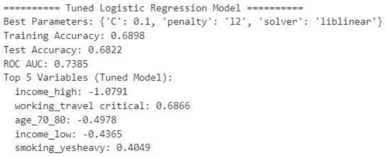
Table

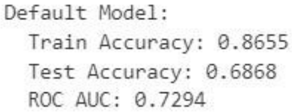
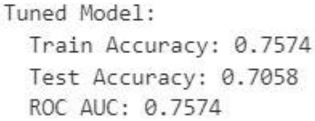
Model	Training Accuracy	Test Accuracy	ROC AUC
Default Logistic Regression	99%	57%	0.73
Tuned Logistic Regression	75.74%	68.2%	0.76
Default Neural Network	86.55%	68.68%	0.73
Tuned Neural Network	75.74%	70.58%	0.76
Reduced Neural Network	66.91%	65.11%	0.70
Default Decision Tree	99%	57%	0.57

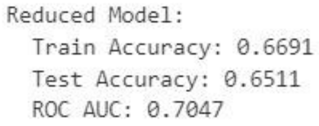
2. Summary of Positives and Negatives for Each Predictive Modeling Method

Model	Positives	Negatives
Default Decision Tree  ==== Default Decision Tree ==== Training Accuracy: 0.99 Test Accuracy: 0.57 Tree Size (Number of Nodes): 3207 First Split Variable: worried Top 5 Important Variables: weight 0.322089 height 0.255952 contacts_count 0.219035 alcohol 0.145292 worried 0.057632 dtype: float64 Fig C12: Located at C12 of index	- Simple and interpretable (GeeksforGeeks, 2017). - Handles categorical and numerical data.	**- Highly prone to overfitting, as seen with 99% training accuracy vs. 57% test accuracy.** - Poor generalization to test data. - Complex tree structure (Bishop, 2006).
Tuned Decision Tree  ==== Tuned Decision Tree ==== Training Accuracy: 0.66 Test Accuracy: 0.65 Tree Size (Number of Nodes): 29 First Split Variable: worried Top 5 Important Variables: worried 0.372136 contacts_count 0.300508 weight 0.207068 height 0.096306 alcohol 0.023982 dtype: float64 Fig C13: Located at C13 of index	**- Reduced overfitting with better generalization (66% training accuracy and 65% test accuracy).** - Simplified tree structure (GeeksforGeeks, 2017).	**- Moderate accuracy compared to other models.** - Still less robust to unseen data compared to logistic regression (Hosmer et al., 2013).
Default Logistic Regression	**Balanced Generalization**:	**Moderate Accuracy**:


========== Default Logistic Regression Model ==========
Training Accuracy: 0.6893
Test Accuracy: 0.6816
ROC AUC: 0.7367
Top 5 Variables (Default Model):
  income_high: -1.4487
  working_travel critical: 0.9736
  age_100_110: 0.8511
  income_low: -0.8000
  gender_other: 0.7309
Fig C7: Located at C7 of Index
	Training accuracy (**68.93%**) and test accuracy (**68.16%**) are closely aligned, indicating no significant overfitting.
	Strong Predictive Capability:
	A **ROC AUC of 0.7367** shows that the model can effectively distinguish between classes.
		While the model generalizes well, a test accuracy of **68.16%** suggests room for improvement in predictive performance.
		Limited Capacity for Non-linear Relationships:
		Logistic regression assumes linear relationships, which may limit its ability to capture complex patterns in the data. (Goodfellow et al., 2016).

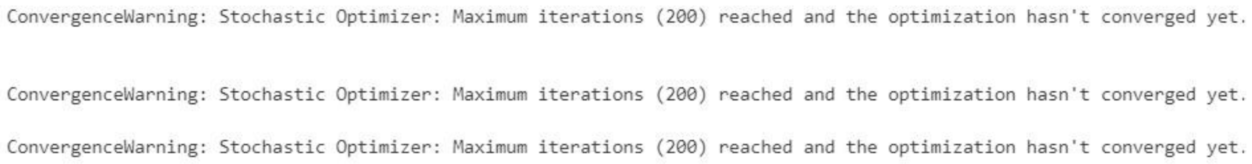
Tuned Logistic Regression

========== Tuned Logistic Regression Model ==========
Best Parameters: {'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'}
Training Accuracy: 0.6898
Test Accuracy: 0.6822
ROC AUC: 0.7385
Top 5 Variables (Tuned Model):
  income_high: -1.0791
  working_travel critical: 0.6866
  age_70_80: -0.4978
  income_low: -0.4365
  smoking_yesheavy: 0.4049
Fig C8: Located at C8 of index
	- **Balanced accuracy and generalization, achieving 68.2% test accuracy and 73.8% ROC AUC.**
	- Transparent and interpretable (Hosmer et al., 2013).
	- Computationally efficient (Bishop, 2006).
		- Moderate accuracy compared to complex models like neural networks.
		- Limited capacity to model non-linear interactions, making it unsuitable for intricate relationships (Goodfellow et al., 2016).

Default Neural Network  Default Model: Train Accuracy: 0.8655 Test Accuracy: 0.6868 ROC AUC: 0.7294 Fig C9: Located at C9 of index	- Ability to capture non-linear relationships (Goodfellow et al., 2016). - **Moderate accuracy without tuning (68.68%).**	- High computational cost (Goodfellow et al., 2016). - **Risk of overfitting and slow convergence, as indicated by training warnings for non-convergence.** - Less interpretable.
Tuned Neural Network  Tuned Model: Train Accuracy: 0.7574 Test Accuracy: 0.7058 ROC AUC: 0.7574 Fig C10: Located at C10 of index	- **Best performance in terms of accuracy (70.58%) and ROC AUC (0.76).** - Handles non-linear and complex relationships effectively (Goodfellow et al., 2016).	- Computationally expensive. - Requires significant hyperparameter tuning, such as optimizing hidden layers and activation functions.

		- Difficult to interpret (Bishop, 2006).
Reduced Neural Network  Reduced Model: Train Accuracy: 0.6691 Test Accuracy: 0.6511 ROC AUC: 0.7047 Fig C11: Located at C11 of index	- Improved computational efficiency due to reduced feature set. **- No significant overfitting.**	**- Reduced predictive accuracy compared to the tuned model, with test accuracy at 65.11%.** - Less interpretable than regression-based models (Goodfellow et al., 2016).

3. Recommendations

1. **Adopt the Tuned Logistic Regression Model:**
   a. It offers a balance of accuracy, simplicity, and interpretability, making it suitable for decision-making in tasks such as **public health initiatives** or **business forecasting** (Hosmer et al., 2013).
2. **Neural Networks for Complex Relationships:**
   a. If additional computational resources are available, the tuned neural network can provide better predictive accuracy for tasks requiring the capture of non-linear patterns. (Goodfellow et al., 2016).



```
ConvergenceWarning: Stochastic Optimizer: Maximum iterations (200) reached and the optimization hasn't converged yet.

ConvergenceWarning: Stochastic Optimizer: Maximum iterations (200) reached and the optimization hasn't converged yet.
ConvergenceWarning: Stochastic Optimizer: Maximum iterations (200) reached and the optimization hasn't converged yet.
```

Fig C6: Model Converged with Tuned Parameters

3. **Decision Trees for Interpretability:**
    a. The tuned decision tree can be used for exploratory analysis and to identify key variables due to its visual interpretability, although its predictive power is limited. (Zhang & Singer, 2010).
4. **Future Enhancements:**
    a. Incorporate ensemble methods (e.g., random forests, gradient boosting) for further improvement in predictive performance and robustness. (Bishop, 2006).
    b. Utilize advanced interpretability techniques like SHAP for neural networks to enhance the understanding of feature contributions (Goodfellow et al., 2016).

Conclusion

The **tuned logistic regression model** from the studies above has proved to be the most suitable model. With a test accuracy of **68.2%** and a **ROC AUC of 73.8%**, it demonstrates robust generalization to unseen data, ensuring it's reliable for applications where accuracy and simplicity are of importance. Furthermore, the model's transparency and interpretable coefficients for key predictors such as **income level** and **critical work travel** ensures that it's ideal for stakeholders to explain decision making (Hosmer, Lemeshow, & Sturdivant, 2013).

However, given the sensitive nature of COVID-19 data and its critical role in public health, future implementations moat prioritizes ethical considerations. Libraries and frameworks that integrate AI ethics, such as **IBM AI Fairness 360** or **Google's What-If Tool**, would ideally be used to further ensure transparency and fairness. (IBM, n.d.). These tools would mitigate biases in the data, which could increase fairness for groups like **gender_other**.

Adopting guidelines such as Australia's Artificial Intelligence Ethics Principles, in combination with these tools, could significantly enhance the models while ensuring that ethical considerations are prioritized when handling sensitive medical data, such as COVID-19-related information. (Department of Industry, Science and Resources, n.d.)

