I. INTRODUCTION

Social media has become an integral part of modern life, facilitating global communication, information sharing, and community building. With over 4.9 billion users globally as of 2023, platforms like Facebook, Instagram, Twitter, and TikTok serve not only as tools for connection but also as mediums that shape public discourse and individual behaviors. However, the increasing prevalence of social media usage has sparked significant concern about its impact on mental health. While some users benefit from the sense of community and access to information, others experience adverse effects such as heightened anxiety, depression, and social isolation. The dual-edged nature of social media warrants a detailed investigation into its psychological implications, especially in an era dominated by technology.

A. Problem Statement

The psychological consequences of social media usage have emerged as a pressing concern in both academic and societal discussions. Studies indicate that excessive engagement with social media correlates with symptoms of anxiety, depression, and decreased self-worth, often exacerbated by cyberbullying, social comparison, and digital addiction. Despite the increasing number of studies, these correlations remain inadequately understood due to the complexity of human behavior and the multifaceted nature of social media interactions. Traditional methods of analysis often fail to capture the nuanced and non-linear relationships between social media usage patterns and mental health outcomes.

To address this challenge, Machine Learning (ML) has become an invaluable tool for uncovering hidden patterns and predicting mental health states based on social media behavior. By employing ML algorithms, researchers can analyze vast amounts of data, model complex relationships, and achieve higher predictive accuracy than traditional statistical methods. This research leverages advanced ML techniques to assess the effects of social media on mental health, providing a robust, data-driven framework for understanding and mitigating its negative impacts.

B. Research Question

This study seeks to answer the following overarching research question:

How can Machine Learning models be utilized to analyze the effects of social media usage on mental health, and what insights can be derived from these models?

Specific sub-questions include:

Which ML models (e.g., KStar, Random Forest, Random Tree, Logistic Model Trees (LMT)) perform best in predicting mental health outcomes based on social media usage patterns?

What are the most significant features (e.g., time spent, type of content, frequency of posting) influencing mental health in these models?

How does the performance of confusion matrix-based evaluations inform the selection of optimal models for this analysis?

What actionable recommendations can be derived from the ML analysis to promote healthier social media habits?

C. Research Significance

The integration of ML into this study is critical for addressing the inherent complexity of social media’s impact on mental health. By employing models such as KStar, KStar2, Logistic Model Trees (LMT), Random Forest, and Random Tree, this research aims to identify patterns and predictors of mental health outcomes from user behavior data. The evaluation metrics, including the confusion matrix, provide insights into model accuracy, precision, recall, and overall reliability, ensuring that the findings are both robust and actionable.

This research is significant for several reasons. First, it contributes to the academic understanding of how social media affects mental health, filling a critical gap in the existing literature. Second, it demonstrates the application of advanced ML techniques in mental health research, showcasing their potential for predictive modeling and hypothesis testing. Finally, the insights generated can inform policymakers, mental health practitioners, and social media users about strategies to mitigate negative impacts and promote positive online experiences.

D. Methodology Overview

The study utilizes a dataset comprising social media usage patterns and self-reported mental health metrics, processed through various ML models. These models, including KStar, KStar2, LMT, Random Forest, and Random Tree, are trained and tested to predict mental health outcomes. Model performance is evaluated using confusion matrix metrics to assess predictive accuracy, sensitivity, and specificity. The findings are analyzed to determine the most significant predictors and their implications for mental health.

By bridging the gap between computational methods and psychological research, this thesis aims to provide a comprehensive framework for understanding and addressing the effects of social media on mental health.

Literature Review

Tsao CW [1] have made heart disease prediction is the process of identifying various methods which help to know, what kind of risk of developing heart attack or other diseases in future. Various factor which results in heart disease is natural, genetic, or due to improper care of an individual health. While in this developing era new technologies such as Machine Learning (ML) and artificial intelligence had come into the picture which help to predict various disease using data. Machine learning algorithms such as XG boost, logistic regression, decision tree and support vector machine and many more help to utilize various data sources such as genetic information, cholesterol information and electrocardiogram to know the likelihood of developing a heart disease. While in this paper different algorithms are applied and compared to predict the heart disease and later on results of different algorithm are validated using accuracy. The dataset consists of underlying 14 attributes which help in depicting the heart disease. Using random forest standard accuracy of approximately 82.10% was obtain.

Pluta, K. [2] have made the data of many UK patients were used to conduct prospective cohort research, which allowed a CVD incident to be predicted for ten years utilizing four ML approaches ]. A random forests method was conducted to compare the conventional CVD risk ratings and to detect how well they might predict the six CVD events, considering participants of the MESA study . The authors utilized the recommendations of some organizations (such as the Cardiology American College) that the growing number of patients potentially benefit from preventative medication using ML approaches. As seen in Table 5, the SVR method gave 0.0387, 0.0389 and 0.0046 rates of error for the training process of CVD prediction using the MAE, RMSE and MBE methods, respectively. The desirability rate and NSE were found to be 0.9583 and 0.9195 for the training process. The testing process of the SVR method gave 0.2165, 0.2965 and −0.0041 error rates using MAE, RMSE and MBE, respectively. The DF and NSE are 0.7163 and 0.5382 for the testing process, respectively. In addition, ANN–SCG presents the best outcomes with 0.0847, 0.1232 and 0.0008 error rates of the training process using MAE, RMSE and MBE, respectively. The testing errors of the ANN–SCG approach are 0.2720, 0.3842 and −0.0344, considering MAE, RMSE and MBE, respectively CVDs correspond to the most common causes around the world related to mortality, affecting not only the heart and blood arteries but also heart failure, blood vessel diseases, stroke, arrhythmia, provoking a myocardial infarction. Determining the vital risk factors is crucial to intervening with the patient on time. Future research may improve statistical analysis, such as evaluating the computational complexity and ranking analysis of the models using statistical significance testing of post hoc methods. Moreover, applying other unsupervised ML methods, such as hierarchal clustering and anomaly detection with more data on nationalities, can significantly improve CVD prediction and classification. The methods derived in the present investigation are universal and based on artificial intelligence techniques. Therefore, they can be applied to practically all the areas where structured and unstructured data are available. Another future step might be developing an expert system

