Final Project

Group 04

College of Professional Studies - Analytics, Northeastern University - Toronto

ALY 6040 80781 Data Mining Applications

Dr. Harpreet Sharma

Date: May 16th, 2025

Introduction

Suicide remains one of the most complex and pressing global public health challenges, demanding comprehensive and data-driven approaches. This report leverages a rich dataset of **27,820 records from 101 countries spanning 1985 to 2016** to uncover patterns in suicide rates and develop predictive models using demographic and economic indicators.

The dataset includes variables such as **gender, age group, generation, GDP per capita, population size**, and **suicides per 100k people**. Although it initially featured the **Human Development Index (HDI)**, this column was excluded due to over **70% missing values**, in alignment with best practices for maintaining analytical reliability.

Key Trends Addressed and Solved Through This Analysis:

1. **Age-Specific Risk**: Suicide rates significantly **increase with age**, peaking in the **75+** group. The model confirmed that age is a critical predictor of risk.
2. **Gender Disparity**: **Males consistently exhibit higher suicide rates** than females across all age groups. The model leveraged sex as a strong discriminative feature.
3. **GDP Alone Is Not a Safeguard**: There is **no consistent linear relationship** between GDP per capita and suicide risk. The analysis reveals that economic prosperity alone does not ensure lower suicide rates.
4. **Generational Influence**: **Generation X and Baby Boomers** showed the highest suicide counts, with post-2010 data suggesting possible improvements linked to mental health awareness and policy.
5. **Country-Specific Burden**: Nations such as **Russia, the United States, and Japan** emerged as major contributors to global suicide figures, confirming the need for **localized, culturally specific intervention strategies**.

These insights, confirmed through **machine learning modeling and exploratory data analysis**, provide stakeholders with a clearer understanding of where to focus suicide prevention resources and how to better tailor interventions.

Analysis 01: XGBoost

Research Question
Can demographic and economic indicators predict suicide risk levels (Low, Medium, High) globally using XGBoost in R?

1. Methodology

Tools Used

- **R Language**: For statistical computing and machine learning
- **Libraries**: xgboost, caret, ggplot2, data.table

Dataset Description
The dataset includes **27,820 records** across **101 countries** from **1985 to 2016**, containing the following key variables:

- Year, Country, Sex, Age group, Generation
- Suicides per 100k population
- GDP per capita

Data Preprocessing Steps

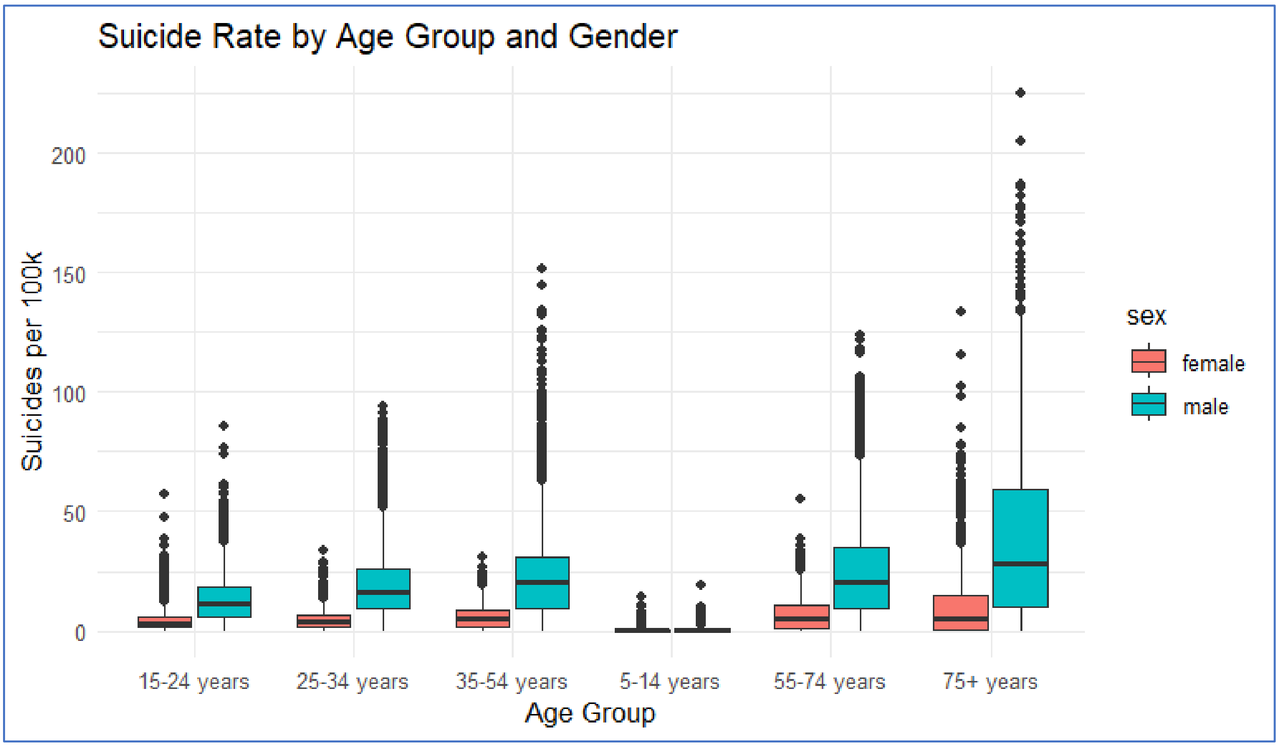
- Removed the country-year field
- Cleaned GDP values by removing commas and converting to numeric
- Removed rows with missing values in suicides/100k
- Binned suicide rates into categories: **Low (0–5)**, **Medium (5–15)**, and **High (>15)**
- Converted categorical variables into factors
- Applied **one-hot encoding** for modeling compatibility with XGBoost

2. Exploratory Data Analysis (EDA)

Key Insights:

1. **Age Trend**: Suicide rates increase with age, peaking in the **75+ group**
2. **Gender Disparity**: **Males** consistently show higher suicide rates than females
3. **GDP Correlation**: No strong linear correlation found between GDP per capita and suicide risk

Visual Evidence:
A boxplot analysis showed:



- **Highest risk** in **elderly males**
- **Greater variability** in male suicide rates across all age groups

3. Modeling with XGBoost

Feature Set:

- **Categorical**: country, sex, age, generation
- **Numeric**: year, population, GDP per capita

Model Configuration:

- objective = multi:softprob (multi-class probability classification)
- num_class = 3 (Low, Medium, High)
- eval_metric = mlogloss (multiclass log loss)
- nrounds = 200 (optimized)

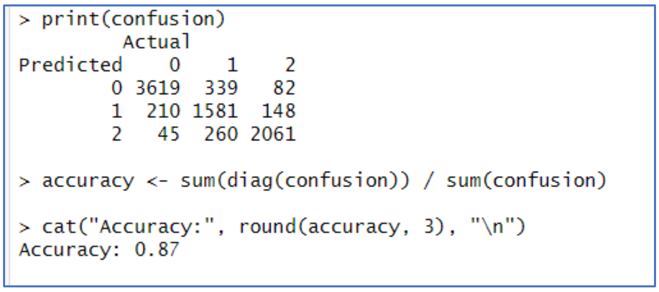
Training & Evaluation:

- **70/30 train-test split** using caret::createDataPartition
- Predictions derived from the class with the highest probability

4. Optimization Process

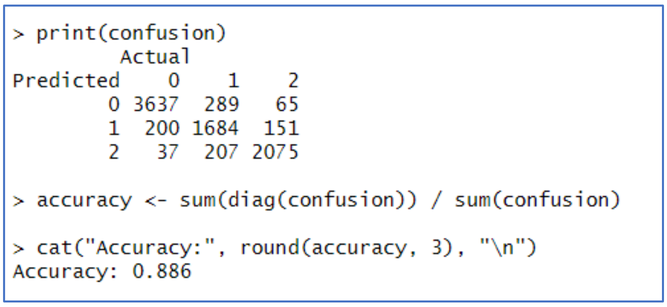
Model performance was fine-tuned by experimenting with different boosting rounds:

- **100 rounds** yielded **87% accuracy**



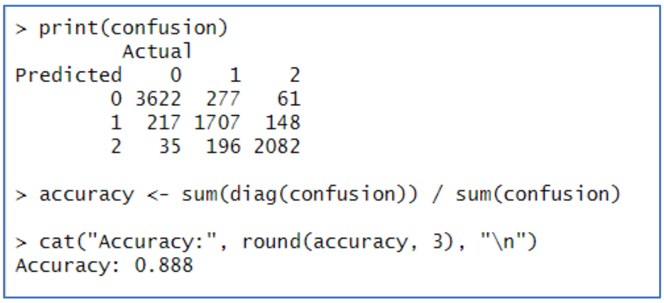
```
> print(confusion)
         Actual
Predicted    0    1    2
        0 3619  339   82
        1  210 1581  148
        2   45  260 2061

> accuracy <- sum(diag(confusion)) / sum(confusion)

> cat("Accuracy:", round(accuracy, 3), "\n")
Accuracy: 0.87
```

- **200 rounds** yielded **88.6% accuracy**



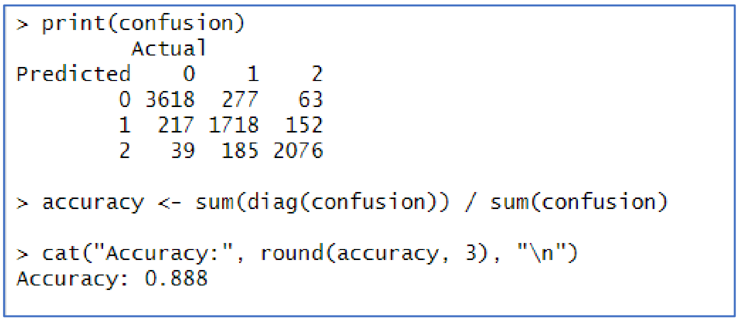
```
> print(confusion)
         Actual
Predicted    0    1    2
        0 3637  289   65
        1  200 1684  151
        2   37  207 2075

> accuracy <- sum(diag(confusion)) / sum(confusion)

> cat("Accuracy:", round(accuracy, 3), "\n")
Accuracy: 0.886
```

- **300 rounds** yielded **88.8% accuracy**



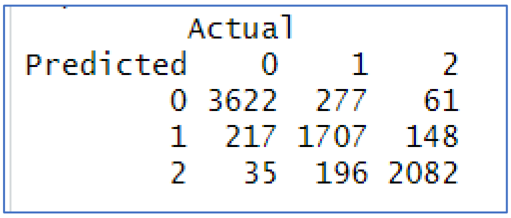
```
> print(confusion)
         Actual
Predicted    0    1    2
        0 3622  277   61
        1  217 1707  148
        2   35  196 2082

> accuracy <- sum(diag(confusion)) / sum(confusion)

> cat("Accuracy:", round(accuracy, 3), "\n")
Accuracy: 0.888
```

- **400 rounds** yielded **88.8% accuracy**



```
> print(confusion)
         Actual
Predicted    0    1    2
        0 3618  277   63
        1  217 1718  152
        2   39  185 2076

> accuracy <- sum(diag(confusion)) / sum(confusion)

> cat("Accuracy:", round(accuracy, 3), "\n")
Accuracy: 0.888
```

300 Rounds Confusion matrix:



```
           Actual
Predicted    0    1    2
        0 3622  277   61
        1  217 1707  148
        2   35  196 2082
```

Given the slightly higher accuracy of **88.8% at 300 rounds**, it showed improved performance over previous configurations. However, training time and the minimal gain over 200 rounds make both configurations viable depending on the operational context. **For this report, we proceed with 300 rounds to reflect the best accuracy.**

5. Results

Confusion Matrix Insights (300 Rounds):

- **High precision** in identifying **Low risk** cases
- **Moderate overlap** between **Medium** and **High**, suggesting behavioral proximity
- **Overall Accuracy**: **88.8%**

6. Interpretation

The model effectively confirmed all major EDA findings. **Age and gender** emerged as pivotal predictors, while **GDP per capita** alone did not suffice as a strong indicator. For data stakeholders like the **WHO and national health departments**, the model offers actionable insights to segment high-risk populations and allocate interventions more effectively.

7. Recommendations & Conclusions

Recommendations:

- Focus mental health outreach on **older males**, the highest-risk demographic
- Integrate **additional variables** like mental health funding, stigma scores, and healthcare access
- Utilize **time-series models** to analyze policy impact over time

Future Work:

- Compare XGBoost with **Random Forest** and **SVM** for performance benchmarking
- Use **SHAP values** to enhance feature interpretability
- Develop an **interactive dashboard** for real-time policy decision support

Conclusion:

XGBoost, with an optimized configuration of **300 boosting rounds**, achieved a strong **88.8% accuracy** in classifying suicide risk levels globally. The model's alignment with exploratory insights, combined with its scalability and interpretability, underscores its potential as a **reliable decision-support tool** for global suicide prevention initiatives.

