1. **Project Title: Sleep Efficiency**
2. **Data Set Name: Sleep Efficiency**
3. **Data Set Source: Kaggle**
4. **Data set Link: https:** [**Sleep Efficiency Dataset**](https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency)
5. **Data Set Description:** The dataset contains information about a group of test subjects and their sleep patterns.

*The dataset consists of 15 columns and 618 rows.*

*Variable description along with data type:*

* *Independent Variables:*

1. Subject ID: Unique identifier for each test subject (categorical)
2. Age: Age of each subject (numeric)
3. Gender: Gender of each subject (categorical)
4. Bedtime: Time when the subject goes to bed each day (datetime)
5. Wakeup time: Time when the subject wakes up each day (datetime)
6. Sleep duration: Total amount of time each subject slept in hours (numeric)
7. REM sleep percentage: Percentage of time spent in REM sleep (numeric)
8. Deep sleep percentage: Percentage of time spent in deep sleep (numeric)
9. Light sleep percentage: Percentage of time spent in light sleep (numeric)
10. Awakenings: Number of times each subject wakes up during the night (numeric)
11. Caffeine Consumption: Information about each subject's caffeine consumption in the 24 hours prior to bedtime (categorical)
12. Alcohol Consumption: Information about each subject's alcohol consumption in the 24 hours prior to bedtime (categorical)
13. Smoking Status: Information about each subject's smoking status (categorical)
14. Exercise Frequency: Information about each subject's exercise frequency (categorical)

* *Dependent Variable:*

1. Sleep efficiency: Proportion of time spent in bed that is actually spent asleep (numeric)

*Missing Values:*

1. Total missing values: 91
2. Missing values in Caffeine Consumption: 36
3. Missing values in Alcohol Consumption: 21
4. Missing values in Awakenings: 27
5. Missing values in Exercise: 7

1. **Description of Work Done:**

*The project involves predicting sleep efficiency based on various factors using different feature selection techniques and regression algorithms. Initially, the dataset is preprocessed by removing unnecessary columns and imputing missing values. Categorical variables are converted to factors, and the data is split into training and testing sets.*