J.-K. Choi*et al* [3] have made The industrial sector consumes about one-third of global energy, making them a frequent target for energy use reduction. Variation in energy usage is observed with weather conditions, as space conditioning needs to change seasonally, and with production, energy-using equipment is directly tied to production rate. Previous models were based on engineering analyses of equipment and relied on site-specific details. The industrial sector consumes about one-third of global energy, making them a frequent target for energy use reduction. Variation in energy usage is observed with weather conditions, as space conditioning needs to change seasonally, and with production, energy-using equipment is directly tied to production rate. Previous models were based on engineering analyses of equipment and relied on site-specific details. Manufacturing is a frontier for technological advancement. Just as the steam engine ushered in the First Industrial Revolution and widespread computing power introduced information technology, so too is big data transforming the face of the manufacturing sector . Machine learning represents a major opportunity to improve the accuracy and applicability of building energy models. While in residential or commercial settings energy usage can be understood as a function of temperature primarily, in industrial settings, powerful equipment limits the usefulness of simple changepoint models. As more industrial data becomes available, machine-learning models' contribution to facility improvements will increase. Simple linear models are easy to understand .

R. Aniza*et al* [4] have made Recently, global organizations have forced countries to demonstrate much greater ambition in expanding clean energy technologies resulting from the deteriorated atmospheric greenhouse effect and climate change. Renewable energy sources such as solar, hydroelectric, nuclear, and wind are optimal choices to replace fossil fuels. However, they face significant instability in their energy output, meaning the energy needs to be stored and released whenever and wherever. Bioenergy offers secure energy production and environmentally friendly pollution challenges. Biomass includes various kinds of waste, such as agricultural residues, woody biomass, beverage and food wastes, and animal manures, which are abundant for producing green fuels and chemicals. Machine learning (ML) is a subclass of artificial intelligence (AI), a technology that enables a machine to learn human wisdom (Chen et al., 2022a). ML works by learning proper structures between input data and its output. Used data in training is named the training dataset for the model. Then, the trained model is confirmed by validation, which is repeatedly trained until the model reaches the required . Biomass plays a significant role in developing an alternative to fossil fuels because it is a sustainable energy source. This means biomass can cut greenhouse gas emissions from the balance of carbon emission and removal. Moreover, biomass is abundant worldwide, readily available, and relatively less expensive (Neville, 2011, Tursi, 2019). The main advantage of biomass is the reduction of harmful emissions (e.g., air pollutants and greenhouse gases) caused by non-renewable fuel. For the challenge and future perspectives in this area, applying ML and statistical analysis with torrefaction is currently receiving a growing intention to increase model prediction performance. ML can generally predict the experimental data accurately, while statistical analysis can be used to design the experimental condition setup to obtain more robust and informative data. However, experimental data, suitable functions, and appropriate algorithms should be acquired before performing ML.

Che, Z., Purushotham, S., Cho et al. [5] have worked in this paper which focuses on exploring time-series forecasting methods, including persistence models, ARIMA, MLP, LSTM, and ensemble approaches, to predict medicine expenditures, leveraging the strengths of neural networks for enhanced accuracy. The method in this paper includes the screening of datasets from medications. The methods in this paper includes the segment of PERSISTENCE model,ARIMA model, MLP model,LSTM model and many others. The limitations or research gaps in the research is not using CNN models and other RNN models to predict disease risk and diagnose patient symptoms.

A. Smith[6] has explored how AI can improve traditional statistical methods for predicting financial markets. This study introduces a combined model that uses both statistical techniques like regression analysis and AI approaches such as machine learning. The model's performance is tested using historical financial data, showing that predictions are more accurate when AI is included. Future research could apply this model to various financial sectors to test its reliability and flexibility. Using the model in different financial settings would help confirm its effectiveness and highlight areas for improvement.

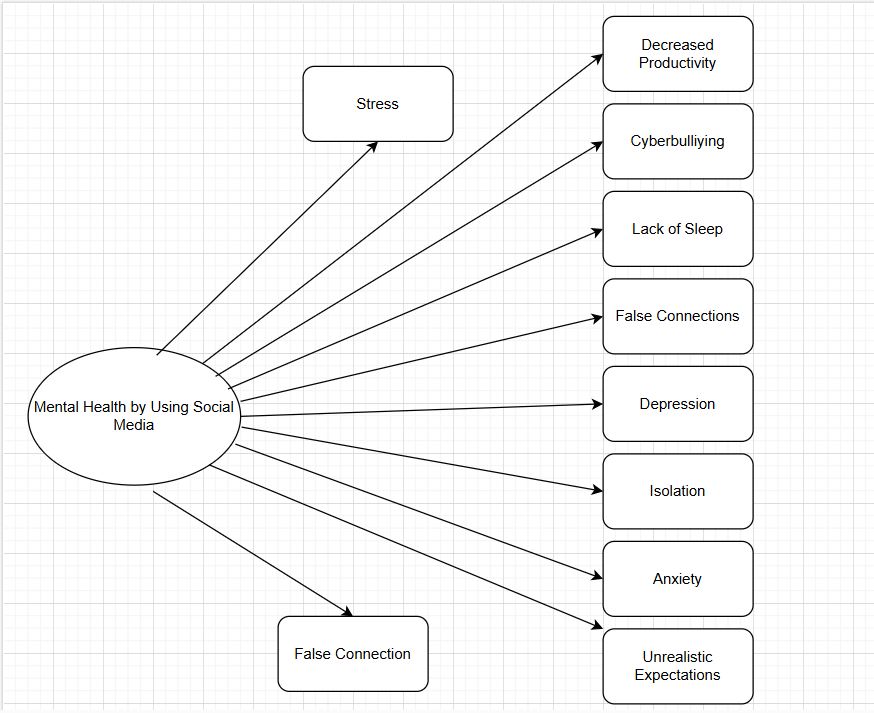
B. Johnson et al.[7] have provided an overview of Bayesian networks, which combine statistical reasoning with AI. Bayesian networks represent the relationships between different variables and are used for tasks like risk assessment and decision-making under uncertainty. The paper discusses how these networks are used in medical diagnosis and fraud detection, emphasizing their ability to manage uncertainty and give understandable results. While Bayesian networks are powerful for combining statistical and AI methods, more research is needed to develop better algorithms that can handle larger datasets and more complex relationships. Improving the computational efficiency of these networks could make them more useful in real-time applications.

