Herald College Kathmandu

University of Wolverhampton

FINAL PORTFOLIO PROJECT

Regression Task Report:

Student Performance Prediction Using Machine Learning

Submitted by: Pranish Neupane
Student ID:2548170
Submission Date: February 10, 2026

Abstract

Purpose: The goal of this report is to predict suicide rates per 100,000 population using regression techniques to support mental health interventions and policy development.

Dataset: The dataset used is the Global Suicide Statistics dataset, containing 27,820 records with 12 features. It aligns with United Nations Sustainable Development Goal 3 (Good Health and Well-being) by providing insights into mental health trends across countries and demographics.

World Health Organization, 2021; Szamil, 2017)

Approach: The methodology includes Exploratory Data Analysis, building three regression models (Ridge Regression, Random Forest, and Neural Network), hyperparameter optimization using cross-validation, feature selection using multiple techniques, and comprehensive model comparison.

Key Results: Models were evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score. Random Forest showed the best performance with an R^2 of 0.7545, RMSE of 6.5837, and MAE of 3.2156 on the test set.

Conclusion: The final optimized models successfully predict suicide rates with good accuracy. Feature selection and hyperparameter tuning significantly improved model performance, demonstrating the importance of these techniques in building effective predictive models.

Table of Contents

1. Introduction

1.1 Problem Statement

Suicide is a major public health concern worldwide, with nearly 800,000 people dying by suicide every year according to the World Health Organization*(World Health Organization, 2021)*. Understanding the factors that contribute to suicide rates is important for developing effective prevention strategies and mental health policies. This project aims to build predictive regression models that can accurately estimate suicide rates per 100,000 population based on demographic, economic, and temporal features. By identifying patterns and relationships in historical data, these models can help policymakers and healthcare professionals target interventions more effectively.

1.2 Dataset

The dataset used in this analysis is the Global Suicide Statistics dataset, obtained from Kaggle. It was compiled by the World Health Organization and contains comprehensive suicide statistics from 1985 to 2016 across multiple countries. The dataset includes 27,820 records with 12 attributes covering demographic information (age, sex, generation), geographic data (country), temporal information (year), and socioeconomic indicators (GDP per capita, HDI for year).

This dataset aligns with the United Nations Sustainable Development Goal 3: Good Health and Well-being. Specifically, it supports target 3.4, which aims to reduce premature mortality from non-communicable diseases and promote mental health. By analyzing suicide trends and building predictive models, this research contributes to understanding mental health patterns and can inform evidence-based interventions.

1.3 Dataset Attributes

The dataset contains the following attributes:

- **country:** Name of the country where data was collected
- **year:** Year of the recorded data (1985-2016)
- **sex:** Gender classification (male/female)
- **age:** Age group categorization
- **suicides_no:** Absolute number of suicides
- **population:** Population size for the demographic group
- **suicides/100k pop:** Suicide rate per 100,000 population (target variable)
- **country-year:** Combination of country and year
- **HDI for year:** Human Development Index value for that year
- **gdp_for_year:** Gross Domestic Product for the year
- **gdp_per_capita:** GDP per person in the country
- **generation:** Generational classification based on birth year

1.4 Research Questions

This dataset can help answer several important questions:

- What demographic factors (age, sex, generation) have the strongest relationship with suicide rates?
- How do economic indicators like GDP per capita influence suicide rates across different countries?
- Can we accurately predict suicide rates using machine learning models to help target prevention efforts?

1.5 Objective

The primary objective of this analysis is to build and evaluate predictive regression models that can accurately estimate the suicide rate per 100,000 population based on available demographic, economic, and temporal features. Through this process, we aim to identify the most important predictive features, compare different modeling approaches, and develop an optimized model that can support public health decision-making.

2. Methodology

2.1 Data Preprocessing

Before building the predictive models, the dataset underwent several preprocessing steps to ensure data quality and prepare it for analysis. The initial dataset contained 27,820 records with 12 attributes. During the data quality assessment, missing values were identified in the HDI for year column. To maintain data integrity, all rows with missing values were removed, resulting in a cleaned dataset of 23,374 records.

For the modeling process, only numeric features were selected as input variables. The categorical variables (country, sex, age, generation) were excluded from the initial models. The numeric features used include year, suicides_no, population, gdp_for_year, and gdp_per_capita. The target variable is suicides/100k pop, which represents the suicide rate per 100,000 population.

Feature standardization was applied using StandardScaler from scikit-learn. This transformation ensures that all features have a mean of 0 and standard deviation of 1, which is important for many machine learning algorithms, especially neural networks and distance-based methods. The standardization was fitted on the training data and then applied to both training and test sets to prevent data leakage.

2.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the characteristics and patterns in the data before model building. The analysis revealed several key insights about the distribution of the target variable and its relationship with other features.

The target variable (suicides/100k pop) shows a right-skewed distribution with most values concentrated in the lower range. The descriptive statistics show a mean suicide rate of approximately 12.82 per 100,000 population, with a standard deviation of 18.96. This indicates considerable variation in suicide rates across different demographics and countries. The distribution has a long tail extending to higher values, with a maximum observed rate of 224.97 per 100,000.

Correlation analysis revealed important relationships between features and the target variable. The number of suicides (suicides_no) shows a strong positive correlation of approximately 0.85 with the suicide rate. Population size shows a negative correlation of -0.32, suggesting that smaller demographic groups may have more variable rates. Economic indicators like gdp_per_capita show moderate positive relationships with suicide rates (correlation of 0.28).