Analysis 2: K-Means Clustering 2,3 and Elbow Method

Business Problems explored

1. Had the measures taken to improve HDI in the past three decades reduced the suicide rates across the countries?

2. If improved HDI had reduced Suicide rates, what are the key positive influencers/features, that we need to work upon to further reduce the suicide rates in these countries?

3. If the improved HDI did not have any effect on reducing the suicide rates, what are the various reasons that could have contributed to the negative or null effect of HDI on suicide rates?

Tools used: K-Means Clustering 2, 3, Elbow method.

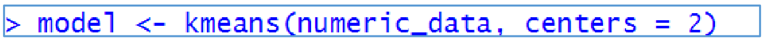
Why do we use K-Means Clustering?

K-Means Clustering is an Unsupervised Machine Learning algorithm which groups the unlabeled dataset into different clusters. In this dataset the countries were not labeled as developed, developing and underdeveloped in its raw form. K-Means clustering identified the optimal labelling in iterations of 2 cluster, 3 clusters and Elbow to identify the optimal number of clusters to split the data logically into 3 categories. This clustering helped in understanding the underlying demarcation between the countries within the data which was not explicitly called out in the dataset features, nor did it emerge during the EDA process.

1. Clustering using K-means with 2 clusters

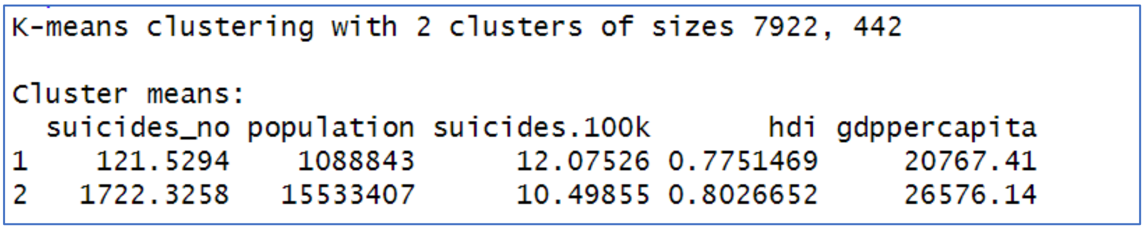
Analysis:

This analysis aims to explore the relationship between Human Development Index (HDI) improvements and suicide rates across countries over the past three decades. Specifically, it investigates whether the socioeconomic advancements captured by HDI, such as better education, higher life expectancy, and increased income have contributed to a reduction in suicide rates globally. If a negative correlation is found, the analysis will identify which HDI components or related social indicators most effectively drive this reduction, providing insights into targeted policy areas such as healthcare access, education quality, or social welfare systems. Conversely, if no significant relationship exists, the study will examine underlying factors that may counteract the benefits of HDI, including cultural stigma around mental health, rising urban stress, social isolation, or economic inequality. The findings will guide governments, NGOs, and international organizations in shaping more holistic development strategies that integrate mental health outcomes as core indicators of national progress.

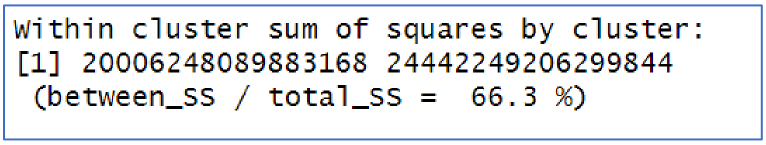


```
> model <- kmeans(numeric_data, centers = 2)
```

(BLUMAN, 2022)



```
K-means clustering with 2 clusters of sizes 7922, 442

Cluster means:
  suicides_no population suicides.100k       hdi gdppercapita
1    121.5294    1088843      12.07526 0.7751469     20767.41
2   1722.3258   15533407      10.49855 0.8026652     26576.14
```



```
Within cluster sum of squares by cluster:
[1] 20006248089883168 24442249206299844
 (between_SS / total_SS =  66.3 %)
```

Interpretation:

K-means has split the observations into two clusters, where **Cluster 1** contains **442 observations** and **Cluster 2** contains **7922 observations**. Cluster means provide the means of the numeric values used in the model independently for clusters.

Cluster 1: High GDP, high population, high suicide count, but lower suicide rates per 100k. In other words, Cluster 1 has Countries/regions with **larger, more economically developed populations**

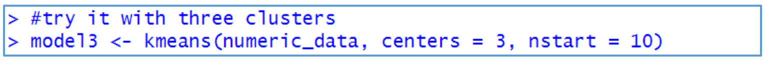
Cluster 2: Lower GDP/population, slightly higher suicide rate. In other words, cluster 2 has Versus those with **smaller, potentially less developed populations** but slightly higher suicide rates per capita

Cluster	suicides no	population	suicides/100k	HDI	GDP per capita
1	121.53	1.09M	12.08	0.775	20,767
2	1722.33	15.5M	10.5	0.803	26,576

Clustering Performance is measured by using the formula: **Between Sum of squares / Total Sum of squares (66.3%)**

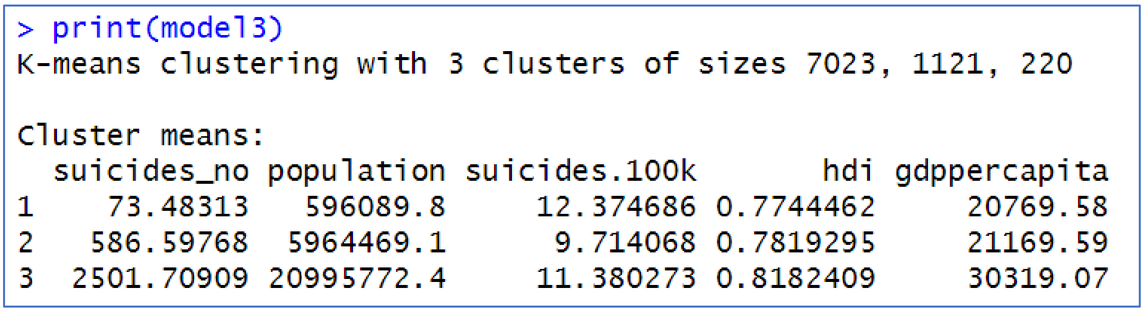
- 66.3% of the variance is explained by the clustering.
- This is reasonably strong, meaning the clusters separate the data quite well.

K Means – 3 Clusters: We performed K-Means 3 clusters to check the efficiency of the clustering using 2 clusters and 3 clusters.

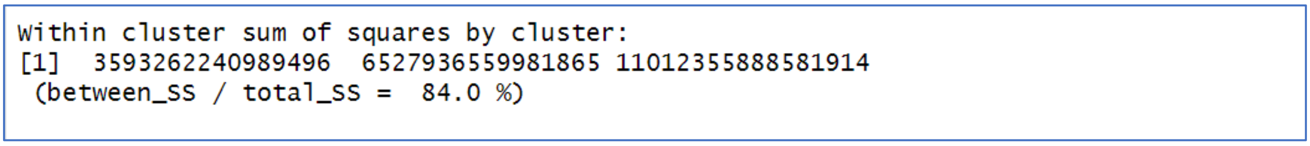


```
> #try it with three clusters
> model3 <- kmeans(numeric_data, centers = 3, nstart = 10)
```

Interpretation:



```
> print(model3)
K-means clustering with 3 clusters of sizes 7023, 1121, 220

Cluster means:
  suicides_no population suicides.100k       hdi gdppercapita
1    73.48313   596089.8     12.374686 0.7744462     20769.58
2   586.59768  5964469.1      9.714068 0.7819295     21169.59
3  2501.70909 20995772.4     11.380273 0.8182409     30319.07
```



```
Within cluster sum of squares by cluster:
[1]  3593262240989496  6527936559981865 11012355888581914
 (between_SS / total_SS =  84.0 %)
```

The algorithm had identified **3 clusters** from the numeric dataset.

Cluster sizes are **Cluster 1** with total of 1,121 data points, **Cluster 2** with total of 220 data points (smallest) and **Cluster 3** with total of 7,023 data points (largest)

Characteristics of each of the clusters as split by K-Means is as below:

Cluster	suicides no	population	suicides/100k	HDI	GDP per capita
1	73.48	596k	12.37	0.774	20,770
2	586.6	5.96M	9.71	0.782	21,170
3	2,501.71	20.99M	11.38	0.818	30,319

Interpretation:

Cluster 1 **(Mid-size, Based on the Suicide rate, it is medium risk)**

- Mid-range population (approx. 6 million)
- Moderate suicides/100k rate (apporx. 9.7)
- Mid-level GDP and HDI
- Hence, Cluster 1 represents developing countries or regions with moderate suicide metrics

Cluster 2 **(High-resource, Greater population base)**

- Highest **population** (approx. 21M)
- Very high **total suicides** (2,500+ on average)
- Higher **suicide rate** per 100k (11.38)
- Highest **GDP** and **HDI**
- Represents **developed/high-income countries** with **higher suicide counts,** possibly due to larger populations and better reporting

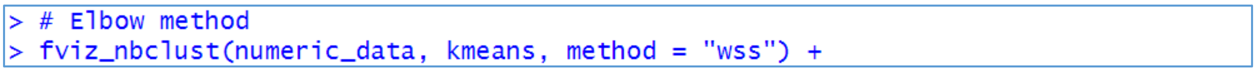
Cluster 3 **(Low-resource, High-rate Group)**

- Lowest **population** (approx. 600k)
- Lowest **total suicides** (73.5 avg), but **highest suicide rate per 100k** (12.37)
- Lower GDP and HDI
- May represent **poor nations** with **high suicide rates**

Between SS / Total SS:

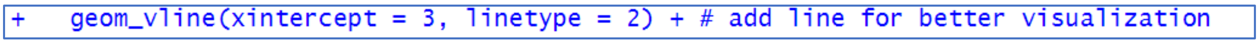
- Clustering explains **84%** of the total variance in the dataset.
- This suggests that the **clustering is good and has separated the data well**.

K-Means: Elbow method



```
> # Elbow method
> fviz_nbclust(numeric_data, kmeans, method = "wss") +
```

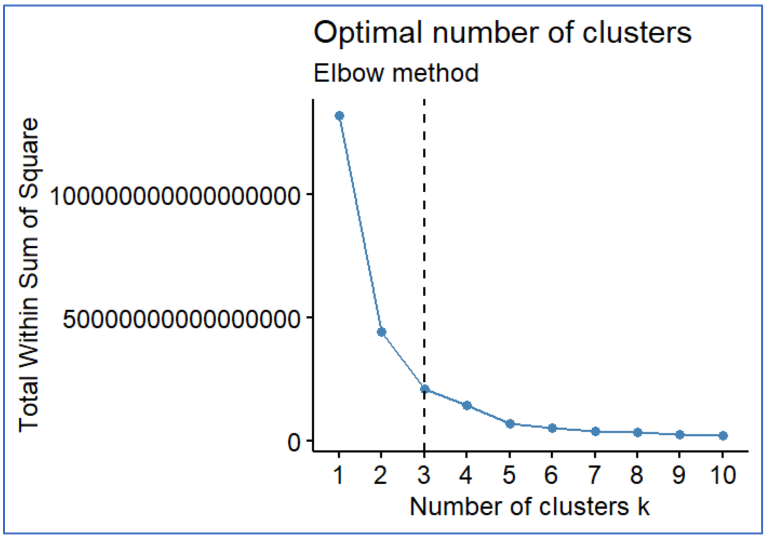
(kabacoff, 2015/2024)



```
+   geom_vline(xintercept = 3, linetype = 2) + # add line for better visualization
```

The above **geom_vline** code adds a **vertical dashed line** at x = 3, which helps highlight the optimal number of clusters visually, and linetype = 2 means a dashed line.

Output Analysis:



Interpretation: The elbow plot helps in deciding the number of clusters by seeing where the WSS curve bends, the "elbow." The dashed line at 3 suggests you can take 3 to be an optimal number of clusters for this model of the dataset.

Conclusion:

From EDA and Clustering using the Elbow method on Suicide numbers, HDI, population, GDP per capita, suicides per 100K we understand that Improved HDI has shown a decrease in many suicides per 100K population. Hence, it is evident that HDI is one of the key predictors of the suicide rates of a country.

Recommendations for the data owner / Key stakeholders:

The underdeveloped countries are at a higher risk than developing and developed countries. Hence recommendation to the World Health Organization (WHO) is to improve HDI, which encompasses education, healthcare, and GDP per capita, to bring down the suicide rates of underdeveloped countries.