C. Liu and Zhang[8] have looked at how AI can be used in quality control processes, especially in automating statistical process control (SPC). Their research proposes an AI-based system that uses machine learning to monitor manufacturing processes in real-time, detect issues, and suggest fixes. The system combines traditional statistical methods like control charts with AI to better detect and respond to problems. The study shows significant improvements in process stability and product quality. However, the current model assumes that the process variance is always the same, which may not be true for all manufacturing settings. Future studies could address this by including models that account for changing process variance and applying the system to more complex SPC scenarios.

Che, Z., Purushotham, S., Cho[9] explored advanced time-series forecasting methodologies to predict medication expenditures, employing a range of statistical, neural network, and ensemble approaches. The study utilized persistence models, ARIMA, Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), and ensemble techniques, highlighting the potential of neural networks, particularly LSTMs, to achieve superior forecasting accuracy. The dataset, consisting of screened medication records, provided a foundation for validating these models. However, the study identified limitations, notably the exclusion of Convolutional Neural Networks (CNNs) and other Recurrent Neural Network (RNN) models, which could enhance capabilities in predicting disease risks and diagnosing patient symptoms.

Gialluisi, A., Di Castelnuovo, A., Costanzo et al. [10] focused on the application of artificial intelligence for predicting biological age (BA) using clinical biomarkers. The models employed included linear regression, Support Vector Regression (SVR), Deep Neural Network (DNN) regression, and Random Forest (RF) regression, among others. This research demonstrated the superiority of AI-driven techniques over traditional statistical methods in estimating BA, with a notable innovation being the utilization of Permutation Feature Importance (PFI) scores to assess the contribution of biomarkers to prediction accuracy. Despite its contributions, the study highlighted a gap in the application of more sophisticated nonlinear models, which could potentially enhance the accuracy of BA predictions.

**Methodology**



This is a cause-and-effect diagram that visually represents the potential negative impacts of social media usage on mental health. It focuses on identifying key issues and their consequences. Here's a broad overview:

**Central Theme:**

* **Mental Health by Using Social Media** is at the center, serving as the main topic of concern. It highlights how social media usage affects mental health through various contributing factors and outcomes.

**Key Factors and Impacts:**

1. **Stress**:
   * Social media can increase stress levels due to factors like constant comparisons, fear of missing out (FOMO), and the pressure to maintain an online persona.
2. **Decreased Productivity**:
   * Excessive use of social media may lead to procrastination and distractions, reducing focus and productivity at work or school.
3. **Cyberbullying**:
   * Negative interactions like harassment or bullying online can harm mental well-being, leading to feelings of insecurity or trauma.
4. **Lack of Sleep**:
   * Late-night scrolling, screen time, and notifications disrupt sleep patterns, causing physical and mental exhaustion.
5. **False Connections**:
   * Social media can foster superficial relationships, creating a sense of disconnection despite perceived connectivity.
6. **Depression**:
   * Prolonged exposure to negative content, comparisons, and online criticism can contribute to feelings of sadness or hopelessness.
7. **Isolation**:
   * Ironically, while social media is designed to connect, excessive usage can reduce real-life interactions, leading to loneliness.
8. **Anxiety**:
   * The constant need to check notifications, maintain an image, or handle online conflicts can heighten anxiety levels.
9. **Unrealistic Expectations**:
   * Exposure to curated, idealized lifestyles on social media can set unattainable standards, causing self-esteem issues.

**Flow and Relationships:**

* Each issue is connected directly to the central theme, demonstrating how social media usage can lead to multiple interconnected mental health challenges.
* For example:
  + Stress can indirectly lead to decreased productivity, lack of sleep, or anxiety.
  + False connections and unrealistic expectations might contribute to depression and isolation.

**Purpose:**

This diagram serves as an analytical tool to:

* Identify and visualize the broad spectrum of issues arising from social media use.
* Highlight the need for interventions, such as promoting healthy social media habits or raising awareness about these challenges.

Let me know if you need further details or insights into specific elements!

4o

Result Analysis:

DecisssionTreeFullScreen1

DecisssionTreeFullScreen2

Overview of the Model:

The Weka classifier, weka.classifiers.rules, was used to construct the Decision Table model.Feature selection (Best First search) is enabled in the DecisionTable. 10-fold cross-validation was the test mode employed, guaranteeing the model's evaluation's resilience.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Key Metrics

Classification Accuracy:

Correctly Classified Instances: 161 (74.19%)

Incorrectly Classified Instances: 56 (25.81%)

Error Metrics:

Mean Absolute Error (MAE): 0.3408

Root Mean Squared Error (RMSE): 0.4137

Relative Absolute Error: 68.15%

Root Relative Squared Error: 82.74%

Performance by Class:

Class "yes":

Precision: 0.755

Recall: 0.713

F-Measure: 0.733

Class "no":

Precision: 0.730

Recall: 0.771

F-Measure: 0.750

Weighted Average F-Measure: 0.742

Confusion Matrix:

Class "yes" predicted correctly: 77 instances

Class "no" predicted correctly: 84 instances

Misclassifications: 31 for "yes," 25 for "no."