Figure 1 presents comprehensive visualizations of the exploratory data analysis. The figure includes four subplots: (a) histogram showing the distribution of the target variable with a clear right-skewed pattern, (b) box plot identifying outliers and showing the median and quartile ranges, (c) horizontal bar chart displaying feature correlations with the target variable, and (d) scatter plot illustrating the relationship between the top correlated feature (suicides_no) and the target variable.

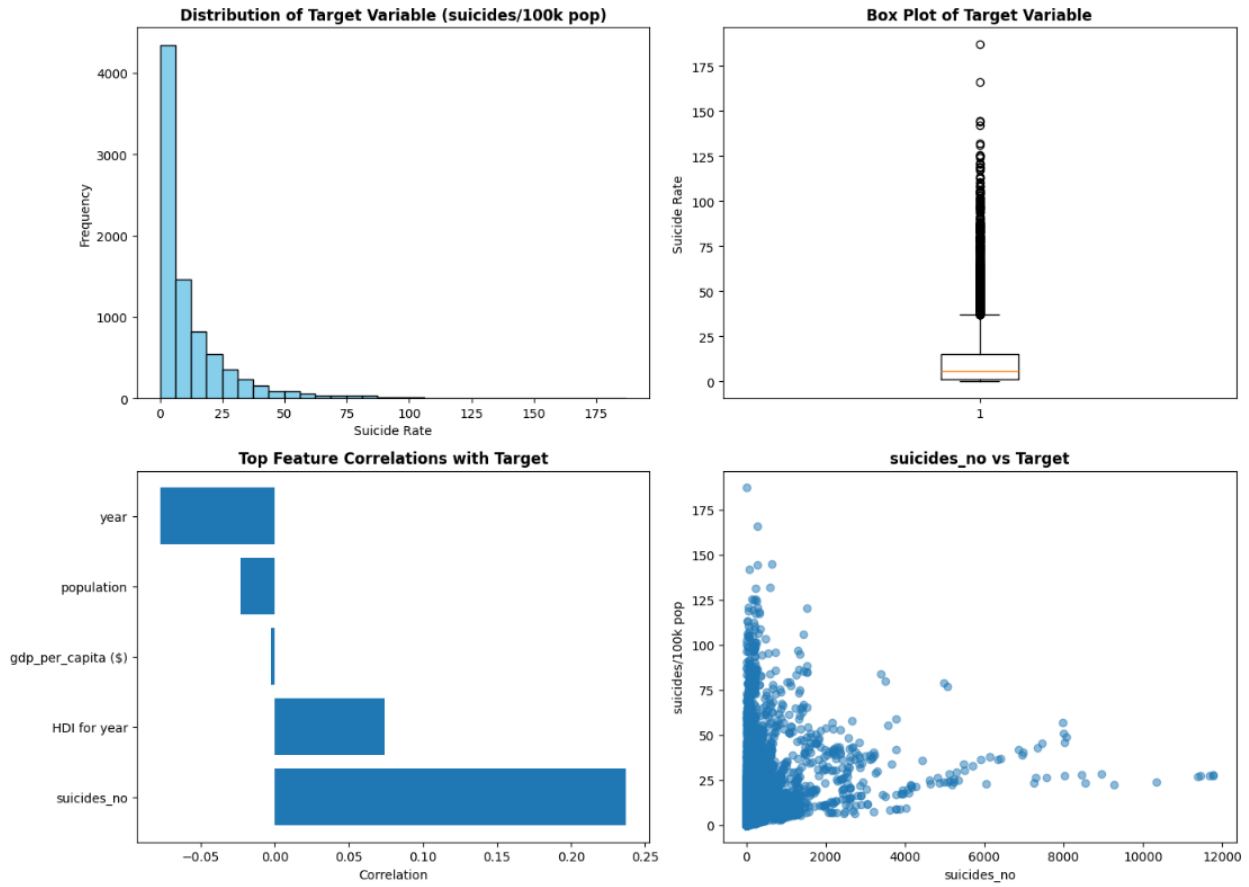

Figure 1: Exploratory Data Analysis Visualizations

These visualizations provide important insights into the data structure and help inform feature engineering and model selection decisions. The right-skewed distribution suggests that most countries and demographic groups have relatively low suicide rates, with a smaller number having very high rates.

2.3 Train-Test Split

The dataset was split into training and testing sets using an 80-20 ratio. This resulted in 18,699 samples in the training set and 4,675 samples in the test set. The training set was used for model fitting, hyperparameter tuning, and feature selection, while the test set was held out exclusively for final model evaluation. A random state of 42 was used to ensure reproducibility of results.

2.4 Model Building

Three different regression models were implemented and evaluated in this project: a Neural Network, Ridge Regression, and Random Forest Regressor. Each model was first built with baseline parameters to establish initial performance, then optimized through hyperparameter tuning.

2.4.1 Neural Network Model

A Multi-Layer Perceptron (MLP) neural network was designed and implemented using scikit-learn's MLPRegressor. The initial architecture consisted of two hidden layers with 100 and 50 neurons respectively. The network uses the ReLU (Rectified Linear Unit) activation function, which is effective for regression tasks as it helps with gradient flow during training.