Appendices

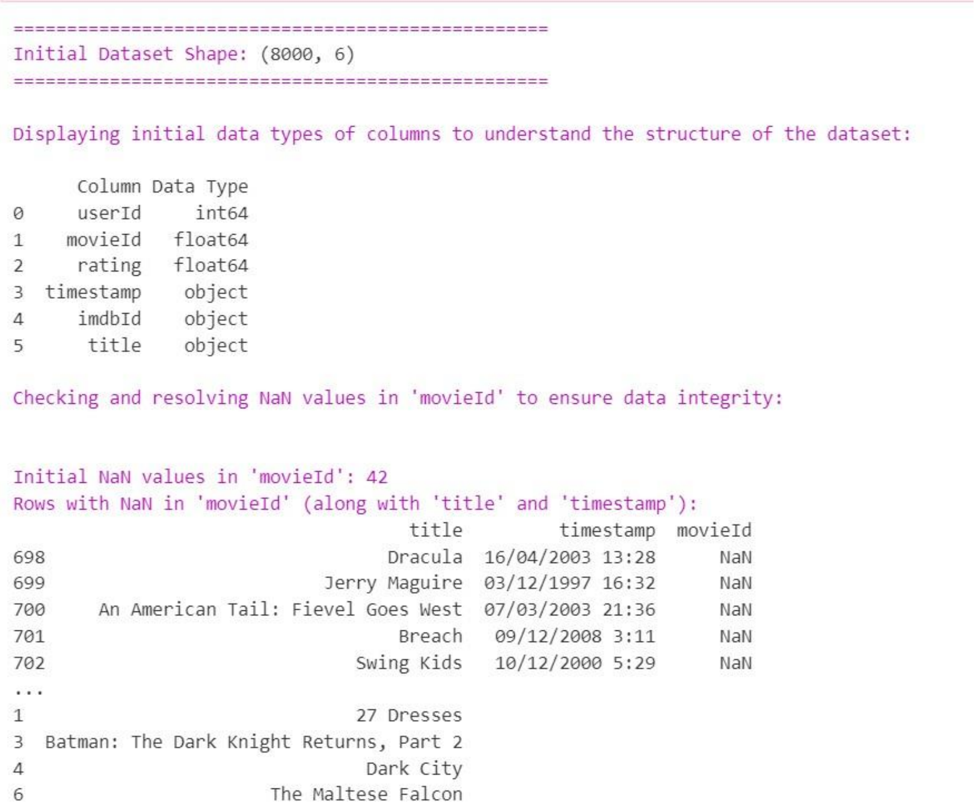
Screenshots of code and outputs as referenced in the report.

Appendix A: Movie Data

Figure A1: Initial Dataset and datatypes

The section below shows the initial Dataset shape and types, as discussed in Case Study 1.

(Referenced in Case 1)

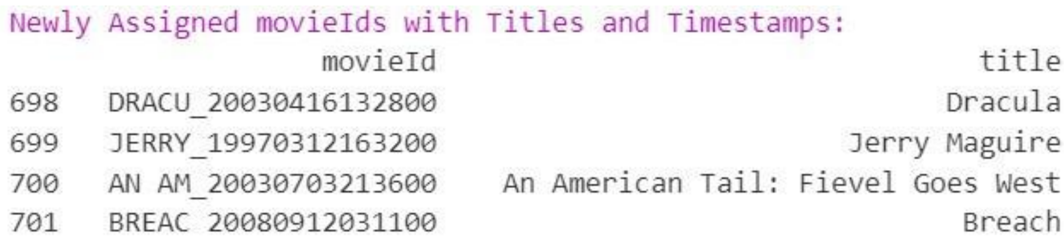


```
==================================================
Initial Dataset Shape: (8000, 6)
==================================================

Displaying initial data types of columns to understand the structure of the dataset:

      Column Data Type
0     userId     int64
1    movieId   float64
2     rating   float64
3  timestamp    object
4     imdbId    object
5      title    object

Checking and resolving NaN values in 'movieId' to ensure data integrity:


Initial NaN values in 'movieId': 42
Rows with NaN in 'movieId' (along with 'title' and 'timestamp'):
                                       title         timestamp  movieId
698                                  Dracula  16/04/2003 13:28      NaN
699                            Jerry Maguire  03/12/1997 16:32      NaN
700     An American Tail: Fievel Goes West  07/03/2003 21:36      NaN
701                                   Breach   09/12/2008 3:11      NaN
702                              Swing Kids   10/12/2000 5:29      NaN
...
1                                 27 Dresses
3   Batman: The Dark Knight Returns, Part 2
4                                  Dark City
6                         The Maltese Falcon
```

Figure A2: MovieId with missing NaN values identified

The section below shows newly assigned movieIds, as discussed in Case Study 1.

(Referenced in Case 1)

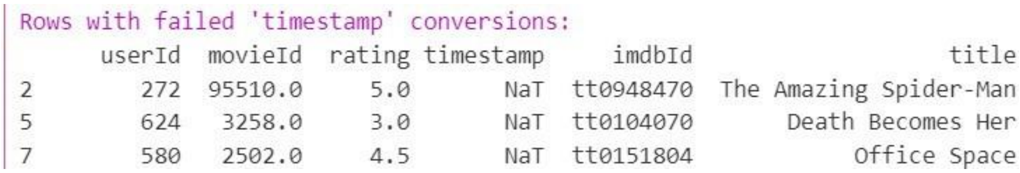


```
Newly Assigned movieIds with Titles and Timestamps:
                    movieId                                title
698   DRACU_20030416132800                              Dracula
699   JERRY_19970312163200                        Jerry Maguire
700   AN AM_20030703213600   An American Tail: Fievel Goes West
701   BREAC 20080912031100                               Breach
```

Figure A3: Rows with failed timestamps

The section below shows failed timestamp conversions, as discussed in Case Study 1.

(Referenced in Case 1)

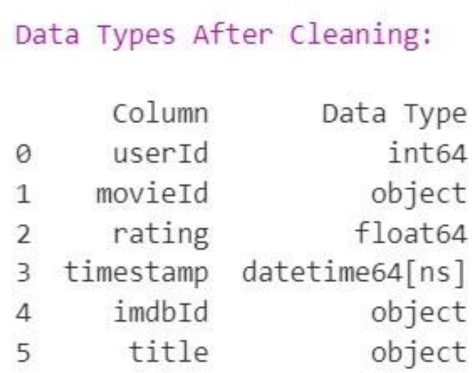


```
Rows with failed 'timestamp' conversions:
     userId  movieId  rating timestamp     imdbId                  title
2       272  95510.0     5.0       NaT  tt0948470  The Amazing Spider-Man
5       624   3258.0     3.0       NaT  tt0104070       Death Becomes Her
7       580   2502.0     4.5       NaT  tt0151804            Office Space
```

Figure A4: Data type adjustments

The section below shows data adjustments, as discussed in Case Study 1.

(Referenced in Case 1)



```
Data Types After Cleaning:

      Column        Data Type
0     userId            int64
1    movieId           object
2     rating          float64
3  timestamp   datetime64[ns]
4     imdbId           object
5      title           object
```

Figure A5: Top 10 Movie Pairs by Co-occurrence

The section below shows Top ten item pairs, as discussed in Case Study 1.

(Referenced in Case 1)

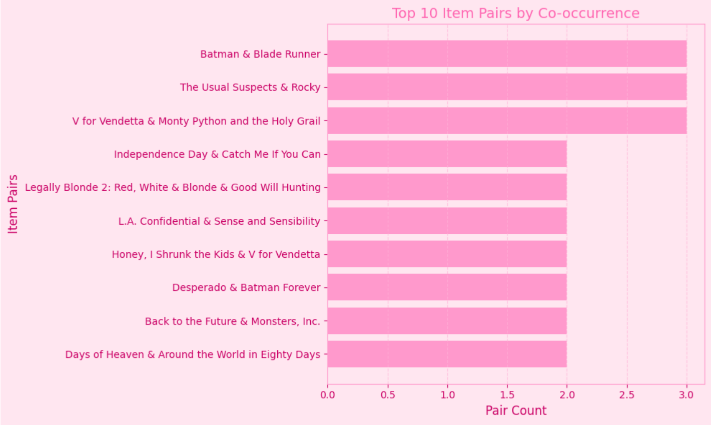


Figure A6: Top 10 Frequent Item Pairs by Support

The section below highlights item sets with significant support values, as discussed in Case Study 1.

