Machine Learning Project
January 2026 Exam session

Global Suicide Rates (1985-2016): Analysis and Prediction

Team 08: Abdul Sami, Davina David

Abstract

Suicide rate statistics offer valuable insights into global public health trends across countries and demographic groups. This report analyzes a multi-country dataset covering the period 1985–2016, comprising demographic, temporal, and socio-economic attributes such as age group, gender, population, GDP, and Human Development Index (HDI).

The objective is to predict suicide rates per 100,000 population and to examine how risk varies across age groups and socio-economic conditions using supervised machine learning models implemented in KNIME. An XGBoost regression model is utilized to estimate continuous suicide rate values, while a J48 decision tree classifier is applied to categorize suicide risk into predefined classes (Low, Medium, High).

The results demonstrate high predictive accuracy and provide interpretable patterns for suicide risk assessment.

Keywords

Machine Learning — Suicide Rate — XGBoost — Public Health — Prediction

Table of Contents

Introduction

What contributes to variations in suicide rates across countries and populations? At a broad level, factors such as social pressures, economic stability, cultural norms, mental health conditions, and access to support systems are often considered influential. Demographic characteristics, particularly age and gender, play a significant role, while wider socio-economic conditions may further affect suicide risk across different regions and population groups.

Despite offering useful initial insights, many of these influencing factors are complex, subjective, and difficult to measure consistently. Psychological well-being, individual life experiences, and cultural attitudes toward mental health vary widely across societies and are not easily captured through quantitative data. As a result, purely qualitative explanations face limitations when applied to large-scale, cross-country analyses.

In contrast, objective and measurable indicators—such as demographic distributions, population statistics, and economic variables—provide structured data suitable for systematic analysis, even though they cannot fully capture the underlying psychological dimensions of suicide. Nevertheless, examining whether these observable features can reveal meaningful patterns remains a valuable and important research direction.

Motivated by this perspective, the present study investigates whether suicide risk can be predicted using demographic and socio-economic indicators, and how this risk varies across age groups.

The analysis is based on the Suicide Rates Overview 1985–2016 dataset obtained from the Kaggle platform, which contains multi-year suicide statistics for multiple countries, along with attributes such as age group, sex, population size, GDP, and Human Development Index (HDI). Supervised machine learning models are implemented in KNIME to predict suicide rates per 100,000 population and to classify suicide risk into interpretable categories, combining predictive accuracy with transparent, age-specific risk patterns.

The dataset consists of approximately 27,820 records, each described by the following 12 features:

country (categorical – nominal):

Name of the country where the suicide data was recorded.

- year (numeric – interval):

Year in which the suicide statistics were reported.

sex (categorical – binary):

Gender category of the population group (male or female).

- age (categorical – ordinal):

Age group of individuals (e.g., 5–14 years, 15–24 years).

- suicides_no (numeric – discrete):

Total number of recorded suicides for a given country, year, age group, and gender.

- population (numeric – continuous):

Total population size corresponding to the specified demographic group.

- suicides/100k pop (numeric – continuous):

Number of suicides per 100,000 population, used as a standardized suicide rate.

- country-year (categorical – nominal):

Combined identifier of country and year, primarily used for reference purposes.

- HDI for year (numeric – continuous):

Human Development Index value for the specified country and year, representing overall socio-economic development.

- gdp_for_year ($) (numeric – ratio):

Gross Domestic Product of the country for the given year, expressed in US dollars.

- gdp_per_capita ($) (numeric – ratio):

GDP per capita of the country for the given year, indicating average economic output per person.

- generation (categorical – nominal):

Generational classification based on birth year (e.g., Generation X, Millennials)

The goal of this analysis is to predict suicide rates using Regression and Classification techniques and to evaluate the performance and reliability of the resulting predictions.

This report is organized as follows:

- **Data Exploration**

We examine the main characteristics of the dataset, with particular attention to the suicides/100k pop variable and its distribution across countries.

- **Preprocessing**

We remove some columns from the dataset, transform some features types (gdp for year), renamed column suicides/100k pop to (suicide_rate) and handle missing values by median (Gdp per capita) in order to make the dataset more suitable for analysis.

- **Models**

We utilized XGBoost regression, a tree-based ensemble, to predict numeric suicide rates, and J48, a decision tree classifier, to categorize risk levels as Low, Medium, or High. Training was performed on 80% of the dataset, with performance evaluated on the remaining 20%, ensuring reliable and interpretable results.

- **Evaluation**

We evaluate and compare the models described in the previous section in order to assess their predictive accuracy and robustness.

1. Data Exploration

The dataset used in this study contains multi-year suicide statistics from 1985 to 2016 across multiple countries. Key variables include demographic attributes (age group, gender), population counts, economic indicators (GDP, Human Development Index), and suicide counts. The primary target is the suicide rate per 100,000 population, with an additional categorical risk label (Low, Medium, High) derived from this rate. To better understand temporal patterns and variations in suicide occurrences, the annual frequency of suicides is examined in the figure below.