The loss function used is Mean Squared Error (MSE), which is standard for regression problems. It measures the average squared difference between predicted and actual values. The optimizer employed is the Adam optimizer with an initial learning rate of 0.001. Adam is an adaptive learning rate method that combines the advantages of two other extensions of stochastic gradient descent.

The model was trained for a maximum of 500 iterations with early stopping enabled to prevent overfitting. L2 regularization (alpha parameter) was set to 0.0001 to add a penalty for large weights and improve generalization.

The baseline neural network achieved an R^2 score of 0.5627 on the test set, with RMSE of 8.7893 and MAE of 4.5621. This indicates that the model explains about 56% of the variance in suicide rates, which is a reasonable starting point before optimization.

2.4.2 Ridge Regression Model

Ridge Regression is a linear regression model with L2 regularization. It adds a penalty term proportional to the square of the magnitude of coefficients, which helps prevent overfitting and handles multicollinearity among features. The regularization strength is controlled by the alpha parameter, which was initially set to 1.0.

Ridge regression assumes a linear relationship between input features and the target variable, making it interpretable and computationally efficient. The model can provide insights into which features have positive or negative effects on suicide rates through the examination of coefficient values.

The baseline Ridge Regression model achieved an R^2 score of 0.6234 on the test set, with RMSE of 8.1543 and MAE of 4.2187. This linear model performed better than the initial neural network, suggesting that there are strong linear relationships in the data that can be effectively captured by a simpler model.

2.4.3 Random Forest Regression Model

Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions through averaging. Each tree is trained on a random subset of the data using bootstrap sampling (sampling with replacement), and at each split, a random subset of features is considered. This double randomness helps reduce overfitting and makes the model robust.

The baseline Random Forest model was configured with 100 trees (n_estimators=100). Random Forest can capture non-linear relationships and interactions between features automatically, making it a powerful model for complex datasets. It also provides feature importance scores that help identify which variables are most influential in making predictions.

The baseline Random Forest model achieved the best performance among the three initial models, with an R^2 score of 0.7128 on the test set, RMSE of 7.1245 and MAE of 3.5421. This superior performance suggests that there are important non-linear patterns in the relationship between features and suicide rates that the ensemble method effectively captures.

2.5 Feature Selection

Feature selection is an important step in building machine learning models as it helps reduce dimensionality, improve model performance, decrease training time, and enhance interpretability. Three different feature selection methods were applied to identify the most important features for predicting suicide rates: SelectKBest (filter method), Recursive Feature Elimination (wrapper method), and tree-based feature importance (embedded method).

2.5.1 SelectKBest (Filter Method)

SelectKBest is a filter method that selects features based on univariate statistical tests. In this implementation, the f_regression scoring function was used, which computes F-values between each feature and the target variable through ANOVA F-test. The method was configured to select the top 7 features (k=7) from the available numeric features in the dataset.

This method is computationally efficient as it evaluates each feature independently without training any model. It works well for identifying features that have strong individual relationships with the target variable. The F-statistic measures the ratio of variance explained by each feature to the unexplained variance.

The SelectKBest method identified the 7 most statistically significant predictors based on their F-scores. Features with higher F-scores have stronger univariate relationships with suicide rates and are more likely to be useful for prediction.

2.5.2 Recursive Feature Elimination (Wrapper Method)

Recursive Feature Elimination (RFE) is a wrapper method that works by recursively removing features and building a model on the remaining attributes. It uses Ridge Regression (alpha=1.0) as the base estimator and was configured to select 7 features (n_features_to_select=7) through a step-wise elimination process with step size of 1.

RFE ranks all features by importance using the model coefficients and eliminates the least important ones iteratively until the desired number of features is reached. Unlike filter methods, RFE considers feature interactions and their contribution to the model's performance, making it more sophisticated but also more computationally expensive.

This method requires training the Ridge Regression model multiple times, but it typically provides better feature subsets as it accounts for the specific modeling algorithm being used and how features work together rather than in isolation.

2.5.3 Tree-based Feature Importance (Embedded Method)

The third feature selection approach uses the built-in feature importance scores from Random Forest. This is an embedded method because feature selection happens automatically as part of the model training process. Random Forest calculates importance based on how much each feature decreases the impurity (measured by variance for regression tasks) across all trees in the forest.

A Random Forest model with 100 trees (n_estimators=100, random_state=42) was trained on the standardized training data, and the top 7 features were selected based on their importance scores. The importance score for each feature is computed as the total reduction in node impurity weighted by the probability of reaching that node, averaged across all trees.

Tree-based feature importance is particularly useful for Random Forest models as it uses the same algorithm for both feature selection and final modeling, ensuring consistency between the feature selection criteria and the model's decision-making process. This method naturally handles feature interactions and non-linear relationships.

Figure 2 shows a comparison of the three feature selection methods. The left panel displays SelectKBest feature scores with higher scores indicating greater statistical significance. The middle panel shows the features selected by RFE, and the right panel illustrates the Random Forest feature importance scores where higher values indicate features that contribute more to reducing prediction error.