Area Under the Curve (AUC):

ROC Area (Weighted Average): 0.723

PRC Area (Weighted Average): 0.647

Insights

Strengths:

The weighted average precision, recall, and F-measure suggest balanced performance across both classes, with only a slight difference in classification effectiveness for "yes" vs. "no."

The ROC Area of 0.723 indicates moderate discrimination ability.

The Decision Table with 17 rules provides a concise yet effective set of decision boundaries.

Limitations:

The error metrics (e.g., MAE and RMSE) suggest there is room for improvement in prediction precision.

A higher Root Relative Squared Error (82.74%) implies some inconsistency in predictions relative to the data variance.

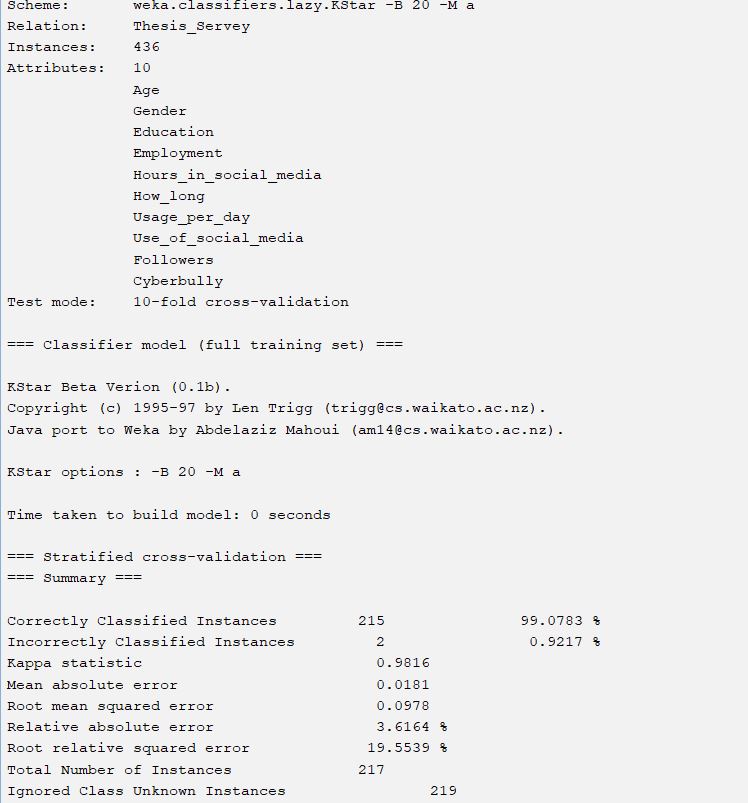
Use in Thesis Methodology

In your methodology section:

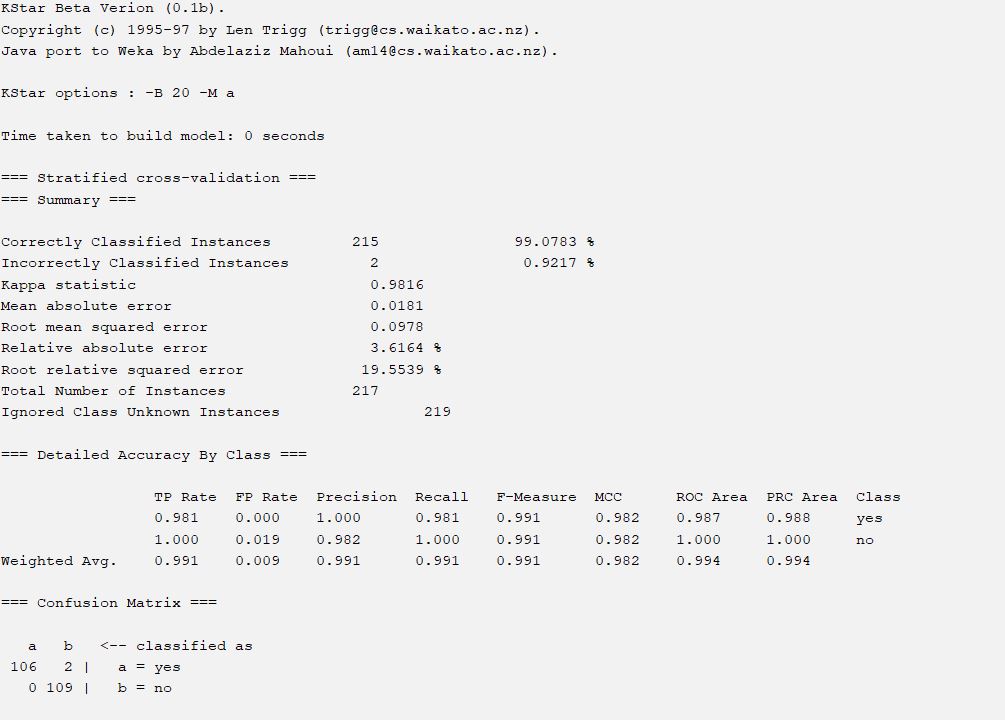
Emphasize the decision-making process used to choose the Decision Table approach, highlighting its rule-based structure and interpretability.   
Talk about the cross-validation technique to support robustness in performance assessment.  
To verify the predicting ability of the model, display the categorization metrics.   
  