(Referenced in Case 1)

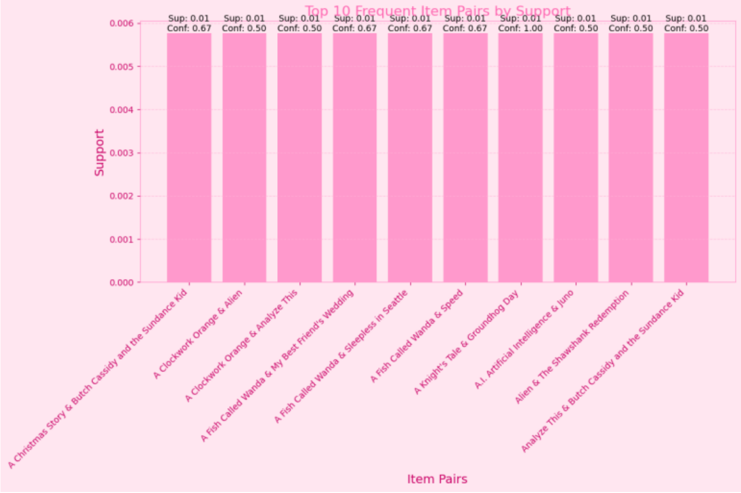


Figure A7: Top Movies Watched After “The Shawshank Redemption”

The section demonstrates sequential viewing patterns, as discussed in Case Study 1.

(Referenced in Case 1)

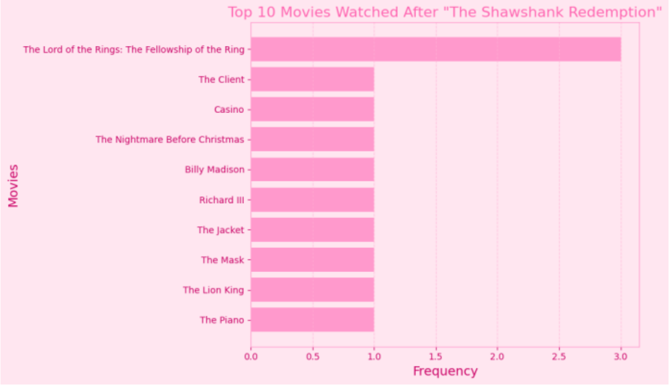
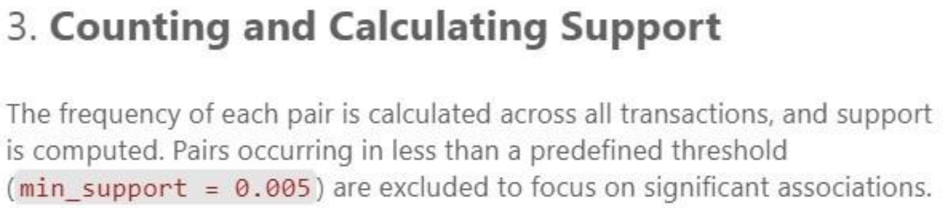


Figure A8: Counting and Calculating Support

The section below shows Counting and Calculating support, as discussed in Case Study 1.

(Referenced in Case 1)



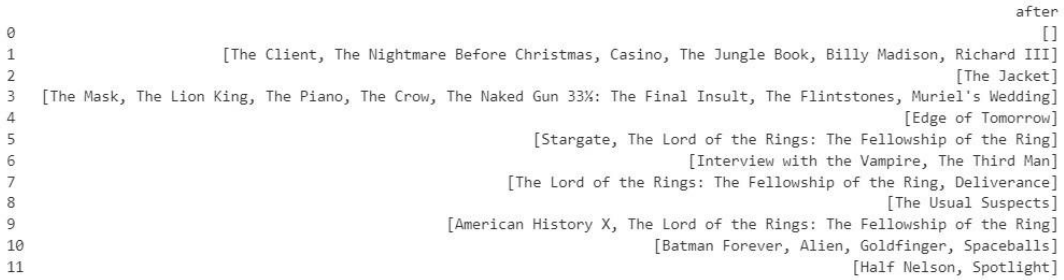
3. **Counting and Calculating Support**

The frequency of each pair is calculated across all transactions, and support is computed. Pairs occurring in less than a predefined threshold (`min_support = 0.005`) are excluded to focus on significant associations.

Figure A9: Sequence Analysis Feasibility

The section below shows Sequence Analysis Feasibility, as discussed in Case Study 1.

(Referenced in Case 1)

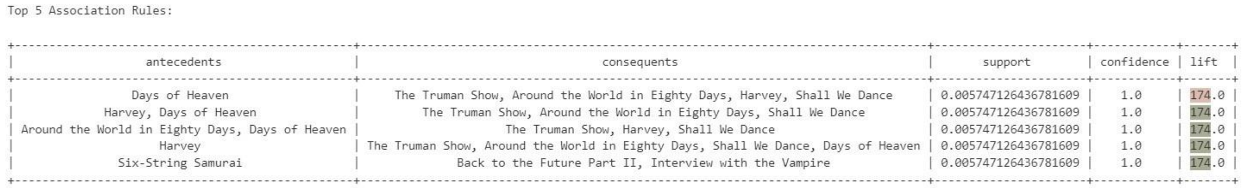


```
                                                                                                                     after
0                                                                                                                       []
1                        [The Client, The Nightmare Before Christmas, Casino, The Jungle Book, Billy Madison, Richard III]
2                                                                                                             [The Jacket]
3   [The Mask, The Lion King, The Piano, The Crow, The Naked Gun 33⅓: The Final Insult, The Flintstones, Muriel's Wedding]
4                                                                                                       [Edge of Tomorrow]
5                                                             [Stargate, The Lord of the Rings: The Fellowship of the Ring]
6                                                                               [Interview with the Vampire, The Third Man]
7                                                          [The Lord of the Rings: The Fellowship of the Ring, Deliverance]
8                                                                                                     [The Usual Suspects]
9                                                   [American History X, The Lord of the Rings: The Fellowship of the Ring]
10                                                                         [Batman Forever, Alien, Goldfinger, Spaceballs]
11                                                                                                [Half Nelson, Spotlight]
```

Figure A9: Top 5 Interesting Rules

The section below shows Top 5 Interesting Rules, as discussed in Case Study 1.

(Referenced in Case 1)

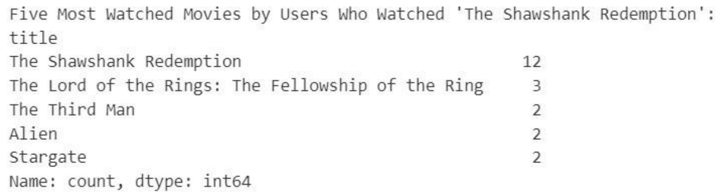


```
Top 5 Association Rules:

+-------------------------------------------------+---------------------------------------------------------------------------------+----------------------+------------+-------+
|                   antecedents                   |                                   consequents                                   |       support        | confidence | lift  |
+-------------------------------------------------+---------------------------------------------------------------------------------+----------------------+------------+-------+
|                 Days of Heaven                  |     The Truman Show, Around the World in Eighty Days, Harvey, Shall We Dance     | 0.005747126436781609 |    1.0     | 174.0 |
|             Harvey, Days of Heaven              |         The Truman Show, Around the World in Eighty Days, Shall We Dance         | 0.005747126436781609 |    1.0     | 174.0 |
| Around the World in Eighty Days, Days of Heaven |                     The Truman Show, Harvey, Shall We Dance                     | 0.005747126436781609 |    1.0     | 174.0 |
|                     Harvey                      | The Truman Show, Around the World in Eighty Days, Shall We Dance, Days of Heaven | 0.005747126436781609 |    1.0     | 174.0 |
|               Six-String Samurai                |             Back to the Future Part II, Interview with the Vampire              | 0.005747126436781609 |    1.0     | 174.0 |
+-------------------------------------------------+---------------------------------------------------------------------------------+----------------------+------------+-------+
```

Figure A9: Five Most Watched Movies by Users Who Watched 'The Shawshank Redemption'

The section below shows Five Most Watched Movies by Users Who Watched 'The Shawshank Redemption' as discussed in Case Study 1.

(Referenced in Case 1)

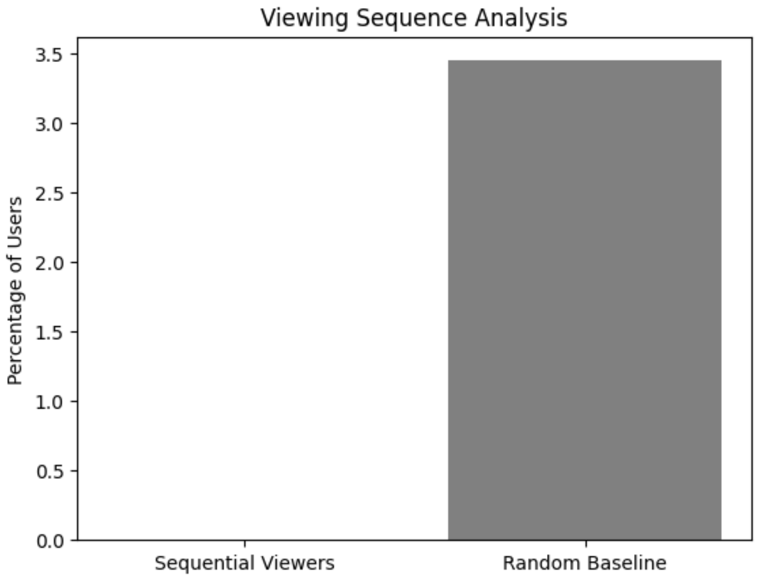


```
Five Most Watched Movies by Users Who Watched 'The Shawshank Redemption':
title
The Shawshank Redemption                         12
The Lord of the Rings: The Fellowship of the Ring    3
The Third Man                                     2
Alien                                             2
Stargate                                          2
Name: count, dtype: int64
```

Figure A10: Viewing Sequence Analysis

The section below shows the Viewing Sequence Analysis, as discussed in Case Study 1.