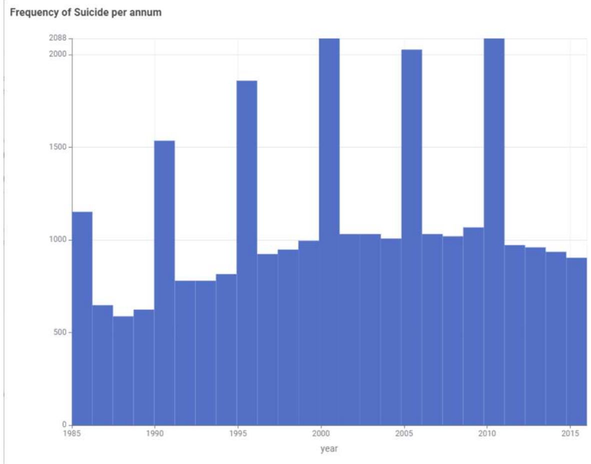


Figure 1.1: Overall frequency of Suicides per Annum

The Figure 1.1 illustrates the annual frequency of suicides from 1985 to 2015 using histogram, highlighting overall trends and significant deviations. Over this 30-year period, the data exhibits considerable fluctuations, with certain years showing pronounced spikes. From the late 1990s onward, the distribution shows a positive skew, indicating that most years have moderate suicide counts, while a few years register exceptionally high frequencies. Notably, the year 2000 represents the peak, with 2088 suicides, standing out as an extreme outlier. Other high-frequency years include 1990 (1536 suicides) and 2005 (2028 suicides), whereas 2015 records the lowest frequency at 744 suicides, marking a sharp decline. Despite these fluctuations, a gradual downward trend is observable after 2010, although year-to-year variation remains significant.
Building on this temporal analysis, the second figure examines the average suicide rates by gender.

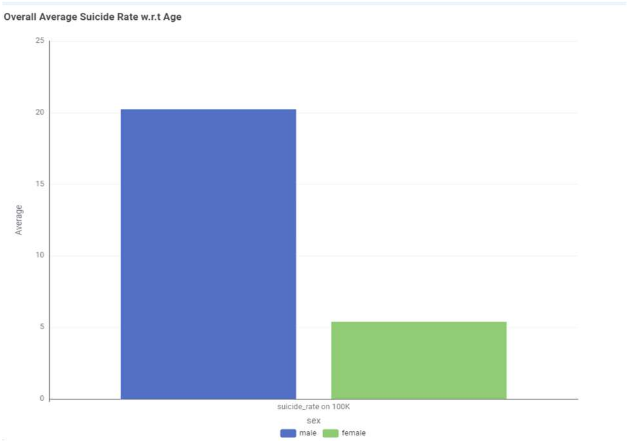


Figure 1.2: Overall Average Suicide rate w.r.t Gender

The above Figure(1.2) clearly represents the Bar chart of average suicide rate contributed by both (Male and Female) gender. The male suicide rate is substantially higher, averaging around 20 per 100,000, compared to approximately 5 per 100,000 for females. This pronounced difference indicates a skewed distribution toward higher male suicide rates. The peak observed in the male category underscores that males are significantly more affected by suicide, whereas the female rate represents the lower end of the spectrum. These observations highlight the importance of considering gender as a key factor in analyzing suicide trends, reflecting both social and demographic influences on suicide risk.

2. Preprocessing

In order to make the dataset more suitable for analysis we have implemented the following changes.

2.1 Feature selection

Out of all the variables in the raw dataset, we discarded those that intuitively should have no effect on the dependent variable, suicide_rate, such as HDI for year.
or those that were duplicates of existing information.
We retained the variables most relevant for the analysis, grouped as follows:
demographic attributes (country, year, sex, age, generation)
outcome and population measures (suicides_no, population, suicide_rate),

economic indicators (gdp_for_year, gdp_per_capita, HDI for year).

The column originally named suicides/100k pop, which serves as the primary measure of suicide incidence, has been renamed to suicide_rate to reflect its central role in the analysis.

2.2 Feature transformation

Several quantitative variables in the dataset needed to be transformed in order to highlight significant underlying patterns and make them appropriate for machine learning models. To accomplish this, a number of operations were applied.

Cleaning and renaming numeric features:

The variables gdp_for_year ($) and gdp_per_capita ($) originally contained currency symbols and comma separators, which prevented them from being recognized as numeric. These non-numeric characters were removed, and the features were converted to numeric variables renamed as gdp_for_year and gdp_per_capita, respectively. This ensured consistency and proper interpretability for subsequent modeling steps.

Logarithmic transformations:

Several key features exhibit skewed distributions that could affect modeling.