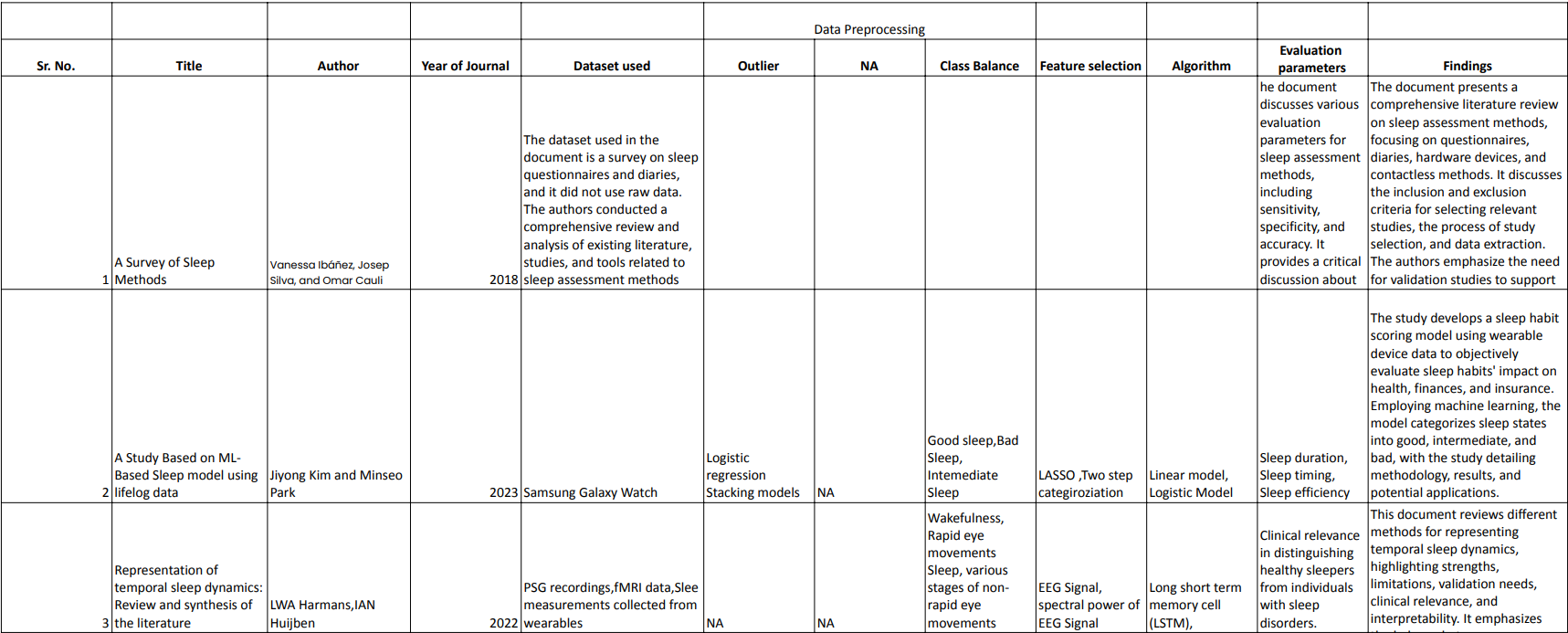
1. *Feature Selection Techniques:*

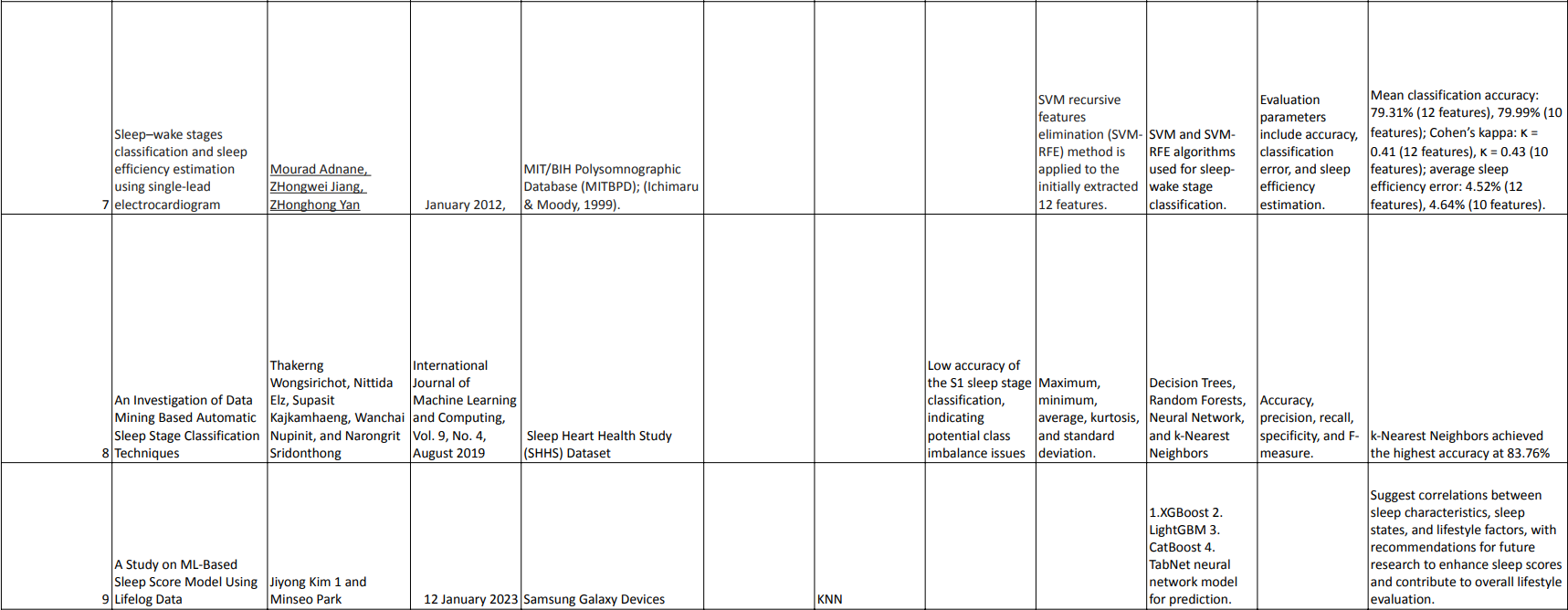
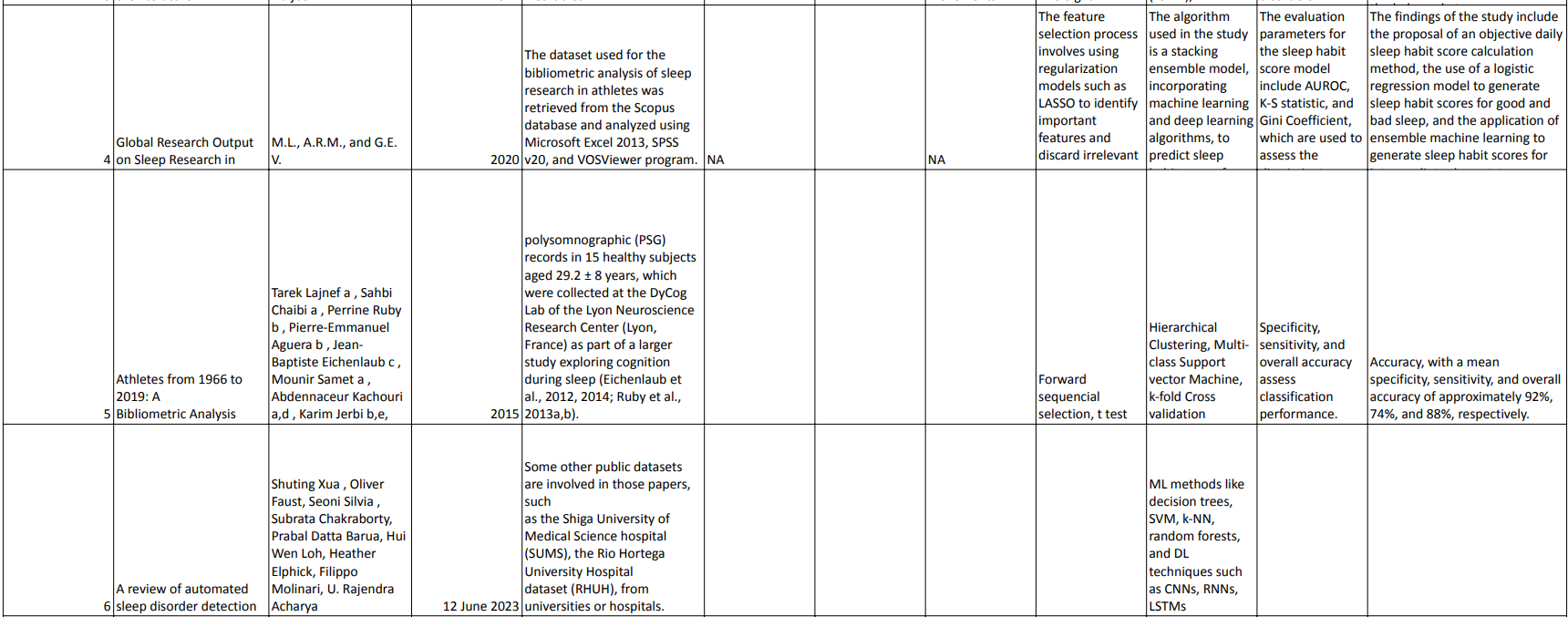
*The project explores multiple feature selection methods, including Lasso Regression, ANOVA, Recursive Feature Elimination (RFE), and Forward Feature Selection. These techniques help identify the most relevant predictors for the sleep efficiency prediction model.*

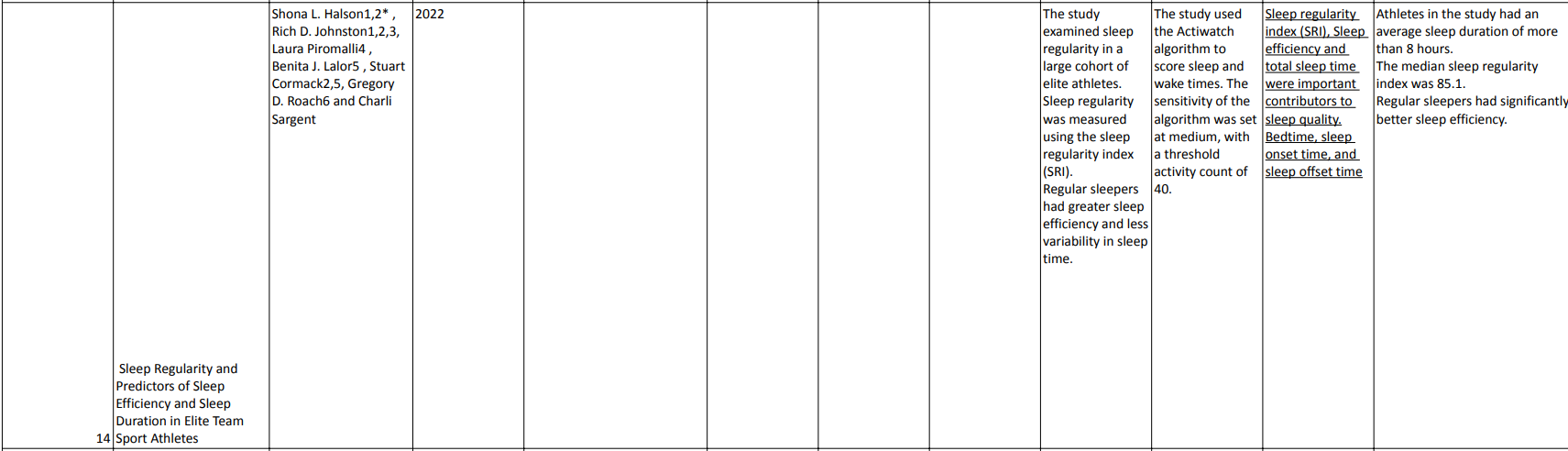
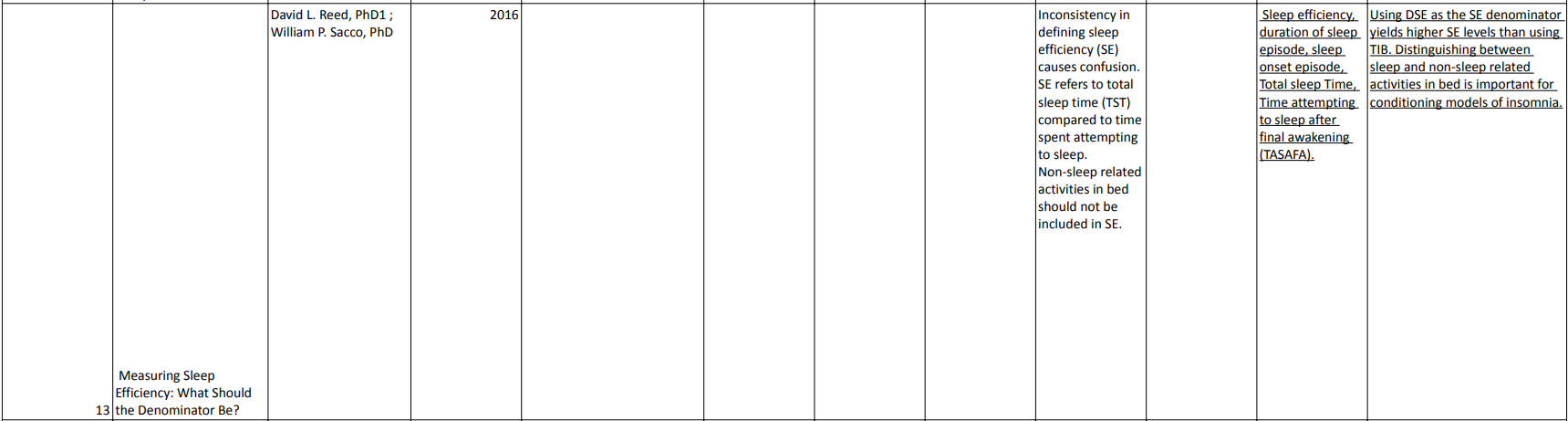
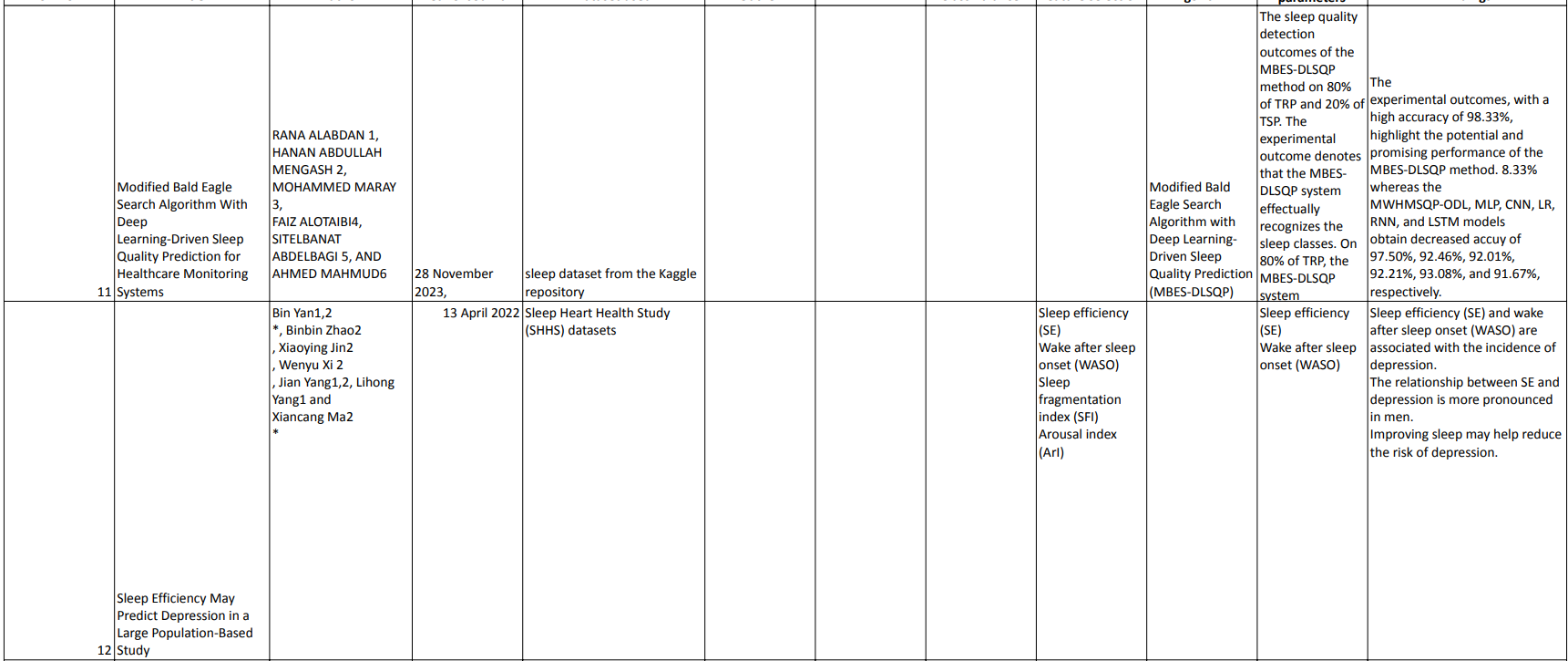
1. *Regression Algorithms: Several regression algorithms are employed to build predictive models:*
   * *Random Forest: A robust ensemble learning method capable of capturing complex interactions and nonlinear relationships in the data.*
   * *Gradient Boosting: Another ensemble method that builds multiple weak learners sequentially, focusing on the errors of the previous models.*
   * *SVR (Support Vector Regression): A regression model that uses support vector machines to map input data to high-dimensional feature spaces.*
   * *Decision Trees: A simple yet powerful non-parametric model that partitions the feature space into segments to make predictions.*
2. *Model Training and Evaluation:*