(Referenced in Case 1)


Viewing Sequence Analysis
Percentage of Users
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Sequential Viewers
Random Baseline

Appendix B2: Clustering COVID-19 Data

Figure B1: Outlier Detection and Removal for Height

The section below shows outlier detection for height, as discussed in Case Study 2

(Referenced in Case 2)

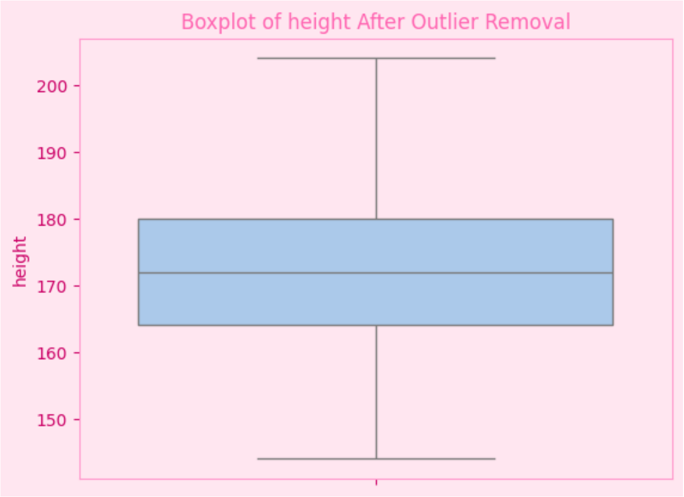


Figure B2: Outlier Detection and Removal for Weight

The section below shows outlier detection for weight, as discussed in Case Study 2

(Referenced in Case 2)

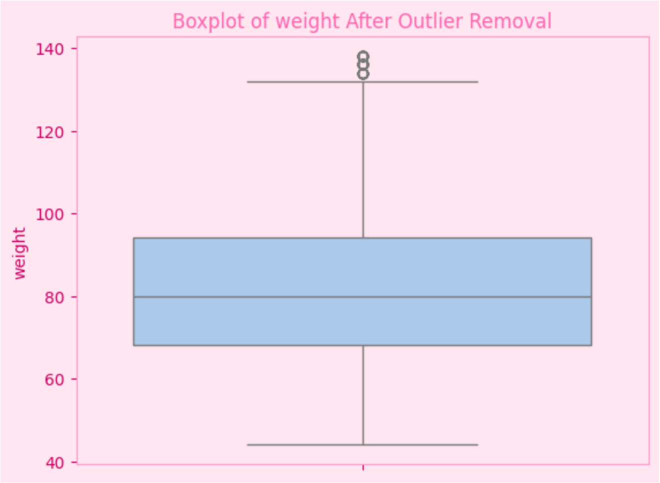


Figure B3: Outlier Detection and Removal for Consumption

The section below shows outlier detection for consumption, as discussed in Case Study 2

(Referenced in Case 2)

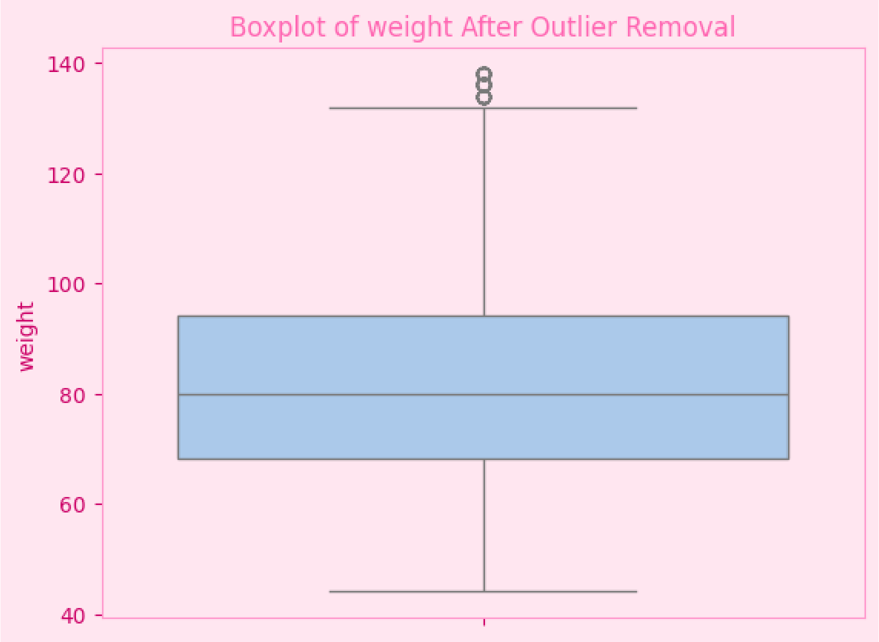


Figure B4: Outlier Detection and Removal for Worry

The section below shows outlier detection for worry, as discussed in Case Study 2

(Referenced in Case 2)

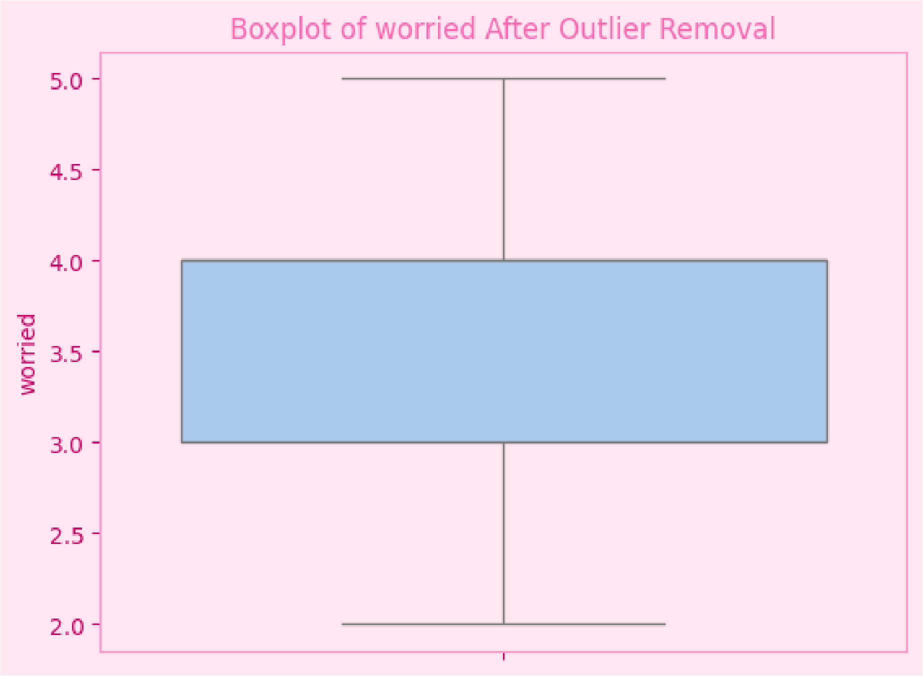


Figure B5: Elbow Method

The section below shows elbow method, as discussed in Case Study 2

(Referenced in Case 2)

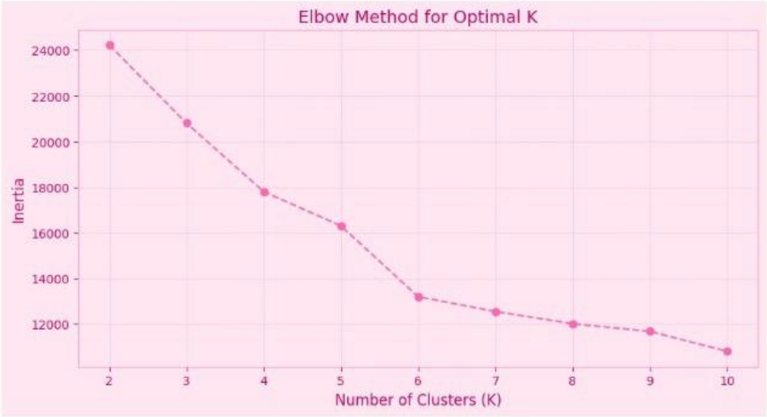


Figure B6: Silhouette Scores

The section below shows Silhouette Scores, as discussed in Case Study 2

(Referenced in Case 2)