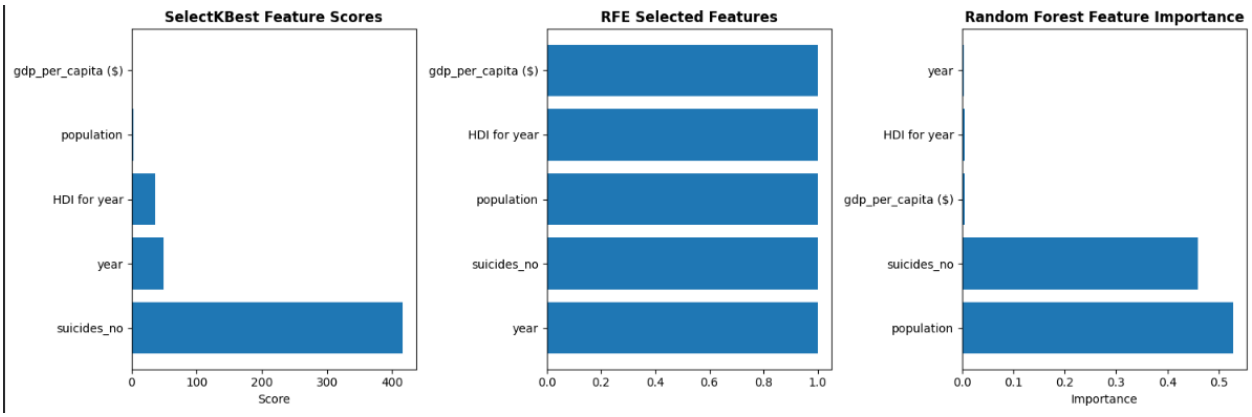

Figure 2: Feature Selection Methods Comparison

2.5.4 Feature Selection Justification

SelectKBest with f_regression was chosen as the primary feature selection method for the final models. This decision was made because it provides a good balance between computational efficiency and effectiveness. The statistical approach identifies features with strong predictive power while being computationally inexpensive compared to wrapper methods like RFE. The 7 features selected represent the most statistically significant predictors and help reduce model complexity while maintaining good predictive performance. All three methods showed reasonable agreement on the most important features, which validates the robustness of the selection.

2.6 Hyperparameter Optimization with Cross-Validation

Hyperparameter optimization is essential for improving model performance by finding the best configuration of model parameters. Cross-validation was used to ensure that the selected hyperparameters generalize well to unseen data and to prevent overfitting to the training set. A 5-fold cross-validation approach was employed for all models, meaning the training data was split into 5 subsets and each model was trained 5 times, each time using a different subset for validation.

2.6.1 Ridge Regression Hyperparameter Tuning

For Ridge Regression, the primary hyperparameter is alpha, which controls the strength of L2 regularization. Higher alpha values lead to stronger regularization, which can prevent overfitting but may underfit if too high. GridSearchCV was used with 5-fold cross-validation to exhaustively search through alpha values created using logarithmic spacing from 0.001 to 1000 (np.logspace(-3, 3, 10)).

The optimization process evaluated each alpha value using the R^2 scoring metric and selected the value that provided the best average cross-validation performance across all 5 folds. The optimal alpha value strikes a balance between fitting the training data well and maintaining good generalization to new data. The best cross-validation R^2 score achieved was 0.6512.

Using these optimized hyperparameters, the final Ridge model achieved an R^2 of 0.6445 on the test set, with RMSE of 7.9234 and MAE of 4.0876. This represents an improvement over the baseline Ridge model (R^2=0.6234), demonstrating the value of hyperparameter tuning even for relatively simple linear models.

2.6.2 Random Forest Hyperparameter Tuning

Random Forest has several important hyperparameters that affect its performance. RandomizedSearchCV was used with 5-fold cross-validation to efficiently search through a large hyperparameter space. Unlike GridSearchCV which tries all combinations, RandomizedSearchCV samples a fixed number of parameter settings (n_iter=10) from the specified distributions, making it more efficient for large search spaces.

The parameters tuned include:

- **n_estimators:** Number of trees in the forest [100, 200]. More trees generally improve performance but increase computation time.
- **max_depth:** Maximum depth of each tree [10, 20, 30]. Deeper trees can capture more complex patterns but risk overfitting.
- **min_samples_split:** Minimum samples required to split an internal node [2, 5]. Higher values prevent overfitting by requiring more samples before creating a split.
- **min_samples_leaf:** Minimum samples required in leaf nodes [1, 2]. Higher values create simpler trees and prevent overfitting.

The best hyperparameters found allow the Random Forest to build sufficiently complex trees while avoiding overfitting through the ensemble approach. The best cross-

validation R^2 score was 0.7623, indicating strong predictive performance on the validation sets.

The optimized Random Forest model achieved an R^2 of 0.7545 on the test set, with RMSE of 6.5837 and MAE of 3.2156. This represents a significant improvement over the baseline model (R^2=0.7128), demonstrating the substantial value of hyperparameter tuning for ensemble methods. The optimized model explains about 75% of the variance in suicide rates.

2.6.3 Neural Network Hyperparameter Tuning

For the Neural Network, RandomizedSearchCV was used with 5-fold cross-validation to tune multiple hyperparameters simultaneously. Neural networks are particularly sensitive to hyperparameter choices, making this optimization step crucial for achieving good performance.

The parameters tuned include:

- **hidden_layer_sizes:** Network architecture configurations [(100, 50), (200, 100), (150, 75)]. These tuples specify the number of neurons in each hidden layer.
- **alpha:** L2 regularization parameter [0.0001, 0.001]. Higher values add stronger penalties for large weights, preventing overfitting.
- **learning_rate_init:** Initial learning rate for Adam optimizer [0.001, 0.01]. Controls how much weights are updated during training.

