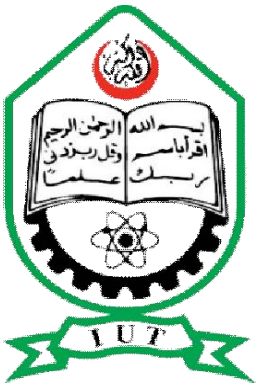

Islamic University of Technology (IUT)

A machine learning approach for analyzing and predicting suicidal thoughts and behaviors.

by
Kazi Raine Raihan (180021102)
Sayed Rakibul Hasan Shanto (180021118)
Mahmudul Hasan (180021325)

Supervised by
Mr. Fahim Faisal
Assistant Professor
Dept. of Electrical and Electronic Engineering (EEE)

A Dissertation
Submitted in Partial Fulfillment of the Requirement for the
Bachelor of Science in Electrical and Electronic Engineering
Academic Year: 2021-2022

Department of Electrical and Electronic Engineering (EEE)
Islamic University of Technology (IUT)
Organization of Islamic Cooperation (OIC)
Gazipur – 1704, Dhaka, Bangladesh

A machine learning approach for analyzing and predicting suicidal thoughts and behaviors.

A thesis
presented to
The Academic
Faculty by

Kazi Raine Raihan (180021102)
Sayed Rakibul Hasan Shanto (180021118)
Mahmudul Hasan (180021325)

Approved by
Mr. Fahim Faisal

…….
Mr. Fahim Faisal
Thesis supervisor
Dept. of Electrical and Electronic Engineering

Islamic University of Technology
The Organization of Islamic Co-operation (OIC)
Gazipur-1704, Dhaka, Bangladesh

Declaration

This is to certify that the project entitled "**A Machine Learning approach for analyzing and predicting suicidal behaviors and thoughts** is supervised by Mr. Fahim Faisal. This project work has not been Submitted anywhere for a degree.

…….
Mr. Fahim Faisal
Thesis supervisor
Dept. of Electrical and Electronic Engineering

…….
Kazi Raine Raihan
(180021102)

…….
Sayed Rakibul Hasan Shanto
(180021118)

…….
Mahmudul Hasan
(180021325)

Acknowledgements

First and foremost, the authors would like to express their profound gratitude and appreciation to Allah, the Almighty, without Whom the study would not have been possible.

We want to express our gratitude to everyone who assisted us in completing this assignment and recognize the university's significant contribution as well as the department's exceptional kindness to us during the entire process.

We owe a debt of gratitude to our esteemed thesis supervisor manager, Mr. Fahim Faisal, assistant professor in the department of electrical and electronic engineering and Mr. Mirza Muntasir Nishat, assistant professor in the department of electrical and electronic engineering, for their unfailing encouragement, inspiration, patience, passion, and comprehensive understanding of the subject matter. Throughout the entirety of our work period, their constant direction and close monitoring kept us safe.

We also want to thank our parents for keeping us on the right track by giving us motivation and their love.

Abstract

In the field of public health, suicide is a problem of the utmost significance that demands immediate attention and successful preventative measures. There has been an increase in interest in using machine learning to predict and identify people who are at a high risk of suicide as society struggles with the tremendous effects suicide has on individuals, families, and communities. In this work, we provide a complete evaluation of the state-of-the-art machine learning algorithms for suicide prediction, with the goal of highlighting the achievements made thus far and outlining potential avenues for future research.

Examining the various aspects and data sources used in prior studies is essential if one wants to comprehend the complicated environment of suicide prediction. As people frequently convey their feelings, problems, and distress signals through written communication, researchers have realized the enormous utility of harnessing text-based data from social media sites. Machine learning algorithms can find patterns and signs that can point to a higher risk of suicide by examining these textual data sources. Electronic health records have also proven to be a useful tool since they include important details regarding a person's medical background, mental health diagnoses, and previous interactions with healthcare systems.

The use of machine learning techniques is critical in converting a large amount of data into useful insights for suicide prevention. To evaluate the obtained data, a variety of algorithms have been used, with neural networks emerging as a major technique. Neural networks can understand complicated patterns and correlations in data, allowing them to make accurate forecasts and identify people who are suicidal. Other machine learning approaches, such as support vector machines, decision trees, and ensemble methods, have also shown promising results, demonstrating the wide range of tools available for suicide prediction.

While machine learning has the potential to significantly improve suicide prevention efforts, it is critical to address the ethical considerations related to putting such models into practice. To secure individuals' sensitive information, privacy and data security problems must be properly managed. Furthermore, the potential for bias and prejudice within machine learning models must be.

carefully analyzed and reduced to provide fair and equal results. Researchers and practitioners may strive toward establishing responsible and ethical suicide prediction algorithms by actively engaging with these ethical factors.

This thesis focuses on the considerable advances achieved in suicide prediction via the use of machine learning techniques. Researchers have made significant progress in detecting patients at high risk of suicide by using multiple data sources such as social media, electronic health records, and demographic information, as well as employing machine learning algorithms such as neural networks. Looking ahead, machine learning has enormous potential to improve suicide prevention efforts, opening new avenues for tailored treatments and support. However, it is critical that these advances be achieved responsibly and ethically, with privacy, fairness, and equity being valued in the creation and implementation of these models.

Keywords:

Suicide prediction, Machine Learning, styling, Neural network, Kaggle dataset, Predictive analysis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Suicide is a major public health issue that affects millions of people globally. Because of the increasing frequency of mental health concerns, it is necessary to investigate machine learning algorithms for detecting people at risk of suicide. This field has the potential to improve suicide prevention efforts while also addressing the worldwide mental health epidemic. To assess data and predict suicide risk, machine learning techniques such as neural networks, support vector machines, decision trees, and ensemble approaches are used. These models capture complex patterns and correlations in data, which improves accuracy. Machine learning has the potential to improve suicide prevention efforts by providing tailored treatments and improving resource allocation. Integrating machine learning into current mental health care systems can enhance screening procedures and clinical decision-making. Our ability to predict suicide attempts has been near chance levels for several decades [1]. There have been various attempts to pinpoint suicidality risk factors. Suicide affects individuals, families, communities, and even entire countries [2]. For young individuals, suicide is the second leading cause of death, killing more people than diabetes, liver disease, stroke, or infection [3]. The stigma attached to mental illnesses prevents more than 40% of people from seeking primary care because they are hesitant to discuss their pertinent symptoms. There is no effective method for handling, evaluating, or preventing suicide, making immediate intervention into suicidal thoughts and actions necessary [3]. A crucial field of research called suicidal ideation detection (SID) looks to determine whether a person is having suicidal thoughts or ideas. Analyzing behavioral data and looking at text that the subject of the investigation produced are also steps in this process [5]. The historical, environmental, and health-related components of suicide have been divided into three separate categories by the American Foundation for Suicide Prevention

(AFSP). These classifications provide a framework for comprehending and evaluating the numerous variables that affect suicide risk [6].

Potential suicide victims frequently act out their self-harming impulses through a variety of activities, such as role-playing, transient thoughts, or even the creation of intricate suicide plans. To avoid catastrophic results, it is crucial to understand and proactively identify the dangers connected to these actions and intentions linked to suicide thoughts. It is a complex and comprehensive attempt to comprehend the causes of suicide. It is critical to understand that suicide rarely results from a single cause, but rather through intricate interactions between several different circumstances. These influencing factors might include biological components like underlying mental health issues or genetic predispositions. A large part is also played by psychological issues, such as emotional anguish, hopelessness, or poor coping skills. Furthermore, it is important to consider the social and environmental factors because they might affect the total risk of suicide. These factors include social isolation, bad life events, interpersonal disputes, and availability to fatal weapons. Mental health professionals and society at large can strive toward putting into practice effective preventive methods and interventions that holistically address the intricacies surrounding suicide by thoroughly analyzing and addressing these varied causes. Effective early suicidal ideation detection can reveal a person's suicide thoughts, enabling proactive support and intervention. It is crucial to recognize that there are many different elements at play, all of which interact intricately to cause suicide [7],[8]. Recently, there has been an increase in interest in using machine learning, often known as ML, to predict suicide. This topic has been the subject of extensive research, illuminating the potential for ML systems to recognize and predict suicide risk. The risk factors for suicide can be broadly categorized into traits or states, which include both constant qualities and dynamic variations. The implementation of machine learning algorithms for reliable suicide risk assessment has been the subject of numerous studies and investigations thanks to this complex understanding of suicide risk. Due to the increased visibility and interest in this area of research, machine learning models created specifically for suicide prediction have continued to evolve and be improved. For instance, academics have employed machine learning to sift through social media messages and find people who could be suicidal or have self-harm tendencies [9]. Researchers have made tremendous progress in creating prediction models that combine a variety of criteria, including

clinical, demographic, and data from electronic health records (EHRs), in order to identify people who are at risk of suicide. These thorough prediction models make use of a multitude of data to offer a full evaluation of suicide risk. These models gain important insights about a person's mental health status and treatment trajectory by utilizing clinical data, such as psychiatric diagnoses, medication history, and prior suicide attempts. To account for the potential impact of societal and environmental variables, demographic parameters including age, gender, socioeconomic position, and geography are also incorporated into these models. Additionally, by including data from electronic health records, which collect essential details about medical history, comorbidities, the use of healthcare services, and current interactions with healthcare practitioners, these models' prediction accuracy is improved. Researchers want to create reliable prediction models that could recognize people who are at risk of suicide and open the door for targeted interventions and assistance by combining these many data sources.