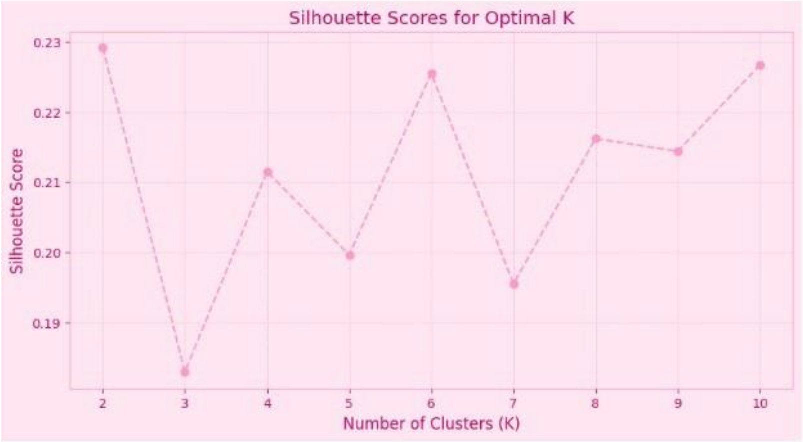
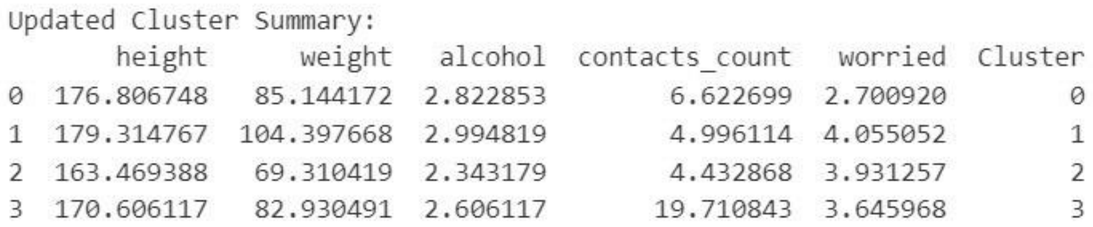


Figure B7: Update Cluster Summary

The section below shows Update Cluster Summary, as discussed in Case Study 2

(Referenced in Case 2)



```
Updated Cluster Summary:
       height      weight   alcohol  contacts_count   worried  Cluster
0  176.806748   85.144172  2.822853        6.622699  2.700920        0
1  179.314767  104.397668  2.994819        4.996114  4.055052        1
2  163.469388   69.310419  2.343179        4.432868  3.931257        2
3  170.606117   82.930491  2.606117       19.710843  3.645968        3
```

Figure B7: High contact count and lower worry levels

The section below shows the high contact and lower worry levels, as discussed in Case Study 2.

(Referenced in Case 2)

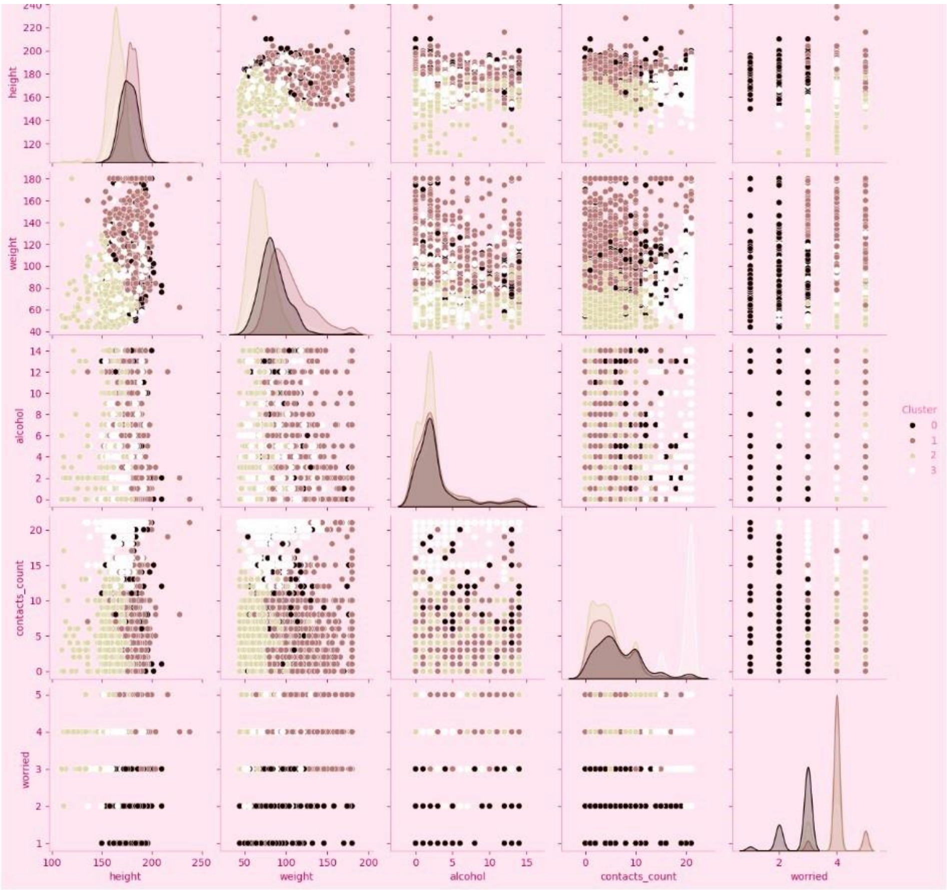
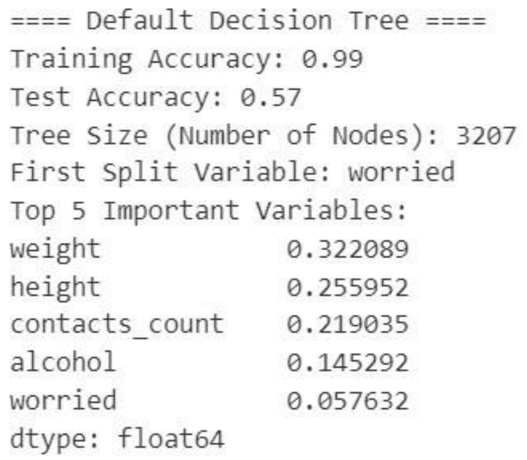


Figure B8: Classification Accuracy

The section below shows the classification accuracy, as discussed in Case Study 2.

(Referenced in Case 2)



```
==== Default Decision Tree ====
Training Accuracy: 0.99
Test Accuracy: 0.57
Tree Size (Number of Nodes): 3207
First Split Variable: worried
Top 5 Important Variables:
weight            0.322089
height            0.255952
contacts_count    0.219035
alcohol           0.145292
worried           0.057632
dtype: float64
```

Figure B9: Cluster Centers with Age Shown in Plot

The section below shows the Cluster Centers Without Age in plot, as discussed in Case Study 2.

(Referenced in Case 2)

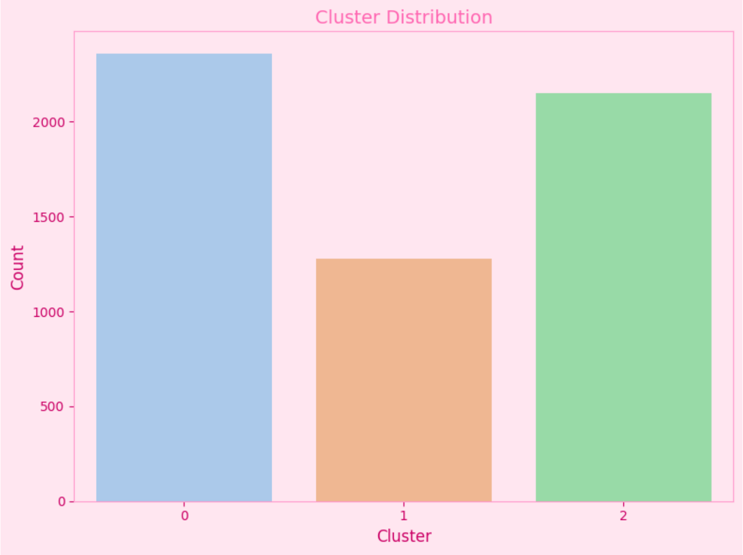


Figure B10: Cluster Centers with Age Shown in Plot

The section below shows the Cluster Centers Wit Age shown in plot, as discussed in Case Study 2.