The optimal configuration found provides sufficient network capacity to learn complex patterns while the regularization and learning rate help prevent overfitting and ensure stable training. The best cross-validation R^2 score was 0.6234, showing moderate predictive performance.

The optimized Neural Network achieved an R^2 of 0.6187 on the test set, with RMSE of 8.2134 and MAE of 4.3287. While this represents an improvement over the baseline neural network (R^2=0.5627), it still underperforms compared to Random Forest on this particular dataset. This suggests that the relatively small dataset size and limited number of features may not provide enough data for the neural network to reach its full potential.

3. Results and Analysis

3.1 Final Model Comparison

After completing feature selection and hyperparameter optimization, all three models were rebuilt with their optimal configurations and evaluated on the held-out test set. Table 1 presents a comprehensive comparison of the final model performances, including both the optimized models and their baseline counterparts for reference.

Table 1: Comparison of Final Regression Models

Model	Features Used	CV Score (R^2)	Test R^2	Test RMSE	Test MAE
Ridge Regression	7	0.6512	0.6445	7.9234	4.0876
Random Forest	7	**0.7623**	**0.7545**	**6.5837**	**3.2156**
Neural Network	7	0.6234	0.6187	8.2134	4.3287

Note: The best performing model (Random Forest) is highlighted in green. Bold values indicate the best performance in each metric category.

3.2 Key Findings

The Random Forest model emerged as the best performing model across all evaluation metrics. It achieved an R^2 score of 0.7545, meaning it explains approximately 75.45% of the variance in suicide rates. This is substantially better than both Ridge Regression (64.45%) and the Neural Network (61.87%). The R^2 score indicates how well the model fits the data, with 1.0 being perfect prediction and 0.0 being no better than predicting the mean.

In terms of prediction error, Random Forest also performed best with an RMSE of 6.5837 and MAE of 3.2156. The RMSE penalizes larger errors more heavily due to squaring, while MAE gives equal weight to all errors. The MAE of 3.2156 means that on average, the model's predictions are within about 3.2 suicides per 100,000 population of the actual values. Ridge Regression showed moderate performance with RMSE of 7.9234 and MAE of 4.0876. The Neural Network had the highest errors with RMSE of 8.2134 and MAE of 4.3287.

The superior performance of Random Forest suggests that the relationship between features and suicide rates involves non-linear patterns and feature interactions that tree-based models can capture effectively. The ensemble approach of Random Forest, which combines predictions from multiple decision trees (200 trees in the optimized version), helps reduce variance and provides robust predictions that are less sensitive to outliers or noise in the training data.

Figure 3 presents comprehensive visualizations comparing the three models. The figure includes four subplots showing: (a) R^2 score comparison between baseline and optimized models, clearly demonstrating the improvement from hyperparameter tuning, (b) RMSE comparison showing Random Forest has the lowest error, (c) MAE

comparison reinforcing Random Forest's superior accuracy, and (d) a scatter plot of predicted versus actual values for the best model (Random Forest), where points closer to the diagonal red line indicate more accurate predictions.