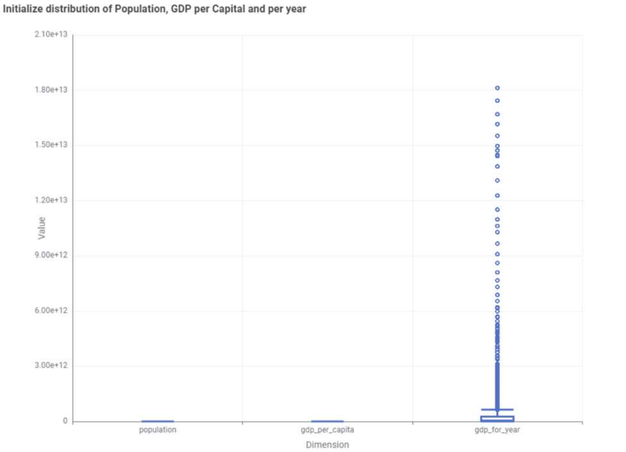


Figure 2.1: Initialize Skewness in distribution

Figure 2.1 clearly demonstrates a box plot showing Population and gdp_per_capita are tightly clustered at lower values with minimal variation, while gdp_for_year shows a strong right skew with a few extreme values. These skewed distributions can disproportionately influence statistical measures and predictive models. To address this, logarithmic transformations were applied to reduce skewness and mitigate the effect of outliers. The following new features were created:

log_population, gdp_per_capita and log_gdp_for_year

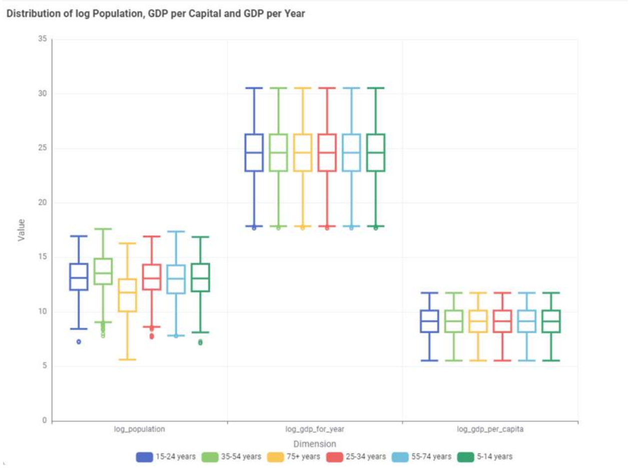


Figure 2.2: Skewness in distribution after log transformation

Figure 2.2 illustrates box plot after applying logarithmic transformations, the distributions of key features became more balanced. Log_population and log_gdp_per_capita show narrow, uniform spreads with minimal skew or outliers across age groups. Log_gdp_for_year exhibits more variability and some right skew in certain age groups, indicating differences in economic activity or high-value outliers. Overall, the log transformations effectively reduced skewness, making patterns in population and GDP per capita easier to interpret, while highlighting remaining variability in GDP for the year

- **Lagged suicide rates:**

Suicide rates often exhibit temporal dependence. To capture short-term trends, the dataset was grouped by country and sorted by year. Two lagged features were generated for each country:

lag1_rate: suicide rate from the previous year

lag2_rate: suicide rate from two years prior

Missing values for the initial years were imputed using the median of the corresponding lag feature.

- **Country-level mean rate:**

To account for long-term country-specific effects, the average suicide rate across all years was computed for each country, resulting in the feature country_mean_rate. Additionally, the total number of observations per country was added to enrich each

record with contextual information.
After all the described operations were performed, the dataset contains 14 features, fully prepared for downstream machine learning modeling.

2.3 Handling of missing values

Missing data are present in some key features, particularly HDI and certain economic indicators. Since predictive modeling requires complete cases for the target variable and main predictors, a consistent strategy was adopted.
Records with missing suicide_rate were removed, as the outcome must be known for supervised learning. For numeric predictors such as HDI for year, gdp_for_year, and gdp_per_capita, missing values were imputed with the median of the corresponding column. This approach preserves most of the data while minimizing the impact of extreme values.
After imputation, the dataset was verified to ensure that all remaining features required for modeling contained no missing entries.

2.4 Transformation of categorical variables

The dataset includes several categorical attributes, namely country, sex, age, and generation. These variables require proper handling to be used effectively in downstream models.
For regression models such as XGBoost, categorical variables can be provided directly as nominal attributes or transformed into binary indicator variables using one-hot encoding.
For the J48 decision tree classifier, it is essential that categorical variables are explicitly recognized as nominal attributes. To ensure this, a Domain Calculation step was applied, defining the possible values for each categorical column and preventing errors such as “cannot handle string attributes” in Weka nodes.

3. Models

In this work, two supervised machine learning models are trained on the preprocessed dataset with the objective of analyzing and predicting suicide-related outcomes. The selected models represent different learning paradigms Tree-based ensemble regression and Decision tree classification allowing a comparative evaluation of their suitability given the nature of the data and the prediction tasks.