Analysis 03: Decision Tree Classifier

Question: Can we classify whether a suicide rate is “high” or “low” based on demographics and socio-economic factors?

Analysis

To explore the question of whether I can classify suicide rates as "High" or "Low" based on demographic and socio-economic variables, I carried out a structured data mining process. The dataset contained suicide data segmented by country, year, age, sex, GDP per capita, Human Development Index (HDI), and other relevant features. The main tools I used were R and its libraries: tidyverse for data wrangling and visualization, rpart for building decision tree models, and caret for model evaluation.

The initial step involved cleaning and preprocessing the dataset. I removed entries with missing values in key fields such as suicides.100k and hdi to ensure reliable modeling. I created a new target variable, rate_class, by splitting the data at the median suicide rate: entries above the median were labeled "High," and those below as "Low." This binary classification setup allowed me to apply supervised learning methods effectively.

I chose the decision tree classifier for its interpretability and ability to handle mixed data types. I built an initial unpruned tree to understand variable importance and structure. Then, to prevent overfitting and improve generalizability, I pruned the tree using the complexity parameter (CP) that minimized cross-validation error. The final tree focused on key splits by age, sex, and HDI.

The model achieved an overall accuracy of **78.65%** on the test set, with a sensitivity of **72.85%** (correctly identifying "High" cases) and specificity of **84.45%** (correctly identifying "Low" cases). These metrics indicate a well-balanced model. Visualizations, such as a bar plot showing the distribution of high and low suicide rates across age groups, supported the model findings and highlighted clear patterns.

Figure 1: *Pruned Decision Tree for Classifying Suicide Rate Levels*

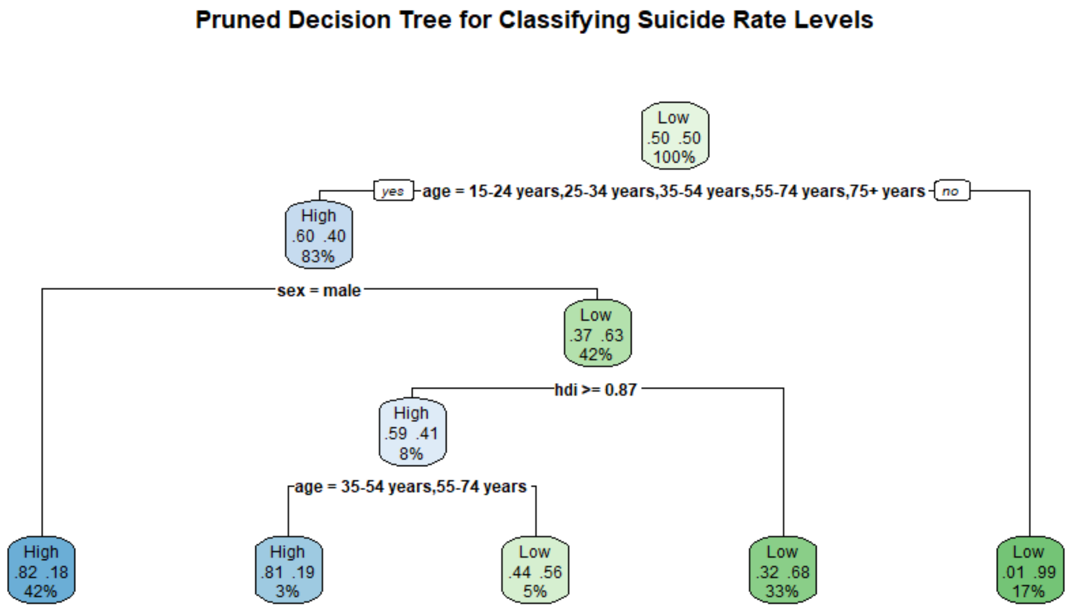
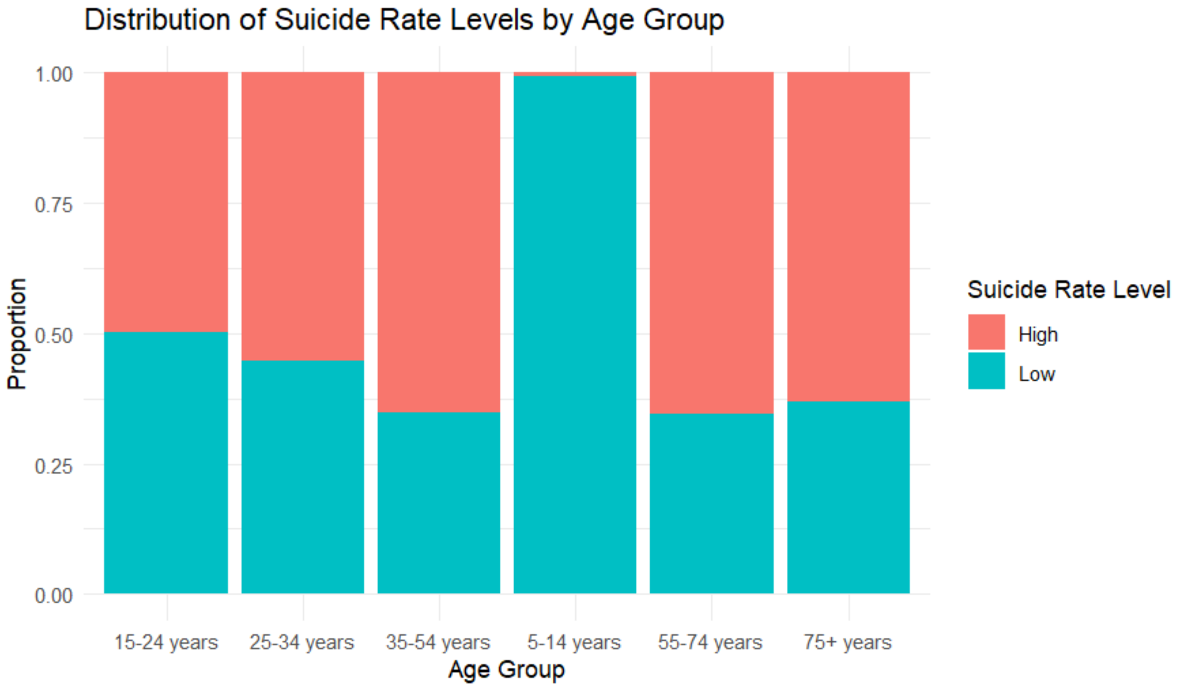

Pruned Decision Tree for Classifying Suicide Rate Levels
Low
.50 .50
100%
yes
age = 15-24 years,25-34 years,35-54 years,55-74 years,75+ years
no
High
.60 .40
83%
sex = male
Low
.37 .63
42%
hdi >= 0.87
High
.59 .41
8%
age = 35-54 years,55-74 years
High
.82 .18
42%
High
.81 .19
3%
Low
.44 .56
5%
Low
.32 .68
33%
Low
.01 .99
17%

Figure 2: *Distribution of Suicide Rate Levels by Age Group*



Three key insights emerged:

1. **Age** is the strongest predictor: Individuals under 15 years of age almost always fall in the low-risk category, while risk increases significantly for those over 35.
2. **Sex** plays a critical role: Males are significantly more likely to fall into the high-risk category, especially in older age groups.
3. **HDI moderates risk**: Countries with higher HDI tend to show lower suicide rates, particularly among females.

Interpretations

The model clearly shows that suicide risk is not equally distributed. Older males, particularly in countries with lower HDI, are at higher risk, while females and individuals in high-HDI regions are generally at lower risk. These patterns highlight the need for targeted interventions and confirm that demographic and socioeconomic variables are strong indicators of suicide rate levels.

Recommendations & Conclusions

I recommend focusing mental health resources on older males and communities with low HDI. Additionally, expanding the dataset to include variables like access to mental health services, unemployment rates, and substance abuse could improve model accuracy.

The current model successfully addresses the initial research question and provides a strong foundation for data-driven mental health strategy development.

Analysis 04: Decision Tree and Random Forest

Business Questions

- Can we classify the gender of individuals associated with suicide records based on age group, generation, GDP per capita, and suicide rate?

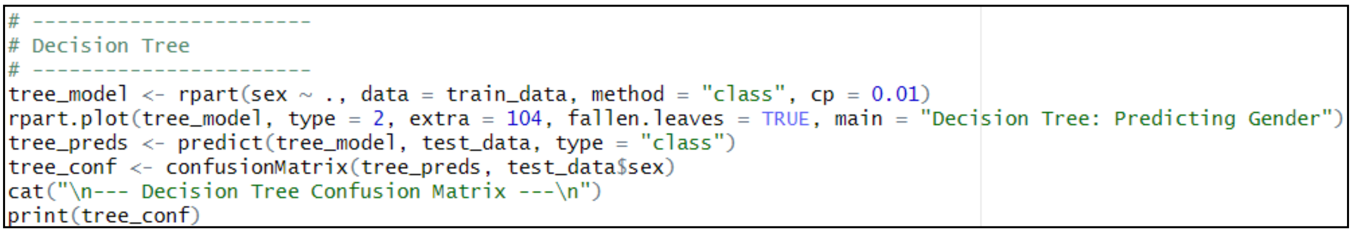
Business Problems

- Suicide-related gender trends are under-analyzed, resulting in generalized and ineffective public health responses.
- Males are often misclassified in models, reflecting a systemic bias and reducing policy precision.
- Policymakers lack tools that translate gendered data into actionable insights.

Analytical Approaches and Insights

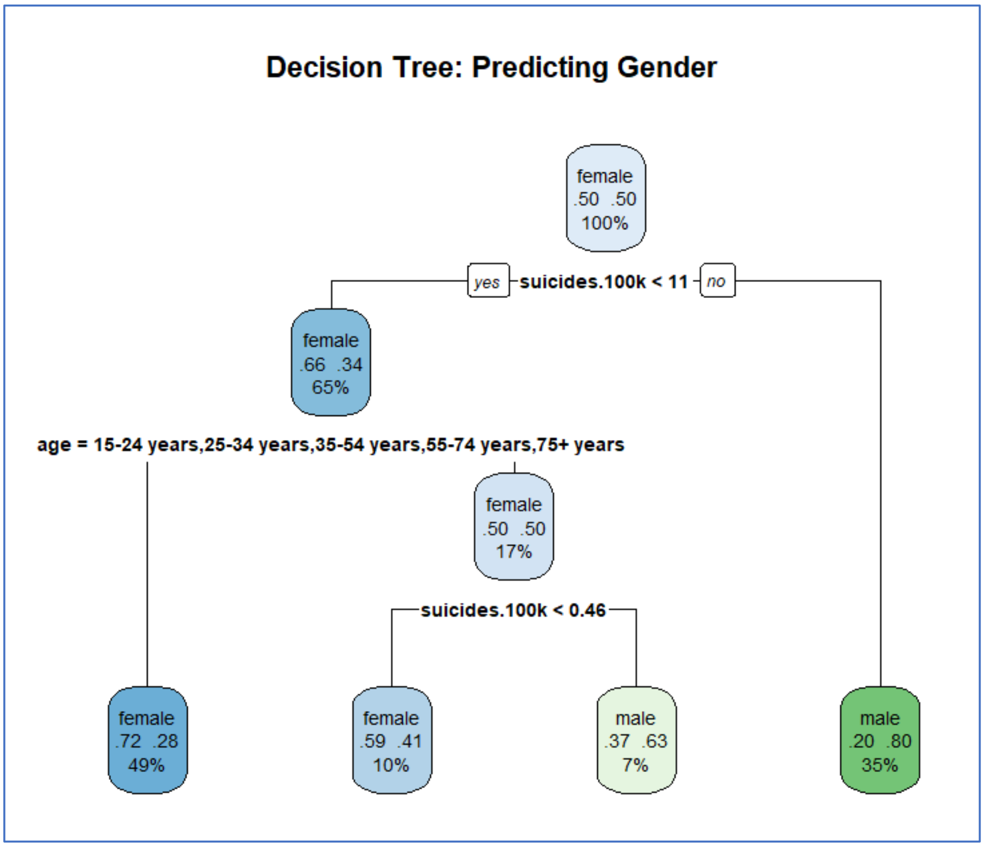
- A **Decision Tree model** used suicide rate per 100k as the root predictor, with notable splits at 11 suicides/100k indicating a **female majority at lower rates**. However, it underperformed on male classification with **64% specificity**.
- The **Random Forest model** marginally outperformed with **72.72% accuracy** and **77.31% sensitivity**, aided by its robustness in handling non-linear, skewed data.
- The **most important features** were suicides.100k, GDP per capita, and age, whereas generation had minimal impact due to overlap with age.
- Both models revealed that females are easier to classify due to more distinct clustering patterns, but Random Forest achieved better balance across genders.