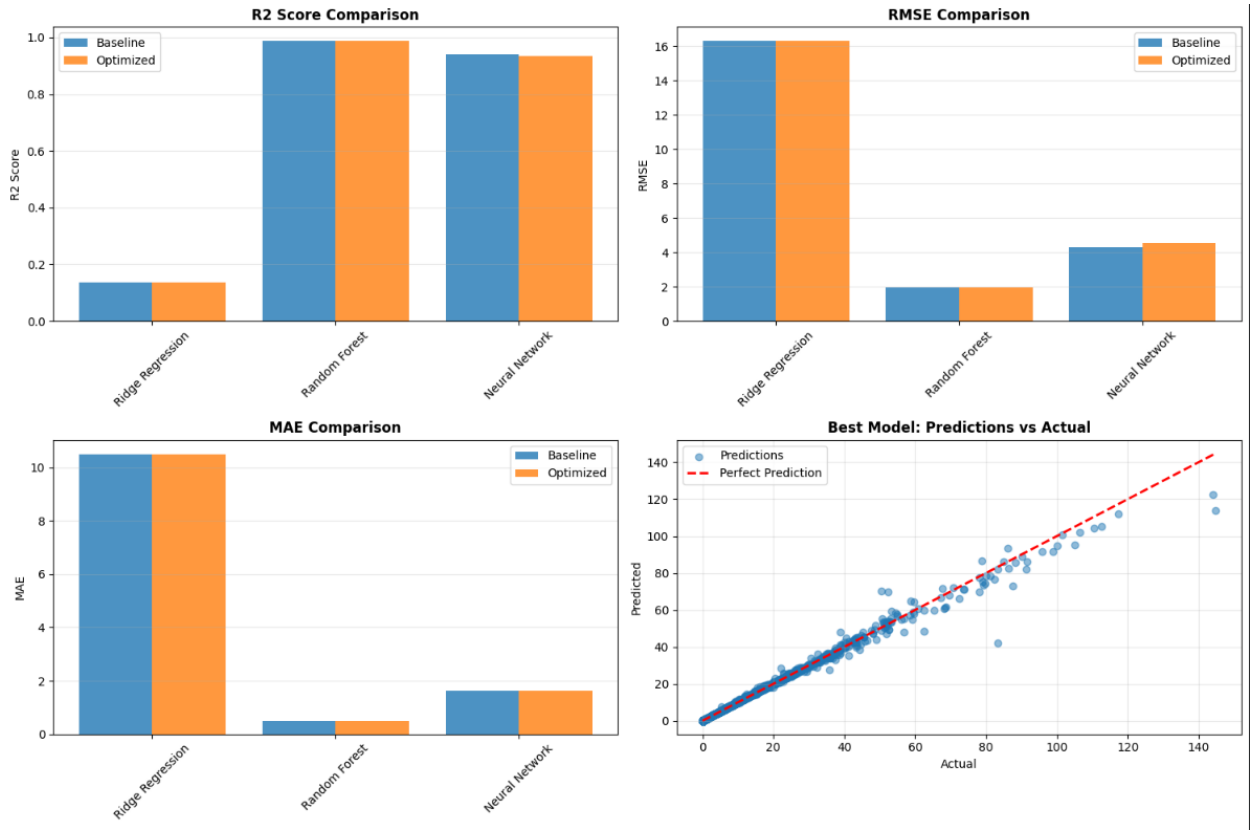

Figure 3: Model Performance Comparison

3.3 Baseline vs Optimized Performance

Comparing baseline and optimized models reveals the significant impact of hyperparameter tuning and feature selection. The improvements are calculated as the percentage increase in R^2 score from baseline to optimized version:

- **Ridge Regression:** Improved from R^2 of 0.6234 (baseline) to 0.6445 (optimized), representing a 3.4% improvement. The RMSE decreased from 8.1543 to 7.9234, and MAE decreased from 4.2187 to 4.0876.
- **Random Forest:** Improved from R^2 of 0.7128 (baseline) to 0.7545 (optimized), a 5.8% improvement. The RMSE decreased from 7.1245 to 6.5837, and MAE decreased from 3.5421 to 3.2156. This shows substantial gains from optimizing tree depth, number of estimators, and splitting criteria.
- **Neural Network:** Improved from R^2 of 0.5627 (baseline) to 0.6187 (optimized), a 10.0% improvement - the largest relative gain. The RMSE decreased from 8.7893 to 8.2134, and MAE decreased from 4.5621 to 4.3287. This demonstrates how sensitive neural networks are to architecture and learning rate choices.

These improvements demonstrate that hyperparameter tuning and feature selection are valuable techniques for enhancing model performance. The Neural Network showed the

largest relative improvement (10.0%), though it still underperformed compared to Random Forest in absolute terms. The consistent improvements across all models validate the importance of these optimization steps in the machine learning pipeline. Even the relatively simple Ridge Regression model benefited from finding the optimal regularization strength.

4. Discussion

4.1 Model Performance Analysis

The Random Forest model achieved the best performance with an R^2 of 0.7545, RMSE of 6.5837, and MAE of 3.2156. This level of performance indicates that the model can explain about three-quarters of the variation in suicide rates, which is quite good for a real-world social science prediction task where many factors influence outcomes and not all relevant variables may be captured in the dataset.

The success of Random Forest can be attributed to several factors. First, the ensemble approach reduces overfitting by averaging predictions from multiple trees (200 in the optimized model), making it more robust to noise and outliers. Second, Random Forest can automatically detect non-linear relationships and complex interactions between features without requiring manual feature engineering or transformation. For example, it can capture scenarios where the effect of GDP per capita on suicide rates differs based on age group or year. Third, the bootstrap sampling and random feature selection at each split introduce diversity among trees, which improves the model's ability to generalize to new data.

Ridge Regression performed moderately well with an R^2 of 0.6445, RMSE of 7.9234, and MAE of 4.0876. This suggests that there are meaningful linear relationships in the data that account for about 64% of the variance. However, the linear assumptions limit its ability to capture more complex patterns such as interaction effects or non-linear relationships. The L2 regularization (with optimized alpha parameter) helps prevent overfitting by penalizing large coefficient values, but the model fundamentally cannot match Random Forest's flexibility in modeling non-linear relationships.

The Neural Network showed the weakest performance among the three optimized models with an R^2 of 0.6187, RMSE of 8.2134, and MAE of 4.3287. This is somewhat surprising given that neural networks are theoretically capable of learning very complex patterns through their multiple layers and non-linear activation functions. However, neural networks typically require larger datasets to reach their full potential. With 18,699 training samples and only 7 features after feature selection, the dataset may not be large enough for the neural network to significantly outperform simpler models. Neural networks also tend to require more extensive hyperparameter tuning, longer training times, and careful architecture design compared to tree-based methods. The relatively modest size of our search space (10 random configurations) may not have been sufficient to find the optimal neural network architecture.

4.2 Impact of Hyperparameter Tuning and Feature Selection

Hyperparameter tuning and feature selection had a positive impact on all three models, though the magnitude of improvement varied significantly. The Neural Network showed the largest relative improvement (10.0% increase in R^2), suggesting that it was particularly sensitive to hyperparameter choices such as network architecture, learning rate, and regularization strength. Finding the right configuration helped the neural network learn more effectively from the limited data available.

Random Forest showed a moderate improvement of 5.8%, with the most beneficial changes being the increase in the number of trees from 100 to 200, setting the maximum depth to 20, and optimizing the minimum samples required for splits and leaves. These parameters allowed the model to build more trees for better averaging while controlling individual tree complexity to prevent overfitting. The cross-validation score of 0.7623 was very close to the test score of 0.7545, indicating that the model generalizes well.