CHAPTER 2

BACKGROUND

In certain research endeavors, machine learning techniques have been employed to predict suicide attempts within a sizable sample of patients, spanning a duration of five years [10]. By leveraging the power of machine learning, these studies aimed to analyze an extensive array of data and extract patterns or signals that could help identify individuals at heightened risk of engaging in suicide attempts. This approach allowed researchers to gain valuable insights into long-term trends and potential risk factors associated with suicidal behavior, facilitating the development of proactive interventions and targeted preventive measures. Studies have used machine learning models to predict suicidal behavior in patients using longitudinal electronic health information. Researchers sought to create reliable machine learning algorithms capable of identifying patterns, trends, and risk factors that contribute to suicide behavior by utilizing these extensive and rich datasets. This ground-breaking method made it possible to identify people who might be at a higher risk of attempting suicide or self-harm, giving healthcare professionals the opportunity to act quickly and offer the right kind of resources and support [11]. Numerous academic works stress the need to recognize and treat risk factors linked to suicide tendencies. These studies shed light on the complex interactions between multiple risk and protective factors for teenage suicide behavior. They emphasize how crucial early detection and intervention are to successfully preventing these behaviors and reducing the risks they pose. Additionally, according to these scientists, machine learning algorithms show potential in anticipating suicide attempts, allowing for the creation of customized countermeasures to foil such attempts. The potential of machine learning in this situation provides a ray of hope for putting preemptive measures into

place and providing targeted support to people who are at risk, ultimately helping to lower suicide rates and promote mental health [12]. Numerous studies have focused on developing predictive models and tools for suicide prediction. These models include techniques based on clinical and demographic data, electronic health record (EHR) data, and demographic and clinical data designed specifically for individuals with schizophrenia. Furthermore, a large majority of the available procedures for detecting suicidal ideation are based on interactions between social workers, specialists, or mental health experts and the individuals under investigation. Furthermore, machine learning approaches play an important role, employing techniques such as feature engineering or deep learning to automatically detect symptoms of suicidal ideation in social media posts. Several studies provide in-depth assessments of cutting-edge machine learning technologies and their applications in detecting suicidal ideation. These evaluations investigate the potential of these strategies in improving suicide prevention and intervention efforts, emphasizing their importance in tackling this critical public health issue [13]. Furthermore, much research has examined the limitations of current suicide risk assessment methodologies. These studies underline the importance of taking a complete strategy to suicide risk management, one that considers individual risk factors as well as clinical judgment. Recognizing the complexities inherent in assessing suicide risk, our findings urge for a more nuanced and individualized strategy that considers a variety of factors. A more holistic and effective framework for suicide risk management can be built by embracing both objective risk variables and the expertise of mental health specialists [14]. Furthermore, several study publications dig into the ethical issues of suicide prevention, specifically the usage of social media platforms. These debates critically analyze the ethical implications of using social media data and content to identify people at danger of suicide. They investigate the issues of privacy, consent, and the possibility of unintended consequences when using these platforms as a source of information for preventive interventions. By addressing these ethical concerns, researchers hope to ensure that suicide prevention programs strike a careful balance between exploiting useful data while respecting individual rights and safeguarding the trust and well-being of vulnerable persons [15]. Chapter 03 of the paper presents the methodology. The results are analyzed in Chapter 4 and the conclusion is exhibited in Chapter 5.

CHAPTER 3

METHODOLOGY

3.1 Research Objective:

Because of the alarming rise in suicide rates in recent years, there has been a surge in interest among suicide detection researchers. This topic has received a lot of interest and has been looked at from numerous angles. Machine learning approaches have emerged as significant tools in the quest for autonomous detection. Researchers have thoroughly examined the available datasets, conducting empirical analyses to get insights into the underlying patterns and causes linked to suicide.

A correlation heatmap was created as part of the research to visually show the correlations and dependencies between different attributes in the dataset [16]. Figure 1 is a heatmap that shows the relationships between numerous variables, offering information on their interconnections and potential influence on suicidal tendencies. Researchers can detect significant relationships and uncover critical aspects that lead to the prediction of suicidal actions by evaluating the heatmap [17].

Table I also contains numerical facts about the dataset, such as its characteristics and properties. This table is a thorough reference that provides critical information for further research and modeling. The types of variables, their ranges, and any unique concerns that might come up during the modeling process are all listed in this table to aid researchers in understanding the dataset's structure.

In general, the empirical analysis, correlation heatmap, and dataset details shown in Figure 1 and Table I help to investigate and understand suicide detection using machine learning. Insights into the dataset's structure, attribute correlations, and potential predictive factors are vitally revealed by these graphics and numerical representations, guiding researchers toward the creation of workable models and actions to combat the worrisome rise in suicide rates.

TABLE I: Different Attributes

No.	Attributes	Min-Max	Mean	Standard Deviation
1	Currently_Drink_Alcohol	1.4-548.0	102.571148	182.775312
2	Really_Get_Drunk	0.8-106.0	35.646794	37.184987
3	Overwieght	3.3-106.0	35.551052	34.948497
4	Use_Marijuana	0.0-106.0	22.688249	36.318479
5	Have_Understanding_Parents	5.6-106.0	40.040622	32.095851
6	Missed_classes_without_permssion	6.5-106.0	37.963494	32.363091
7	Had_sexual_relation	2.5-106.0	37.31632	35.07681
8	Smoke_cig_currently	1.2-106.0	27.342978	34.255473
9	Had_fights	3.5-106.0	40.855727	34.156889
10	Bullied	9.9-106.0	40.767906	33.939727
11	Got_Seriously_injured	14.81-106.0	49.597762	32.425293
12	No_close_friends	1.5-106.0	20.694949	35.190736
13	Attempted_suicide	2.7-106.0	29.912681	36.769804

The GSHS dataset includes numerical values for a variety of characteristics related to risk factors and health-related activities among teenagers to aid in analysis. certain statistics provide important information on the prevalence and distribution of certain factors. For instance, characteristics like "Currently_Drink_Alcohol," "Overweight," and "Use_Marijuana" show that the teens in the survey exhibit

worrying habits. To effectively address these concerns, these findings highlight the significance of focused treatments and prevention measures.

The GSHS dataset also provides insight on adolescents' social and mental wellbeing. Teenagers who were polled were asked about their emotional and social experiences, and attributes like "Understanding_Parent," "No_close_friends," and "Attempted_suicide" reveal important details about these experiences. These findings highlight the need for complete support systems, which should include healthy peer connections, positive parent-child relationships, and readily available mental health resources.

The findings from the GSHS dataset have a big impact on public health programs and regulations. Adolescent-specific therapies can be created by recognizing risk factors, comprehending health practices, and addressing mental health issues. The GSHS dataset is an essential tool for advancing evidence-based strategies meant to enhance the general health and wellbeing of teens around the world.

In conclusion, the Global School-based Student Health Survey is essential for gaining an understanding of the risk factors and health-related behaviors that affect adolescents around the world. The GSHS promotes ethical decision-making and evidence-based treatments by using a consistent methodology, upholding ethical standards, and providing thorough data. The knowledge gained from the GSHS dataset has the potential to influence worldwide policies and initiatives that support the health and wellbeing of teenagers.

3.2 Data Source:

The Global School-based Student Health Survey (GSHS) is a cross-sectional surveillance study carried out in 26 countries with the objective of gathering in-depth data on teenagers all over the world [10]. In order to make meaningful comparisons between nations, its main objective is to quantify the prevalence of risk factors and healthy behaviors among teenagers. The survey uses a mix of core-expanded and chosen core questions that are adjusted to the context of each participating country. The Global Study of Adolescent Health (GSHS) offers important insights into the

health status of teenagers internationally and informs evidence-based interventions and policies. It does this through standardized data gathering procedures and stringent ethical considerations.

The GSHS's standardized methodology, which ensures uniform questions and answer choices across nations, is its strongest point. This homogeneity allows for trustworthy cross-national comparisons that consider variables like population distribution, non-response rates, and classroom selection. The representativeness of prevalence rates is improved by the weighted survey design.

A big focus of the Global School-based Student Health Survey (GSHS) is on data protection and ethical considerations. The poll is thoroughly reviewed and approved by institutional ethics committees or boards as well as national governmental agencies in each country to assure its integrity. Obtaining participant and parental consent ensures the privacy and confidentiality of the data acquired. This dedication to moral behavior builds confidence in the survey and preserves its integrity.

The GSHS dataset plays a pivotal role in facilitating analysis by categorizing adolescents into two age groups: younger (13-15 years old) and older (16-17 years old). These age ranges have been carefully selected to align with publicly reported question response prevalence, enabling meaningful comparisons of risk factors and health behaviors across countries. This segmentation allows researchers to identify age-specific trends and patterns, providing valuable insights into the unique challenges faced by adolescents at different developmental stages.

The GSHS dataset covers a wide range of health-related sectors and consists of 10 major components and numerous auxiliary modules. These include crucial markers including alcohol intake, physical exercise, mental health, and other crucial elements important to adolescent wellbeing. The GSHS provides useful insights that support evidence-based policies and treatments by collecting data on three essential facets of adolescent health. This multidimensional strategy supports efforts to improve the general wellbeing and quality of life for teenagers around the world by customizing interventions to address particular risk factors and health behaviors common among this age group.