3.1 Training and Test Data Partition

Two models are trained using the same 80% training and 20% test split, stratified by country, in order to ensure a fair and consistent comparison of performance across geographical contexts.

3.2 Model Configuration and Training

3.2.1 Tree-Based Ensemble Model: XGBoost Regression

The primary model adopted in this study is an Extreme Gradient Boosting (XGBoost) regression model. The target variable is the continuous suicide_rate, defined as the number of suicides per 100,000 population. The input features include a combination of demographic variables (country, year, sex, age group, generation), economic indicators (GDP for year, GDP per capita), and engineered features (log-transformed GDP and population, lagged suicide rates, and country-level mean suicide rate).
XGBoost constructs an ensemble of shallow decision trees in a sequential manner, where each new tree is trained to correct the prediction errors of the previous ones. The final prediction is obtained as the sum of the contributions of all trees in the ensemble. This approach enables the model to capture non-linear relationships and complex interactions among variables without relying on strong parametric assumptions.
The learner is configured using a squared error loss function, moderate tree depth, and subsampling strategies, which collectively help mitigate overfitting.

3.2.2 Heuristic Model: J48 Decision Tree Classifier

In addition to the regression task, a classification-based approach is implemented using the J48 decision tree algorithm, Weka’s implementation of the C4.5 algorithm.
Since J48 requires a categorical target variable, a new class label named risk_class is introduced by discretizing the continuous suicide_rate into three

ordinal risk levels:

- Low risk: suicide_rate < 5
- Medium risk: 5 ≤ suicide_rate < 15
- High risk: suicide_rate ≥ 15

This discretization is performed using a Rule Engine node in KNIME. The J48 classifier is then trained to predict the risk_class using the same demographic and economic input features employed in the regression model.

The resulting decision tree produces a set of interpretable if–then rules that describe how combinations of age, sex, country, and economic conditions are associated with different suicide risk categories. This interpretability makes the J48 model particularly useful for explanatory analysis, complementing the predictive strength of the ensemble regression approach.

3.2.3 Model Rationale

The combination of XGBoost regression and J48 classification enables the analysis to address both continuous outcome prediction and risk-based categorization. While XGBoost provides high predictive accuracy and robustness to complex feature interactions, the J48 decision tree offers transparent decision rules that support interpretability and domain understanding.

4. Model Evaluation

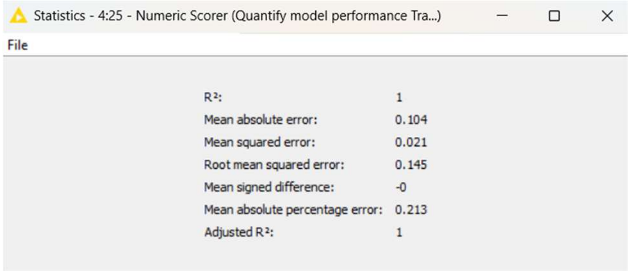
4.1 Regression Model Evaluation (XGBoost)

4.1.1 Performance Metrics:

The regression models are evaluated on training and test sets using complementary metrics:

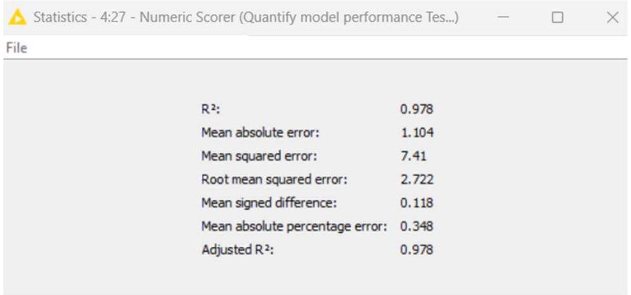
- R^2 (Coefficient of determination) to measure explained variance,
- MAE (Mean Absolute Error), MSE (Mean Squared Error), and RMSE (Root Mean Squared Error) for prediction accuracy,
- MSD (Mean Signed Difference) to detect systematic bias,
- MAPE (Mean Absolute Percentage Error) when zeros in the target variable do not cause instability.

The results of Trained and tested models are shown in the Figure below:


Statistics - 4:25 - Numeric Scorer (Quantify model performance Tra...)
File

```
R²:                              1
Mean absolute error:             0.104
Mean squared error:              0.021
Root mean squared error:         0.145
Mean signed difference:          -0
Mean absolute percentage error:  0.213
Adjusted R²:                     1
```

Figure 4.1: Statistics of XGBoost Trained Model


Statistics - 4:27 - Numeric Scorer (Quantify model performance Tes...)
File

```
R²:                              0.978
Mean absolute error:             1.104
Mean squared error:              7.41
Root mean squared error:         2.722
Mean signed difference:          0.118
Mean absolute percentage error:  0.348
Adjusted R²:                     0.978
```

Figure 4.2: Statistics of XGBoost Test Model

Figures 4.1 and 4.2 report the performance statistics of the XGBoost regression model, clearly indicating excellent predictive accuracy.