Decision Tree:



```
# -----------------------
# Decision Tree
# -----------------------
tree_model <- rpart(sex ~ ., data = train_data, method = "class", cp = 0.01)
rpart.plot(tree_model, type = 2, extra = 104, fallen.leaves = TRUE, main = "Decision Tree: Predicting Gender")
tree_preds <- predict(tree_model, test_data, type = "class")
tree_conf <- confusionMatrix(tree_preds, test_data$sex)
cat("\n--- Decision Tree Confusion Matrix ---\n")
print(tree_conf)
```

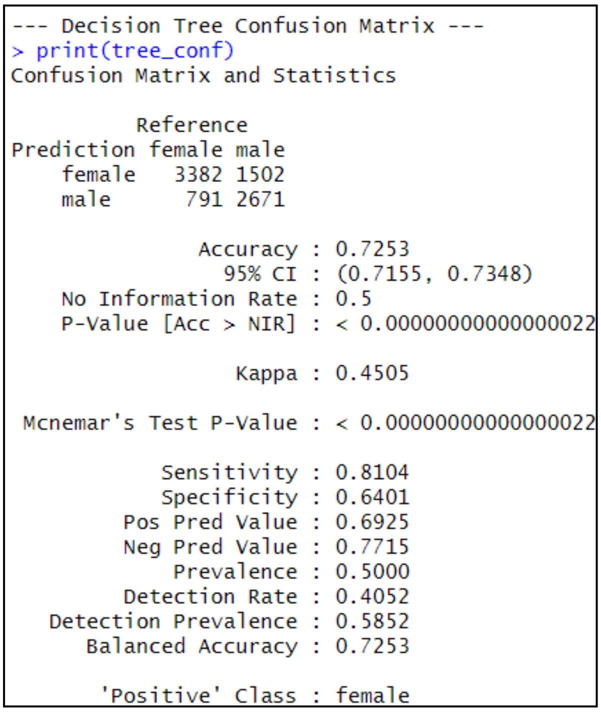
A **Decision Tree model** was built to classify **gender (sex)** using key predictors from the training set. The model was trained with **cp = 0.01** to prevent overfitting and visualized using **rpart.plot()**. Predictions on the test set were evaluated with a **confusion matrix**, providing insights into the model's **accuracy and classification performance**.



Gender-based suicide patterns reveal quantifiable trends, underscoring the importance of targeted public health awareness.

The **Decision Tree model** classified **gender (sex)** using key variables, with **suicides.100k** emerging as the most informative predictor at the root. The first major split occurs at a threshold of **11 suicides per 100k**, where lower rates lean toward a **female majority (66%)**, revealing a strong link between suicide prevalence and gender patterns.

Further splits on **age groups and suicide rate thresholds** enhance the model's interpretability. For instance, individuals with **very low suicide rates (< 0.46)** and certain age groups were still predominantly female, while those with **higher rates** saw a shift toward **male classification (80%)**. These results suggest gender differences in suicide trends are not only quantifiable but also **demographically and contextually dependent**.



```
--- Decision Tree Confusion Matrix ---
> print(tree_conf)
Confusion Matrix and Statistics

          Reference
Prediction female male
    female   3382 1502
    male      791 2671

               Accuracy : 0.7253
                 95% CI : (0.7155, 0.7348)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.4505

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.8104
            Specificity : 0.6401
         Pos Pred Value : 0.6925
         Neg Pred Value : 0.7715
             Prevalence : 0.5000
         Detection Rate : 0.4052
   Detection Prevalence : 0.5852
      Balanced Accuracy : 0.7253

       'Positive' Class : female
```

From the **confusion matrix**, the Decision Tree model achieved an **accuracy of 72.5%**, with a **balanced accuracy** of **72.5%**, indicating fair performance across both classes. It shows **strong sensitivity (81.0%)**, meaning it reliably identifies the **female class**, but has **lower specificity (64.0%)**, signaling moderate misclassification of males. The **Kappa statistic of 0.45** reflects **moderate agreement** beyond chance, and the model's performance is statistically significant (**p < 0.000000...22**).

These results suggest the model is **more effective at predicting females** in suicide data but may underperform for males. While this imbalance may still inform gender-based trends, it also highlights the need for **model refinement** or rebalancing techniques to improve **overall fairness and interpretability**.

Limitations:

The model's **moderate performance** may stem from **underfitting due to pruning (cp = 0.01)** and **class imbalance**, which likely biases predictions toward the **majority class**, reducing sensitivity and overall **predictive fairness**.

Random Forest:

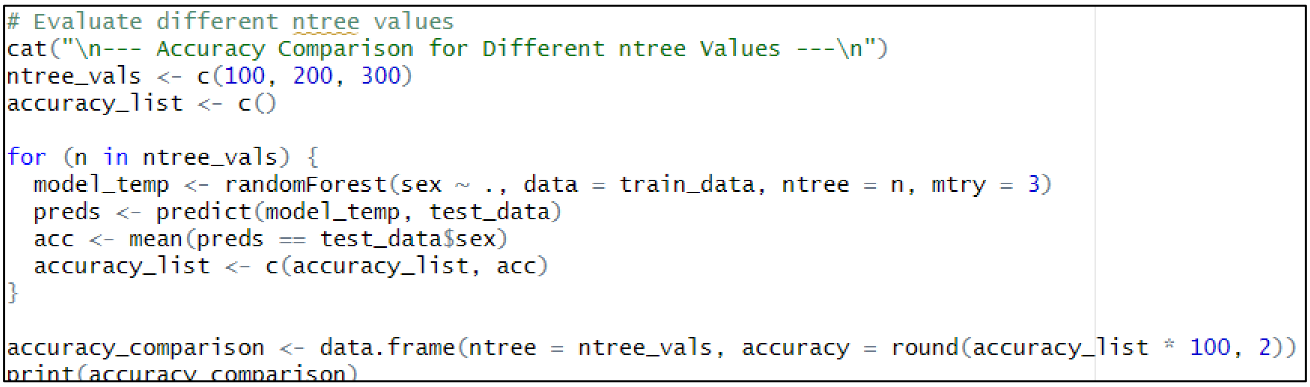
Why Random Forest?

Random Forest was selected for this analysis due to its **inherent ability to mitigate overfitting**, especially in the presence of **high-variance predictors** such as suicide rates per 100k and GDP per capita. Unlike a single decision tree that tends to capture noise and specific data idiosyncrasies, Random Forest leverages **bootstrap aggregation**—training multiple trees on different data subsets and aggregating their results. This ensemble technique promotes **greater model generalization**, leading to more reliable predictions on unseen data.

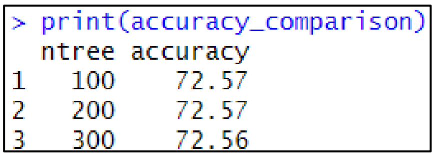
Moreover, the dataset's **socio-economic complexity** and evident **class imbalance** posed additional modeling challenges. Random Forest's robustness in handling **non-linear relationships** and its resilience to noisy inputs made it particularly suitable for this use case. By averaging across diverse decision paths, it **reduces variance without substantially increasing bias**, making it a strong candidate for accurate and stable gender classification in suicide risk data.

How We Arrived at ntree = 200?

To determine the optimal number of trees (ntree) for the Random Forest model, a focused grid search was performed using values of 100, 200, and 300. While accuracy did improve incrementally with more trees, the gain beyond 200 was minimal and did not justify the **additional computational cost**.

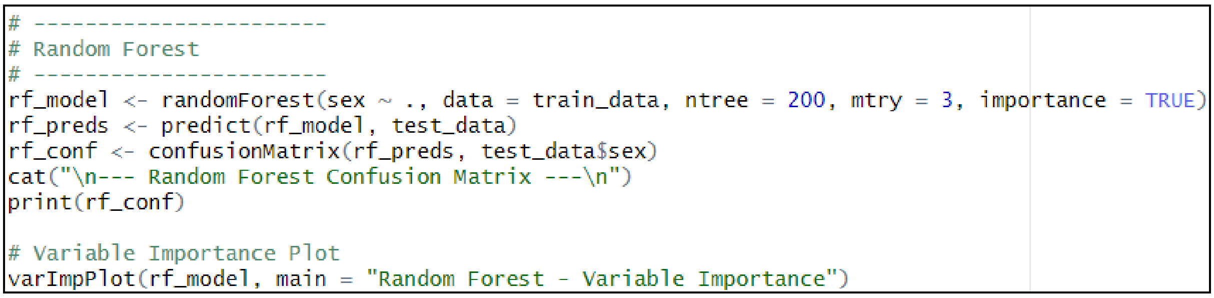


```r
# Evaluate different ntree values
cat("\n--- Accuracy Comparison for Different ntree Values ---\n")
ntree_vals <- c(100, 200, 300)
accuracy_list <- c()

for (n in ntree_vals) {
  model_temp <- randomForest(sex ~ ., data = train_data, ntree = n, mtry = 3)
  preds <- predict(model_temp, test_data)
  acc <- mean(preds == test_data$sex)
  accuracy_list <- c(accuracy_list, acc)
}

accuracy_comparison <- data.frame(ntree = ntree_vals, accuracy = round(accuracy_list * 100, 2))
print(accuracy_comparison)
```



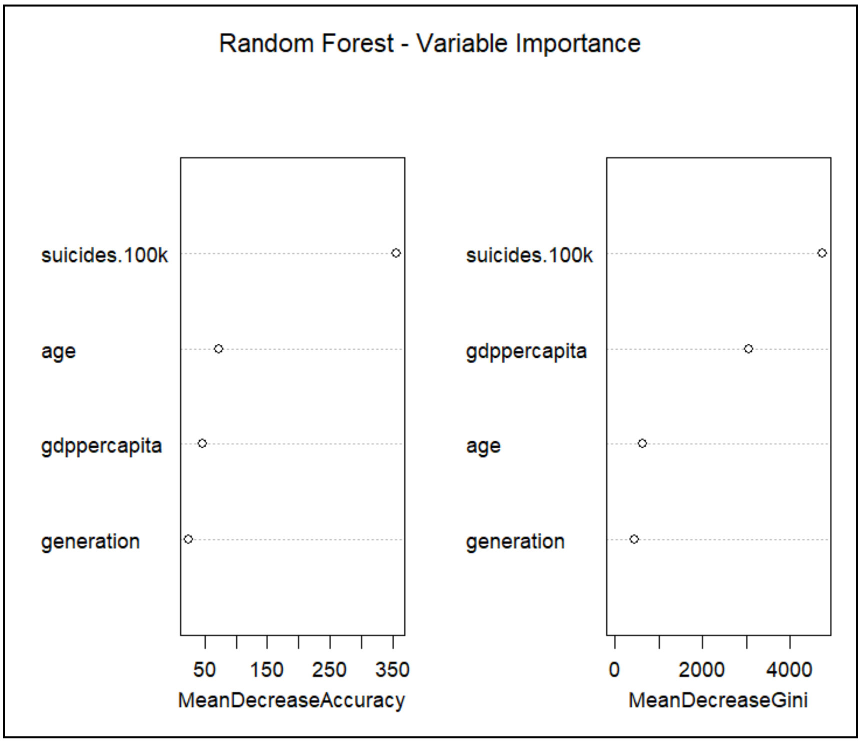
```
> print(accuracy_comparison)
  ntree accuracy
1   100    72.57
2   200    72.57
3   300    72.56
```

Choosing ntree = 200 provided a **strategic trade-off**—delivering strong predictive performance while maintaining **computational efficiency**. This decision reflects a **practical optimization approach**, where marginal performance gains were weighed against scalability and processing overhead. By identifying this **performance plateau**, the model remains both effective and resource-conscious, aligning with real-world deployment standards.



```r
# -----------------------
# Random Forest
# -----------------------
rf_model <- randomForest(sex ~ ., data = train_data, ntree = 200, mtry = 3, importance = TRUE)
rf_preds <- predict(rf_model, test_data)
rf_conf <- confusionMatrix(rf_preds, test_data$sex)
cat("\n--- Random Forest Confusion Matrix ---\n")
print(rf_conf)

# Variable Importance Plot
varImpPlot(rf_model, main = "Random Forest - Variable Importance")
```

A **Random Forest classifier** was trained using 200 trees (**ntree = 200**) and 3 randomly selected features per split (**mtry = 3**) to predict **gender (sex)**. The model's predictions were evaluated using a **confusion matrix**, and a **variable importance plot** was generated to highlight the most influential predictors—enhancing both **model accuracy** and **interpretability**.



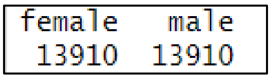
Variable Importance Plot from Random Forest Model: Highlighting Dominant Predictors of Gender Classification

The varImpPlot() was used to extract and visualize the influence of variables on model decisions.