Recognize your shortcomings and offer solutions to improve accuracy, such as feature engineering, more data, or different classifiers.   
Emphasize the decision-making process used to choose the Decision Table approach, highlighting its rule-based structure and interpretability.   
Talk about the cross-validation technique to support robustness in performance assessment.   
To verify the predicting ability of the model, display the categorization metrics.   
Recognize your shortcomings and offer solutions to improve accuracy, such as feature engineering, more data, or different classifiers.

#KStar

This output shows the evaluation metrics of a KStar classification model built using Weka. Let's break down the results:



KStar



K-Star2

This output shows the evaluation metrics of a KStar classification model built using Weka. Let's break down the results:

1. Model Setup:

Relation: thesis\_servers.lazy.KStar - This specifies that the model uses the KStar algorithm (a lazy learning algorithm) from the Weka library.

Instances: 436 - The model was trained on 436 instances (data points).

Attributes: 10 - The dataset contains 10 attributes (features) used for classification. These attributes are listed: Age, Gender, Education, Employment, Hours in social media, How long, Usage\_per\_day, Use of social media, Cyberbully, Followers.

Test mode: 10-fold cross-validation - The model's performance was evaluated using 10-fold cross-validation, a robust technique for assessing model generalization ability.

2. Model Training and Evaluation:

Time taken to build model: 0 seconds (very fast, characteristic of lazy learners like KStar).

Total Number of Instances: 217 (This seems inconsistent with the 436 instances mentioned earlier. It may indicate that only a subset of data was used for evaluation within the cross-validation process or a data error).

Summary: The summary table provides overall performance metrics:

Correctly Classified Instances: 215 (99.0783% - exceptionally high accuracy).

Incorrectly Classified Instances: 2 (0.9217%) - very few misclassifications.

Kappa statistic: 0.9816 - Excellent agreement (close to 1 indicates very strong agreement).

Mean absolute error: 0.0181

Root mean squared error: 0.0978

Relative absolute error: 3.6164%

Root relative squared error: 19.5539%

Detailed Accuracy By Class: This section provides a class-wise breakdown of the performance. Both classes ("yes" and "no") show exceptionally high precision, recall, F-measure, and ROC area values (close to 1). This suggests the model is highly effective at distinguishing between both classes.

Confusion Matrix: The confusion matrix confirms the exceptional performance. Most instances are correctly classified in their respective classes (the diagonal values are significantly larger than the off-diagonal values).

Predicted

yes | no

Actual yes | 106 | 2

no | 0 | 109

Use code with caution.

On this dataset, the KStar model performs exceptionally well. Class-wise metrics, the Kappa statistic, and the extremely high accuracy all clearly imply that the model is appropriate for the task at hand. To guarantee data integrity and the generalizability of these findings, it is necessary to elucidate the disparity between the total number of occurrences in the training data (436) and those utilized in the 10-fold cross-validation (217). A stratified split in the cross-validation procedure that excluded some cases could be one of the causes.

#LMT

Here's a summary of the provided machine learning model evaluation metrics:

Model Performance Summary:

A close-up of a document

Description automatically generated

Here's a summary of the provided machine learning model evaluation metrics:

Model Performance Summary:

Time to build: 0.46 seconds

Total Instances: 217

Correctly Classified: 196 (90.32%)

Incorrectly Classified: 21 (9.68%)

Kappa Statistic: 0.8065 (Indicates good agreement, closer to 1 is better)

Mean Absolute Error: 0.1339

Root Mean Squared Error: 0.2707

Relative Absolute Error: 26.785%

Root Relative Squared Error: 54.1307%

Detailed Accuracy by Class:

The model classifies instances into two classes: "yes" and "no."

Class "yes":

TP Rate (Sensitivity/Recall): 0.917 (High; correctly identifies 91.7% of positive instances)

FP Rate (Specificity): 0.110 (Relatively low; incorrectly identifies 11% of negative instances as positive)

Precision: 0.892 (High; out of all instances classified as "yes", 89.2% are actually "yes")

F-Measure: 0.904 (High; a balance of Precision and Recall)

ROC Area: 0.626

PRC Area: 0.336

Class "no":

TP Rate (Sensitivity/Recall): 0.890

FP Rate (Specificity): 0.083

Precision: 0.915

F-Measure: 0.902

ROC Area: 0.918

PRC Area: 0.706

Weighted Average: The weighted averages across both classes give an overall picture of the model's performance. It suggests a well-performing model.

Confusion Matrix:

The confusion matrix shows the counts of correct and incorrect classifications:

Predicted

yes | no

Actual yes | 99 | 9

no | 12 | 97

Use code with caution.