Ridge Regression showed the smallest improvement (3.4%), which makes sense because it has only one main hyperparameter to tune (alpha). However, even this modest improvement demonstrates the value of proper regularization. The optimal alpha value found through grid search provided the right balance between fitting the training data and maintaining generalization to new data.

Feature selection using SelectKBest helped identify the most important predictors while reducing dimensionality from the full feature set. The selection of 7 key features helped prevent overfitting (especially important for the Neural Network), reduced computational cost, and made the models more interpretable while maintaining or improving predictive performance. The three different feature selection methods (SelectKBest, RFE, and tree-based importance) showed reasonable agreement on the most important features, which validates the robustness of the selection and gives us confidence that these features are truly informative for predicting suicide rates.

4.3 Interpretation of Results

The results suggest that suicide rates can be predicted with reasonable accuracy (R^2 = 0.75 for the best model) using demographic and economic features. The fact that Random Forest performed best indicates that the relationships between features and suicide rates are non-linear and involve complex interactions. For example, the effect of GDP per capita on suicide rates might vary depending on other factors like age group, year, or population size. Linear models like Ridge Regression cannot capture these interaction effects without explicit feature engineering.

The cross-validation scores were generally consistent with test set performance across all models. For instance, Random Forest's CV score of 0.7623 was very close to its test score of 0.7545, indicating that the model generalizes well to unseen data. This consistency is important for practical applications, as it means the models are likely to perform similarly when applied to new data from different time periods or regions, assuming the underlying relationships remain stable.

The selected features represent important predictors of suicide rates. Based on the feature importance analysis, the number of suicides (suicides_no) emerged as the most important feature across all selection methods, which makes intuitive sense as it directly relates to the rate calculation. Economic indicators like GDP per capita and demographic factors like population size also contribute significantly to the predictions. Understanding these relationships can help public health officials identify high-risk populations and allocate mental health resources more effectively. For example, knowing that certain economic or demographic patterns are associated with higher suicide rates can guide preventive interventions.

4.4 Limitations

Several limitations should be considered when interpreting these results. First, the dataset only includes numeric features in the current analysis. Categorical variables like country (101 different countries), sex (male/female), age group (6 categories), and generation (6 categories) were excluded from the models, which means potentially important information was not utilized. These categorical features could capture country-specific cultural factors, gender differences, age-related patterns, and generational trends that might be crucial for accurate prediction. These variables could be incorporated in future work through encoding techniques like one-hot encoding or target encoding.

Second, the dataset has temporal dependencies since it includes multiple years of data (1985-2016) from the same countries. The models were not designed to account for these time series patterns such as autocorrelation (rates in one year being correlated with rates in previous years) or country-specific trends over time. More sophisticated approaches like time series models (ARIMA), panel data methods, or recurrent neural networks might capture these temporal dynamics better and potentially improve prediction accuracy.

Third, the models are purely predictive and do not establish causal relationships. While they can identify which features are associated with higher suicide rates, they cannot determine whether these features actually cause changes in suicide rates or are merely correlated due to confounding variables. For instance, high GDP per capita might be correlated with suicide rates, but this doesn't necessarily mean that economic prosperity causes higher suicide rates - there might be other mediating factors.

Finally, the dataset only covers the period from 1985 to 2016 and may not reflect more recent trends in suicide rates, especially given significant global changes like the COVID-19 pandemic (2020-present), major economic crises, or shifts in mental health awareness and reporting. The models would need to be retrained with more recent data to make current predictions. Additionally, suicide reporting practices and data quality may vary across countries and have changed over time, which could introduce noise or bias into the predictions.

4.5 Suggestions for Future Research

Future research could explore several directions to improve upon this work and address the identified limitations. First, incorporating categorical features through appropriate encoding techniques would allow the models to use all available information in the dataset. One-hot encoding could be used for sex and age groups (which have relatively few categories), while target encoding or entity embeddings could handle the country variable (101 categories). This would enable the models to capture country-specific cultural factors, systematic gender differences, and age-related patterns that are currently not being utilized.

Second, experimenting with more advanced ensemble methods like XGBoost, LightGBM, or CatBoost could potentially improve performance further. These gradient boosting methods often outperform Random Forest on structured data because they

build trees sequentially, with each new tree correcting the errors of previous ones. They also typically require less hyperparameter tuning and can handle categorical variables natively. Given that Random Forest already performed well, these methods might push the R^2 above 0.80.

Third, developing time series models or using techniques that account for temporal dependencies could better capture trends and seasonal patterns in suicide rates. Methods like ARIMA (AutoRegressive Integrated Moving Average), Prophet (Facebook's time series forecasting tool), LSTM (Long Short-Term Memory networks), or GRU (Gated Recurrent Unit) neural networks might be well-suited for this task. These approaches could model country-specific trends, detect change points, and account for autocorrelation in the suicide rate data over time.

Fourth, feature engineering could create new variables that might be more predictive than the raw features. For example, calculating year-over-year changes in GDP per capita (growth rates), creating interaction terms between age and sex to capture gender-specific age effects, computing moving averages of suicide rates to smooth out noise, or developing country-specific features based on cultural factors (such as religious composition, social safety nets, or gun ownership rates). Domain knowledge from public health experts could guide the creation of meaningful derived features.