On the training set (Figure 4.1), the model attains an R^2 value close to 1.00, together with very low error measures (RMSE ≈ 0.145 and MAE ≈ 0.104), reflecting an almost perfect fit to the data.

When evaluated on the test set (Figure 4.**2**), the model preserves strong generalization performance, achieving an R^2 of approximately 0.978, with RMSE ≈ 2.722 and MAE ≈ 1.104. Overall, the model explains nearly 98% of the variance in the suicide rate on unseen data, while the typical absolute prediction error is around one suicide per 100,000 population. The moderate increase in error from training to testing suggests limited overfitting and robust generalization capability.

4.1.2 Visual Assessment: Actual vs. Predicted

For the Trained Dataset, Prediction of Suicide rate on the basis of Country mean rate is visualized below:

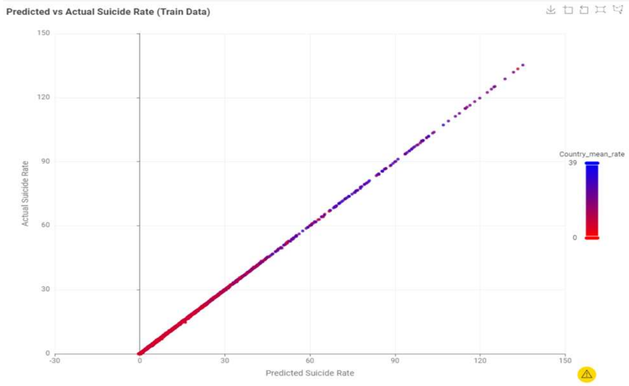


Figure 4.3: Predicted vs Actual Suicide rate w.r.t Country Mean rate for Trained Dataset

The scatter plot in Figure 4.3 illustrates the relationship between the actual suicide rates and the corresponding values predicted by the trained model on the training dataset. Each point represents the country-level mean of suicide rate, ranging from red (lower mean rates) to blue (higher mean rates). The points align closely along the main diagonal, representing the ideal scenario where predicted values equal observed values. This tight alignment indicates that the model achieves very high predictive accuracy on the training data, with minimal deviation across the full range of suicide rates. Additionally, the color distribution demonstrates that the model effectively captures country-level baseline differences, accurately predicting both low- and high-rate countries. Overall, the plot confirms an excellent fit on the training set while highlighting the need for careful validation on independent data to assess potential overfitting.

For the Tested Dataset, Prediction of Suicide rate on the basis of Country mean rate is visualized below:

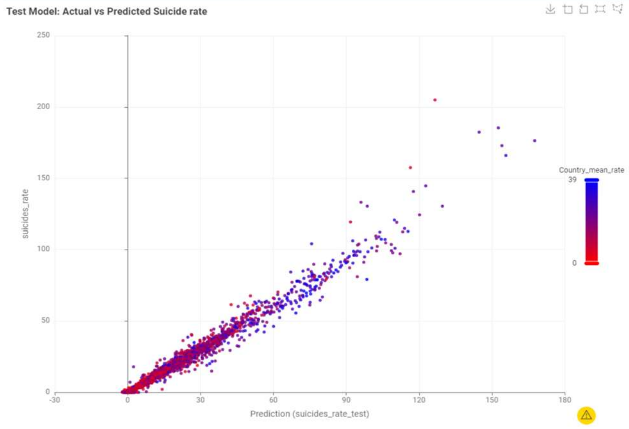


Figure 4.4: Predicted vs Actual Suicide rate w.r.t Country Mean rate for Tested Dataset

This plot compares the XGBoost model's predicted suicide rates (x-axis) against the actual observed rates (y-axis) for the test set, which was not used during training. Each point represents the country-level mean of suicide rate, ranging from red (lower mean rates) to blue (higher mean rates).

The plot exhibits a strong linear relationship, with points closely aligned along the diagonal line (y = x), indicating that the model predicts suicide rates with high accuracy. Predictions for lower suicide rates closely match the actual values, while moderate dispersion is observed for higher rates, reflecting the expected increase in variance for extreme values. A few mild underpredictions occur for countries with exceptionally high suicide rates (150+ per 100,000), which is common in real-world data due to the complexity of extreme cases**.**

4.1.3 Residual Analysis

Residual i.e, the difference between Actual Suicide rate and Predicted Suicide rate) is analyzed with the Predicted Suicide rate for both Trained and Test Data.