Understanding the prevalence of risk factors and healthy behaviors among teenagers around the world is made possible in large part by the Global School-based Student Health Survey (GSHS). The GSHS provides meaningful cross-national comparisons, producing evidence-based insights by using a standardized approach, abiding by ethical rules, and using representative sampling. The GSHS makes a substantial contribution to the promotion of teen health and well-being globally through its thorough survey design and data analysis, making it a crucial tool in tackling the issues this population faces. The correlation Heatmap:

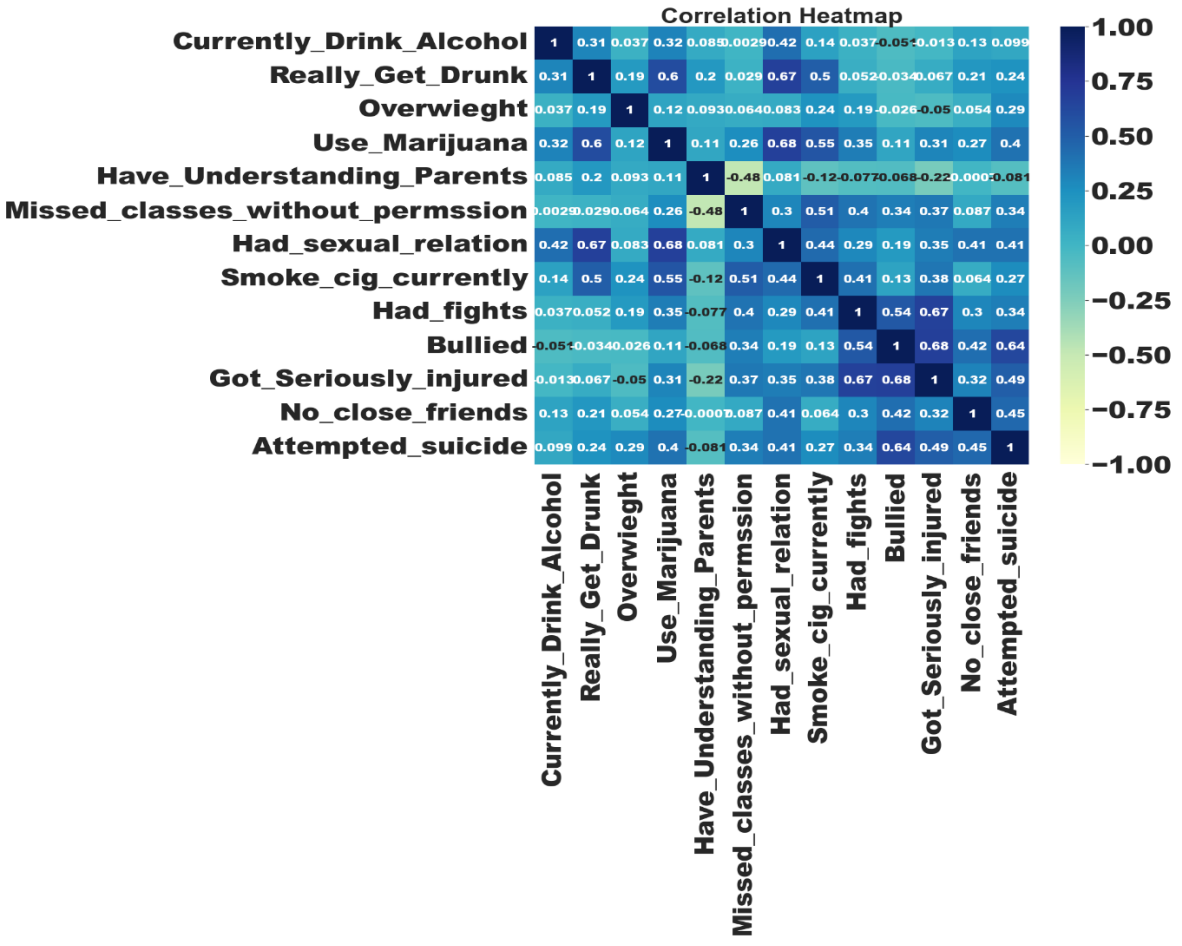

Fig. 3.1: Correlation Heatmap

-1: Perfect negative correlation. The variables tend to move in opposite directions when one variable increases, the other variable decreases.
0: No correlation between two variables

1: Perfect positive correlation. The variables tend to move in the same direction when one variable increases, the other variable also increases.

According to the results of this analysis, the correlations between these behaviors are:

- There is a strong possibility that when someone has sexual relations, they use marijuana.
- Someone who got bullied had a strong possibility of getting seriously injured.
- Someone who had sexual relations is possible to have really drunk behaviors.

But, in this case, we are going to conclude the behaviors correlation that is considered "Suicidal Behaviors".

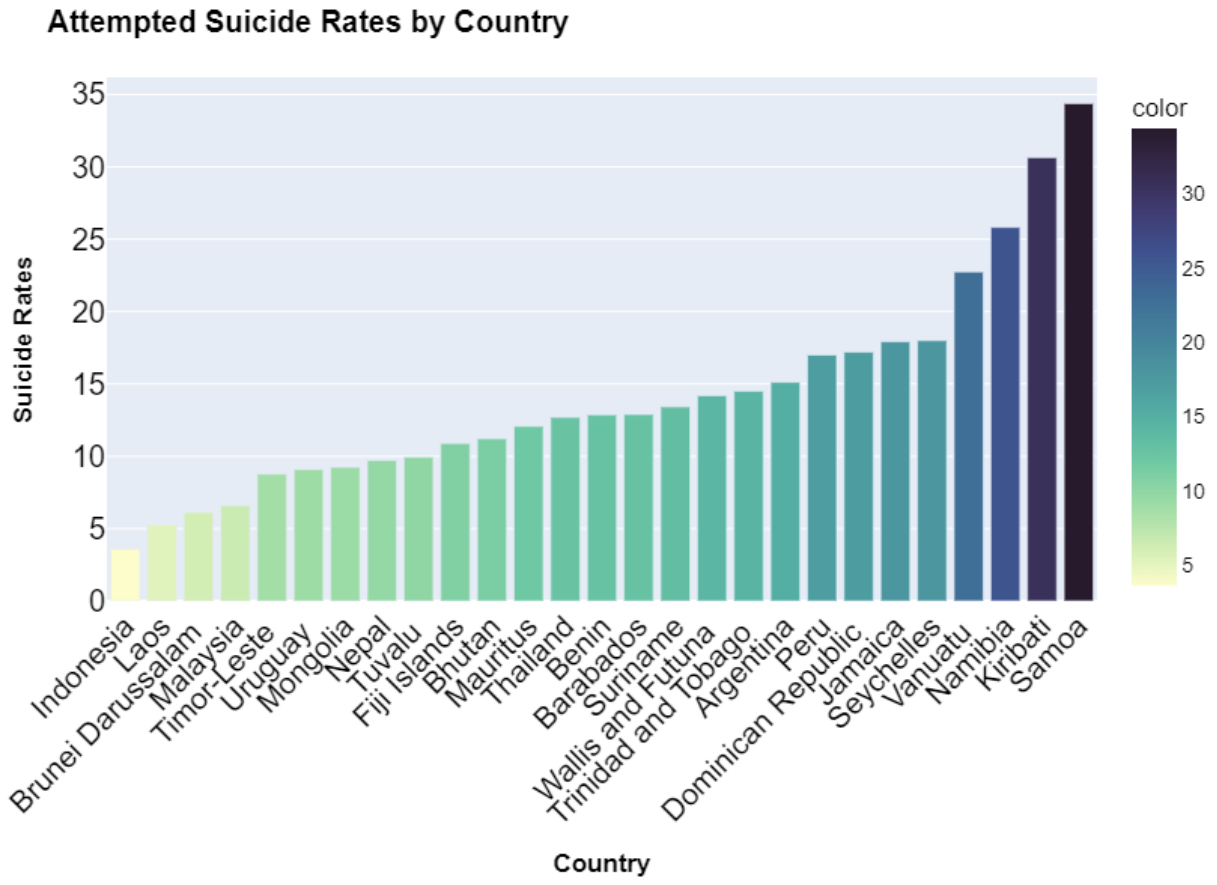

Fig. 3.2: Suicide Rates vs Country

According to studies, Indonesia has the lowest suicide rate, with a reported rate of 3.4833, while Samoa has the highest suicide rate at 34.3833 (as shown in Fig 2). Notably, the top four nations with the highest suicide rates are Samoa (34.3833), Kiribati (30.65), Namibia (25.825), and Vanuatu (22.75), all of which exhibit significantly higher rates compared to other countries. These findings underscore the alarming disparity in suicide rates among nations, highlighting the urgent need for targeted interventions and support systems to address mental health issues and mitigate the risk of suicide.