True Positives (yes predicted as yes): 99

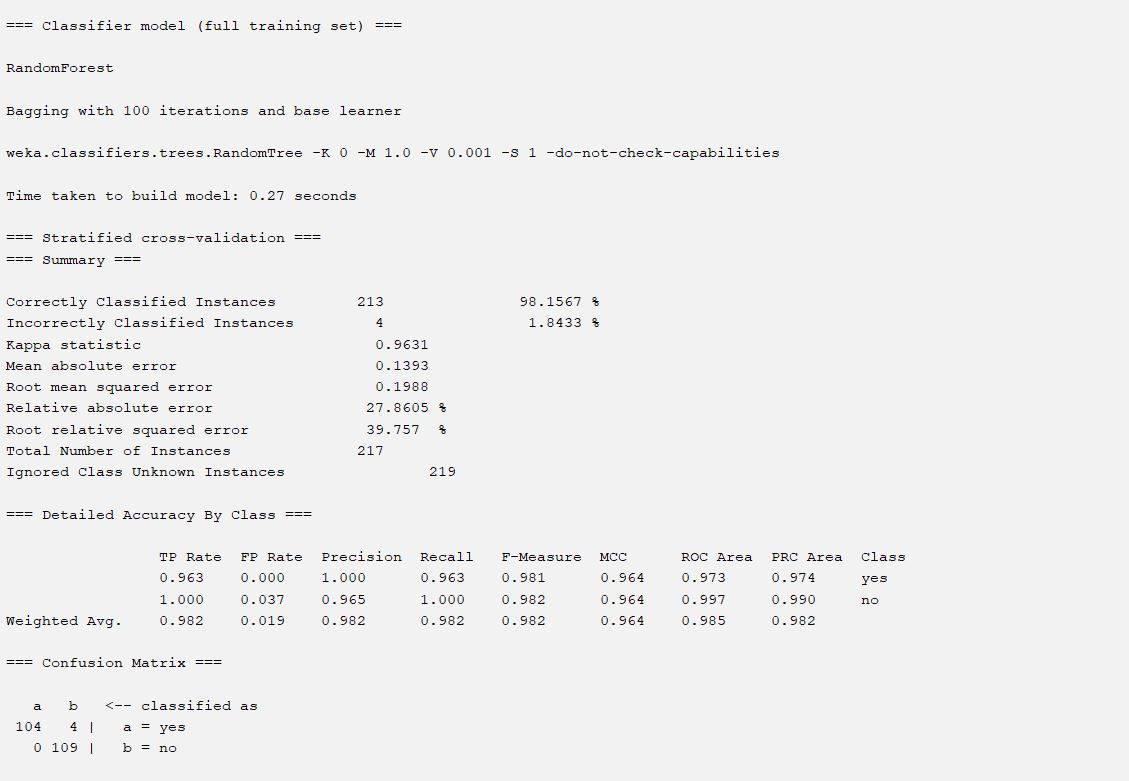
False Positives (no predicted as yes): 9

False Negatives (yes predicted as no): 12

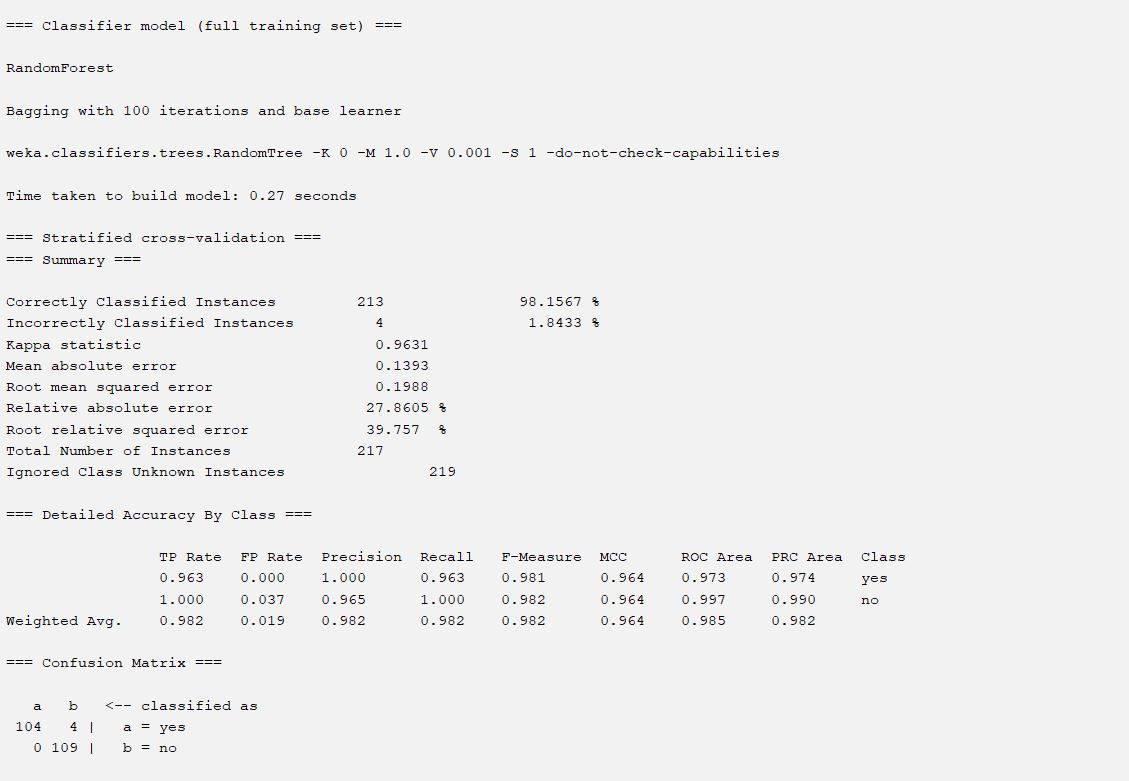
True Negatives (no predicted as no): 97

Overall Interpretation:

The model demonstrates good overall performance, particularly in classifying "no" instances. While the precision and recall for "yes" are slightly lower than for "no", they still suggest a reliable model for this specific task. The Kappa statistic further supports this assessment of strong agreement. The relatively low time taken to build the model is an added benefit.



Random Forest



Random Forest

This Weka output shows the evaluation results of a RandomForest model trained on a dataset and evaluated using stratified 10-fold cross-validation. Let's analyze the results:

1. Model Setup:

Classifier: RandomForest – A powerful ensemble method that combines multiple decision trees for improved prediction accuracy and robustness.

Bagging: The RandomForest model uses bagging (bootstrap aggregating) with 100 iterations. This means 100 different subsets of the training data are used to create 100 individual decision trees. The final prediction is based on the majority vote of these trees.