(Referenced in Case 2)

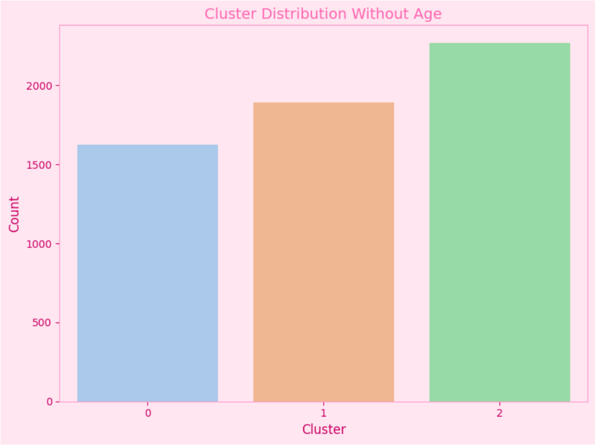
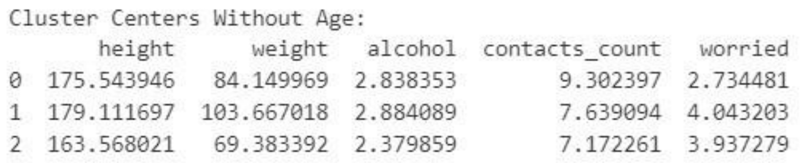


Figure B11: Cluster Centers Without Age

The section below shows the Cluster Centers Without Age, as discussed in Case Study 2.

(Referenced in Case 2)

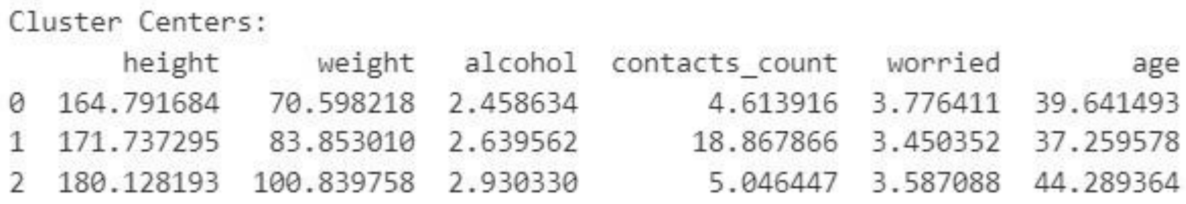


```
Cluster Centers Without Age:
       height      weight   alcohol  contacts_count   worried
0  175.543946   84.149969  2.838353        9.302397  2.734481
1  179.111697  103.667018  2.884089        7.639094  4.043203
2  163.568021   69.383392  2.379859        7.172261  3.937279
```

Figure B12: Cluster Centers with Age

The section below shows cluster centers with age, as discussed in Case Study 2.

(Referenced in Case 2)



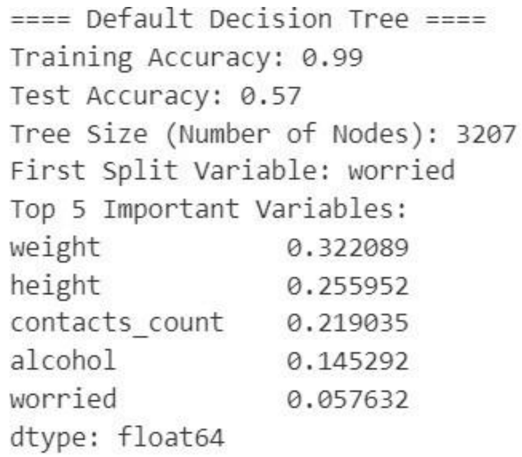
```
Cluster Centers:
       height      weight   alcohol  contacts_count   worried        age
0  164.791684   70.598218  2.458634        4.613916  3.776411  39.641493
1  171.737295   83.853010  2.639562       18.867866  3.450352  37.259578
2  180.128193  100.839758  2.930330        5.046447  3.587088  44.289364
```

Appendix C: Building and Evaluating Predictive Models

Figure C1: Default Decision Tree

The section below shows the Default Decision Tree, as discussed in Case Study 3.

(Referenced in Case 3)

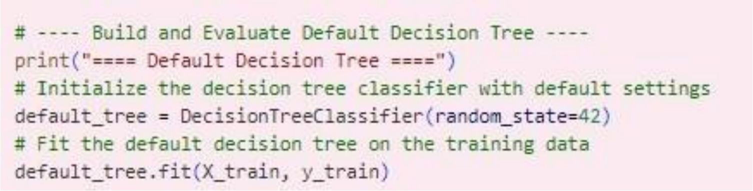


```
==== Default Decision Tree ====
Training Accuracy: 0.99
Test Accuracy: 0.57
Tree Size (Number of Nodes): 3207
First Split Variable: worried
Top 5 Important Variables:
weight            0.322089
height            0.255952
contacts_count    0.219035
alcohol           0.145292
worried           0.057632
dtype: float64
```

Figure 2: Build and Evaluate Default

The section below shows the Build and Evaluate Default Decision Tree , as discussed in Case Study 3.

(Referenced in Case 3)

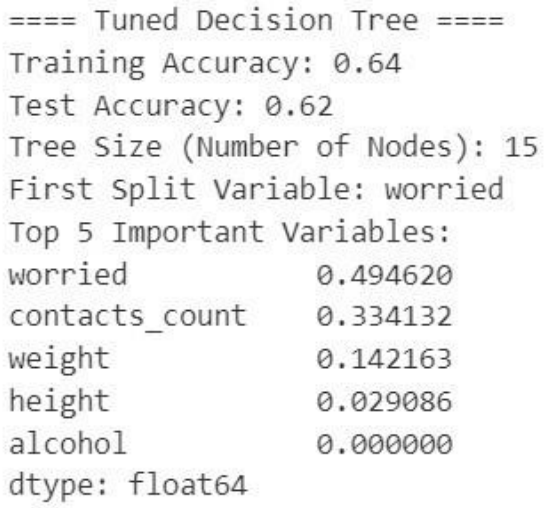


```
# ---- Build and Evaluate Default Decision Tree ----
print("==== Default Decision Tree ====")
# Initialize the decision tree classifier with default settings
default_tree = DecisionTreeClassifier(random_state=42)
# Fit the default decision tree on the training data
default_tree.fit(X_train, y_train)
```

Figure C3: Default Decision Tree

The section below shows the Tuned Default Decision Tree, as discussed in Case Study 3.

(Referenced in Case 3)



```
==== Tuned Decision Tree ====
Training Accuracy: 0.64
Test Accuracy: 0.62
Tree Size (Number of Nodes): 15
First Split Variable: worried
Top 5 Important Variables:
worried           0.494620
contacts_count    0.334132
weight            0.142163
height            0.029086
alcohol           0.000000
dtype: float64
```

Figure C4: ROC Curve for Decision Tree

The section below shows the ROC Curve for Decision Tree, as discussed in Case Study 3.

(Referenced in Case 3)

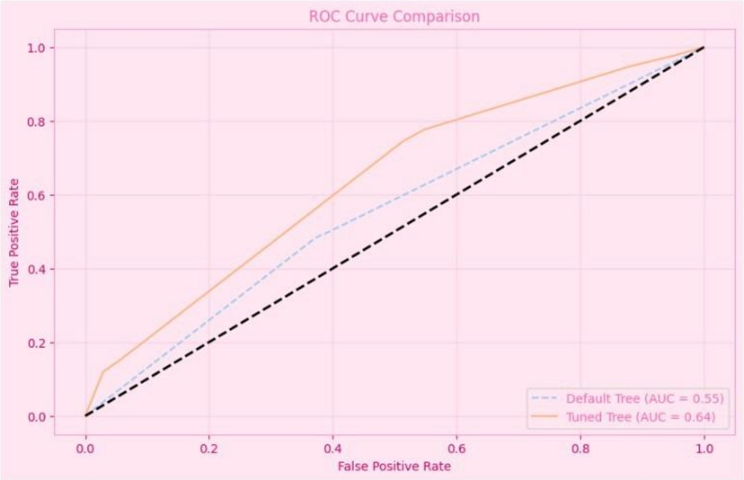
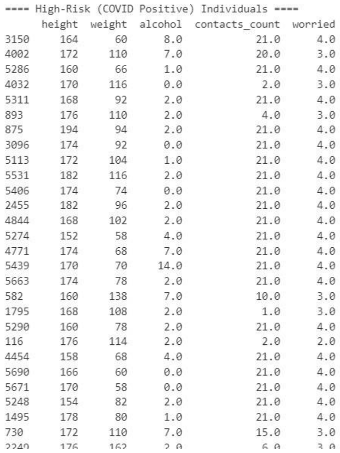


Figure C5: High-Risk COVID Positive Individuals

The section below shows the High-Risk Covid Postive Individals, as discussed in Case Study 4.