3.3 Data Preprocessing and Feature Engineering:

Using the Jupyter Notebook provided by the Python Navigator platform, several preprocessing processes were carried out to prepare the dataset for the training of various algorithms on it. These steps were taken to prepare the data for optimal analysis and training of the model. The first phase involved distinguishing between the data that was input and the data that was output. Input data consisted of the history of performance on 16 attributes for suicide attempters, as well as demographic information and their history concerning features which might create suicidal thought. Additionally, the data included their behavior. The output data showed the likelihood of attempting suicide at some point in the future. The dataset was organized into different input and output variables thanks to the separation of these components, which made future analysis much simpler. Some columns were eliminated so that the dataset may be improved in terms of both its quality and its usefulness. This stage was carried out to improve the values of various metrics and to refine the dataset as a whole. Eliminating columns that are superfluous or duplicative is an effective way to streamline the data and direct attention to the most vital aspects. Using label encoding, the alphabetical data were turned into numerical values. The method of label encoding involves giving categorical variables to their distinct numeric labels to render them understandable by computers. The data can then be used by machine learning algorithms, which only accept quantitative inputs after the categorical data have been converted into their numerical equivalents. Following this stage, you can rest assured that the regressions will be able to read

and analyze the data accurately. Following the completion of the data preprocessing step, feature engineering strategies were implemented to further improve the dataset before the training of the classification machine learning models. In the beginning, the demographic information, particularly the country, year, age group elements were left out of the equation. Excluding this information helped to streamline the dataset and minimize potential biases because the age range and years were generally comparable, and country can't be converted into numerical values. Additionally, the sex section was left out of the Jupyter Notebook as it is superfluous to the error calculation. It was important to handle null values in the any component while dealing with inputs. This approach was required to accomplish the task of error calculation. The dataset was modified to exclude null values in all components and assure its consistency. Because of this change, the machine learning models were able to discern between the attributes that cause suicidal thoughts to arise. The output variable was produced by taking into account the 12 different kinds of attributes and the effect that each of those attributes has on the occurrence of suicide. The dataset was divided into a training set and a test set. By dividing the data in this way, we were able to ensure that the models were trained on a specific section of the data and evaluated using the remaining data. The effectiveness of the models and their ability to generalize may be properly evaluated by testing them on data that they had not before encountered. The dataset was suitably prepared for the training of the regression machine learning models by conducting these preprocessing processes and feature engineering techniques. These methods ensured the quality of the data, made it compatible with algorithms, and made it suitable for accurate predictions of the likelihood of a suicide occurring based on the identified attributes.

3.4 Hyperparameter Optimization:

For hyperparameter optimization in all six models, the GridSearchCV technique was chosen over RandomizedSearchCV because to its superior performance. GridSearchCV is the best option for fine-tuning since, unlike the parameters that are automatically learned by the models, the hyperparameter values need to be manually modified for each dataset separately. GridSearchCV is a systematic approach that explores a broad range of parameter tunings and then evaluates the outcomes through cross-validation to identify the optimal settings that yield the best results. The

objective of this process is to find the hyperparameters that maximize the model's performance. GridSearchCV conducts a thorough search over a given set of hyperparameters, exhaustively considering all possible combinations in a grid-like structure. By employing such a comprehensive search strategy, one can confidently discover the optimal combination of hyperparameters. Through cross-validation testing on each potential input combination, GridSearchCV can determine the hyperparameters that lead to the highest level of model performance. It offers a reliable and efficient technique for hyperparameter optimization, enabling data scientists to fine-tune their models and achieve superior results in various tasks. However, RandomizedSearchCV adopts a fundamentally different approach. Over the distributions that have already been established for each hyperparameter, it conducts a random search. The RandomizedSearchCV algorithm randomly selects the combinations of hyperparameters to employ before training the model with those particular values. The search technique is repeated numerous times, and after each iteration, the model's performance is assessed to identify which set of hyperparameters delivers the best outcomes.

In contrast to RandomizedSearchCV, which only considers a limited number of these options, GridSearchCV explores every possible combination of random hyperparameters. Due to this important distinction, RandomizedSearchCV may occasionally be unable to reach the same level of optimization as GridSearchCV in terms of finding the hyperparameter combination that yields the best results for the given dataset. After considering the previously described regression dataset, the best option was determined to be GridSearchCV. For a thorough search of the dataset, there were enough computer resources and time available. GridSearchCV ensures that no potentially perfect hyperparameter combination will be missed due to its exhaustive nature. GridSearchCV extensively and comprehensively searches the whole hyperparameter space, giving confidence in finding the biggest potential hyperparameters for increasing model performance. On the other hand, it's important to note that using RandomizedSearchCV would have been a better option if there had been a limit on the length of time or the number of computational resources available. When a thorough search is impractical, RandomizedSearchCV performs a random sampling of possible hyperparameter combinations. This method requires less work and may produce useful results. In conclusion, GridSearchCV is the method of choice for datasets with ample computational time and resources since it

guarantees the identification of the ideal collection of hyperparameters. However, when time and computer resources are limited, RandomizedSearchCV is a better option. This is due to the fact that it just looks at a portion of all hyperparameter combinations in order to approximate the best answer.

3.5 Regression Models:

The regression models subcategory of machine learning algorithms plays a crucial role in the field by enabling the prediction of continuous outcomes. These models analyze the input data and estimate a numerical value based on the underlying patterns and relationships within the data. In machine learning, the primary objective of regression models is to accurately predict unknown values for a given set of features.

Regression models are extensively employed in various domains, such as finance, economics, healthcare, and engineering, where understanding and forecasting numerical quantities are essential. These models can capture complex associations between input variables and output values, enabling the exploration of intricate data patterns and the prediction of continuous outcomes with reasonable accuracy.

There exists a wide range of regression algorithms, each with its own strengths and weaknesses. The choice of a regression model depends on several factors, including the nature of the problem, the characteristics of the dataset, and the desired interpretability of the model.

3.5.1 Random Forest Classifier (RF):

The concept of random forest is expanded from classification to regression issues using the potent machine learning technique known as random forest regression. Random forest regression models are made to predict continuous numerical values, whereas random forest classifiers are used to predict outcomes that fall into one of several categories. The main features and uses of random forest regression will be

covered in this article. Like its classification counterpart, random forest regression is an ensemble approach and supervised learning algorithm. To reach the final regression prediction, it combines the predictions of various decision trees. A portion of the data, randomly picked with replacement from the original dataset, is used to train each decision tree. Several decision trees are trained as part of the ensemble-building process, and their predictions are then combined to produce the final regression result.

Three basic parts make up a decision tree in random forest regression: the root node, decision nodes (also called internal nodes), and leaf nodes (also called terminal nodes). The decision nodes of the decision tree start at the root node and base their binary decisions on the input features. The data is divided into two subsets at each decision node according to a certain characteristic or attribute, and this procedure is repeated recursively until the data reaches the leaf nodes.

The final predictions are represented by the leaf nodes in a random forest regression model. The regression prediction for a specific data point is represented numerically in each leaf node. The final prediction in random forest regression is calculated by taking the mean or average value of the forecasts from several decision trees included in the ensemble. This procedure of averaging aids in decreasing volatility and boosting forecast stability.

The capacity of random forest regression to accommodate non-linear correlations between input factors and the target variable is one of its significant benefits. In order to make more precise predictions, the ensemble of decision trees captures complex interactions and patterns in the data. Random forest regression models are suited for real-world datasets that could contain inaccurate or noisy observations because they are resistant against outliers and data noise.

The number of decision trees in the ensemble, their depth, and the amount of input features taken into account at each split point are a few variables that might affect how accurate random forest regression is. Up until a certain point of diminishing returns, adding more decision trees to the ensemble tends to enhance the accuracy. By increasing diversity and capturing various facets of the data, additional trees can decrease overfitting and enhance generalization.

Numerous fields, including finance, economics, healthcare, and environmental research, find extensive use for random forest regression models. They can be applied to modeling environmental variables, forecasting product demand, calculating property prices, and predicting stock prices. Random forest regression has a lesser interpretability than linear regression models frequently, but it is more flexible in capturing complicated correlations and managing high-dimensional data.[19]

3.5.2 Linear Regression (LR):

For predicting continuous numerical values, linear regression is a commonly used statistical technique in machine learning and data analysis. For its simplicity, interpretability, and capacity to capture linear correlations between input variables and the target variable, it serves as the foundation for numerous prediction models. We shall discuss the idea of linear regression, its essential components, and its various applications in this post. Linear regression is a supervised learning algorithm that aims to establish a linear relationship between one or more independent variables (also known as features or predictors) and a dependent variable (the target variable). It assumes that the relationship between the variables can be represented by a straight line in a multidimensional space. The objective is to find the best-fitting line that minimizes the difference between the predicted values and the actual values.

The regression equation and the error term are the two key elements of the linear regression model. The linear relationship between the input variables and the target variable is represented by the regression equation. The interpretability of linear regression is among its many noteworthy benefits. Understanding the amount and direction of the correlations between the variables is possible thanks to the coefficients. A positive coefficient denotes a positive correlation, whereas a negative coefficient denotes a negative correlation. The higher the correlation between an independent variable's influence and the target variable, the larger the coefficient.

The use of linear regression is widespread across many different fields. It is used in finance to predict stock prices, compute asset returns, and identify potential risks.

For modeling demand and supply, forecasting economic indicators, and analyzing the results of policy changes, it is helpful in economics. Healthcare professionals employ linear regression to identify illness risk factors, assess the efficacy of therapies, and predict patient outcomes. In addition, many other industries, such as engineering, marketing research, and social sciences, can benefit from using linear regression.