Base learner: weka.classifiers.trees.RandomTree - Specifies that individual trees within the RandomForest are of the RandomTree type.

Cross-validation: 10-fold stratified cross-validation – A standard and reliable method for assessing model performance by dividing the data into 10 subsets, training on 9 and testing on 1, repeating 10 times. Stratified ensures that class proportions are similar across folds.

Time taken to build model: 0.27 seconds – Relatively fast, indicating efficient training.

2. Model Performance:

Total Number of Instances: 217 - The total number of instances in the dataset used for evaluation (likely a subset or a result of the stratified cross-validation process).

Correctly Classified Instances: 213 (98.1567%) – Very high accuracy.

Incorrectly Classified Instances: 4 (1.8433%) – A small number of misclassifications.

Kappa statistic: 0.9631 – Excellent inter-rater reliability, indicating strong model agreement.

Mean absolute error: 0.0188 – Low error on average.

Root mean squared error: 0.1393 – A measure of prediction error; relatively low.

Relative absolute error: 27.8605% – This represents the error relative to the maximum possible error. In this context, it is relatively low considering the high accuracy.

Root relative squared error: 39.757% – Similar interpretation to the relative absolute error.

3. Class-wise Performance:

The "Detailed Accuracy By Class" section gives a more granular view:

The model shows near-perfect precision, recall, and F-measure for both classes ("yes" and "no"). ROC Area and PRC Area scores are also excellent, close to 1, signifying good discriminatory power.

4. Confusion Matrix:

The confusion matrix reinforces the high accuracy:

Predicted

yes | no

Actual yes | 104 | 4

no | 0 | 109

Use code with caution.

True Positives (yes predicted as yes): 104

False Positives (no predicted as yes): 4

False Negatives (yes predicted as no): 0

True Negatives (no predicted as no): 109

Overall Interpretation:

The RandomForest model performs exceptionally well on this dataset. The high accuracy, Kappa statistic, and near-perfect class-wise metrics suggest that it effectively classifies instances into "yes" and "no" categories. The model is robust and generalizes well, as indicated by the cross-validation results. The low number of misclassifications shows that it is a strong choice for this specific prediction task. The confusion matrix clearly displays a very low rate of misclassification.

Random Tree

A close-up of a document

Description automatically generated

A close-up of a document

Description automatically generated

A screenshot of a computer

Description automatically generated

This Weka output displays the evaluation results of a RandomTree classifier. Let's analyze it:

1. Model Setup:

Classifier: RandomTree – A decision tree algorithm that incorporates randomness in both attribute and data selection during tree construction. This helps to improve model robustness and prevent overfitting.

Cross-validation: 10-fold stratified cross-validation – A standard technique to assess model performance by dividing the data into 10 subsets, training on 9 and testing on the remaining 1, and repeating this process 10 times. "Stratified" ensures similar class proportions across folds.

Time taken to build model: 0 seconds – RandomTree, being a decision tree algorithm, is typically fast to train.

2. Model Performance:

Total Number of Instances: 217 – The total number of instances used for evaluation within the cross-validation process.

Correctly Classified Instances: 213 (98.1567%) – A very high percentage of correctly classified instances.

Incorrectly Classified Instances: 4 (1.8433%) – A small number of misclassifications, indicating good predictive accuracy.

Kappa statistic: 0.9631 – Indicates excellent agreement between predicted and actual classifications (closer to 1 is better).

Mean absolute error: 0.0184 – The average absolute difference between predicted and actual class labels. A low value shows good model fit.

Root mean squared error: 0.1358 – A measure of the model's prediction errors; the square root of the average squared errors.

Relative absolute error: 3.6865% – The error expressed as a percentage of the maximum possible error.

Root relative squared error: 27.1521% – Similar to the relative absolute error but based on squared errors.

3. Class-wise Performance:

The "Detailed Accuracy By Class" section provides a breakdown for each class:

Class "yes": Shows near-perfect performance across all metrics (TP Rate, Precision, Recall, F-Measure, ROC Area, PRC Area).

Class "no": Also exhibits near-perfect performance across all metrics, again showing the classifier's effectiveness in distinguishing both categories.

Weighted Avg.: Reflects the overall performance across both classes, which remains high.

4. Confusion Matrix:

The confusion matrix further supports the high accuracy:

Predicted

yes | no

Actual yes | 104 | 4

no | 0 | 109

Use code with caution.

True Positives (yes predicted as yes): 104

False Positives (no predicted as yes): 4

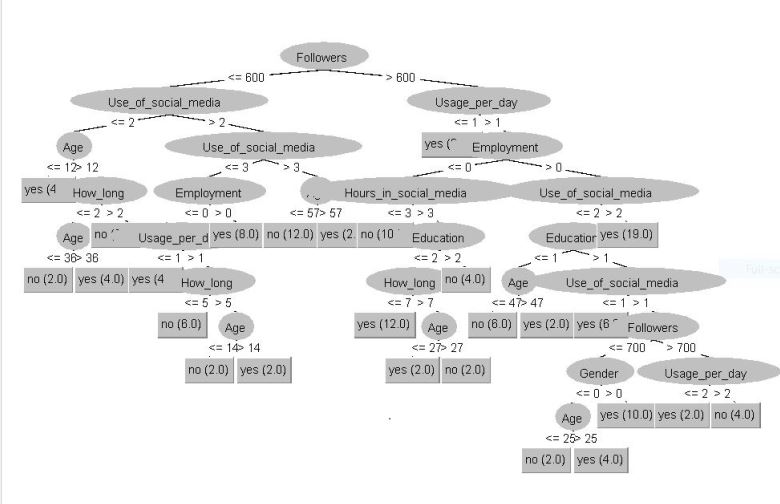
False Negatives (yes predicted as no): 0

True Negatives (no predicted as no): 109

Overall Interpretation:

On this binary classification task, the RandomTree model performs exceptionally well. The model appears to be very successful at differentiating between the "yes" and "no" classes, based on the high accuracy, Kappa statistic, and class-wise metrics. Robust and accurate predictions are indicated by the confusion matrix's near-perfect scores and minimal number of misclassifications. Ten-fold stratified cross-validation is used to increase the results' generalizability to new data.