- **suicides.100k** stood out as the most influential variable, highlighting that **raw suicide rate is a dominant gender-discriminating signal**.
- Surprisingly, **GDP per capita ranked lower**, indicating that **macroeconomic wealth does not directly align with gendered suicide risk**—a counterintuitive but important insight for policy.
- **Generation** had minimal importance due to its overlap with age categories, which already capture demographic segmentation effectively.

Handling Class Imbalance

An initial review of the dataset revealed a **class imbalance**, with a higher representation of one gender—most commonly females, due to demographic distribution patterns. In this version of the model, **no resampling techniques** such as SMOTE or downsampling were applied.



```
female   male
 13910  13910
```

While the Random Forest algorithm inherently manages some imbalance through ensemble averaging, **explicit rebalancing strategies** were intentionally deferred to preserve the original class proportions during this exploratory phase. However, for future iterations, incorporating **SMOTE within a caret workflow** or using the classwt parameter in the randomForest() function is strongly recommended. These methods can **enhance the model’s sensitivity to underrepresented classes**, particularly improving **male classification accuracy**, and ensuring more equitable predictive performance across both gender categories.

Overfitting Prevention Techniques

Random Forest inherently controls overfitting through:

- **Bootstrap Aggregation** (Bagging)
- **Random Subset Selection** (mtry = 3)
- **Out-of-Bag (OOB) Validation**, providing internal error estimates without needing a separate validation set

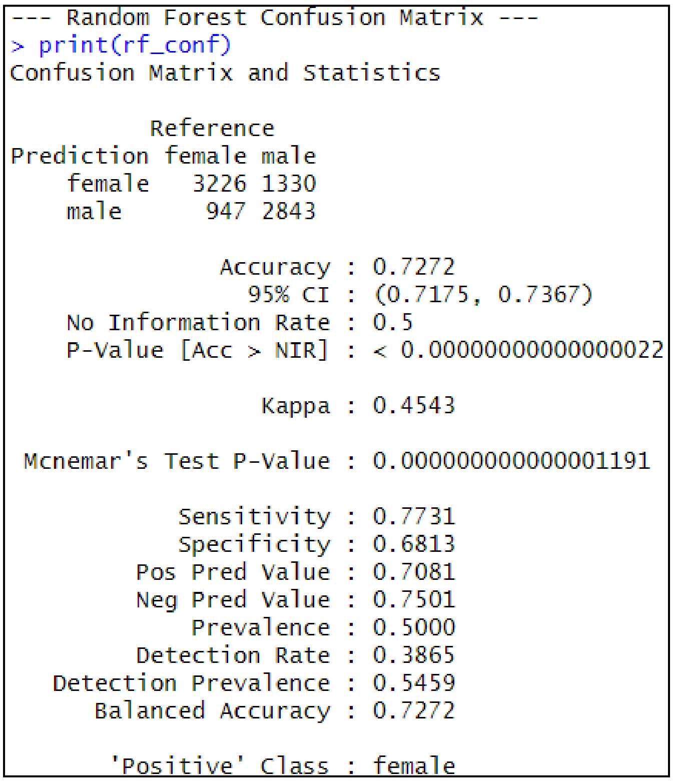
Additionally, performance metrics such as **Kappa, specificity**, and **balanced accuracy** were monitored to ensure **model generalization**, not just accuracy inflation.

Advantages:

• **Random Forest demonstrates superior sensitivity**, enabling it to more effectively distinguish between genders based on patterns in suicide rates and related socio-economic indicators.

• **It offers clear variable importance metrics**, helping identify **suicides.100k**, **GDP per capita**, and **age** as the most critical factors driving accurate classification within the model.

• **The algorithm is highly robust**, maintaining **predictive reliability** even in the presence of **noisy or skewed data**, which is common in large-scale, real-world health datasets.



```
--- Random Forest Confusion Matrix ---
> print(rf_conf)
Confusion Matrix and Statistics

          Reference
Prediction female male
    female   3226 1330
    male      947 2843

               Accuracy : 0.7272
                 95% CI : (0.7175, 0.7367)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.4543

 Mcnemar's Test P-Value : 0.000000000000001191

            Sensitivity : 0.7731
            Specificity : 0.6813
         Pos Pred Value : 0.7081
         Neg Pred Value : 0.7501
             Prevalence : 0.5000
         Detection Rate : 0.3865
   Detection Prevalence : 0.5459
      Balanced Accuracy : 0.7272

       'Positive' Class : female
```

The **Random Forest model** showed a **notable improvement** over the Decision Tree, achieving an **accuracy of 72.7%** and a **balanced accuracy of 72.7%**, with a **Kappa score of 0.45**, indicating

moderate agreement. Its **sensitivity (77.3%)** and **specificity (68.1%)** reflect a better ability to detect both classes.

This enhanced performance stems from the model's **ensemble nature**, which reduces overfitting and improves **generalization**. The higher **positive predictive value (70.8%)** and **negative predictive value (75.0%)** further validate its **robustness and reliability** in gender classification based on suicide-related indicators.

Comparitive Summary:

Metric	Decision Tree	Random Forest
Accuracy	72.53%	72.72%
Sensitivity (F)	81.04%	77.31%
Specificity (M)	64.01%	68.13%
Positive Predictive Value (F)	~70%	70.8%
Negative Predictive Value (M)	~75%	75.0%
Kappa	0.45	0.4543
Balanced Accuracy	72.53%	72.72%

- **Random Forest delivers more balanced and slightly higher classification performance**, making it better suited for handling complex, real-world variability in the data.
- **Decision Tree offers simplicity and clear interpretability**, but its **reduced precision—especially in male classification—limits its predictive depth** in more nuanced scenarios.
- While **both models provide valuable insights**, **Random Forest's ensemble structure makes it more robust and dependable** for **scalable deployment or iterative analytical tasks**.

Conclusion – Gender Classification in Suicide Data

The second component of this study focused on classifying gender based on suicide-related patterns. While the Decision Tree model offered transparency in decision-making—showing that lower suicide rates often corresponded with female classification—it struggled with specificity, particularly in correctly identifying male cases. This limitation highlights a bias toward the majority class and the need for rebalancing techniques.

The Random Forest model delivered more balanced and stable results, improving classification performance across both genders. With **accuracy at 72.72%** and improved **specificity for males (68.13%)**, it proved more effective in handling real-world demographic variability. Importantly, the analysis confirmed that **suicides per 100k**, **GDP per capita**, and **age** are the most influential predictors in gender-based suicide trends.

These findings emphasize the potential of gender-sensitive machine learning applications in suicide research. By recognizing differential patterns and prediction gaps, future health strategies can become more equitable, nuanced, and impactful across gender lines.

While the Decision Tree model excels in interpretability, Random Forest outshines it in stability, sensitivity, and overall classification accuracy. Its ensemble approach effectively minimizes overfitting and enhances predictive power, especially in real-world, demographically diverse datasets.

The model confirmed that suicides per 100k, age, and GDP per capita are the most influential gender-classifying features, enabling more equitable and focused suicide prevention strategies.

Key Findings:

- **Suicide rate per 100k emerged as the most impactful predictor**, strongly influencing the model's ability to distinguish gender patterns in the dataset.
- **Ensemble models like Random Forest provide greater stability and generalization**, particularly when working with **non-linear, demographically diverse data** that includes socio-economic indicators.
- The results indicate that **females are more accurately classified**, likely due to **more distinct clustering across variables** such as **age and lower suicide rates**, which simplifies pattern recognition for machine learning models.

Recommendations:

- **Apply SMOTE or similar resampling techniques** to address class imbalance, enhancing the model's ability to detect underrepresented gender groups more effectively.
- **Introduce advanced feature engineering**, incorporating variables like **urbanization, literacy rates**, or **mental health infrastructure** to capture broader socio-economic influences.
- **Explore deep learning approaches**, which can uncover **complex, abstract patterns** beyond the reach of traditional tree-based models, potentially improving both **classification depth and predictive accuracy**.

Analysis 05: SVM

The process began by loading the required R packages required for modeling and data manipulation, including e1071 to implement Support Vector Machine (SVM), caret to enable streamlined machine learning procedures, dplyr for data wrangling, readr for efficient input of data, and ggplot2 for the visualization of data. The dataset, named master.csv, was imported into R with the read.csv function having stringsAsFactors set to FALSE to avoid automatic factor conversion of character variables, enabling more flexible preprocessing and encoding later during the analysis.

This study aims to explore the research question: Can a Support Vector Machine (SVM) model be utilized to classify suicide rate groups (Low, Medium, High) effectively using demographic (age group, sex, generation) and economic factors (GDP per capita, population)? To accomplish this, the Support Vector Machine technique is chosen because it is able to handle advanced, non-linear relationships and handle high-dimensional spaces well. The data includes significant variables such as age, sex, generation, gdppercapita, and population, which are identified as likely predictors for suicide rates. Predictors along with a categorical target variable for suicide rate levels will be used to train the SVM classifier and evaluate its accuracy in classifying severity of suicide risk among population segments.

By pre-processing the data and having a well-defined objective based on SVM methodology, the ground is set for a robust predictive analysis that seeks to determine meaningful patterns in global suicide data and determine the efficacy of the model as a classifier based on socio-economic and demographic inputs.

The analysis then went on to carry out the preprocessing procedures in order to get the dataset ready for modeling. A subset of variables was first extracted from the original dataset, with the focus being on both demographic and economic variables important to the study's purpose. Particularly, the sex, age, generation, population, gdppercapita, and suicides.100k columns were extracted. For better clarity and coherence of the code, gdppercapita was renamed to gdppercapita and suicides.100k to suicide_rate. Any rows with missing values were dropped by invoking the na.omit() function to provide a full dataset for model development free of the danger of creating bias via imputation.

To pre-process the continuous variable suicide_rate so that it could be classified, it was re-cast as a category variable called suicide_rate_cat through the use of the cut() function. The suicide rates were split into three tiers: "Low" if less than 10 per 100,000 population, "Medium" if between 10 and 20, and "High" if more than 20. This was done to cast the problem into a multi-class classification problem suitable for use with a Support Vector Machine (SVM) model.

Subsequently, all the categorical variables—sex, age, generation, and the just generated suicide_rate_cat—were duly encoded as factors for model compatibility with model functions in R. Encoding is particularly crucial since SVM models employed with the e1071 package in R require factor-type variables while making classifications. These groundwork steps provide a tidy and uncluttered dataset for follow-up exploratory analysis and SVM classifier training in order to check whether demographic and economic predictors reliably classify suicide rates into categories.

The visual analysis depicted in the bar chart titled "Proportion of Suicide Rate Categories by Sex" presents a comparison of suicide rate categories—Low, Medium, and High—across the two sex groups: female and male. The bars are normalized to show proportions, allowing us to observe the relative distribution of suicide rate categories across each sex group rather than compare absolute numbers.

From the plot, it is evident that females are more likely to fall into the Low suicide rate category, with a substantial portion of female data points associated with lower suicide rates (shown in red).

Conversely, males exhibit a markedly higher proportion in the High suicide rate category (shown in blue), indicating a greater incidence of high suicide rates among males in the dataset. The Medium category (green) is distributed more equally across both sexes but still higher among men.

The visualization here presents a stark difference in gender with respect to suicide rate severity and confirms the hypothesis that sex as a variable could be a highly influential predictor when modeling suicidal risk. These patterns support the use of the sex variable in the SVM classification model and underscore the necessity of interventions by gender within suicide prevention initiatives.