Although simple and easy to comprehend, linear regression has significant disadvantages. The variables are assumed to be linearly connected, which may not necessarily be the case in intricate real-world scenarios. Additionally, it is subject to outliers and might be affected by the presence of important data points. Additionally, linear regression could have trouble identifying nonlinear correlations or interactions between variables.

In order to get over these restrictions and capture complex correlations, researchers have recently created more sophisticated regression approaches. Nevertheless, due to its simplicity in usage, readability, and applicability for situations involving linear correlations, linear regression continues to be a fundamental and often employed tool.

In conclusion, linear regression is a strong and adaptable predictive modeling method. Making predictions and gaining understanding of the variables that influence the target variable are all made possible by it. We can also discover and quantify the correlations between variables. Even though it has some drawbacks, its clarity and ease of use make it a valuable tool in many different sectors, helping researchers, analysts, and decision-makers comprehend and forecast continuous numerical outcomes.[20]

3.5.3 Support Vector Regression (SVR):

Support Vector Regression (SVR) is an effective supervised learning method that enhances Support Vector Machines' (SVM) capacity to address regression issues. SVR uses the same fundamental concepts to estimate continuous target

variables rather of discrete class labels, despite the fact that SVM is well known for its efficiency in classification problems. The kernel trick, margin, and hyperplanes are all included in SVR, which offers a strong and adaptable framework for dealing with regression problems.

Regression and Supervised Learning: SVR and SVM both fall under the umbrella of supervised learning, which involves training a model using labeled data to generate predictions. SVR, on the other hand, focuses on estimating a continuous target variable rather than forecasting discrete class labels. Given that the objective of a regression problem is to predict numerical values rather than class memberships, SVR can be used to solve these types of problems.

Regression and hyperplanes: SVR uses the idea of hyperplanes to depict the connections between the input variables and the target variable. A line or hyperplane (in higher dimensions) that best matches the data can be used to represent a hyperplane in regression tasks. Finding an ideal hyperplane that maximizes the margin while minimizing errors or departures from the genuine values is the goal of SVR.

Margin and Robustness: SVR strives to maximize the margin between the training data points and the hyperplane, similar to how SVM does. The margin denotes the area where the model may tolerate certain deviations or inaccuracies. SVR achieves higher generalization and robustness to manage noise or outliers in the data by allowing some points to fall within the margin. With the help of this margin control technique, the model is better able to generalize to new data and avoid overfitting.

Non-linear correlations between the input variables and the target variable are handled by SVR, like SVM, using the kernel approach. The input data can be implicitly mapped into a higher-dimensional feature space by the kernel function, allowing SVR to effectively segregate the changed data using a linear hyperplane. With the help of this method, SVR is able to capture complex non-linear patterns without the requirement for explicit feature engineering.

SVR supports several different types of kernel functions, including sigmoid, linear, polynomial, and radial basis function (RBF) kernels. The particular problem and the

properties of the data will determine the kernel to use. For instance, the RBF kernel is suitable for capturing non-linear and complicated patterns, but the linear kernel is beneficial when the connection between the input variables and the target variable is anticipated to be linear.

Efficiency and Computational Complexity: Even in high-dimensional areas, SVR is intended to provide a computationally efficient learning method. In order to choose the optimum hyperplane that minimizes the errors, it makes use of optimization strategies and linear algebraic operations. The kernel approach additionally lessens the computing load by enabling SVR to function in the modified feature space without explicitly mapping each data point. Because of these efficiency factors, SVR is a good choice for complex regression issues.

SVR provides a number of hyperparameters that can be tweaked in order to enhance performance. The regularization parameter, commonly abbreviated as C, regulates the trade-off between minimizing errors and obtaining a narrow margin. A greater value of C results in a smaller margin but maybe lesser training data mistakes. To accurately capture the proper level of complexity in the data, the kernel parameters, such as the degree in polynomial kernels or the bandwidth in RBF kernels, can be changed.

Benefits of SVR: SVR has a number of benefits that contribute to its acceptance and efficiency in regression tasks. First off, it uses kernel functions to handle non-linear data and capture intricate relationships between the input variables and the target variable. Second, SVR demonstrates robustness by tolerating errors and effectively handling outliers.[21]

3.5.4 K- Nearest Neighbors (KNN):

K-nearest-neighbor regression, sometimes referred to as KNN regression, is a basic and uncomplicated regression technique applied in supervised learning. As it does not make any assumptions about the distribution of the underlying data, it falls within the category of non-parametric regression. KNN regression attempts to estimate the target value of incoming data points by analyzing how closely they

resemble the training dataset. By averaging the goal values of the K closest neighbors, KNN regression estimates the target values of the new points rather than classifying them. The Euclidean distance is frequently used to calculate how close a new point is to previously recorded data points. The selection of K, or the number of nearby sites to take into account, is essential in determining the accuracy of the forecasts. To avoid ties in voting, it is frequently advised to choose an odd value for K. KNN regression provides a straightforward yet effective method for predicting target values for new instances by utilizing the idea of similarity and neighbor-based prediction. The Euclidean distance is written as

$$Euclidean\ Distance = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

Here are some further details:

Non-parametric Regression and Supervised Learning: The KNN regression technique uses supervised learning to create predictions using labeled training data. It is a non-parametric regression technique, similar to KNN classification, that does not make any assumptions about the data distribution or teach explicit parameters during training.

The KNN regression algorithm predicts the target value of a new data point by evaluating how similar or unlike its features are to those of the training data. The target value of the new point in the feature space is determined by averaging the target values of its K closest neighbors. By taking into account the neighbors' target values and combining them with a weighted average or another appropriate aggregation approach, the forecast is made.

Euclidean Distance and Distance Metrics: The Euclidean distance is the most widely used distance metric in KNN. It determines the straight-line distance in the feature space between any two places. Depending on the type of data and the issue at hand, various distance measures, such as Manhattan distance, Makowski distance, or cosine similarity, can also be utilized. The idea of proximity between data points is impacted by the chosen distance metric.

The value of K in a KNN defines how many surrounding points are taken into account when producing a prediction. Choosing an acceptable value for K is crucial. The model could be sensitive to noise and outliers if K is set too low. If K is too big, the model could lose local patterns and oversimplify the decision boundary. To prevent tied voting, where many classes have an identical number of votes among their neighbors, it is frequently advised to use an odd value for K.KNN is renowned for being straightforward and simple to use. Both binary and multi-class regression issues can benefit from it. KNN must calculate distances between the new point and each training point, which can be computationally expensive when working with huge datasets.[22]

3.5.5 XGBoost Regression (XGB):

XGBoost has become a very powerful algorithm in the field of machine learning for a variety of problems, including both classification and regression. While the fundamentals of XGBoost have been covered in relation to classification, this article will focus on XGBoost regression and examine how this potent ensemble learning method can be used to address regression-related issues. The gradient boosting framework is tailored for use in regression tasks and is used in XGBoost regression. To produce incredibly accurate predictions on continuous target variables, it combines weak learners, often decision trees. XGBoost regression delivers greater efficiency, scalability, and regularization abilities by taking advantage of gradient boosting's advantages.

Combining Weak Learners:
XGBoost regression, like XGBoost classification, combines a number of weak learners to produce a strong learner. Regression decision trees with low individual predictive power are referred to be weak learners. However, XGBoost regression creates a more precise and reliable model that can capture complicated correlations in the data by integrating the predictions of these weak learners.

Sequential Learning with Feedback:

XGBoost regression uses the gradient boosting approach, adding weak learners to the ensemble in a sequential fashion so they can learn from the errors of the preceding models. Each decision tree aims to minimize the ensemble's residual errors, which indicate the discrepancies between the true and projected values. The prediction performance of the XGBoost regression is steadily enhanced by iteratively updating the model based on the residuals.

Gradient optimization and the objective function:
In XGBoost regression, the objective function is a key component in assessing the model's efficacy. The objective function, which is frequently expressed as a loss function, calculates the difference between the true values and the anticipated values. By minimizing the loss and lowering the residuals throughout the training process, XGBoost regression achieves the optimal performance for this objective function. With this iterative process, the model may continuously improve its predictions and move closer to the ideal outcome.

Training Stopping Criteria:
XGBoost regression uses stopping criteria to decide when to terminate the training process and choose the final model, just like XGBoost classification. The intended degree of residual error or the number of trees might serve as the basis for the halting criteria. The training procedure is terminated, and the current model is selected as the final one if the number of decision trees falls below a predetermined threshold before the residual error reaches the desired level.

Regularization Strategies:
Overfitting, a frequent problem in regression tasks, is when the model gets too complicated and struggles to generalize to new data. This problem is addressed by XGBoost regression, which uses regularization methods. By imposing penalties or limitations on the objective function, regularization reduces the model's complexity. Regularization increases the generalizability and performance of the model by preventing the construction of excessively complicated trees.

Scalability through Parallel Execution:

XGBoost regression uses parallel processing, much like XGBoost classification, to ensure efficiency and scalability. To handle massive datasets and drastically shorten training durations, it makes use of parallelism at different levels, including remote computation, parallel tree construction, and column blocking. Because of this capabilities, XGBoost regression can be used in practical situations where it is necessary to process large volumes of data quickly.