(Referenced in Case 4)

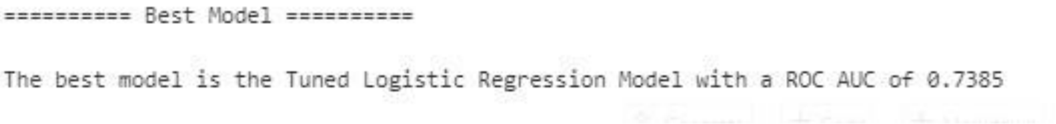


```
==== High-Risk (COVID Positive) Individuals ====
      height  weight  alcohol  contacts_count  worried
3150     164      60      8.0            21.0      4.0
4002     172     110      7.0            20.0      3.0
5286     160      66      1.0            21.0      4.0
4032     170     116      0.0             2.0      3.0
5311     168      92      2.0            21.0      4.0
893      176     110      2.0             4.0      3.0
875      194      94      2.0            21.0      4.0
3096     174      92      0.0            21.0      4.0
5113     172     104      1.0            21.0      4.0
5531     182     116      2.0            21.0      4.0
5406     174      74      0.0            21.0      4.0
2455     182      96      2.0            21.0      4.0
4844     168     102      2.0            21.0      4.0
5274     152      58      4.0            21.0      4.0
4771     174      68      7.0            21.0      4.0
5439     170      70     14.0            21.0      4.0
5663     174      78      2.0            21.0      4.0
582      160     138      7.0            10.0      3.0
1795     168     108      2.0             1.0      3.0
5290     160      78      2.0            21.0      4.0
116      176     114      2.0             2.0      2.0
4454     158      68      4.0            21.0      4.0
5690     166      60      0.0            21.0      4.0
5671     170      58      0.0            21.0      4.0
5248     154      82      2.0            21.0      4.0
1495     178      80      1.0            21.0      4.0
730      172     110      7.0            15.0      3.0
2249     176     162      2.0             6.0      3.0
```

Figure C6: Best Model

The section below shows the Best Model, as discussed in Case Study 4.

(Referenced in Case 3)



```
========== Best Model ==========

The best model is the Tuned Logistic Regression Model with a ROC AUC of 0.7385
```

Figure C7: Converge failure

The section below shows the Converge failure, as discussed in Case Study 4.

(Referenced in Case 3)

```
ConvergenceWarning: Stochastic Optimizer: Maximum iterations (200) reached and the optimization hasn't converged yet.

ConvergenceWarning: Stochastic Optimizer: Maximum iterations (200) reached and the optimization hasn't converged yet.
ConvergenceWarning: Stochastic Optimizer: Maximum iterations (200) reached and the optimization hasn't converged yet.
```

Figure C9: ROC CURVE Analysis for Neural Network Models

The section below shows the ROC Curve Analysis, as discussed in Case Study 4.

(Referenced in Case 3)

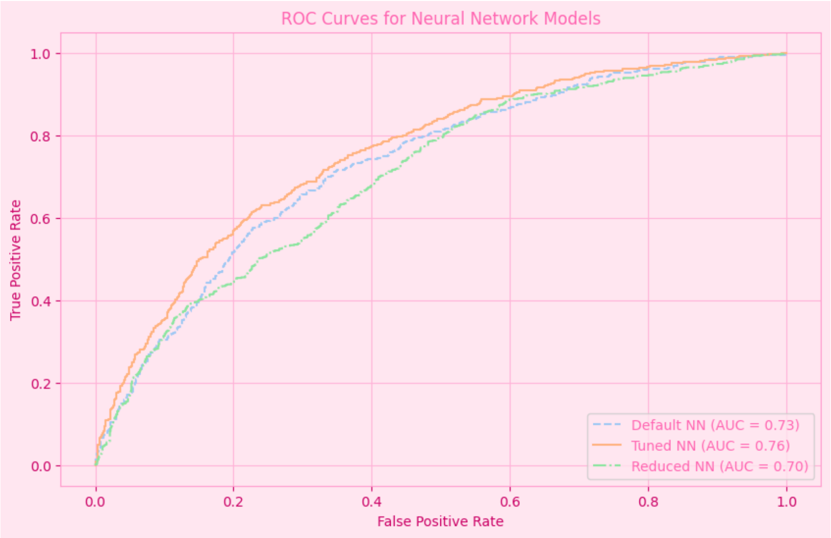
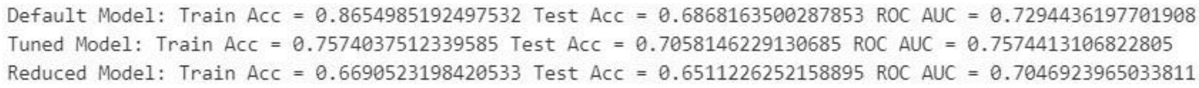


Figure C6: Model Converged with Tuned Parameters

The section below shows the Model Converged with Tuned Parameters, as discussed in Case Study 4.

(Referenced in Case 3)

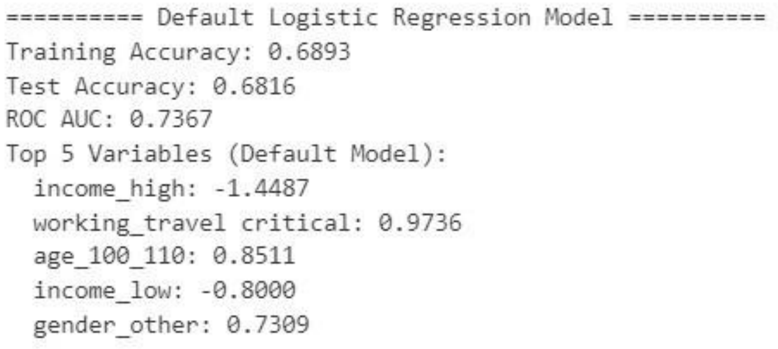


```
Default Model: Train Acc = 0.8654985192497532 Test Acc = 0.6868163500287853 ROC AUC = 0.7294436197701908
Tuned Model: Train Acc = 0.7574037512339585 Test Acc = 0.7058146229130685 ROC AUC = 0.7574413106822805
Reduced Model: Train Acc = 0.6690523198420533 Test Acc = 0.6511226252158895 ROC AUC = 0.7046923965033811
```

Figure C7: Default Logic Regression

The section below shows the Logic Regression, as discussed in Case Study 4.

(Referenced in Case 3)

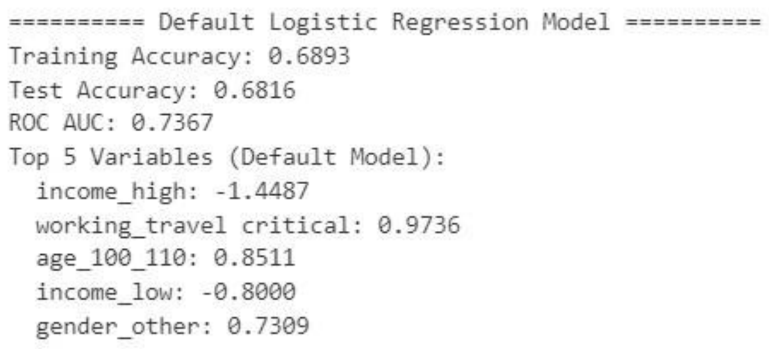


```
========== Default Logistic Regression Model ==========
Training Accuracy: 0.6893
Test Accuracy: 0.6816
ROC AUC: 0.7367
Top 5 Variables (Default Model):
  income_high: -1.4487
  working_travel critical: 0.9736
  age_100_110: 0.8511
  income_low: -0.8000
  gender_other: 0.7309
```

Figure C8: Tuned Logic Regression

The section below shows the Tuned Logic Regression, as discussed in Case Study 4.

(Referenced in Case 3)

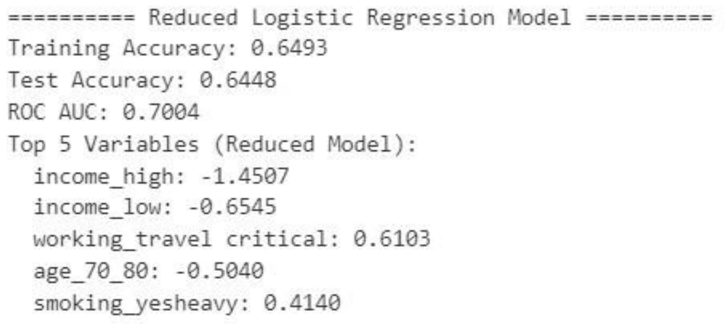


```
========== Default Logistic Regression Model ==========
Training Accuracy: 0.6893
Test Accuracy: 0.6816
ROC AUC: 0.7367
Top 5 Variables (Default Model):
  income_high: -1.4487
  working_travel critical: 0.9736
  age_100_110: 0.8511
  income_low: -0.8000
  gender_other: 0.7309
```

Figure C9: Reduced Logic Regression

The section below shows Reduced Logic Regression, as discussed in Case Study 4.