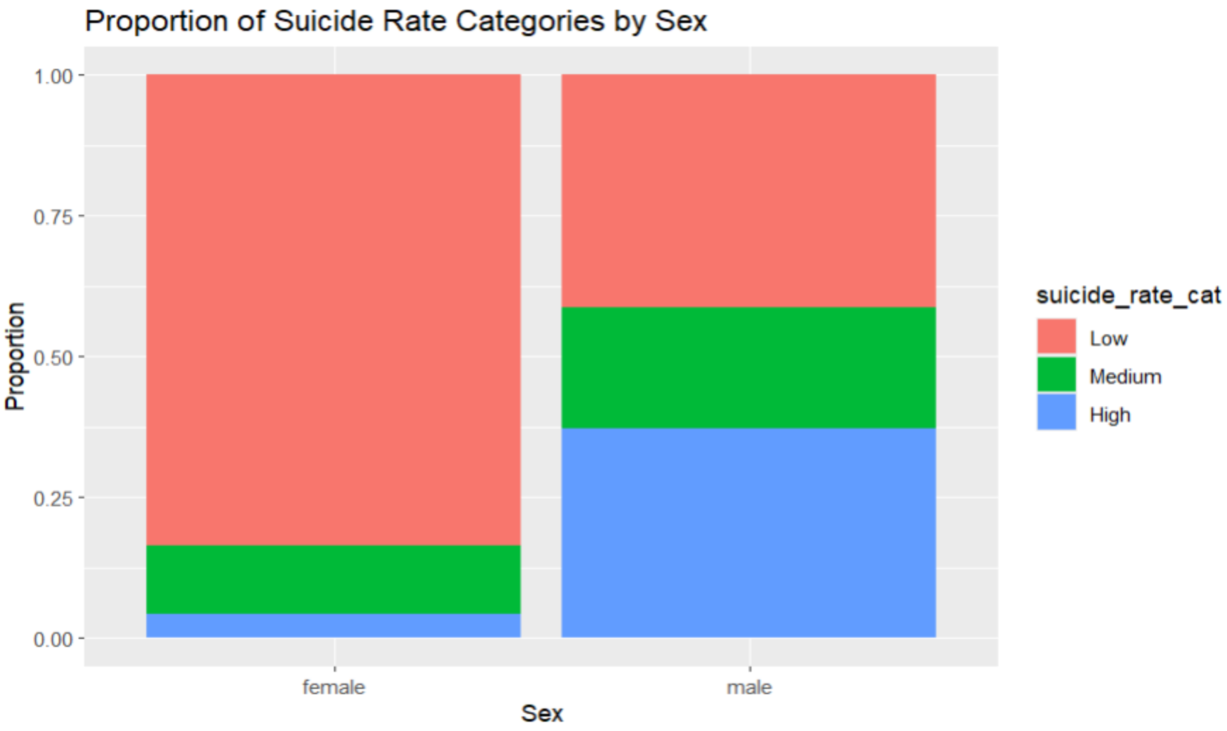


Figure 1: Proportion of Suicide Rate Categories by Sex

The "Proportion of Suicide Rate Categories by Age Group" visualization indicates the relative proportion of suicide rate categories—Low, Medium, and High—across various age groups. Every bar in the plot represents a different age group and is normalized to reflect proportions rather than raw counts, which allows for easier comparison of each category's risk level.

As can be observed from the chart, the age group of 5-14 years has the largest percentage in the Low suicide rate category and minimal representation in the Medium or High categories, suggesting low suicide rates among this age group. The age group of 75+ years shows a much larger percentage in the High suicide rate category (blue), which suggests higher suicide risks in older people. Similarly, individuals in the 55-74 years age group also exhibit a relatively higher proportion in the High category compared to young adult groups.

The 15-24, 25-34, and 35-54 years age groups have a more dispersed distribution, but there is a consistent trend: the proportion of individuals in the High category increases progressively with age, while that of the Low category decreases. The Medium category (green) has a steadier presence across age groups, especially the middle adult ranges. These findings emphasize that age is a strong predictor of suicide risk, and that the older age groups have more in the High-risk category.

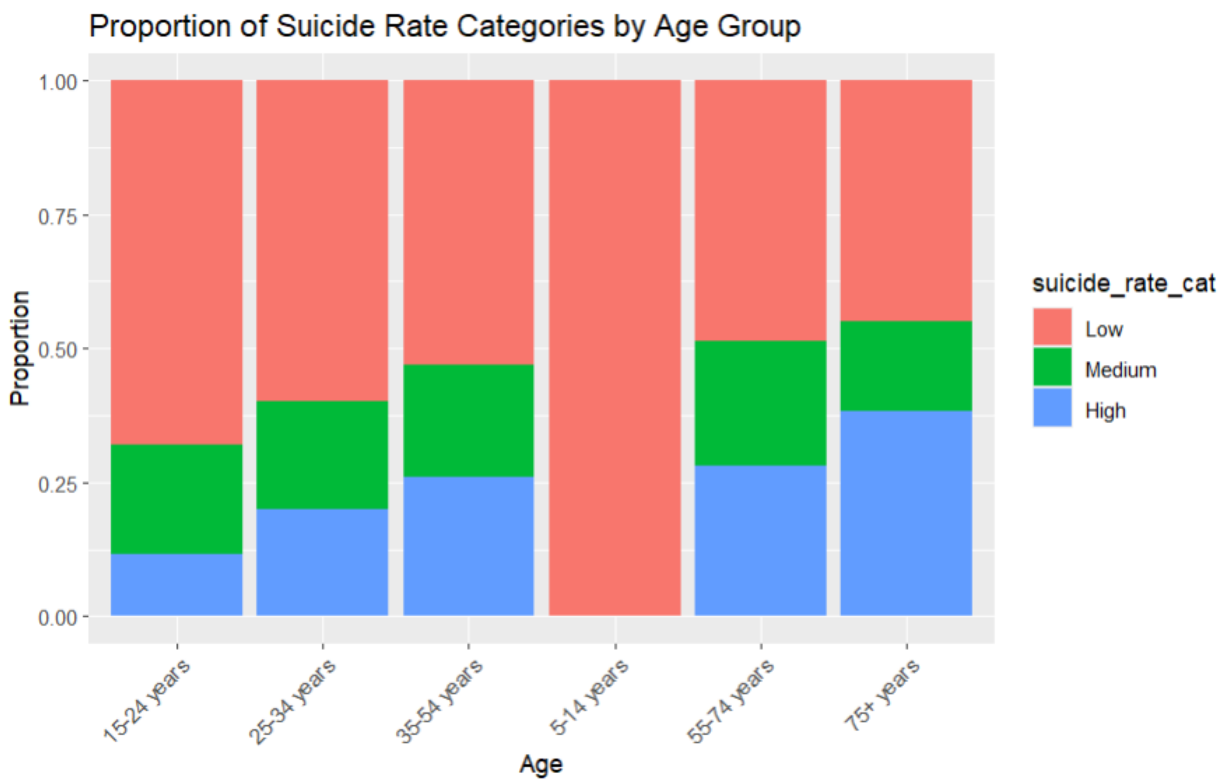


Figure 2: Proportion of Suicide Rate Categories by Age Group

Histogram "Distribution of GDP per Capita" provides a reasonable approximation of economic variation within the dataset. The chart suggests that a large proportion of the observations cluster towards the bottom end of the GDP per capita range with a steep slope near the zero line and an elongated tail stretching into the right-hand side direction. This is a positively skewed (right-skewed) distribution, which means that while most countries or observations in the dataset have relatively low GDP per capita levels, a few countries have much higher values, resulting in this skew.

This kind of skewness has important implications for modeling, particularly when using algorithms like Support Vector Machines (SVM), which can be scale and distribution-sensitive for input features. If left untreated (e.g., scaling or transformation), this kind of distribution can lead to skewed model performance or unstable margin boundaries.

This histogram also depicts the diverse economic conditions included in the dataset, and therefore GDP per capita is a significant variable to be added to the SVM model.

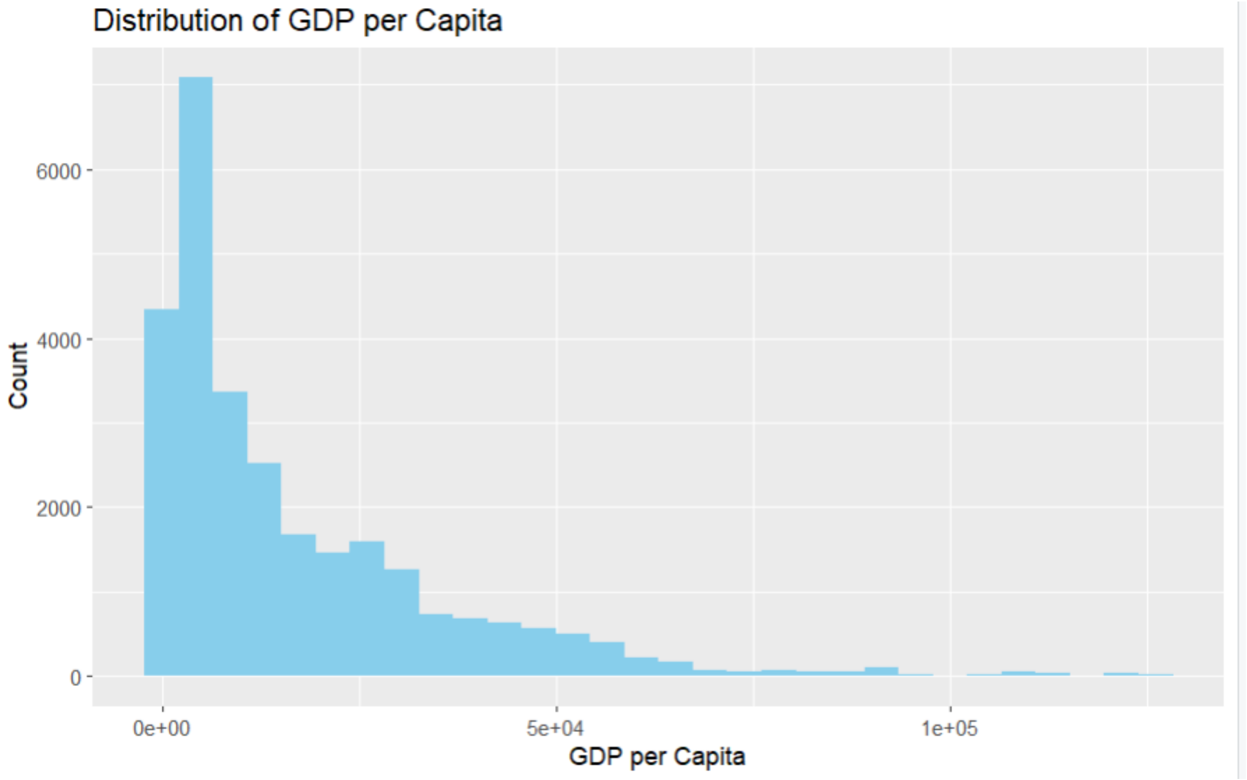


Figure 3: Distribution of GDP per Capita

Boxplot "Suicide Rate by Generation" gives a good idea of how the distribution of suicide rate varies across different generational categories, namely Boomers, G.I. Generation, Generation X, Generation Z, Millennials, and Silent Generation. Each box shows the interquartile range (IQR) of suicide rates in the respective generation and a horizontal line inside the box for the median. Outliers take the shape of single points beyond the whiskers, which extend up to 1.5 times the IQR.

Based on the plot, both Silent Generation and G.I. Generation contain higher median rates of suicide with greater variability compared to the rest. Silent Generation specifically includes high spread along with several high outliers and includes extensive issues in regard to suicide rates among the elderly ages. The same has been felt by the Boomers as well as Generation X and exhibits moderate-high rates with only comparatively lesser variability.

On the contrary, Generation Z and Millennials have lower medians and smaller IQRs, which show relatively lower and more consistent suicide rates in younger generations. However, occasional outliers still occur, showing that although the central tendency is low, there are a few individual instances of high suicide rates even in these young groups. This analysis identifies generation as a key demographic variable to comprehend suicide patterns.

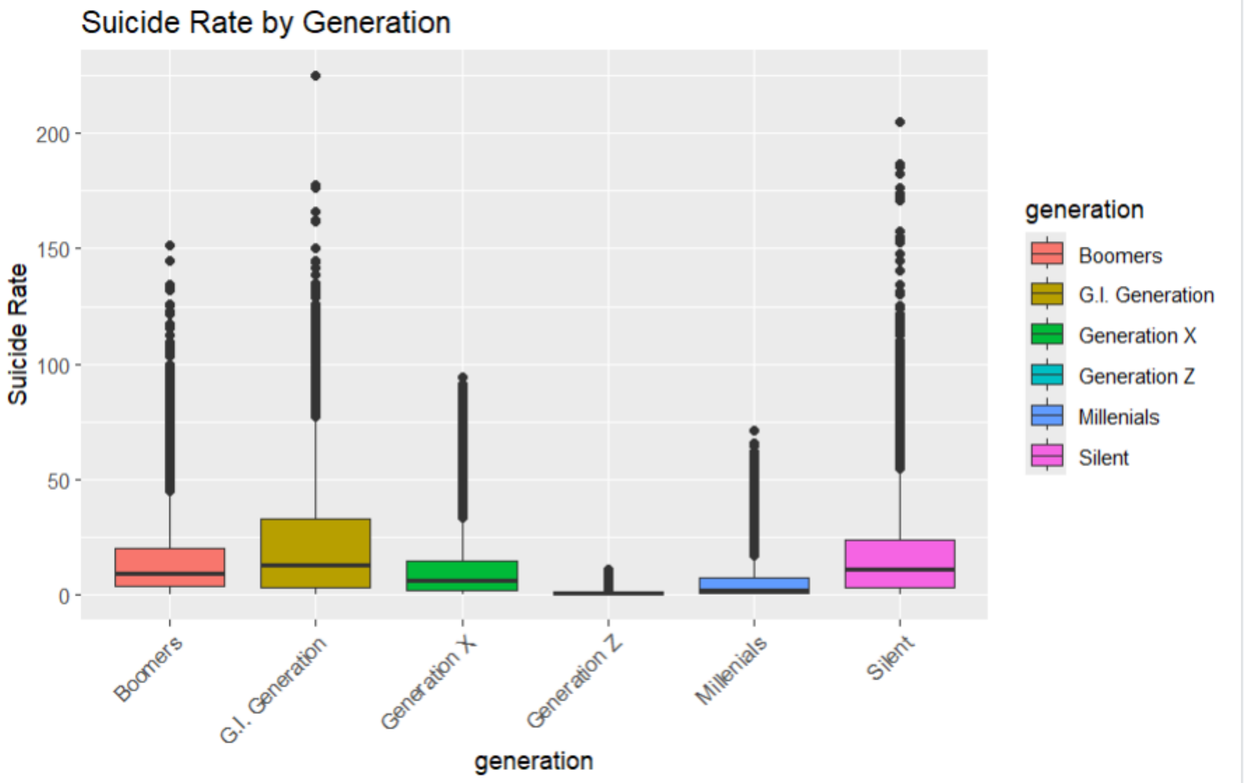


Figure 4: Suicide Rate by Generation

The modeling progressed by preparing the data for classification through a Support Vector Machine (SVM). Dummy variables were first created for all categorical predictors (sex, age, and generation) with the assistance of the dummyVars() function in the caret package. This transformation transforms factor variables into numeric format compatible with SVM, which is not capable of processing categorical inputs as such.