Importance of Features:
Analyzing the value of various input features is crucial for interpreting the model and deriving insightful conclusions. A measure of feature relevance is provided by XGBoost regression, which ranks the features according to how much they helped to lower the loss function during training. Data scientists can decide on feature selection, feature engineering, and further data exploration by examining the value of the features and determining the most important variables.[23]

3.5.6 Ridge Regression (RR):

Ridge regression is a useful improvement to linear regression that tackles some of its drawbacks and improves its ability to predict. It is a regularization method that is particularly helpful when working with high-dimensional datasets because it aids in overcoming problems like multicollinearity and overfitting. We shall examine the idea of ridge regression, its salient characteristics, and its applications in a variety of domains in this post.

Ridge regression, sometimes referred to as Tikhonov regularization, reduces the coefficient estimates by adding a penalty term to the linear regression model. Ridge regression sets a constraint on the coefficients, preventing them from growing too large, by including a regularization term to the objective function of ordinary least squares (OLS).

Ridge regression is useful in many different fields, especially when analyzing datasets with a lot of variables. It is employed in finance for risk management, portfolio optimization, and asset return forecasting. Ridge regression is a technique

used in genetics to find disease-related genetic markers. Along with other areas, it is used in bioinformatics, image processing, and text mining.

Ridge regression has a number of benefits; however it should be noted that in comparison to conventional linear regression, it sacrifices some interpretability. Since the coefficient estimates are approaching zero, it is difficult to interpret their magnitudes directly. Ridge regression is still a useful tool for predictive modeling, and it can be used in conjunction with other approaches to extract key variables and improve interpretability, such as feature selection methods.

Ridge regression is a useful regularization method that enhances the capabilities of linear regression, in conclusion. It solves multicollinearity and overfitting concerns by including a penalty element in the objective function, resulting in more stable and generalizable models. It is used in many different domains, especially when working with high-dimensional datasets. Ridge regression delivers better predictive performance while sacrificing some interpretability, and it helps to draw out important information from complex data. Ridge regression is a useful tool in the data scientist's arsenal because it strikes a balance between simplicity and complexity.[24]

3.5.7 Multilayer Perceptron (MLP):

The Multilayer Perceptron (MLP) algorithm is a well-liked and effective machine learning technique for tackling challenging pattern recognition issues. It is a form of artificial neural network that has several interconnected layers, allowing it to learn and represent complex correlations in the data. In this article, we'll examine the major characteristics and uses of the multilayer perceptron and learn how important it is for dealing with a variety of contemporary problems.

Similar to other neural network architectures, Multilayer Perceptron is inspired by the structure and functioning of the human brain. It is designed to simulate the behavior of interconnected neurons, where each neuron performs a weighted sum of its inputs, applies an activation function, and passes the result to the next layer. By

combining multiple layers, MLP can learn hierarchical representations of the data, allowing it to capture complex patterns and correlations.

An input layer, one or more hidden layers, and an output layer make up the architecture of a multilayer perceptron. The final prediction or classification is provided by the output layer, which also gets the features or attributes of the data. The input data is extracted and transformed by the hidden layers, which are positioned in between the input and output layers, in order to recognize and represent the underlying patterns.

Forward propagation and backpropagation are the two main processes in the Multilayer Perceptron learning process. Forward propagation involves sending input data through the network while computing each neuron's activations up until it reaches the output layer. The algorithm then calculates the error between the expected and actual values during backpropagation and modifies the connection weights to reduce this error. The model goes through this iterative process again and again until it reaches the necessary degree of accuracy.

One of the key advantages of Multilayer Perceptron is its ability to learn non-linear relationships between input features and target variables. Unlike traditional linear models, MLP can model complex and non-linear patterns, making it suitable for a wide range of tasks such as image recognition, natural language processing, speech recognition, and financial forecasting. Its flexibility allows it to handle high-dimensional data and capture intricate correlations that may not be apparent through simple statistical methods.

The Multilayer Perceptron's resistance to erratic data and outliers is another benefit. MLP may generalize successfully even in the presence of noisy or imperfect data by learning from a large number of examples and utilizing the collective wisdom of the network. Because of this, it is especially helpful in practical situations where data quality may be at risk.

However, Multilayer Perceptron also has some considerations. It requires a sufficient amount of labeled training data to learn effectively and avoid overfitting. Additionally, determining the appropriate architecture, including the number of

hidden layers and neurons per layer, is a crucial aspect of building an optimal MLP model. Choosing the right activation functions and regularization techniques is also important to enhance performance and prevent overfitting.

As a strong machine learning method, Multilayer Perceptron can successfully tackle challenging pattern recognition problems. It is an effective tool in many disciplines due to its capacity to learn non-linear correlations, handle high-dimensional data, and robustness against noise. Researchers and practitioners can solve difficult challenges in fields like image processing, natural language interpretation, and financial analysis by utilizing the hierarchical structure and learning capabilities of MLP. To guarantee top performance and prevent overfitting, however, significant consideration should be given to the architecture and training procedure.

3.6 Work Flow Diagram:

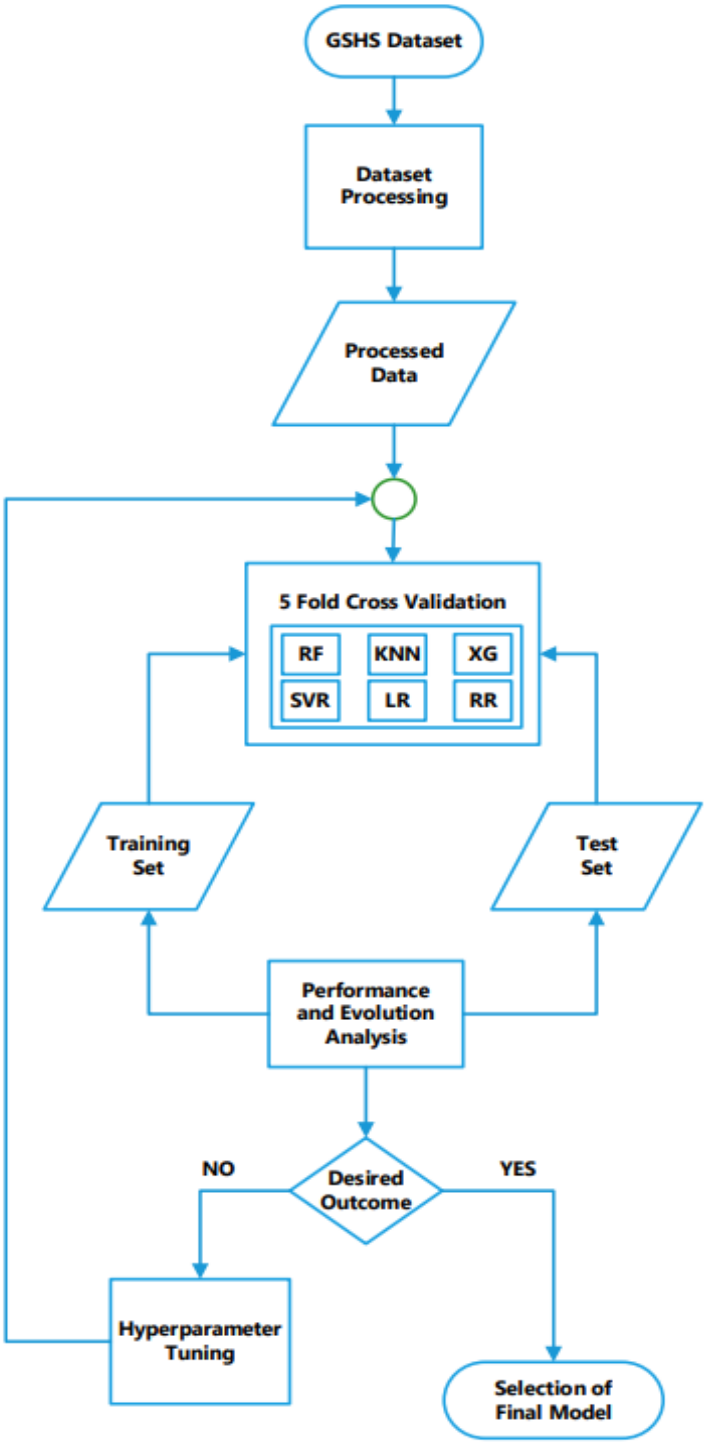

Figure 3.3: Workflow diagram

CHAPTER 4

RESULT

4.1 Performance parameters:

Performance Parameters for Regression Models: Evaluating Predictive Accuracy. When working with regression models, performance parameters are critical in determining the efficacy and accuracy of the predictions. In contrast to classification tasks, where the outputs are classified, regression tasks involve predicting continuous numerical values. In this post, we will look at performance parameters designed expressly for regression models, allowing us to quantify their predictive potential.

Popular performance metrics for regression models include Mean Squared Error (MSE). MSE is a metric that assesses the average squared deviation between expected and actual data. Lower MSE values imply better projected accuracy, and it quantifies how well the model accounts for observed variability. On the other hand, the MSE metric gives more weight to higher errors and is susceptible to outliers.

Another often employed performance indicator is the Root Mean Squared Error (RMSE), which is the MSE's square root. The RMSE is easier to interpret because it is in the same unit as the target variable. By computing the square root, which may be helpful in some circumstances, RMSE penalizes larger errors more severely than smaller errors.