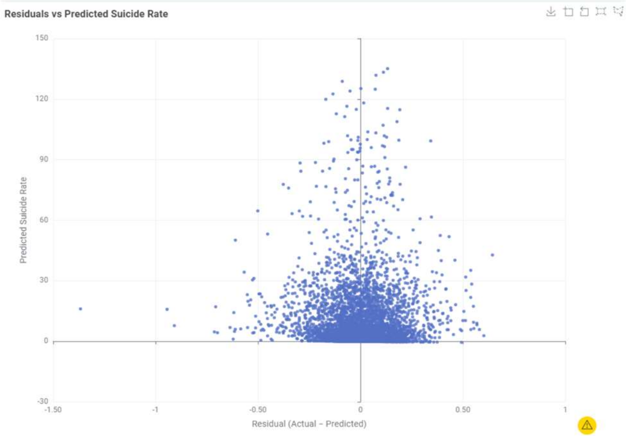


Figure 4.5: Residual vs Predicted Suicide (Train Data)

The following scatter plot (Figure 4.5) shows the Residuals vs Predicted Suicide Rate plot displays each training sample's residual on the x-axis against its predicted suicide rate on the y-axis. This visualization helps assess model performance, bias, and error patterns. In this plot, residuals are densely clustered around zero and symmetrically distributed, indicating that the model does not systematically over- or under-predict. While a mild increase in residual spread is observed at higher predicted values, suggesting slight non-constant variance, the majority of residuals remain small (within ±0.5), reflecting a strong fit. No clear curvature or systematic pattern is visible, confirming that the model structure appropriately captures the underlying relationships. A few outliers appear at high predicted values, which is expected in real-world demographic data and does not compromise overall performance. Overall, the plot demonstrates a well-fitted model with accurate predictions and only minor variance at extreme values

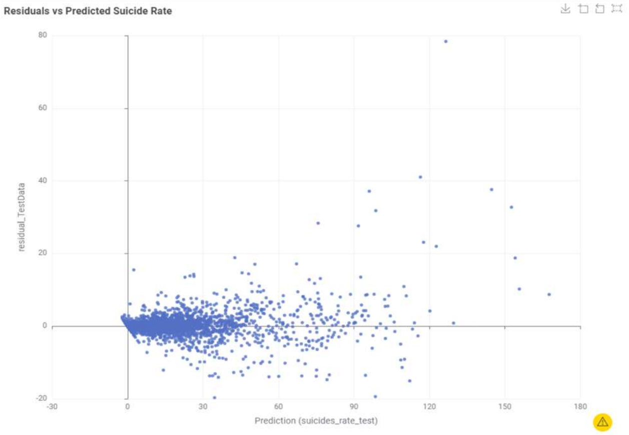


Figure 4.6: Residual vs Predicted Suicide (Test Data)

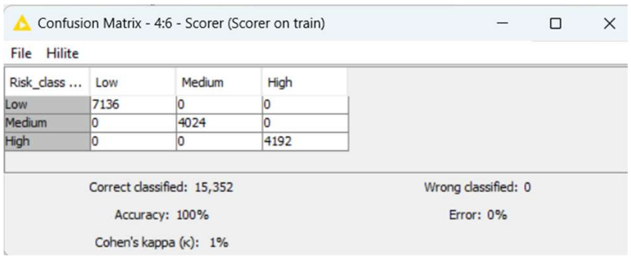
The figure 4.6 Residuals vs Predicted Suicide Rate plot for the test data (Figure 4.6) shows each sample's residual on the x-axis against its predicted value on the y-axis, assessing model generalization. Most residuals cluster around zero, especially for predicted values 0–40, indicating accurate predictions for low-to-moderate suicide rates. Residual spread increases at higher predicted values, showing slight non-constant variance, consistent with the training data. A few large positive residuals at very high predictions (~120–150+) reflect underestimation in rare extreme cases, while small negative residuals at low predictions indicate minor overprediction. No systematic curvature or patterns are observed. Overall, the plot confirms strong generalization, accurate predictions for most cases, and expected variance at extreme values, with underprediction of rare high rates being normal.

4.2 Classification Model Evaluation (J48)

4.2.1 Performance Metrics:

The performance of the J48 decision tree classifier is evaluated using overall accuracy and the confusion matrix, which provide insight into both aggregate performance and class-specific prediction behavior. Using the defined discretization thresholds for the risk_class variable, the classifier achieves perfect performance on both datasets.

Here are the Evaluation results of train and test Models.