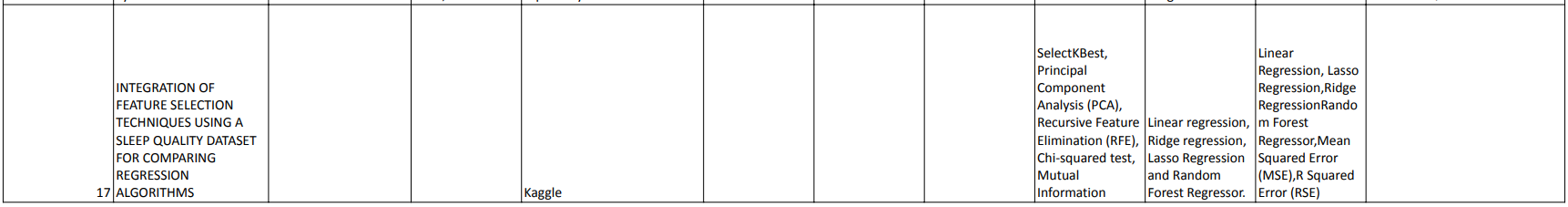
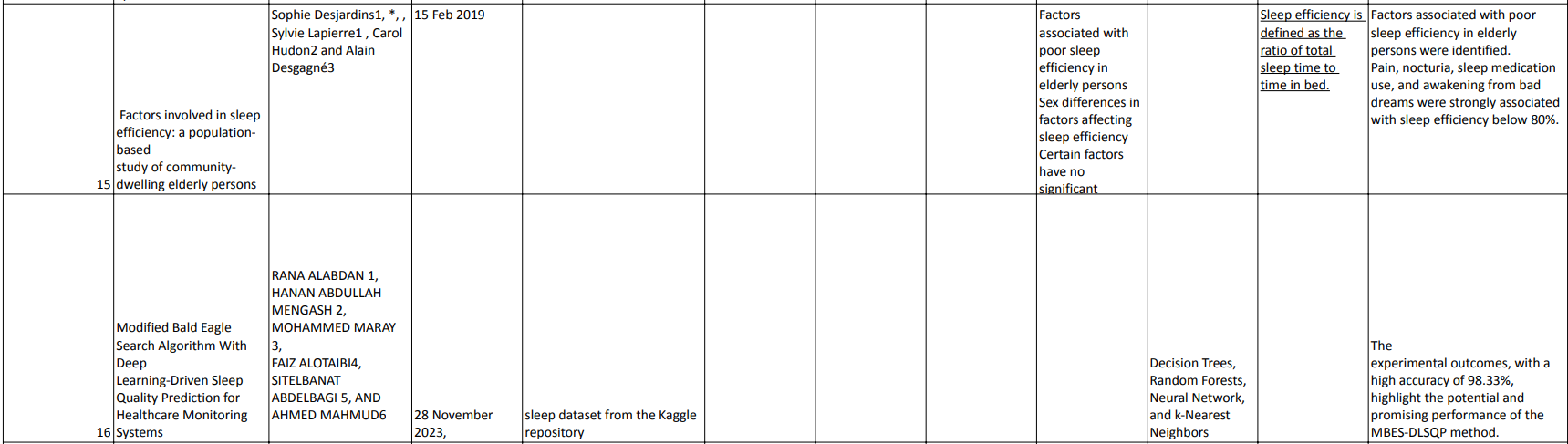
*Each regression model is trained on the selected features and evaluated using performance metrics such as R-squared, RMSE (Root Mean Squared Error), and MAE (Mean Absolute Error). These metrics provide insights into how well the models are capturing the variance in sleep efficiency and making accurate predictions.*

1. **Literature Survey:**









1. **Data Preprocessing:** *Four major preprocessing steps are being followed in the project which include the following:*
2. *Loading dataset:*

*Loading dataset into a file/variable ‘f’ using* ***‘read.csv ()’*** *command.*

1. *Removing unnecessary columns:*

*Removing unnecessary columns includes* "Bedtime", "Wakeup time", "ID" as they are date-time type of variable and don’t affect the Sleep Efficiency factor.

1. *Missing Values:*

*Columns which include missing values are initially found and consists of* "Awakenings", "Caffeine Consumption", "Alcohol Consumption", and "Exercise Frequency".