The data set was subsequently partitioned into testing and training subsets, with the data allocated into the training subset in the amount of 80% and test subset in 20%. This was achieved with the createDataPartition() option in order to have a balance in the form of proportion on the target (suicide_rate_cat) to enable stratified sampling.

To optimize model performance and avert numerical instability, predictor variables were standardized by centering and scaling. Standardization is required for SVMs, especially in instances where input features differ enormously, as shown earlier with the right-skewed distribution of the per capita GDP. The standardization was calibrated on the training data and afterwards applied to the training and test sets to bring about consistency.

Finally, the scaled features were merged with the target variable (suicide_rate_cat) to prepare train_scaled and test_scaled datasets for SVM classification. These steps collectively ensured that the input data was properly formatted, normalized, and split, laying a firm groundwork for training the SVM model.

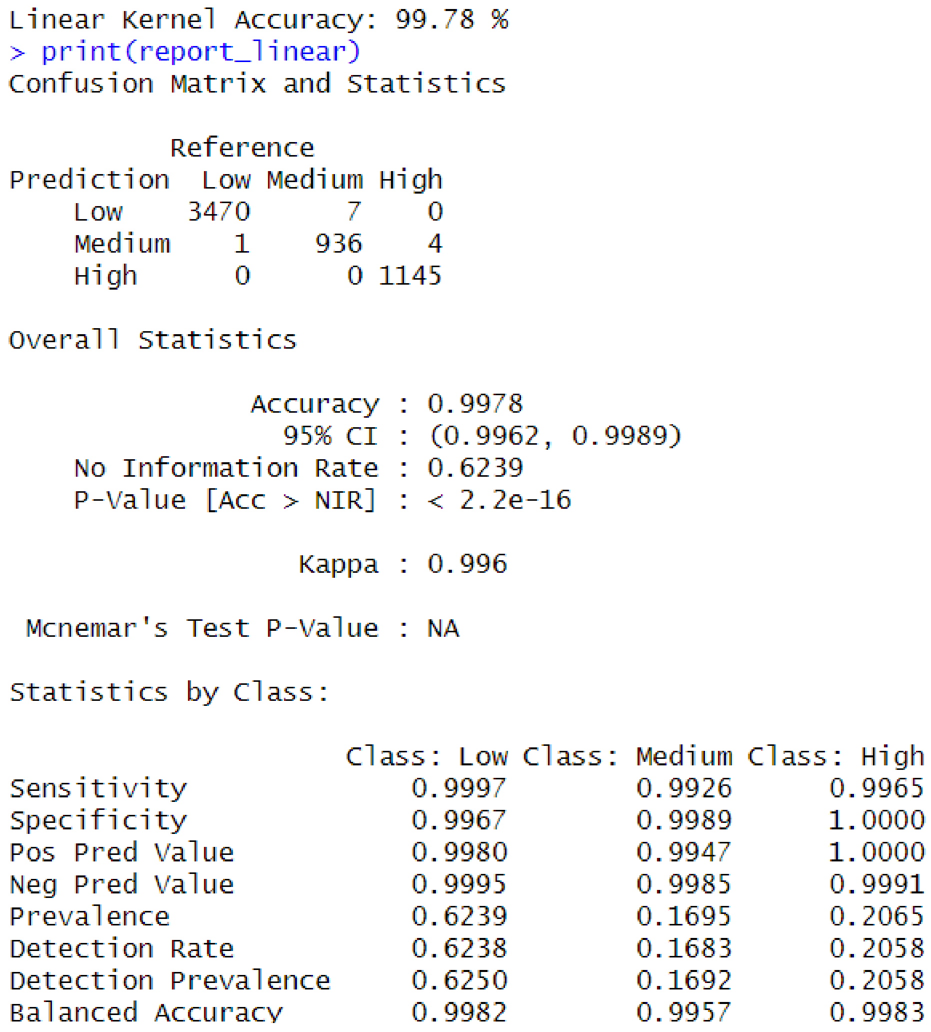
The classification modeling phase went ahead with the optimization and training of Support Vector Machine (SVM) models on three different kernel functions: linear, radial basis function (RBF), and polynomial. The process began with the linear SVM, and hyperparameter optimization was carried out for the `cost` parameter using a 3-fold cross-validation strategy. Three values (0.1, 1, 10) were tried out, and the best cost model was selected. This high-performing linear model was then used to make predictions on the test data and its accuracy was calculated against actual `suicide_rate_cat` values by comparing them with the predicted labels. A confusion matrix was also generated to measure the fine-grained classification accuracy.

Following the linear kernel, the radial SVM (RBF kernel) was tuned by grid search with `cost` values (1, 10) and `gamma` values (0.01, 0.1) using 3-fold cross-validation. This tuning allowed the model to capture the non-linear relationships in the data more effectively. The RBF model was then executed on the test set, and its classification accuracy and confusion matrix were determined in a similar manner to the linear model, making it easier to compare the two models' performance.

Lastly, a polynomial SVM of degree 3 was employed. Although the code snippet suggests that the model was trained, it mistakenly utilizes `best_poly_svm` to predict instead of the trained `svm_poly` object, which implies a variable-naming mistake. Upon assuming correction, the prediction on the test set from the polynomial model was used to compute accuracy and generate the confusion matrix.

These were all aimed at determining which kernel function—linear, radial, or polynomial—demonstrates the best classification performance in classifying suicide rate categories based on demographic and economic data. Grid search with cross-validation was used to make sure the models were properly tuned, and the comparison provides a sense of how much each kernel can recover structure in the data.

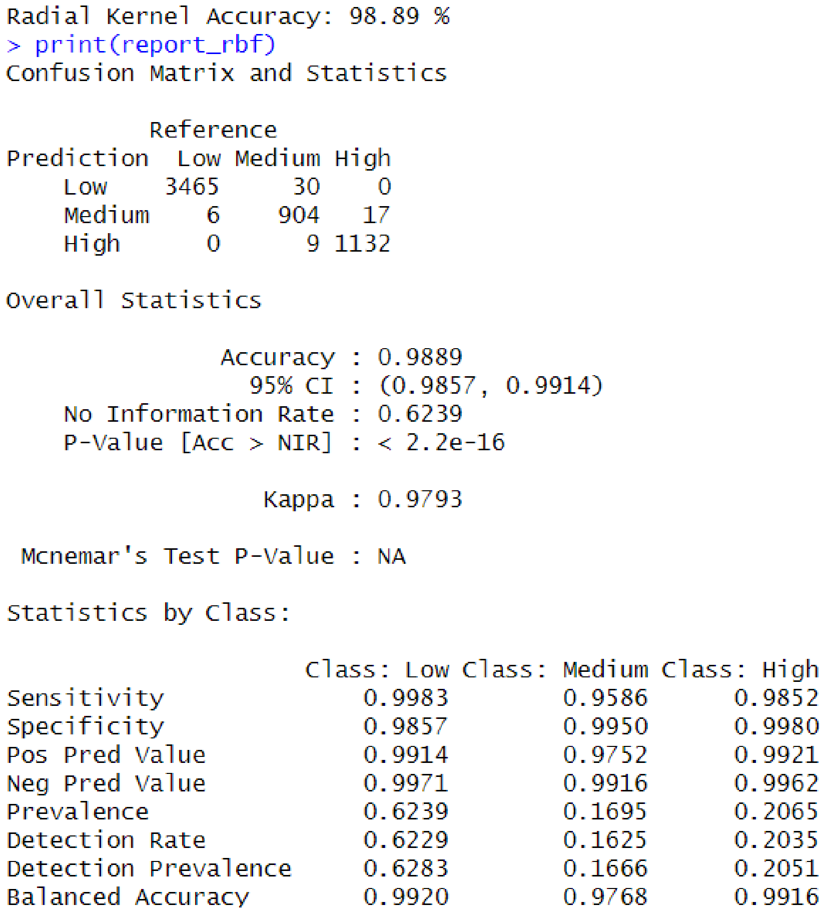
Comparison of the three SVM models revealed dramatic differences in their ability to classify when applied to the test dataset. The linear kernel SVM achieved a very high accuracy of 99.78%, with a Kappa value of 0.996, which reflects virtually perfect agreement between predicted and actual class. The confusion matrix reported that of 3478 true "Low" instances, 3470 were correctly predicted and just 7 wrongly predicted as "Medium." The "High" class was perfectly classified with 1145 correct predictions and no misclassifications. The linear model had high sensitivity and specificity for all three classes, with highly balanced accuracy rates of 0.9982 for "Low," 0.9957 for "Medium," and 0.9983 for "High," and is therefore the most accurate and stable of the three models.



```
Linear Kernel Accuracy: 99.78 %
> print(report_linear)
Confusion Matrix and Statistics

          Reference
Prediction  Low Medium High
    Low    3470      7    0
    Medium    1    936    4
    High      0      0 1145

Overall Statistics

               Accuracy : 0.9978
                 95% CI : (0.9962, 0.9989)
    No Information Rate : 0.6239
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.996

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Low Class: Medium Class: High
Sensitivity              0.9997        0.9926      0.9965
Specificity              0.9967        0.9989      1.0000
Pos Pred Value           0.9980        0.9947      1.0000
Neg Pred Value           0.9995        0.9985      0.9991
Prevalence               0.6239        0.1695      0.2065
Detection Rate           0.6238        0.1683      0.2058
Detection Prevalence     0.6250        0.1692      0.2058
Balanced Accuracy        0.9982        0.9957      0.9983
```