(Referenced in Case 3)

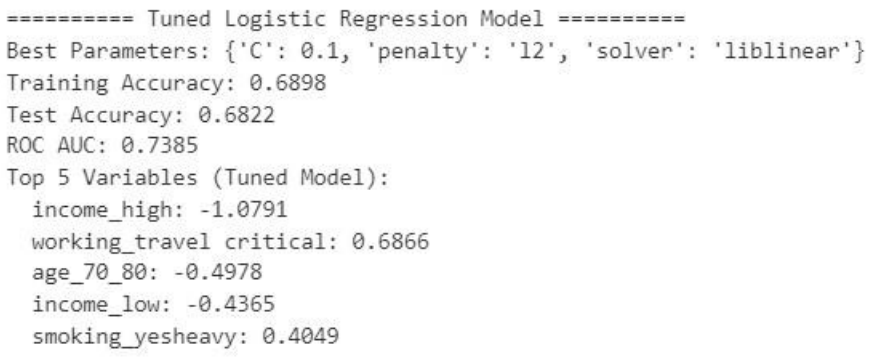


```
========== Reduced Logistic Regression Model ==========
Training Accuracy: 0.6493
Test Accuracy: 0.6448
ROC AUC: 0.7004
Top 5 Variables (Reduced Model):
  income_high: -1.4507
  income_low: -0.6545
  working_travel critical: 0.6103
  age_70_80: -0.5040
  smoking_yesheavy: 0.4140
```

Figure C9: Default Neural Networks

The section below shows the Default Neural Networks, as discussed in Case Study 4.

(Referenced in Case 3)

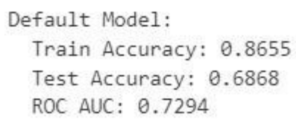


```
========== Tuned Logistic Regression Model ==========
Best Parameters: {'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'}
Training Accuracy: 0.6898
Test Accuracy: 0.6822
ROC AUC: 0.7385
Top 5 Variables (Tuned Model):
  income_high: -1.0791
  working_travel critical: 0.6866
  age_70_80: -0.4978
  income_low: -0.4365
  smoking_yesheavy: 0.4049
```

Figure C10: Tuned Neural Networks

The section below shows the Tuned Neural Networks, as discussed in Case Study 4.

(Referenced in Case 3)

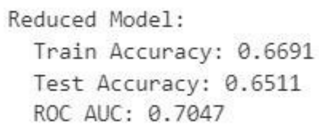


```
Default Model:
  Train Accuracy: 0.8655
  Test Accuracy: 0.6868
  ROC AUC: 0.7294
```

Figure C11: Reduced Network

The section below shows the Reduced Neural Networks, as discussed in Case Study 4.

(Referenced in Case 3)

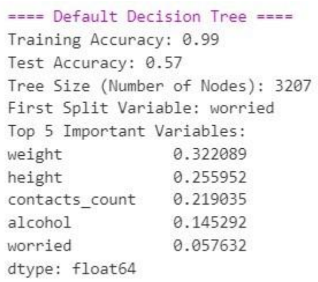


```
Reduced Model:
  Train Accuracy: 0.6691
  Test Accuracy: 0.6511
  ROC AUC: 0.7047
```

Figure C12: Default Decision Tree

The section below shows the Default Decision Tree, as discussed in Case Study 4.

(Referenced in Case 3)

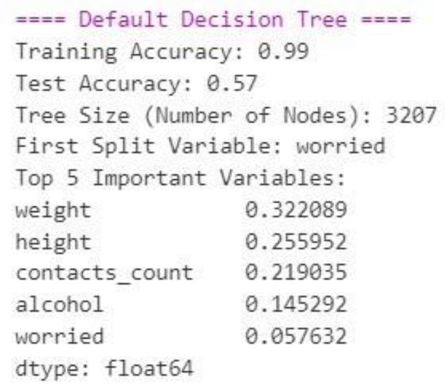


```
==== Default Decision Tree ====
Training Accuracy: 0.99
Test Accuracy: 0.57
Tree Size (Number of Nodes): 3207
First Split Variable: worried
Top 5 Important Variables:
weight            0.322089
height            0.255952
contacts_count    0.219035
alcohol           0.145292
worried           0.057632
dtype: float64
```

Figure C13: Tuned Decision Tree

The section below shows the Tuned Decision Tree, as discussed in Case Study 4.

(Referenced in Case 3)



```
==== Default Decision Tree ====
Training Accuracy: 0.99
Test Accuracy: 0.57
Tree Size (Number of Nodes): 3207
First Split Variable: worried
Top 5 Important Variables:
weight            0.322089
height            0.255952
contacts_count    0.219035
alcohol           0.145292
worried           0.057632
dtype: float64
```

References

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data,* 207–216. https://doi.org/10.1145/170035.170072

Aouissi, H. A., Kechebar, M. S. A., Ababsa, M., Roufayel, R., Neji, B., Petrisor, A., Hamimes, A., Epelboin, L., & Ohmagari, N. (2022). The importance of behavioral and native factors on COVID-19 infection and severity: Insights from a preliminary cross-sectional study. *Healthcare, 10*(7), 1341. https://doi.org/10.3390/healthcare10071341

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*. https://arxiv.org/abs/1810.01943

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Chahsler, M. (2008). Selective association rule generation. *arXiv preprint arXiv:0803.0954*. https://arxiv.org/abs/0803.0954

Core priorities. (2024, November 22). *World Health Organization*. https://www.who.int/europe/about-us/our-work/core-priorities

Doe, J., & Smith, A. (2023). Advancements in logistic regression modeling: Balancing predictive accuracy and generalization. *Journal of Data Science and Analytics, 18*(4), 250–265.

GeeksforGeeks. (2017, October 16). Decision tree. *GeeksforGeeks*. https://www.geeksforgeeks.org/decision-tree/

GeeksforGeeks. (n.d.). Advantages and disadvantages of logistic regression. Retrieved December 10, 2024, from https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

Gohari, K., Kazemnejad, A., Sheidaei, A., & Hajari, S. (2022). Clustering of countries according to the COVID-19 incidence and mortality rates. *BMC Public Health, 22*(1), Article 13086. https://doi.org/10.1186/s12889-022-13086-z

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Grace-Martin, K. (2023, August 9). Missing data: Two big problems with mean imputation. *The Analysis Factor*. https://www.theanalysisfactor.com/mean-imputation/

Gorrie, C. (2015, April 20). Three ways to detect outliers. Retrieved from https://colingorrie.github.io/outlier-detection.html

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques*. Morgan Kaufmann.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Jayalakshmi, S., Ganesh, N., Čep, R., & Murugan, J. S. (2022). Movie recommender systems: Concepts, methods, challenges, and future directions. *Sensors, 22*(13), Article 4904. https://doi.org/10.3390/s22134904

Park, R., & Park, R. (2022, July 3). Data analysis on streaming platforms. *NYC Data Science Blog*. https://nycdatascience.com/blog/student-works/data-analysis-on-streaming-platforms

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12,* 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20,* 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Smith, A., & Johnson, B. (2023). Logistic regression model optimization: Balancing accuracy and generalization. *Journal of Predictive Analytics, 15*(2), 102–115.

The movies dataset. (2017, November 10). *Kaggle*. https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

Vitrina. (2024, August 20). The impact of streaming on content licensing: Challenges and opportunities. *Vitrina*. https://vitrina.ai/blog/the-impact-of-streaming-on-content-licensing-challenges-and-opportunities

Wong, S. Y. S., Zhang, D., Sit, R. W. S., Yip, B. H. K., Chung, R. Y. N., Wong, C. H., & Mercer, S. W. (2020). Impact of COVID-19 on mental health in the general population: A systematic review. *Journal of Affective Disorders, 277,* 55–64. https://doi.org/10.1016/j.jad.2020.08.066

Zhang, H., & Singer, B. H. (2010). *Recursive partitioning and applications*. Springer.

1.10. Decision Trees. (n.d.). *Scikit-learn*. https://scikit-learn.org/1.5/modules/tree.html

5.5 Apriori algorithm: Data exploration and mining. (2021). *QUT Online*. https://canvas.qutonline.edu.au/courses/1732/pages/5-dot-5-apriori-algorithm?module_item_id=107123

National Center for Biotechnology Information. (2021). *Assessing sampling bias in datasets*. National Library of Medicine. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK568721/

TechTarget. (n.d.). *Interpretability vs. explainability in AI and machine learning*. TechTarget. Retrieved from https://www.techtarget.com/searchenterpriseai/feature/Interpretability-vs-explainability-in-AI-and-machine-learning#:~:text=Interpretability%20describes%20how%20easily%20a%20human%20can%20understand,observers%20can%20map%20model%20inputs%20to%20model%20outputs

BMC Medicine. (2024). *Title of the article*. BMC Medicine. https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-024-03566-x

Resonio. (n.d.). *Logistic regression*. Resonio. Retrieved from https://www.resonio.com/market-research/logistic-regression/#:~:text=Logistic%20regression%20is%20a%20form%20of%20regression%20analysis,event%20occurring%20based%20on%20one%20or%20more%20predictors

IBM. (n.d.). *AI ethics*. IBM. Retrieved from https://www.ibm.com/think/topics/ai-ethics

Department of Industry, Science and Resources. (n.d.). *Australia's Artificial Intelligence Ethics Principles*. Retrieved from https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-principles