*The next step involves imputing the missing values with mean of the entire column respectively*

1. *Conversion of categorical variables:*

Categorical variables are converted into factors as 1 and 2 which include "Gender" and "Smoking Status

1. **Feature Selection:** *Feature selection techniques are applied to select the most relevant features for predicting Sleep Efficiency. Feature selection methods comprises of Recursive Feature Selection (RFE), Lasso Regression, Analysis of Variance (ANOVA) and Forward Feature Selection.*
2. *Recursive Feature Elimination (RFE):*

* *Recursive Feature Elimination is a wrapper method that recursively selects subsets of features and evaluates their performance using a machine learning model.*
* *The algorithm starts with all features and evaluates their importance based on some criterion (e.g., coefficients in linear models, feature importance in tree-based models).*
* *It then removes the least important feature(s) and repeats the process until the desired number of features is reached.*
* *RFE provides a ranking of features based on their importance and can be used to identify the optimal subset of features for a given machine learning task.*

|  |  |
| --- | --- |
| Feature | Coefficient |
| Light.sleep.percentage | 2.475934 |
| Deep.sleep.percentage | 2.450623 |
| Awakenings | 1.289569 |
| Alcohol.consumption | 0.4742709 |
| Age | 0.3545051 |
| Smoking.status | 0.2636254 |
| Exercise.frequency | 0.1638683 |
| REM.sleep.percentage | 0.133295 |
| Sleep.duration | 0.1226135 |
| Caffeine.consumption | 0.06987856 |

1. *Lasso Regression:*

* *Lasso (Least Absolute Shrinkage and Selection Operator) is a regularization technique that penalizes the absolute size of feature coefficients. As a result, it pushes less influential features' coefficients towards zero, effectively eliminating them from the model.*
* *Lasso feature selection is particularly useful when dealing with high-dimensional datasets with many potentially irrelevant features.*
* *It automatically selects the most relevant features by shrinking the coefficients of less important features towards zero.*

|  |  |
| --- | --- |
| Feature | Coefficient |
| Age | 0.911025544 |
| Smoking Status | -0.043096079 |
| Awakenings | -0.034987054 |
| Caffeine Sleep Percentage | -0.006090499 |
| Light Sleep Percentage | -0.005803604 |
| Sleep duration | 0.004398362 |
| Exercise frequency | 0.002966852 |
| REM Sleep Percentage | 0.001597183 |
| Gender | 0.000974335 |
| Caffeine Consumption | 0.000103843 |
| Deep Sleep Percentage | 0 |

1. *Analysis of Variance (ANOVA):*

* *ANOVA (Analysis of Variance) is a statistical technique used to determine whether there are statistically significant differences between the means of two or more groups.*
* *In the context of feature selection, ANOVA is applied to each feature individually to assess whether it contributes significantly to the target variable's variability.*
* *Features with high F-statistics and low p-values are considered significant and retained, while those with low F-statistics and high p-values are discarded.*

|  |  |
| --- | --- |
| Feature | Coefficient |
| Deep.sleep.percentage | 9.79E-139 |
| Awakenings | 1.40E-38 |
| Smoking.status | 2.78E-11 |
| Age | 8.11E-08 |
| Alcohol.consumption | 0.006516754 |
| REM.sleep.percentage | 0.02809321 |
| Sleep.duration | 0.1476042 |
| Exercise.frequency | 0.171884 |
| Gender | 0.2355469 |
| Caffeine.consumption | 0.2546618 |

1. *Forward Feature Selection:*

* *Forward Feature Selection is a greedy search algorithm that iteratively builds a model by adding one feature at a time based on some criterion, such as the improvement in model performance.*
* *The algorithm starts with an empty set of features and iteratively adds the feature that provides the greatest improvement in model performance until no further improvement is observed.*
* *Forward FS is computationally efficient and easy to implement, making it suitable for datasets with a relatively small number of features.*

|  |  |
| --- | --- |
| Features | Coefficient |
| Age | 1 |
| Sleep.duration | -0.103164293 |
| REM.sleep.percentage | 0.0694031 |
| Light.sleep.percentage | -0.030106567 |
| Awakenings | -0.0250294 |
| Caffeine.consumption | -0.182968785 |
| Alcohol.consumption | 0.04040126 |
| Exercise.frequency | 0.06991375 |
| Gender Male | 0.22667581 |
| Gender Female | -0.23159756 |
| Smoking.status Yes | 0.04587878 |
| Smoking\_status No | -0.04587878 |

1. **Algorithms Implemented:**
2. *Decision Tree:*

* A decision tree is a flowchart-like tree structure where an internal node represents a feature, the branch represents a decision rule, and each leaf node represents the outcome.
* Decision trees are straightforward to understand and interpret, making them suitable for visualization
* In this project, the rpart package is used to build a Decision Tree model with the specified features.

1. *Random Forest:*

* *Random Forest is an ensemble learning method that constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.*
* *Random Forest improves the accuracy and reduces overfitting compared to a single decision tree by averaging multiple decision trees.*
* *In this project, the randomForest package is used to build a Random Forest model with the specified features.*

1. *Support Vector Regression (SVR):*

* *Support Vector Regression is a regression algorithm that uses the same principles as Support Vector Machines (SVM) for classification but applies them to regression problems.*
* *SVR identifies the hyperplane that best fits the data such that the margin between the hyperplane and the data points is maximized.*
* *In this project, the e1071 package is used to build an SVR model with the specified features.*

1. *Gradient Boosting:*

* *Gradient Boosting is an ensemble learning technique where multiple weak learners (typically decision trees) are sequentially trained, with each subsequent model correcting the errors of its predecessor.*
* *Gradient Boosting iteratively minimizes a loss function by adding weak learners, which makes it less prone to overfitting.*
* *In this project, the gbm package is used to build a Gradient Boosting model with the specified features.*

**Code***:*

*# Load required libraries*

*library(caret)*

*library(rpart)*

*library(randomForest)*

*library(e1071)*

*library(gbm)*

*library(Metrics)*

*# Read the dataset*

*f <- read.csv("Sleep\_Efficiency\_Updated.csv")*

*# Remove unnecessary columns*

*f <- f[, -which(names(f) %in% c("Bedtime", "Wakeup.time", "ID"))]*

*# Check for missing values and impute missing values with mean*

*f$Awakenings[is.na(f$Awakenings)] <- mean(f$Awakenings, na.rm = TRUE)*

*f$Caffeine.consumption[is.na(f$Caffeine.consumption)] <- mean(f$Caffeine.consumption, na.rm = TRUE)*

*f$Alcohol.consumption[is.na(f$Alcohol.consumption)] <- mean(f$Alcohol.consumption, na.rm = TRUE)*

*f$Exercise.frequency[is.na(f$Exercise.frequency)] <- mean(f$Exercise.frequency, na.rm = TRUE)*

*# Convert categorical variables to factors*

*f$Gender <- as.factor(f$Gender) # Male: 1, Female: 0*

*f$Smoking.status <- as.factor(f$Smoking.status) # Yes: 1, No: 0*

*# Train/test split*

*set.seed(123)*

*train\_index <- sample(1:nrow(f), 0.7 \* nrow(f))*

*train\_data <- f[train\_index, ]*

*test\_data <- f[-train\_index, ]*

*#----*

*ctrl <- rfeControl(functions = rfFuncs, method = "cv", number = 5)*

*rfe\_profile <- rfe(x = train\_data[, -which(names(train\_data) == "Sleep.efficiency")],*

*y = train\_data$Sleep.efficiency,*

*sizes = c(1:ncol(train\_data) - 1),*

*rfeControl = ctrl)*

*# Get the selected features*

*selected\_features <- predictors(rfe\_profile)*

*# Train a Random Forest model on the selected features*

*rf\_model <- randomForest(Sleep.efficiency ~ ., data = train\_data[, c(selected\_features, "Sleep.efficiency")])*

*# Extract variable importance scores*

*importance\_scores <- importance(rf\_model)*

*# Sort importance scores in descending order*

*sorted\_importance <- importance\_scores[order(importance\_scores, decreasing = TRUE), ]*

*# Print importance scores*

*# print(as.list(sorted\_importance))*

*selected\_col<- names(sorted\_importance)[1:8]*

*#-------*

*# Random Forest Model*

*rf\_model <- randomForest(Sleep.efficiency ~ ., data = train\_data[, c(selected\_col, "Sleep.efficiency")])*