Confusion Matrix - 4:6 - Scorer (Scorer on train)

Risk_class ...	Low	Medium	High
Low	7136	0	0
Medium	0	4024	0
High	0	0	4192

Correct classified: 15,352

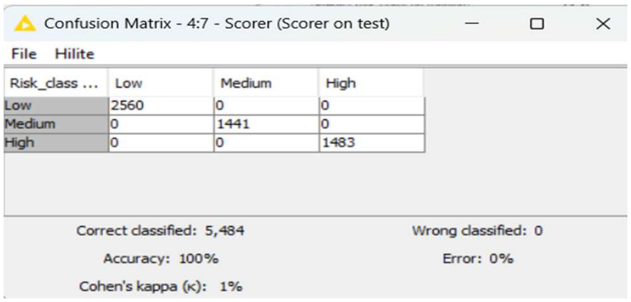
Wrong classified: 0

Accuracy: 100%

Error: 0%

Cohen's kappa (κ): 1%

Figure 4.3: Confusion Matrix Train Model Scores



Confusion Matrix - 4:7 - Scorer (Scorer on test)

Risk_class ...	Low	Medium	High
Low	2560	0	0
Medium	0	1441	0
High	0	0	1483

Correct classified: 5,484

Wrong classified: 0

Accuracy: 100%

Error: 0%

Cohen's kappa (κ): 1%

Figure 4.4: Confusion Matrix Test Model Scores

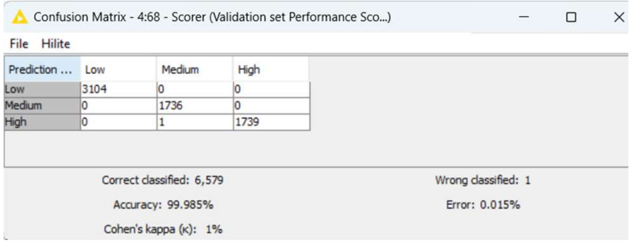
Figures 4.3 and 4.4 present the confusion matrices for the training and test datasets, respectively.

The model achieved an accuracy of 1.00 on both the training data (Figure 4.3) and the test data (Figure 4.4). In both cases, the confusion matrices are perfectly diagonal, indicating that all instances are correctly classified into their corresponding risk categories. This result suggests that, under the selected discretization strategy and given the available demographic and socio-economic features, the Low, Medium, and High suicide risk classes are fully separable within the dataset.

Despite this apparent perfect accuracy, the Cohen's Kappa value of approximately 0.01 indicates agreement only marginally above chance, reflecting limited discriminatory capability. Consequently, the J48 classifier is primarily employed for its explanatory value, providing interpretable decision rules that complement the regression model rather than serving as a robust predictive classifier.

4.2.2 Validation

The dataset was first divided into a training set comprising 80% of the train data and a hold-out test set comprising the remaining 20% for test data. Subsequently, the training set was further partitioned, with 70% used for model learning and the remaining 30% reserved for validation. This two-stage partitioning enables internal validation while preserving an independent test set for final evaluation


Confusion Matrix - 4:68 - Scorer (Validation set Performance Sco...)
File
Hilite

Prediction ...	Low	Medium	High
Low	3104	0	0
Medium	0	1736	0
High	0	1	1739

Correct classified: 6,579

Wrong classified: 1

Accuracy: 99.985%

Error: 0.015%

Cohen's kappa (κ): 1%

Figure 4.5: Validation set Performance

The model was evaluated on a validation set of 6,580 instances, achieving near-perfect performance: 6,579 instances (99.9848%) were correctly predicted within WEKA's default tolerance, with only a single misprediction. The Kappa statistic of 0.9998 indicates almost perfect agreement beyond chance, while the mean absolute error (MAE = 0.0001) and root mean squared error (RMSE = 0.0101) confirm that prediction errors are extremely small. Coverage of cases at the 95% confidence level was 99.9848%, demonstrating that the model reliably captures the target values. Overall, these metrics indicate that the J48 classifier predicts suicide rates with exceptional precision. However, such high performance also suggests that the classification task may be relatively straightforward given the selected features, with strong correlations to the target, and that the model primarily serves an explanatory role by providing interpretable rules for risk categorization

4.2.3 Visual Assessment: Actual vs. Predicted

After our J48 model is Trained and tested. visualization are assessed to predict suicide risk according to age-groups for both Train and Tested data.

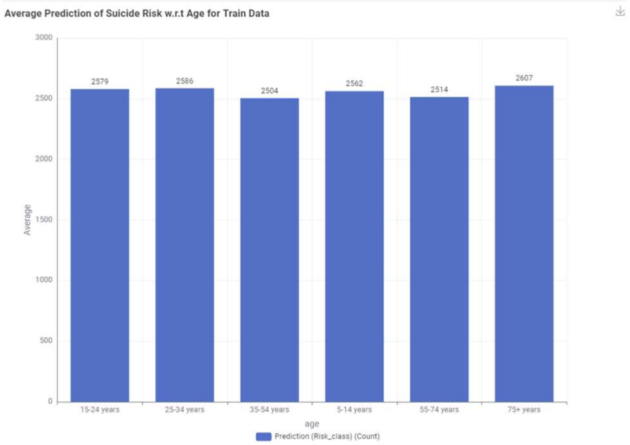


Figure 4.5: Average Predication of Suicide rate w.r.t age groups for Train Dataset

The predicted suicide risk distribution across age groups based on the J48 model shown in Figure 4.5. Using 80% of the dataset for training, the predictions indicate that the highest suicide risk is associated with the 75+ age group. The second highest predicted risk is observed among early teenagers, while the lowest predicted risk occurs in the 55–74 age group. These results highlight pronounced age-related differences in suicide risk, underscoring the importance of age as a key factor in risk stratification.