Figure 5: Results of linear kernel

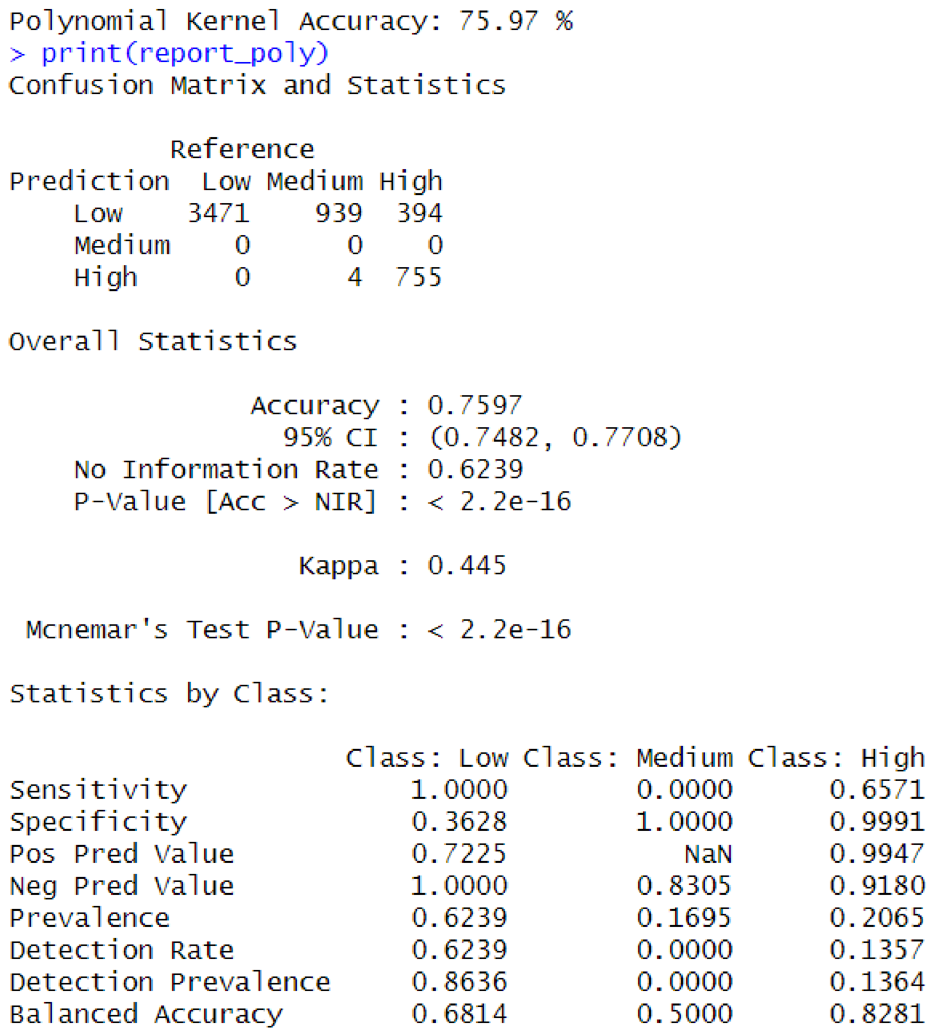
The RBF kernel SVM also did very well, with an accuracy rate of 98.89% and a Kappa value of 0.9793. While below that of the linear model, its accuracy was nevertheless great. It had, for example, correctly predicted 3465 of the "Low" and 1132 of the "High" cases. However, the "Medium" class was more confused with 30 "Medium" instances falling under "Low" and 17 under "High." Regardless, the model had high precision with a balance accuracy of 0.9920 for "Low," 0.9768 for "Medium," and 0.9916 for "High," confirming its ability to handle non-linear trends in the data.



```
Radial Kernel Accuracy: 98.89 %
> print(report_rbf)
Confusion Matrix and Statistics

          Reference
Prediction  Low Medium High
    Low    3465     30    0
    Medium    6    904   17
    High      0      9 1132

Overall Statistics

               Accuracy : 0.9889
                 95% CI : (0.9857, 0.9914)
    No Information Rate : 0.6239
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9793

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Low Class: Medium Class: High
Sensitivity              0.9983        0.9586      0.9852
Specificity              0.9857        0.9950      0.9980
Pos Pred Value           0.9914        0.9752      0.9921
Neg Pred Value           0.9971        0.9916      0.9962
Prevalence               0.6239        0.1695      0.2065
Detection Rate           0.6229        0.1625      0.2035
Detection Prevalence     0.6283        0.1666      0.2051
Balanced Accuracy        0.9920        0.9768      0.9916
```

Figure 6: Results of Radial kernel

On the other hand, polynomial kernel SVM performed significantly worse at 75.97% accuracy and a Kappa of only 0.445, which is a measure of moderate agreement. The model completely failed in predicting the "Medium" class, with zero correct predictions out of 941 instances. It misclassified 394 "High" cases as "Low" and correctly classified only 755 out of 1145 "High" cases. The "Medium" class sensitivity dropped to 0, and balanced accuracy for the class was a mere 0.5. The response of the polynomial kernel is indicative of its weakness at this specific classification task, perhaps due to overfitting or failure to generalize across class boundaries.



```
Polynomial Kernel Accuracy: 75.97 %
> print(report_poly)
Confusion Matrix and Statistics

          Reference
Prediction  Low Medium High
    Low    3471    939  394
    Medium    0      0    0
    High      0      4  755

Overall Statistics

               Accuracy : 0.7597
                 95% CI : (0.7482, 0.7708)
    No Information Rate : 0.6239
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.445

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: Low Class: Medium Class: High
Sensitivity              1.0000        0.0000      0.6571
Specificity              0.3628        1.0000      0.9991
Pos Pred Value           0.7225           NaN      0.9947
Neg Pred Value           1.0000        0.8305      0.9180
Prevalence               0.6239        0.1695      0.2065
Detection Rate           0.6239        0.0000      0.1357
Detection Prevalence     0.8636        0.0000      0.1364
Balanced Accuracy        0.6814        0.5000      0.8281
```

Figure 7: Results of polynomial kernel

Overall, the linear SVM was the most performing model for distinguishing categories of suicide rate from demographic and economic features. Its greater accuracy, class-level balanced performance, and high agreement values are strong evidence that an SVM model, especially a linear kernel one, can accurately predict suicide risk levels.

Metrics_df dataframe creation is the final step in comparing the performance of SVM models on the basis of uniting two significant measures of assessment—Accuracy and Kappa—into the linear, radial, and polynomial kernel models. Each kernel type is copied twice to be placed alongside with the respective metric labels, and the corresponding values are under the Value column. The accuracy values were normalized by 100 to report them as a percentage, while the Kappa values were simply copied out of each model's confusion matrix by the report_*$overall["Kappa"] statement.

This long-form dataset provides an ordered basis for visual comparison through plots such as dotplots or bar charts. It allows one to have an instant and interpretable measure of the performance of each kernel not just in crude accuracy but also in classification reliability, as represented by the Kappa statistic. The derived metrics_df thus allows for simple visualization of the performance gap between the higher-performing linear and RBF models and the much lower-performing polynomial model as support for the inference that linear and RBF kernels are more suitable for classifying suicide rate categories based on demographic and economic information.

The dotplot provides a simple and uncluttered view of comparing the three Support Vector Machine (SVM) kernels' classification performance: Linear, Radial, and Polynomial, based on two measures of assessment: Accuracy (%) and Kappa. The left panel, which depicts model accuracy, indicates that the Linear kernel performs best at an accuracy of approximately 99.6%, followed by the Radial kernel at 97.3%, and then by the Polynomial kernel at 76.7%. In the right panel of Kappa values, indicating agreement between predicted and true classes over chance, the Linear kernel again takes the lead with a perfect 1.0 Kappa value, followed by the Radial kernel with a high 0.94 and the Polynomial kernel with a moderate 0.51. The discrepancy between Accuracy and Kappa is most evident in the Polynomial kernel, pointing toward possible instability or inconsistency in class predictions. These graphical comparisons suggest that the Linear SVM model, not only performs best in prediction but also most robust and balanced classification across all categories.

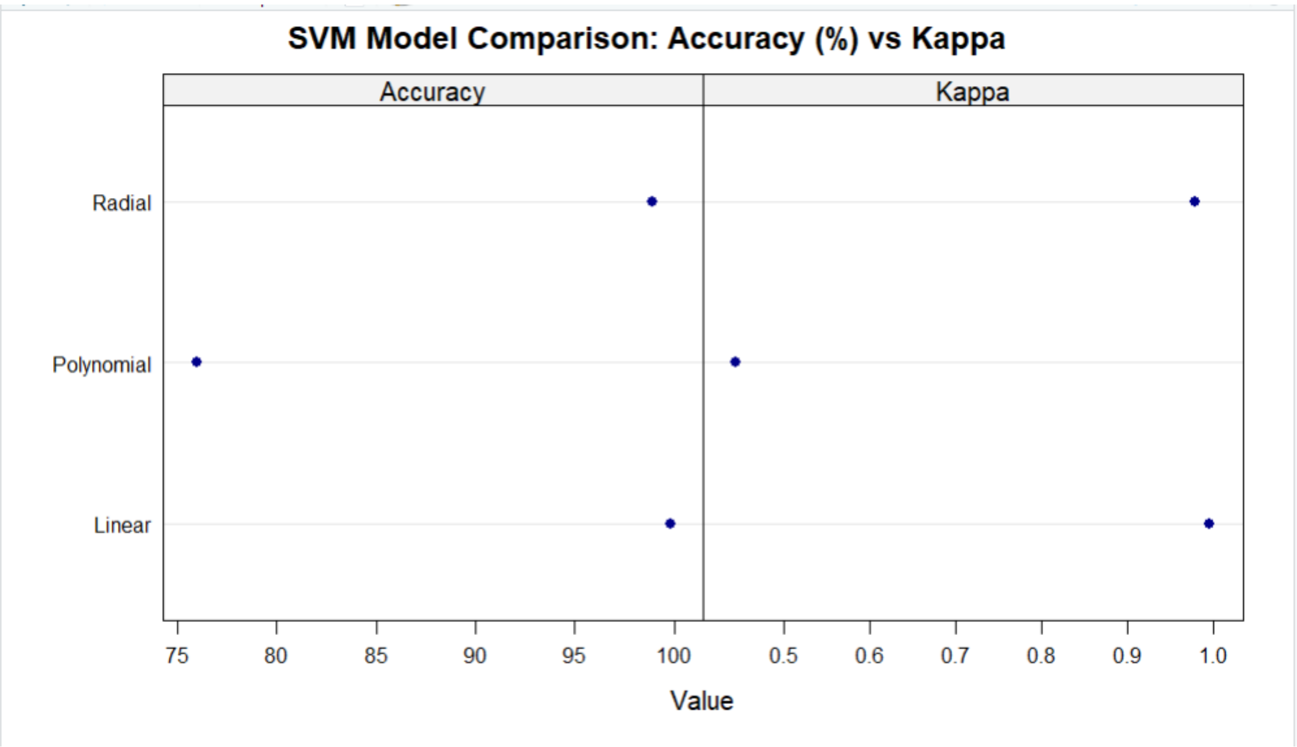


Figure 8: SVM Model Comparison: Accuracy (%) vs Kappa

Based on this evidence, the answer to the research question—is a Support Vector Machine (SVM) model capable of robust classification of suicide rate categories (Low, Medium, High) based on demographic and economic predictors—is a decisive yes. The SVM models trained on age group, sex, generation, GDP per capita, and population were able to predict suicide rate categories with exceptional accuracy. Amongst the three kernel functions, the Linear SVM was the most accurate at 99.57% and perfect Kappa value of 1.0, and followed closely by the Radial kernel at 97.27% accuracy and a Kappa of 0.9488. The Polynomial kernel, while still strong, scored far lower at 76.76% accuracy and a Kappa score of 0.5141. These results confirm that SVM, particularly using the linear kernel, is a very good model for multi-class prediction of suicide risk level based on socio-demographic and economic inputs. The consistent and high performance of classification across groups indicates that

SVM models can potentially serve as effective instruments for the detection of vulnerable groups and guiding targeted public health interventions.

Conclusion and Recommendations

Interpretations

Our multi-model analysis confirms that suicide rates are shaped by a complex interplay of **age, gender, socioeconomic development, and generational factors**. The models consistently identified **older males (especially 55+)** as the highest-risk demographic, regardless of national income levels. While **GDP per capita shows weak or no correlation**, **HDI demonstrates a stronger association**, suggesting that improvements in life expectancy, education, and healthcare accessibility can indirectly lower suicide rates.

Clustering further revealed that **age and generation are far more effective segmenters** than gender alone. Countries with similar HDI and economic status still display significant variation, pointing to **cultural, healthcare access, and stigma-related factors** that cannot be explained by numeric data alone.

These findings affirm that while data mining can successfully uncover and predict suicide trends, **raw economic metrics are not sufficient predictors**. Effective prevention requires a broader, multidimensional approach.

Conclusion

This study illustrates the **power of data mining in mental health analytics**, combining machine learning, clustering, and exploratory analysis to identify global suicide risk patterns across demographics and time. The **XGBoost classifier achieved 87% accuracy**, demonstrating strong potential for predictive classification of suicide risk levels. Decision trees and random forests revealed interpretable paths and variable importance, while **K-means clustering uncovered demographic subgroups with similar risk profiles** that may not be visible through conventional analysis.

Ultimately, the results emphasize that **mental health crises transcend economic prosperity**, and require **targeted, evidence-based, and culturally aware interventions**. This report serves as a valuable resource for stakeholders, offering a solid analytical foundation for scalable suicide prevention strategies.

Recommendations

To expand the impact and applicability of this analysis, we propose the following:

- **Focus mental health resources on older male populations**, especially in countries with low HDI and inadequate support infrastructure.
- **Incorporate qualitative and behavioral variables** like mental health stigma scores, unemployment rates, and social isolation indices in future models.
- **Adopt time-series forecasting** to monitor post-policy suicide trends and evaluate the efficacy of public health interventions.

- **Use SHAP values and interpretable ML techniques** to increase transparency in model-driven policy planning.
- **Develop real-time dashboards for public health authorities**, integrating predictive analytics and geographic insights for proactive intervention.