*# Evaluate model*

*rf\_predictions <- predict(rf\_model, newdata = test\_data[, selected\_col])*

*# Calculate Residual Sum of Squares (RSS) for each model*

*rf\_rss <- sum((rf\_predictions - test\_data$Sleep.efficiency)^2)*

*# Calculate Total Sum of Squares (TSS)*

*mean\_y <- mean(test\_data$Sleep.efficiency)*

*tss <- sum((test\_data$Sleep.efficiency - mean\_y)^2)*

*# Calculate R-squared for each model*

*rf\_r\_squared <- 1 - (rf\_rss / tss)*

*# Print R-squared for each model*

*cat("Random Forest R-squared:", rf\_r\_squared, "\n")*

1. **Shiny app:**

*library(shiny)*

*library(randomForest)*

*f <- read.csv("Sleep\_Efficiency\_Updated.csv")*

*# Load the random forest model from the .rds file*

*rf\_model <- readRDS("rf\_model.rds")*

*f$Smoking.status <- as.factor(f$Smoking.status)*

*# Define UI*

*ui <- fluidPage(*

*titlePanel("Sleep Efficiency Calculator"),*

*sidebarLayout(*

*sidebarPanel(*

*numericInput("light\_sleep", "Light Sleep Percentage:", value = 20, min = 0, max = 100),*

*numericInput("deep\_sleep", "Deep Sleep Percentage:", value = 30, min = 0, max = 100),*

*numericInput("awakenings", "Number of Awakenings:", value = 2, min = 0),*

*numericInput("alcohol\_consumption", "Alcohol Consumption:", value = 2, min = 0),*

*numericInput("age", "Age:", value = 30, min = 0),*

*selectInput("smoking\_status", "Smoking Status:", choices = c("Yes", "No")),*

*numericInput("exercise\_frequency", "Exercise Frequency:", value = 3, min = 0),*

*numericInput("rem\_sleep", "REM Sleep Percentage:", value = 20, min = 0, max = 100),*

*actionButton("calculate\_button", "Calculate Sleep Efficiency")*

*),*

*mainPanel(*

*h3("Results"),*

*verbatimTextOutput("sleep\_efficiency\_result")*

*)*

*)*

*)*

*server <- function(input, output) {*

*observeEvent(input$calculate\_button, {*

*# Convert selectInput choice to 1 or 0*

*# Prepare input data*

*input\_data <- data.frame(*

*Light.sleep.percentage = input$light\_sleep,*

*Deep.sleep.percentage = input$deep\_sleep,*

*Awakenings = input$awakenings,*

*Alcohol.consumption = input$alcohol\_consumption,*

*Age = input$age,*

*Smoking.status = factor(input$smoking\_status, levels = c("Yes", "No")),*

*Exercise.frequency = input$exercise\_frequency,*

*REM.sleep.percentage = input$rem\_sleep*

*)*

*# Print input data for debugging*

*print(input\_data)*

*# Predict sleep efficiency using the random forest model*

*sleep\_efficiency <- predict(rf\_model, newdata = input\_data)*

*# Output sleep efficiency result*

*output$sleep\_efficiency\_result <- renderText({*

*paste("Predicted Sleep Efficiency:", sleep\_efficiency)*

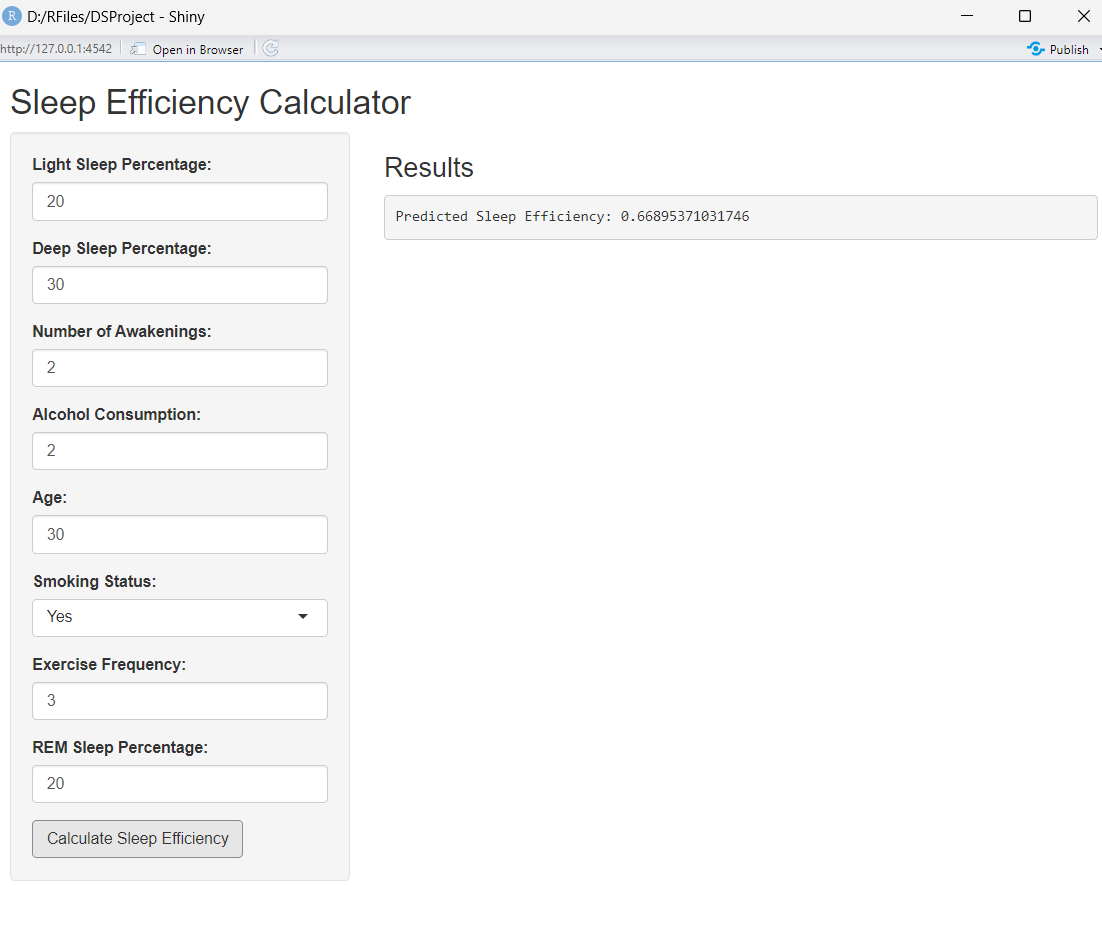
*})*

*})*

*}*

*# Run the application*

*shinyApp(ui = ui, server = server)*

**

*Image: Shiny App UI (along with some calculated values and its output)*

1. **Evaluation Parameters:**

*Evaluation parameters provide different perspectives on model performance. It calculated for each of the machine learning models (Decision Tree, Random Forest, SVR, Gradient Boosting) to compare their performance in predicting sleep efficiency based on the selected features.*

1. *RMSE:*

* *RMSE is a commonly used metric to evaluate the accuracy of regression models.*
* *It measures the average magnitude of the errors between predicted values and actual values.*
* *RMSE is calculated by taking the square root of the mean of the squared differences between predicted and actual values.*
* *Lower RMSE values indicate better model performance, with a value of 0 representing perfect predictions.*

1. *MAE:*

* *MAE is another metric for evaluating the accuracy of regression models.*
* *It measures the average absolute difference between predicted values and actual values.*
* *MAE is calculated by taking the mean of the absolute differences between predicted and actual values.*
* *Lower MAE values indicate better model performance.*