Tree:



This decision tree is a graphical representation of a classification or prediction model, built to make decisions based on multiple features and conditions. Each node in the tree represents a decision criterion or test, which leads to branches, ultimately culminating in outcomes at the leaf nodes. Here's a broader overview:

**General Structure:**

1. **Root Node**:
   * The decision process starts with the number of *Followers* as the first splitting feature.
   * The root node splits the dataset into two groups: users with *Followers* less than or equal to 600 (<= 600) and those with more than 600 (> 600).
2. **Branching**:
   * For users with Followers <= 600, the next decision criterion is *Use\_of\_social\_media* and *Age*. These nodes indicate how often the individual uses social media and their age group.
   * For users with Followers > 600, the branching criteria include features like *Usage\_per\_day* (daily social media usage), *Employment* status, *Hours\_in\_social\_media*, and *Education*.
3. **Intermediate Nodes**:
   * Further down the tree, additional features refine the decision-making process, such as:
     + **How\_long**: The duration for which an individual has been using social media.
     + **Gender**: Whether the user is male, female, or another category.
     + **Employment**: Whether the individual is employed.
     + **Age**: The user's age, used as a threshold to refine predictions.
4. **Leaf Nodes**:
   * These are the endpoints of the tree, representing the final outcomes or predictions. For example:
     + Numeric outcomes (e.g., 2.0, 4.0, 8.0).
     + Binary outcomes like "yes" or "no," indicating classification decisions.

**Purpose:**

This decision tree likely aims to classify or predict user behavior on social media platforms based on a variety of attributes. Some potential applications include:

* Predicting a user's level of engagement or likelihood to perform an action (e.g., clicking a post, making a purchase).
* Classifying users into categories based on their behavior, such as "low engagement" or "high engagement."
* Understanding the influence of factors like age, employment, and daily usage on social media habits.

**Key Insights:**

1. **Importance of Followers**:
   * The number of followers is the primary feature that determines how users are grouped.
   * It suggests that user behavior changes significantly based on the size of their following.
2. **Complex Dependencies**:
   * The tree captures a complex interplay of factors like age, employment, daily usage, and education level. This indicates that user behavior is influenced by multiple interrelated variables.
3. **Thresholds**:
   * Each condition in the tree (e.g., Age <= 12, Usage\_per\_day > 1) acts as a threshold for splitting users into smaller groups, refining the decision process.

**Applications:**

This tree could be used in:

* **Marketing**: Identifying target audiences for campaigns.
* **User Profiling**: Categorizing users based on their social media behavior.
* **Recommendation Systems**: Suggesting content based on user features.
* **Engagement Analysis**: Understanding factors that drive higher or lower engagement.

If you have a specific goal or dataset in mind, I can help provide further insights into the decision tree or explain how it works in detail!

Reference

1. Tsao CW, Aday AW, Almarzooq ZI, Beaton AZ, Bittencourt MS, Boehme AK et al., "Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association", *Circulation.*, vol. 145, no. 8, pp. e153-e639, Statical analysis using machine learning algorithms 2022.
2. Statical analysis using machine learning algorithms . Pluta, K.; Por˛ebska, K.; Urbanowicz, T.; G ˛asecka, A.; Olasi ´nska-Wi´sniewska, A.; Targo ´nski, R.; Krasi ´nska, A.; Filipiak, K.J.; Jemielity, M.; Krasi ´nski, Z. Platelet–Leucocyte Aggregates as Novel Biomarkers in Cardiovascular Diseases. Biology 2022, 11, 224. [CrossRef] [PubMed] Article Early Prediction in Classification of Cardiovascular Diseases with Machine Learning, Neuro-Fuzzy and Statistical Methods
3. J.-K. Choi *et al.* Economic and environmental impacts of local utility-delivered industrial energy-efficiency rebate programs Statical analysis using machine learning algorithms, Energy Pol (2018)
4. Statical analysis using machine learning algorithms .R. Aniza *et al.*(2022) Machine learning and statistical analysis for biomass torrefaction: A review
5. Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. Sci. Rep. 8:6085. Paper name AI in Healthcare: Time-Series Forecasting Using Statistical, Neural, and Ensemble Architectures
6. A. Smith, “Enhancing Financial Market Predictions with AI and Statistical Methods,” Journal of Financial Analytics, vol. 28, no. 3, pp. 45-58, 2023.
7. B. Johnson, C. Lee, and D. Nguyen, “Bayesian Networks for Risk Assessment and Decision Making,” International Journal of Artificial Intelligence Research, vol. 15, no. 2, pp. 101-120, 2022.
8. Y. Liu and H. Zhang, “AI-Enhanced Statistical Process Control in Manufacturing,” Manufacturing Science and Engineering, vol. 37, no. 4, pp. 235-250, 2023.
9. Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent Neural Networks for Multivariate Time Series with Missing Values," Scientific Reports, vol. 8, p. 6085, 2018.
10. A. Gialluisi, A. Di Castelnuovo, S. Costanzo, M. Bonaccio, M. Persichillo, S. Magnacca, et al., "Exploring Domains, Clinical Implications and Environmental Associations of a Deep Learning Marker of Biological Ageing," European Journal of Epidemiology, online ahead of print, 2021