Furthermore, regression programs commonly use the Mean Absolute Error (MAE) performance parameter. MAE is used to determine the average absolute difference between the expected and actual values. Regardless of their direction, errors' average magnitude is calculated. MAE may not capture data variability as well as MSE, although it is less sensitive to outliers than MSE.

R-squared (R2) is another common performance statistic for regression models. It determines how much of the variance of the target variable can be accounted for by the independent variables. R2 is a number between 0 and 1, with larger numbers indicating that the model fits the data more closely. When applied to complex models or datasets with a lot of noise, R2 might, however, be misleading.

Two other performance metrics that are frequently used in regression tasks are Mean Absolute Percentage Error (MAPE), which assesses the average percentage difference between predicted and actual values, and Root Mean Squared Logarithmic Error (RMSLE), which applies a logarithmic transformation to both predicted and actual values before calculating the RMSE.

Regression model performance parameters must be selected after careful consideration of the task's unique objectives and goals. Depending on the nature of the issue, the accuracy of the data, and the objectives of the modeling project, several performance parameters may be selected. A balance must be struck between recording precision, gathering variability, and taking care of any potential biases or sensitivities in the data.

Finally, regression model performance parameters provide quantifiable measurements of predicted accuracy. Metrics for assessing regression model performance include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2). Alternative viewpoints are provided by additional parameters such as MAPE and RMSLE. Choosing the right performance parameters is critical for accurately evaluating and comparing regression models, guiding model modifications, and attaining the desired prediction results. The following are the performance parameters employed in this study, along with their related mathematical equations:

4.1.1 Mean Absolute Error (MAE):

Mean Absolute Error (MAE) is a frequently used performance indicator when assessing the efficiency of regression models. The average magnitude of errors between the expected and actual values is measured intuitively by MAE. The idea

of MAE, its calculation, and its importance in determining the correctness of regression models will all be covered in this article.

The absolute disparities between the anticipated and actual values are averaged to determine the Mean Absolute Error (MAE). You can express the MAE formula as:

$$MAE = \frac{\sum_{i=1}^{n} | y_i - x_i |}{n}$$

Where:
MAE represents the Mean Absolute Error.

- n is the total number of data points.
- i represents each individual data point.
- |i - yi| represents the absolute difference between the predicted value (i) and the actual value (yi).

In order to get the average distance between the projected and actual values, MAE adds all the absolute differences and divides them by the total number of data points. The absolute value makes sure that errors are handled equally for both positive and negative deviations regardless of their direction.

Because MAE can be interpreted in the same unit as the target variable, it is very helpful. The MAE will be expressed in dollars, for instance, if the target variable indicates a product's price in dollars. This makes it simpler to comprehend the size of the faults and contrast them with the scope of the issue that needs to be handled.

The main advantage of MAE over other performance indicators like Mean Squared Error (MSE) is that it is less susceptible to outliers. Regardless of their size, all errors are given the same weight by MAE. The total metric won't be significantly impacted by a single huge inaccuracy as a result, making MAE more resilient in the presence of extreme values or noisy data.

MAE does not distinguish between overestimation and underestimation, which is one of its limitations. Regardless of the direction, each inaccuracy adds the same

amount to the total measure. Depending on the circumstances, this might work to your favor or disadvantage. In some circumstances, it could be more crucial to concentrate on the size of errors rather than their direction.

Another consideration when using MAE is that it treats all errors equally. If certain data points or regions of the dataset are more critical than others, it might be necessary to assign different weights to the errors or focus on specific subsets of the data. Domain knowledge and the specific goals of the regression task should be taken into account when interpreting MAE results.

As a result, Mean Absolute Error (MAE) is a frequently employed performance indicator for assessing the precision of regression models. Its calculation determines the average absolute difference between the expected and actual values. A readable measure of the average magnitude of errors is provided by MAE, which also offers robustness against outliers. It does not, however, distinguish between overestimation and underestimation. To have a thorough grasp of the regression model's accuracy, it's critical to take the context, the specific objectives, and possibly combine MAE with other performance measures into account when evaluating model performance.

4.1.2 Mean Absolute Percentage Error (MAPE):

Mean Absolute Percentage Error (MAPE) is a frequently used performance indicator for assessing the performance of regression models. The average percentage difference between the expected and actual values is measured by MAPE. The idea of MAPE, its formula, and its importance in determining the correctness of regression models will all be covered in this article.

By averaging the absolute percentage discrepancies between the predicted and actual values, the Mean Absolute % Error (MAPE) is determined. One way to express the MAPE formula is as follows:

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

Where:

MAPE represents the Mean Absolute Percentage Error.

- n is the total number of data points.
- i represents each individual data point.
- i is the predicted value, and yi is the actual value.

The MAPE method yields a measurement of the average % deviation between the expected and actual values by computing the absolute percentage difference for each data point, adding them up, and dividing them by the total number of data points. The absolute value makes sure that errors are handled purely based on the relative error's magnitude and not their direction.

Because MAPE expresses the error as a percentage of the true number, it is very helpful. This makes comparisons between different scales of the target variable easier and improves understanding of the relative magnitude of the errors. In evaluating the precision of regression models, it gives a sense of the average amount of error in proportion to the actual values.

The fact that MAPE gives errors the same weight as true values is one of its benefits. This guarantees that the dataset's large or small values will not cause the measure to be biased. Since each data point contributes equitably to the total MAPE, the average percentage variation is fairly represented.

The sensitivity of MAPE to actual values that are zero or almost zero is one of its drawbacks, though. Since the real value serves as the denominator in the percentage computation, the metric may become unstable or undefinable when the actual value is in close proximity to zero. When this occurs, it is crucial to handle these situations carefully, whether by replacing the data points in question with a small constant value or omitting them entirely from the calculation.

Additionally, due to the percentage computation, MAPE could exaggerate significant inaccuracies. Particularly when dealing with extreme numbers or outliers, a single substantial inaccuracy can significantly affect the whole metric. When interpreting the results of the MAPE, it is important to keep in mind this sensitivity to significant mistakes.

In actual practice, it's critical to interpret MAPE in light of the particular issue at hand as well as the objectives of the regression work. For a more thorough analysis, alternative metrics or additional performance characteristics may be utilized in addition to MAPE, depending on the domain and the importance of certain faults.

Finally, Mean Absolute Percentage Error (MAPE) is a performance indicator that is frequently used to assess the efficacy of regression models. The average percentage difference between the expected and actual values is calculated using its algorithm. MAPE enables comparisons between scales and offers insight into the typical extent of relative errors. It may overemphasize significant errors and be sensitive to actual values that are 0 or nearly zero. Understanding the context, the specific aims, and sometimes employing complementing measures are crucial when reading MAPE in order to fully comprehend the precision of the regression model.

4.1.3 Root Mean Square Error (RMSE):

Root Mean Square Error (RMSE) is a performance indicator that is frequently used to assess the efficacy of regression models. Taking into account both the direction and size of the deviations, RMSE provides a measure of the average magnitude of errors between the expected and actual values. The idea of RMSE, its calculation, and its importance in determining the correctness of regression models will all be covered in this article.

The square root of the differences between the expected and actual values is used to determine the Root Mean Square Error (RMSE). One way to express the RMSE formula is as follows:

$$\text{RMSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Where:
RMSE represents the Root Mean Square Error.

- n is the total number of data points.
- i represents each individual data point.

- i is the predicted value, and yi is the actual value.

The RMSE offers a measurement of the average size of errors by computing the squared difference for each data point, adding them up, dividing them by the total number of data points, and taking the square root. Positive and negative deviations are penalized equally by the squared values, which also ensure that larger errors contribute more to the final measure.

Because it provides a measurement of the typical error in the same unit as the goal variable, RMSE is particularly helpful. This makes it simpler to interpret and contrast the size of errors with the scope of the issue being solved. The RMSE will be expressed in dollars, for instance, if the target variable indicates a product's price in dollars.

The fact that RMSE offers a thorough evaluation of the overall model performance is one benefit of the method. It captures the variability and dispersion of the predictions relative to the actual values by accounting for both the direction and magnitude of mistakes. A smaller average deviation and a lower RMSE reflect a more accurate model.

RMSE is, however, susceptible to outliers or extreme values. Large errors will have an outsized effect on the total metric because the squared differences are used to calculate the metric. When analyzing RMSE results, it is important to take into account this sensitivity to outliers, particularly when the evaluation may be significantly impacted by extreme numbers.

Another thing to keep in mind when using RMSE is that it places more emphasis on larger errors than smaller ones. This means that RMSE might not be appropriate in situations where minor errors are more important or require more consideration. In these situations, other performance characteristics or alternative metrics can be employed to round out the evaluation.

Due to its ease of use and interpretability, RMSE is widely used in practice. It enables a simple evaluation of the average error magnitude in relation to the target variable. The specific situation at hand and the objectives of the regression job must

be taken into consideration when interpreting the RMSE results, though. A more thorough assessment of the model's correctness can result from comprehending the implications of the statistic and taking into account other pertinent metrics.

Root Mean Square Error (RMSE) is a frequently employed performance indicator for assessing the precision of regression models. The square root of the mean squared discrepancies between the expected and actual values is computed using this formula. With regard to the target variable's same-unit unit, RMSE sheds light on the average magnitude of mistakes. It captures both the quantity and direction of mistakes, but it also emphasizes greater errors and is sensitive to outliers. To fully comprehend the correctness of the regression model, it is imperative when analyzing RMSE to take the context into account and possibly use other metrics.