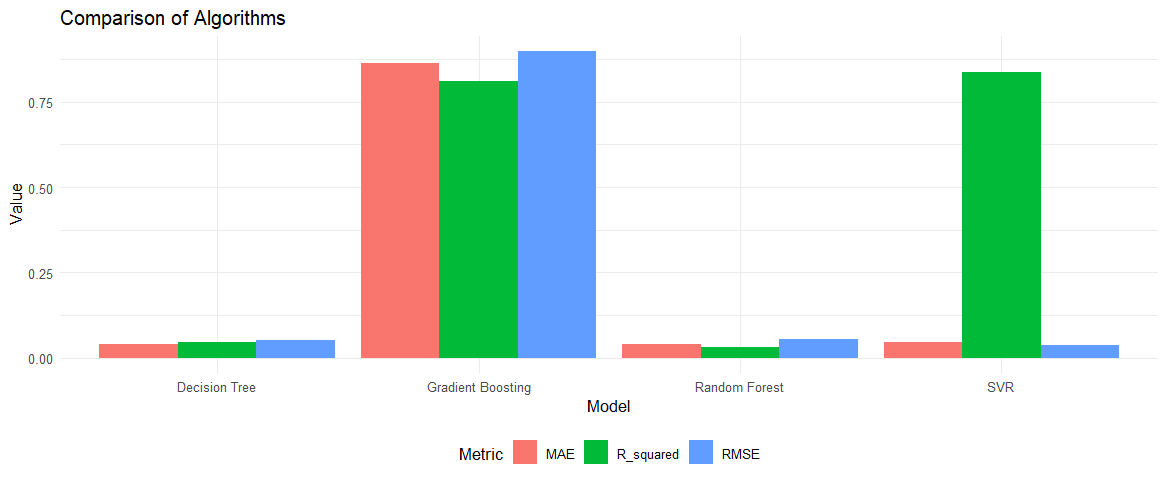
1. *R-Squared:*

* *R-squared is a statistical measure that represents the proportion of variance in the dependent variable (target) that is explained by the independent variables (features) in the model.*
* *It ranges from 0 to 1, where 0 indicates that the model does not explain any variability in the target variable, and indicates that the model perfectly explains the variability.*
* *R-squared values closer to 1 indicate better model fit, while values closer to 0 indicate poor model fit.*
* *R-squared can be interpreted as the percentage of the variance in the target variable that is accounted for by the independent variables.*

Lower RMSE and MAE values and higher R-squared values indicate better predictive performance of the models.

1. **Results and Discussions:**
   1. *Experiment: Algorithms without any feature selection methods.*

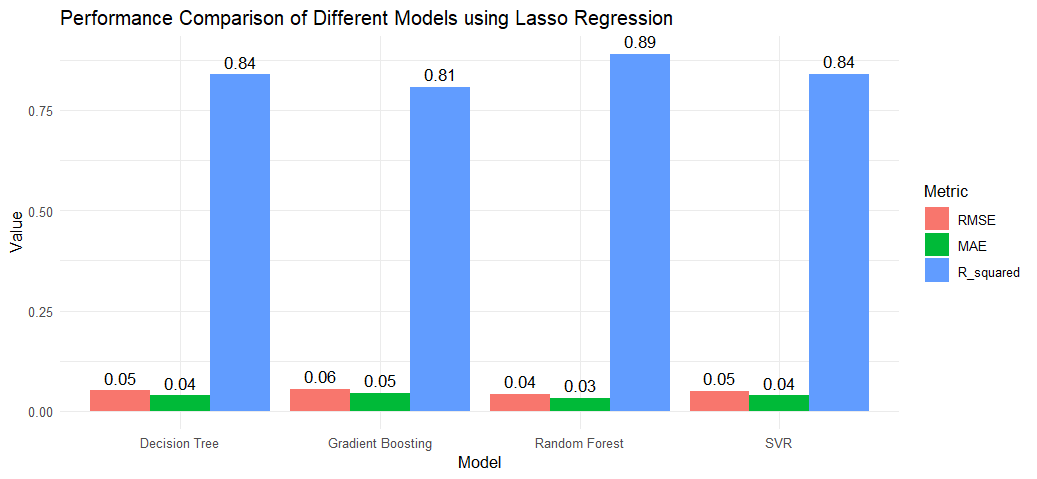
|  |  |  |  |
| --- | --- | --- | --- |
| Algorithms \Evaluation Parameters | R-Squared | RMSE | MAE |
| DT | 0.8386299 | 0.0510072 | 0.0395892 |
| RF | 0.8980748 | 0.0405378 | 0.0315786 |
| SVR | 0.8644608 | 0.0467468 | 0.0360256 |
| GB | 0.8108494 | 0.0552235 | 0.0448295 |



Algorithms are evaluated without any feature selection methods

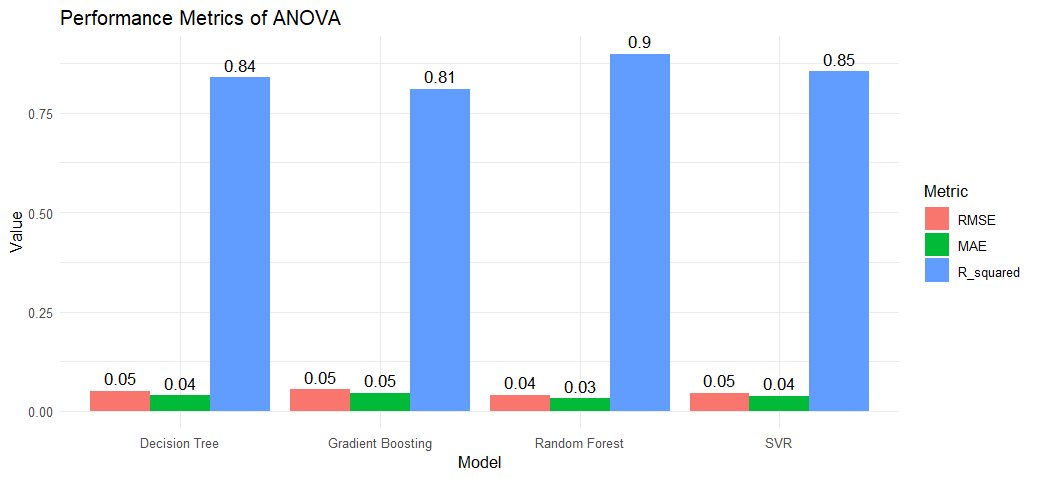
* 1. *Experiment: Algorithms with Lasso Regression feature selection method using all the variables.*

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithms \Evaluation Parameters | R-Squared | RMSE | MAE |
| DT | 0.8386299 | 0.03958917 | 0.05100718 |
| RF | 0.8897944 | 0.03252333 | 0.04215234 |
| SVR | 0.8402972 | 0.03870235 | 0.050743 |
| GB | 0.8077984 | 0.04534769 | 0.05566706 |



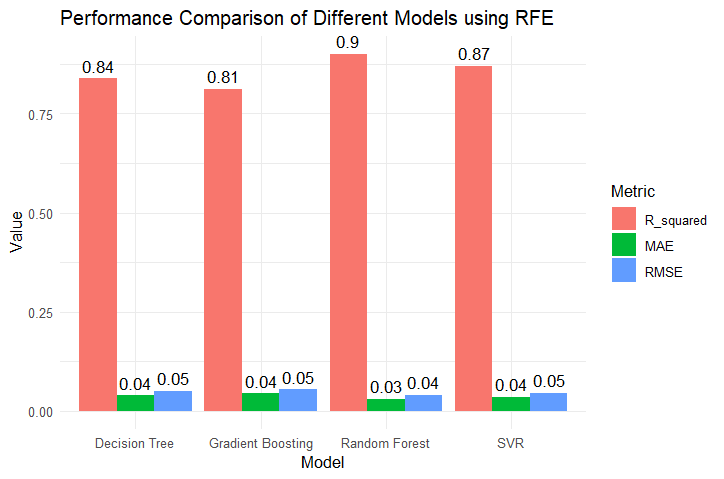
* 1. *Experiment: Algorithms with ANOVA feature selection method using all the variables.*

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithms \Evaluation Parameters | R-Squared | RMSE | MAE |
| DT | 0.8386299 | 0.03958917 | 0.05100718 |
| RF | 0.8982078 | 0.03145493 | 0.04345587 |
| SVR | 0.8546201 | 0.03804044 | 0.05224341 |
| GB | 0.8098832 | 0.04536163 | 0.05580576 |