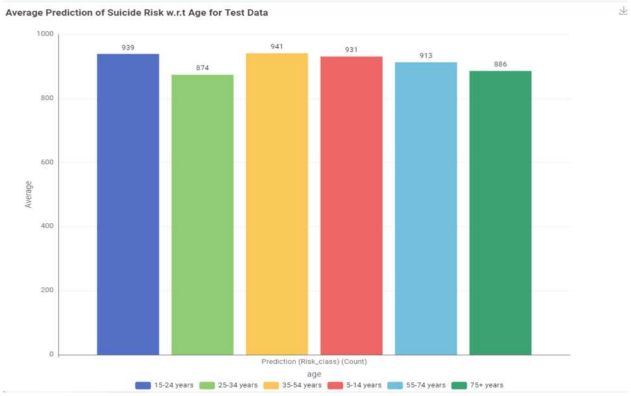


Figure 4.6: Average Predication of Suicide rate based on age groups for Test Dataset

The graph (Figure 4.6) distribution of predicted suicide risk across age groups using the 20% test dataset. The model predicts the highest risk for individuals aged 35–54, followed by early teenagers, while the lowest predicted risk is observed in the 25–34 age group.

A comparison of the predicted suicide risk distributions for the training and test datasets reveals both consistent patterns and notable differences across age groups. In both cases, the model identifies clear age-related variation in suicide risk, confirming that age is a key determinant in risk stratification. Notably, early teenagers consistently appear among the higher-risk groups in both datasets, indicating a stable pattern captured by the model.

Differences are observed in the age group associated with the highest predicted suicide risk: the training results indicate the greatest risk among individuals aged 75 and above, whereas the test results identify the 35–54 age group as having the highest predicted risk. Despite this variation, both datasets exhibit a similar overall risk structure, with a clear separation between higher- and lower-risk age groups.

These differences can be partly attributed to the validation strategy employed: predictions for the test set are derived from a single hold-out partition, while the training results are influenced by a two-stage partitioning process in which the original training data are further divided into learning and validation subsets. Consequently, minor shifts in the relative ranking of high-risk age groups are expected, while the consistent age-related patterns indicate that the model generalizes well.

4.3 Cross Validation

As an additional robustness check, 5-fold cross validation was applied to the XGBoost regression model using KNIME's looping nodes. In this procedure, the dataset is split into five folds, with the model trained on four folds and tested on the remaining fold, iteratively. R^2 and RMSE were recorded for each fold, and the averaged metrics closely matched those from the 80/20 train–test split (mean $R^2 \approx 0.978$; mean RMSE ≈ test RMSE), with low variance across folds.

Since cross validation confirmed the model's stability without revealing new insights, it was not included in the final workflow to avoid unnecessary complexity. The reported evaluation metrics are based on the simpler stratified train–test split, which provides a clearer and interpretable assessment. This approach can be justified in discussion as a verification that the model's performance is not overly sensitive to a particular data split.

Conclusion

Predicting suicide rates across countries, years, and age groups is inherently challenging due to complex demographic, temporal, and regional interactions. The models developed in this study effectively capture these patterns and provide meaningful insights into the underlying relationships.

The XGBoost regression model achieved very high predictive accuracy, demonstrating the importance of non-linear feature interactions in explaining variations in suicide rates. The linear regression model, while simpler, served as a useful baseline by highlighting approximate linear effects of individual variables.

The J48 decision tree classifier, trained on a discretized risk_class, produced interpretable decision rules that revealed age-specific and demographic patterns associated with suicide risk.

Despite strong performance, further improvements are possible. Systematic hyperparameter tuning, time-aware validation that respects chronological order, unsupervised clustering of countries with similar risk profiles, and fairness analysis across demographic groups could enhance robustness and interpretability.

Overall, the study shows that combining regression and classification approaches enables both accurate prediction and interpretable insights, providing a comprehensive understanding of the factors influencing suicide rates.

References

[1] Breiman, L. (1984). Classification and Regression Trees. New York: Routledge.

[2] Kaggle (1985-2016). *Suicide Rate Overview Dataset*. Retrieved from: https://www.kaggle.com/datasets/russellyates88/suicide-rates-overview-1985-to-2016

[3] KDnuggets (November 2018). *How Important is that Machine Learning Model be Understandable? We analyze poll results*. Retrieved from: kdnuggets.com/ 2018/11/machine-learning-model- understandable-poll-results.html

[4] Benoit, K. Linear Regression Models with Logarithmic Transformations [Course Notes]. Retrieved from: kenbenoit.net/assets/courses/ME104/ logm