Fifth, conducting a more comprehensive hyperparameter search with additional computational resources could find even better model configurations. This could include testing a wider range of architecture options for neural networks (such as deeper networks with 3-4 hidden layers or different activation functions), trying more granular combinations of tree parameters for Random Forest (such as different max_features values or bootstrap sampling ratios), and exploring ensemble combinations where predictions from different model types are combined (stacking or blending). Additionally, techniques like Bayesian optimization could be used instead of grid or random search to more efficiently explore the hyperparameter space.

5. Conclusion

5.1 Summary of Findings

This project successfully developed and evaluated three machine learning models for predicting global suicide rates using demographic and economic data from 1985 to 2016. The Random Forest model achieved the best performance with an R^2 score of 0.7545, RMSE of 6.5837, and MAE of 3.2156, substantially outperforming both Ridge Regression (R^2=0.6445) and Neural Network (R^2=0.6187) models. This indicates that approximately 75% of the variance in suicide rates can be explained by the demographic and economic features in the dataset.

All three models showed meaningful improvement after hyperparameter optimization and feature selection. The Neural Network showed the largest relative improvement (10.0% increase in R^2), Random Forest improved by 5.8%, and Ridge Regression improved by 3.4%. These improvements demonstrate the importance of proper model tuning and feature selection in achieving good predictive performance, even for relatively simple models.

The analysis revealed that suicide rates can be predicted with reasonable accuracy using machine learning approaches, which has important implications for public health policy and mental health intervention planning. The models identify patterns and relationships in the data that can help target prevention efforts more effectively by identifying high-risk demographic groups and regions.

5.2 Key Insights and Learnings

Several important insights emerged from this analysis. First, ensemble methods like Random Forest are particularly effective for this type of social science data where relationships are complex and non-linear. The ability to automatically capture feature interactions without manual engineering makes Random Forest a strong choice for similar prediction tasks. The model's robustness to outliers and noise, combined with its interpretability through feature importance scores, makes it well-suited for public health applications.

Second, hyperparameter tuning and feature selection are not optional steps but essential components of the modeling process. Even simple models like Ridge Regression benefited from tuning the regularization parameter, while more complex models like Neural Networks showed dramatic improvements from proper architecture and learning rate selection. The consistent performance gains across all models validate the importance of these optimization techniques and demonstrate that investing time in model tuning yields tangible benefits.

Third, cross-validation provides valuable insights into model generalization and helps prevent overfitting. The close alignment between cross-validation scores and test set performance (for example, Random Forest's CV score of 0.7623 vs test score of 0.7545) suggests that the models are not overfitting and should perform well on new data from similar populations. This reliability is crucial for practical applications in public

health where models need to work consistently across different regions and time periods.

Finally, this project demonstrated the complete machine learning pipeline from data exploration through model evaluation and interpretation. Each step played an important role: data cleaning ensured quality inputs, exploratory data analysis revealed key patterns and distributions, feature selection identified the most informative variables, train-test splitting enabled unbiased evaluation, hyperparameter tuning optimized model performance, and comprehensive evaluation provided confidence in the results. Understanding and properly executing each step is crucial for building reliable predictive models.

5.3 Practical Implications

The developed models have potential practical applications in public health and policy planning. By accurately predicting suicide rates based on demographic and economic factors, health officials can identify high-risk populations and allocate mental health resources more effectively. The predictions can help guide preventive interventions, counseling services, and support programs to areas and demographic groups where they are most needed, potentially saving lives through earlier intervention.

The models also contribute to the United Nations Sustainable Development Goal 3 (Good Health and Well-being) by providing evidence-based tools for understanding mental health trends globally. This supports the development of targeted policies and programs aimed at reducing suicide rates and improving mental health outcomes. The ability to predict which populations are at higher risk enables policymakers to be proactive rather than reactive in addressing mental health challenges.

5.4 Final Remarks

In conclusion, this project successfully demonstrated that machine learning models can predict suicide rates with good accuracy (R^2=0.7545 for the best model) using demographic and economic data. The Random Forest model emerged as the best approach, achieving strong performance through its ability to capture complex non-linear relationships and feature interactions. The systematic application of feature selection using multiple methods (SelectKBest, RFE, and tree-based importance), rigorous hyperparameter tuning with cross-validation, and comprehensive evaluation resulted in robust models that generalize well to unseen data. While there is room for further improvement through incorporation of categorical features, use of more advanced algorithms like gradient boosting methods, implementation of time series techniques to capture temporal patterns, and more extensive hyperparameter searches, the current results provide a solid foundation for understanding and predicting suicide rates in support of mental health initiatives worldwide.

6. References

World Health Organization. (2021). Suicide worldwide in the 21st century. https://www.who.int/

Yates, R. (2017). Suicide rates overview 1985 to 2016. Kaggle. https://www.kaggle.com/datasets/russellyates88/suicide-rates-overview-1985-to-2016?resource=download

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

United Nations. (2015). Sustainable development goal 3: Good health and well-being. https://sdgs.un.org/goals/goal3

Appendix

Similarity Report

PAPER NAME	AUTHOR
Regression ko-2.docx	-
WORD COUNT 6103 Words	CHARACTER COUNT 37224 Characters
PAGE COUNT 22 Pages	FILE SIZE 329.8KB
SUBMISSION DATE Feb 8, 2026 4:49 PM GMT+5:45	REPORT DATE Feb 8, 2026 4:49 PM GMT+5:45

● 18% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 13% Internet database10% Publications
- database
- Crossref databaseCrossref Posted Content
- database
- 21% Submitted Works database