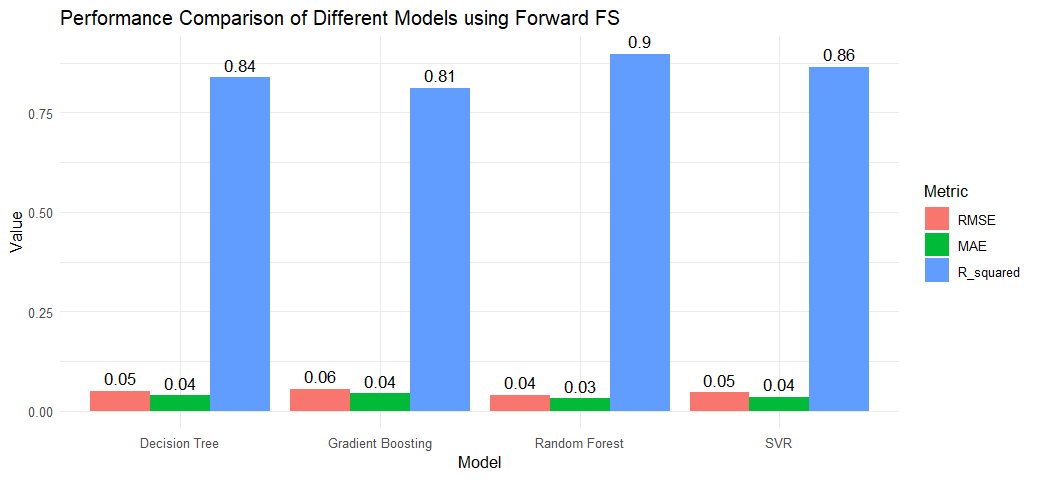
* 1. *Experiment: Algorithms with RFE feature selection method using all the variables.*

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithms \Evaluation Parameters | R-Squared | RMSE | MAE |
| DT | 0.8386299 | 0.03958917 | 0.05100718 |
| RF | 0.9001484 | 0.03097975 | 0.04012337 |
| SVR | 0.8705219 | 0.0355086 | 0.04568967 |
| GB | 0.8126672 | 0.0448281 | 0.05495747 |



* 1. *Experiment: Algorithms with Forward feature selection method using all the variables.*

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithms \Evaluation Parameters | R-Squared | RMSE | MAE |
| DT | 0.8386299 | 0.03958917 | 0.05100718 |
| RF | 0.8971596 | 0.03179588 | 0.04071943 |
| SVR | 0.8644608 | 0.03602563 | 0.04674684 |
| GB | 0.8115416 | 0.04484214 | 0.05512233 |



* 1. *Experiment: Algorithms with Lasso feature selection method using different number of the variables.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Algorithms | R-Squared | MAE | RMSE |
| 5 | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.8206704 | 0.04404145 | 0.05377072 |
| SVR | 0.8341263 | 0.0398274 | 0.05171405 |
| Gradient Boosting | 0.8043706 | 0.04592306 | 0.05616126 |
| 6 | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.8818027 | 0.03391238 | 0.04365395 |
| SVR | 0.8448462 | 0.03788762 | 0.05001509 |
| Gradient Boosting | 0.8058991 | 0.04540295 | 0.05594143 |
| 7 | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.899596 | 0.03172168 | 0.0402342 |
| SVR | 0.8602179 | 0.03619004 | 0.04747288 |
| Gradient Boosting | 0.8099104 | 0.04499346 | 0.05536037 |
| 8 | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.8913242 | 0.03194344 | 0.04185875 |
| SVR | 0.8363151 | 0.03933639 | 0.05137172 |
| Gradient Boosting | 0.8059412 | 0.04524087 | 0.05593536 |
| 9 | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.8878153 | 0.03273276 | 0.04252915 |
| SVR | 0.8287564 | 0.03992389 | 0.05254446 |
| Gradient Boosting | 0.8055061 | 0.04547221 | 0.05599804 |

* 1. *Experiment: Algorithms with ANOVA feature selection method using different number of the variables.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Algorithms | R-Squared | MAE | RMSE |
| 5 | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.8222243 | 0.04339198 | 0.05353724 |
| SVR | 0.05463274 | 0.04148923 | 0.05463274 |
| Gradient Boosting | 0.05660358 | 0.04608853 | 0.05660358 |
| 6 | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.8826169 | 0.03322187 | 0.04033001 |
| SVR | 0.8307131 | 0.03945961 | 0.04598407 |
| Gradient Boosting | 0.8057428 | 0.04520479 | 0.0547953 |
| 7 | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.879896 | 0.03350434 | 0.04115451 |
| SVR | 0.8368838 | 0.03874966 | 0.04530165 |
| Gradient Boosting | 0.7991878 | 0.04599598 | 0.05471283 |
| 8 | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.894505 | 0.03185454 | 0.04033001 |
| SVR | 0.855286 | 0.03770485 | 0.04598407 |
| Gradient Boosting | 0.808464 | 0.04489698 | 0.0547953 |
| 9 | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.8983192 | 0.03102132 | 0.05377072 |
| SVR | 0.8521046 | 0.03816008 | 0.05171405 |
| Gradient Boosting | 0.8128277 | 0.04458619 | 0.05616126 |

* 1. *Experiment: Algorithms with RFE feature selection method using different number of the variables.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Algorithms | R-Squared | MAE | RMSE |
| 5 Variables | Decision Tree | 0.8157928 | 0.04448673 | 0.05449706 |
| Random Forest | 0.846305 | 0.04054931 | 0.04977941 |
| SVR | 0.8013448 | 0.04331001 | 0.05659392 |
| Gradient Boosting | 0.8002754 | 0.04629332 | 0.05674604 |
| 6 Variables | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.880324 | 0.03377912 | 0.04392618 |
| SVR | 0.8448462 | 0.03788762 | 0.05001509 |
| Gradient Boosting | 0.8048293 | 0.04525645 | 0.05609538 |
| 7 Variables | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.8949502 | 0.03223263 | 0.04115451 |
| SVR | 0.8727117 | 0.03558851 | 0.04530165 |
| Gradient Boosting | 0.8143313 | 0.04439003 | 0.05471283 |
| 8 Variables | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.8991172 | 0.03176168 | 0.04033001 |
| SVR | 0.8688479 | 0.03530871 | 0.04598407 |
| Gradient Boosting | 0.8137712 | 0.0441578 | 0.0547953 |
| 9 Variables | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.8996232 | 0.03101658 | 0.04022874 |
| SVR | 0.8704227 | 0.03521038 | 0.04570717 |
| Gradient Boosting | 0.8144273 | 0.0443134 | 0.05469868 |

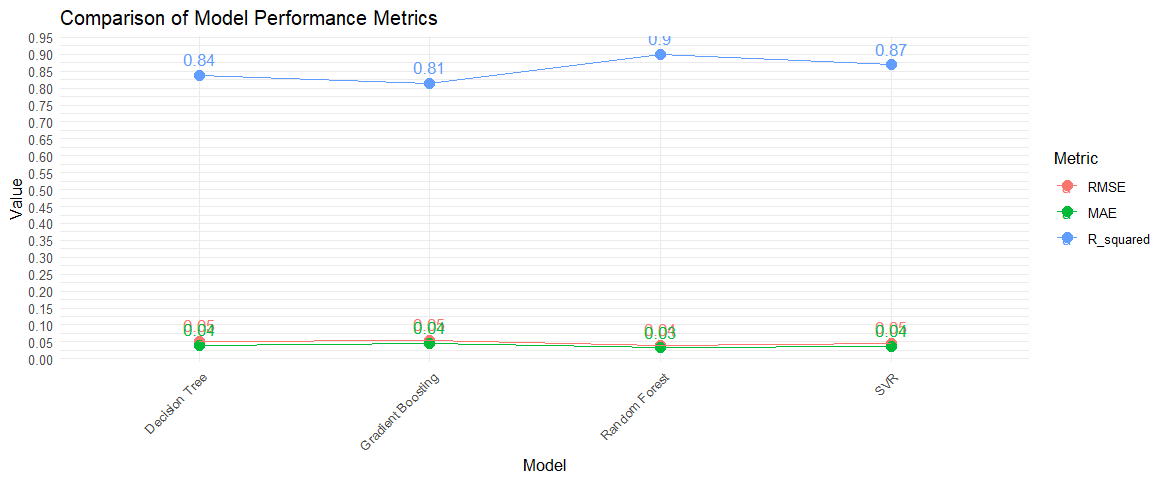
* 1. *Experiment: Algorithms with Forward feature selection method using different number of the variables.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Algorithms | R-Squared | MAE | RMSE |
| 5 | Decision Tree | 0.6450164 | 0.06516263 | 0.07565262 |
| Random Forest | 0.6980616 | 0.05772118 | 0.06977165 |
| SVR | 0.6584758 | 0.05888257 | 0.07420456 |
| Gradient Boosting | 0.6606533 | 0.06300075 | 0.07396762 |
| 6 | Decision Tree | 0.6450164 | 0.06516263 | 0.07565262 |
| Random Forest | 0.748825 | 0.0514089 | 0.06363675 |
| SVR | 0.6624466 | 0.05857591 | 0.07377191 |
| Gradient Boosting | 0.6824886 | 0.06135903 | 0.07154832 |
| 7 | Decision Tree | 0.6450164 | 0.06516263 | 0.07565262 |
| Random Forest | 0.7474923 | 0.05135135 | 0.06380536 |
| SVR | 0.6721603 | 0.05828064 | 0.07270271 |
| Gradient Boosting | 0.6727316 | 0.06187535 | 0.07263934 |
| 8 | Decision Tree | 0.8157928 | 0.04448673 | 0.05449706 |
| Random Forest | 0.8607046 | 0.03657309 | 0.04739016 |
| SVR | 0.7948335 | 0.04414834 | 0.05751393 |
| Gradient Boosting | 0.7971472 | 0.0461 | 0.05718871 |
| 9 | Decision Tree | 0.8157928 | 0.04448673 | 0.05449706 |
| Random Forest | 0.8675632 | 0.03542589 | 0.04620875 |
| SVR | 0.8008699 | 0.0435612 | 0.05666153 |
| Gradient Boosting | 0.7999108 | 0.04557091 | 0.05679782 |