TABLE II: Result Parameters

ML Algorithms	MAE	RMSE	MAPE
RF	**5.4808**	**5.4810**	**0.9363**
XGB	**5.5213**	**5.5210**	**0.6757**
RR	**7.2919**	**7.2920**	**0.9156**
LR	**7.1746**	**7.1764**	**0.9679**
KNN	**6.4200**	**6.4200**	**1.0835**
SVR	**4.6588**	**4.6589**	**0.6561**
MLP	**6.5090**	**6.0590**	**0.7994**

4.2 Analysis of the Result:

In this result analysis, we will examine the performance of several machine learning algorithms on a regression type dataset based on three evaluation metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). The dataset includes results for the following algorithms: Random Forest (RF), XGBoost (XGB), Ridge Regression (RR), Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Regression (SVR), and Multilayer Perceptron (MLP).

First, let's examine the MAE scores. Without respect to the direction of the errors, the MAE calculates the average size of the errors caused by the models. The model performs better in terms of accuracy the lower the MAE value. SVR has the lowest MAE score (4.6588) among the algorithms, demonstrating its capacity to predict outcomes more accurately and with fewer errors. With MAE values of 5.4808 and 5.5213, respectively, RF and XGB closely followed. This shows that these models did a good job of minimizing mistakes as well.

The models will then be evaluated using the RMSE method, which takes both the size and the direction of errors into account. It gives an indication of how well the model fits the data. SVR once more received the lowest RMSE score (4.6589), demonstrating its better effectiveness in identifying the underlying trends in the data. The RMSE values of 5.4810 and 5.5210, respectively, produced by RF and XGB show that their forecasts were also reasonably accurate.

We may learn more about the relative effectiveness of the models in terms of accuracy on a percentage scale by using MAPE, which assesses the average percentage difference between the anticipated and actual values. SVR had the lowest MAPE score (0.6561), which demonstrated its capacity to predict outcomes with a lower percentage error. XGB did well, earning a MAPE score of 0.6757, while RF came in second with 0.9363. These results indicate that the best algorithms for correctly forecasting the target variable are SVR, XGB, and RF.

SVR regularly beat other algorithms in terms of accuracy and precision, according to the examination of these three assessment metrics. It has the strongest predictive abilities on the provided regression dataset, as seen by the fact that it received the lowest MAE, RMSE, and MAPE scores. In terms of all criteria, RF and XGB also did well, showcasing their propensity for making precise predictions.

It is crucial to remember that the effectiveness of machine learning algorithms can vary based on the particular dataset and the type of issue at hand. To ensure robustness and generalizability, it is advised to further assess and contrast the models using other metrics and methodologies, such as cross-validation or feature significance analysis.

In conclusion, based on the provided results, SVR, RF, and XGB emerged as the top-performing algorithms for the given regression type dataset. These models showcased their ability to make accurate predictions with lower errors and smaller percentage differences. It is essential to consider these findings when selecting the most suitable algorithm for similar regression tasks and to further explore their performance characteristics on different datasets to ensure optimal results.

CHAPTER 5

CONCLUSION

"Suicide prevention still requires a lot of work in today's culture. Suicide prediction has been the subject of a number of research, which has shown that machine learning algorithms have the ability to recognize and predict suicidal thoughts and behaviors. These algorithms have shown encouraging results, proving they are efficient at detecting people who are at risk. Despite the development of numerous high accuracy models, there is still potential for development and advancement. It is hoped that further research and development in this area will lead to the development of more complex and precise predictive models, which will help to lower suicide rates and improve mental health services.

When determining a person's likelihood of participating in suicidal behavior, machine learning algorithms have clear advantages. These algorithms provide a thorough and objective review by analyzing numerous data sources, considering various risk factors and their interconnections. However, it is essential to use machine learning algorithms carefully and in conjunction with experts in mental health and those who have first-hand knowledge. The algorithms are applied responsibly and with regard for the complexity of human behavior and mental health thanks to this collaborative approach. Suicide prevention initiatives can be greatly improved by combining the power of machine learning with the knowledge of specialists.

It is impossible to exaggerate the potential of machine learning to have a big impact on suicide prevention. These algorithms have the potential to save lives and lead to better mental health outcomes with additional research and development. Machine learning can aid in the development of a more thorough and pro-active strategy for preventing suicide by iteratively improving the models, including fresh data sources, and taking ethical issues into account. We can significantly reduce the tragic effects of suicide on people and communities by combining the powers of machine learning and human skill.

REFERENCES

[1] J. C. Franklin et al., ―Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research, ‖ Psychol Bull, vol. 143, no. 2, p. 187, 2017.

[2] D. Liu et al., ―Suicidal ideation cause extraction from social texts, ‖ Ieee Access, vol. 8, pp. 169333–169351, 2020.

[3] A. N. Weber, M. Michail, A. Thompson, and J. G. Fiedorowicz, ―Psychiatric emergencies: assessing and managing suicidal ideation, ‖ Medical Clinics, vol. 101, no. 3, pp. 553–571, 2017.

[4] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, ―Suicidal ideation detection: A review of machine learning methods and applications, ‖ IEEE Trans Comput Soc Syst, vol. 8, no. 1, pp. 214–226, 2020.

[5] R. Mahadik, S. Salunkhe, P. Sneha, and V. Sagvekar, ―Analysis of Suicide Attempts and Its Prediction, ‖ 2021.

[6] M. J. Vioules, B. Moulahi, J. Azé, and S. Bringay, ―Detection of suicide-related posts in Twitter data streams, ‖ IBM J Res Dev, vol. 62, no. 1, pp. 1–7, 2018.

[7] R. C. O'Connor and M. K. Nock, ―The psychology of suicidal behavior, ‖ Lancet Psychiatry, vol. 1, no. 1, pp. 73–85, 2014.

[8] G. Kassen, A. Kudaibergenova, A. Mukasheva, D. Yertargynkyzy, and K. Moldassan, ―Behavioral risk factors for suicide among adolescent schoolchildren.,‖ Ilkogretim Online, vol. 19, no. 1, 2020.

[9] G. Castillo-Sánchez, G. Marques, E. Dorronzoro, O. Rivera-Romero, M. Franco-Martín, and I. De la TorreDíez, ―Suicide risk assessment using machine learning and social networks: a scoping review, ‖ J Med Syst, vol. 44, no. 12, p. 205, 2020.

[10] https://www.who.int/teams/noncommunicablediseases/surveillance/systems-tools/global-school-basedstudent-health-survey/questionnaire.‖

[11] C. Su, R. Aseltine, R. Doshi, K. Chen, S. C. Rogers, and F. Wang, ―Machine learning for suicide risk prediction in children and adolescents with electronic health records, ‖ Transl Psychiatry, vol. 10, no. 1, p. 413, 2020.

[12] L. Zheng et al., ―Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records, ‖ Transl Psychiatry, vol. 10, no. 1, p. 72, 2020.

[13] R. Mishra, P. P. Sinha, R. Sawhney, D. Mahata, P. Mathur, and R. R. Shah, ―SNAP-BATNET: Cascading author profiling and social network graphs for suicide ideation detection on social media, ‖ in Proceedings of the 2019 conference of the North American Chapter of the association for computational linguistics: student research workshop, 2019, pp. 147–156.

[14] H. D. Nelson et al., ―Suicide risk assessment and prevention: a systematic review focusing on veterans, ‖ Psychiatric services, vol. 68, no. 10, pp. 1003–1015, 2017.

[15] K. Lehavot, D. Ben-Zeev, and R. E. Neville, ―Ethical considerations and social media: a case of suicidal postings on Facebook, ‖ J Dual Diagn, vol. 8, no. 4, pp. 341–346, 2012

[16] P. Schneider et al., "Visual analysis of large graphs using (X, Y)-clustering and hybrid visualizations," in IEEE Transactions on Visualization and Computer Graphics, vol. 19, no. 6, pp. 953-962, 2013. doi: 10.1109/TVCG.2013.45

[17] M. Correll and J. Heer, "Colorgorical: Creating discriminable and preferable color palettes for information visualization," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2014. doi: 10.1145/2556288.2557009

[18] L. Breiman, "Random Forests," Machine Learning, vol. 45, pp. 5-32, Oct. 2001.

[19] M. H. Moosavi and M. A. Moosavi, "Ethical considerations in dental research," Journal of Dental Research, vol. 88, no. 8, pp. 672-675, Aug. 2009. doi: 10.1177/0022034509344536.

[20] D. Basak, S. Pal, and D. C. Patranabis, "Support Vector Regression," Neural Information Processing - Letters and Reviews, vol. 11, no. 10, pp. 203, Oct. 2007.

[21] Singh, Aishwarya. "KNN algorithm: Introduction to K-Nearest Neighbors Algorithm for Regression." Analytics Vidhya, August 22, 2018. Accessed on April 24, 2023

[22] Shahani, N.M., Zheng, X., Liu, C., Hassan, F.U., & Li, P. (2021). Developing an XGBoost Regression Model for Predicting Young's Modulus of Intact Sedimentary Rocks for the Stability of Surface and Subsurface Structures. Frontiers in Earth Science, 9, 761990. DOI: 10.3389/feart.2021.761990.

[23] Hoerl, A.E., & Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12(1), 55-67. DOI: 10.1080/00401706.1970.10488634.