* 1. *Experiment: Decision Tree and Random Forest Algorithms with RFE and ANOVA feature selection method using all the variables.*

**ANOVA** **RFE**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Algorithms | R-Squared | MAE | RMSE | R-Squared | MAE | RMSE |
| 5 | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 | 0.8157928 | 0.04448673 | 0.05449706 |
| Random Forest | 0.8222243 | 0.04339198 | 0.05353724 | 0.846305 | 0.04054931 | 0.04977941 |
| 6 | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.8826169 | 0.03322187 | 0.04033001 | 0.880324 | 0.03377912 | 0.04392618 |
| 7 | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.879896 | 0.03350434 | 0.04115451 | 0.8949502 | 0.03223263 | 0.04115451 |
| 8 | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.894505 | 0.03185454 | 0.04033001 | 0.8991172 | 0.03176168 | 0.04033001 |
| 9 | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.8983192 | 0.03102132 | 0.05377072 | 0.8996232 | 0.03101658 | 0.04022874 |
| All | Decision Tree | 0.8386299 | 0.03958917 | 0.05100718 | 0.8386299 | 0.03958917 | 0.05100718 |
| Random Forest | 0.8982078 | 0.03145493 | 0.04345587 | 0.9001484 | 0.03097975 | 0.04012337 |



1. **Conclusions:**

* *In the final model of Sleep Efficiency prediction, we have Random Forest regression model trained on a dataset containing sleep efficiency data after feature selection using Recursive Feature Elimination (RFE) with a Random Forest algorithm. The model aims to predict sleep efficiency based on various predictors such as age, sleep duration, awakenings, caffeine consumption, smoking status, gender, and other factors which affect sleep patterns.*
* *The choice of Random Forest as the algorithm is chosen due to its ability to handle non-linear relationships and interactions between features effectively, making it suitable for capturing complex patterns in the data. R-squared is chosen as the evaluation metric because it provides an indication of the proportion of variance in the dependent variable (sleep efficiency) explained by the independent variables (predictors) and provides near 0.9 values whereas MAE and RMSE need near 0 value and here they fail to do so.*
* *Selecting eight variables through RFE allows for a balance between model complexity and performance, aiming to capture the most relevant predictors as we aim to keep minimum columns and highest prediction accuracy. Additionally, RFE helps in identifying the subset of features that contribute the most to the model's predictive accuracy, leading to improved interpretability and potentially better generalization to unseen data.*
* *Overall, the Random Forest model with RFE-selected features achieves a satisfactory level of performance, as indicated by the obtained R-squared value, providing valuable insights into factors influencing sleep efficiency.*

**References:**

1. Ibáñez, Vanessa, Josep Silva, and Omar Cauli. "A survey on sleep assessment methods." *PeerJ* 6 (2018): e4849.
2. Kim, Jiyong, and Minseo Park. "A Study on ML-Based Sleep Score Model Using Lifelog Data." *Applied Sciences* 13, no. 2 (2023): 1043.
3. Hermans, Lieke WA, Iris AM Huijben, Hans van Gorp, Tim RM Leufkens, Pedro Fonseca, Sebastiaan Overeem, and Merel M. van Gilst. "Representations of temporal sleep dynamics: Review and synthesis of the literature." *Sleep Medicine Reviews* 63 (2022): 101611.
4. Lastella, Michele, Aamir Raoof Memon, and Grace E. Vincent. "Global research output on sleep research in athletes from 1966 to 2019: a bibliometric analysis." *Clocks & sleep* 2, no. 2 (2020): 99-119.
5. Lastella, Michele, Aamir Raoof Memon, and Grace E. Vincent. "Global research output on sleep research in athletes from 1966 to 2019: a bibliometric analysis." *Clocks & sleep* 2, no. 2 (2020): 99-119.
6. Xu, Shuting, Oliver Faust, Silvia Seoni, Subrata Chakraborty, Prabal Datta Barua, Hui Wen Loh, Heather Elphick, Filippo Molinari, and U. Rajendra Acharya. "A review of automated sleep disorder detection." *Computers in Biology and Medicine* 150 (2022): 106100.
7. Adnane, Mourad, Zhongwei Jiang, and Zhonghong Yan. "Sleep–wake stages classification and sleep efficiency estimation using single-lead electrocardiogram." *Expert Systems with Applications* 39, no. 1 (2012): 1401-1413.
8. Wongsirichot, Thakerng, Nittida Elz, Supasit Kajkamhaeng, Wanchai Nupinit, and Narongrit Sridonthong. "An investigation of data mining based Automatic Sleep Stage Classification techniques." *International Journal of Machine Learning and Computing* 9, no. 4 (2019): 520-526.
9. Kim, Jiyong, and Minseo Park. "A Study on ML-Based Sleep Score Model Using Lifelog Data." *Applied Sciences* 13, no. 2 (2023): 1043.
10. Kalintha, Wasin, Takafumi Kato, and Ken–ichi Fukui. "SleepAge: sleep quality assessment from nocturnal sounds in home environment." *Procedia Computer Science* 176 (2020): 898-907.
11. Alabdan, Rana, Hanan Abdullah Mengash, Mohammed Maray, Faiz Alotaibi, Sitelbanat Abdelbagi, and Ahmed Mahmud. "Modified Bald Eagle Search Algorithm With Deep Learning-Driven Sleep Quality Prediction for Healthcare Monitoring Systems." *IEEE Access* 11 (2023): 135385-135393.
12. Yan, Bin, Binbin Zhao, Xiaoying Jin, Wenyu Xi, Jian Yang, Lihong Yang, and Xiancang Ma. "Sleep efficiency may predict depression in a large population-based study." *Frontiers in Psychiatry* 13 (2022): 838907.
13. Reed, David L., and William P. Sacco. "Measuring sleep efficiency: what should the denominator be?." *Journal of clinical sleep medicine* 12, no. 2 (2016): 263-266.
14. Halson, Shona L., Rich D. Johnston, Laura Piromalli, Benita J. Lalor, Stuart Cormack, Gregory D. Roach, and Charli Sargent. "Sleep regularity and predictors of sleep efficiency and sleep duration in elite team sport athletes." *Sports Medicine-Open* 8, no. 1 (2022): 79.
15. Desjardins, Sophie, Sylvie Lapierre, Carol Hudon, and Alain Desgagné. "Factors involved in sleep efficiency: a population-based study of community-dwelling elderly persons." *Sleep* 42, no. 5 (2019): zsz038.
16. Alabdan, Rana, Hanan Abdullah Mengash, Mohammed Maray, Faiz Alotaibi, Sitelbanat Abdelbagi, and Ahmed Mahmud. "Modified Bald Eagle Search Algorithm With Deep Learning-Driven Sleep Quality Prediction for Healthcare Monitoring Systems." *IEEE Access* 11 (2023): 135385-135393.
17. Tanuku, Sai Rohith, and Venkat Tummala. "Integration of Feature Selection Techniques using a Sleep Quality Dataset for Comparing Regression Algorithms." *arXiv preprint arXiv:2303.02467* (2023).