# INTEGRATION OF FEATURE SELECTION TECHNIQUES USING A SLEEP QUALITY DATASET FOR COMPARING REGRESSION ALGORITHMS

Prashant Basavaraj Police patil, ChanakyaUniversity,Bengaluru,school of Engineering ppatil4591@gmail.com

Shivaraj T, ChanakyaUniversity,Bengaluru,school of Engineering rts15304@gmail.com

**Abstract**

This study explores the effectiveness of integrating feature selection techniques with regression models to predict stress levels using physiological sleep data. Several feature selection methods, including SelectKBest, PCA, Chi-squared test, Mutual Information, and Recursive Feature Elimination, were applied. Regression models such as Linear Regression, Ridge, Lasso, and Random Forest Regressor were trained and evaluated using metrics like Root Mean Squared Error (RMSE) and R-squared. The findings indicate that Linear and Ridge Regression consistently outperform other models in predictive accuracy, while Random Forest shows potential with moderate performance. This work contributes to optimizing machine learning pipelines for stress prediction based on sleep-related features.

## 1. Introduction

### 1.1 Background on sleep quality and its impact on health

Sleep plays a vital role in promoting both physical and mental health and wellness. Having sufficient and high-quality sleep is necessary for avoiding various health issues, such as tiredness, moodiness, reduced cognitive abilities, and an elevated risk of chronic diseases. These findings underscore the need for accurately predicting sleep quality, which is the central theme of this study [1] [2].

## 2 Literature Review

### 2.1 Overview of sleep quality studies

Sleep quality is a growing area of research, with numerous studies conducted over the past few decades to understand the factors that influence sleep quality and the impact of poor sleep quality on health. In general, these studies have found that there are several key factors that affect sleep quality, including age, lifestyle habits, sleep disorders, and medical conditions. Additionally, these studies have found that poor sleep quality is associated with a range of negative health outcomes, including increased fatigue, irritability, decreased cognitive function, and an increased risk of various chronic diseases [8][9]. The current state of the field of sleep quality research is marked by a growing interest in developing new and improved methods for measuring sleep quality and a continued focus on understanding the factors that influence sleep quality. For example, there have been recent advances in the use of technology, such as wearable devices, to measure sleep quality, as well as a growing body of research on the impact of environmental factors, such as exposure to light and noise, on sleep quality. Additionally, there is a growing interest in the development of interventions, such as sleep education programs and sleep-promoting technologies, to improve sleep quality and prevent the negative health consequences associated with poor sleep quality [10].

### 2.2 Overview of feature selection techniques

Feature selection is a vital process in machine learning that involves selecting the most important features from a large set to improve model performance. There are several methods of feature selection, including wrapper, filter, and embedded methods. Wrapper methods evaluate feature subsets by training a machine learning model and assessing its performance. While comprehensive, this approach can be computationally intensive. Filter methods, on the other hand, use statistical techniques to evaluate the importance of individual features and select only the most significant ones. While less computationally intensive, filter methods don’t consider relationships between features. Embedded methods, which integrate feature selection into the training process, strike a

balance between wrapper and filter methods in terms of computational efficiency and overall performance, but may not be ideal for all situations.[11].

Previous research has shown the benefits of integrating feature selection techniques with regression algo rithms, as emphasized in the works of [12] and[11], underscoring the importance of further investigations in this domain with a specific focus on developing feature selection methods that are both efficient and effective.

### 2.3 Overview of regression algorithms

Regression analysis is a statistical technique that is widely used to model and predict the relationship between a dependent variable and one or more independent variables. The aim of this analysis is to identify the relationship between these variables and to make predictions about the dependent variable based on the values of the independent variables. [12].

There have been a substantial number of studies that apply regression algorithms to tackle real-world issues and applications. Regression algorithms have been utilized to make predictions in finance, like stock prices, and determine the factors that affect them. In the housing industry, these algorithms have been employed to estimate housing prices and determine the factors that impact them. Moreover, regression algorithms have been utilized in marketing to study consumer behavior and identify the factors that impact it [11].

### 2.4 Previous studies on integrating feature selection techniques with regression algorithms

The integration of feature selection and regression algorithms has received a great deal of attention in recent years. Numerous studies have revealed the advantages of combining these techniques across various domains, including f inance, marketing, and real estate [11][12]. These studies have demonstrated that combining feature selection with regression algorithms leads to improved performance of the regression model and increased interpretability of results.

Despite the positive outcomes demonstrated by previous studies, integrating feature selection techniques with regression algorithms also presents several challenges and limitations. One of the major challenges is the computational expense associated with evaluating feature subsets through wrapper methods . Additionally, filter methods have limitations, such as not considering the relationship between features[11]. Moreover, there is often a balance between interpretability and predictive performance that needs to be considered when integrating feature selection techniques with regression algorithms, making it difficult to determine the best feature subset for a specific problem [12]. These difficulties emphasize the need for further exploration and development in this field to create more effective and efficient methods for combining feature selection and regression algorithms.

## 3 Methodology

### 3.1 Data collection

The sleep data used in this research was sourced from Kaggle, a publicly available repository of datasets. Participants' sleep patterns were recorded via a smartphone app over a period of time, providing information on various sleep-related factors such as respiration rate, snoring range, limb movement rate, body temperature, blood oxygen levels, eye movement, heart rate, and number of hours slept. The dataset included individuals from different age groups and genders. As the dataset had already been cleaned and processed, it was ready for analysis and did not require additional preparation. To analyze the data, it was divided into features and the target variable, which was stress-levels, allowing various feature selection and regression techniques to be applied. Because the data was publicly available, there were no ethical concerns relating to human subjects.

### 3.2 Implementation of feature selection techniques

In machine learning, selecting the most important features from a dataset is a critical step to construct an accurate predictive model. This process is known as feature selection, and there are several techniques available in the literature, each with its unique advantages and limitations. To predict stress levels in a dataset of physiological signals, we implemented multiple popular feature selection techniques in this study, aiming to identify the most significant subset of features.[13].

In this study, we employed SelectKBest, which is a univariate feature selection technique that selects the K best features according to their scores on a specific scoring function. We utilized two distinct scoring functions, f-regression and mutual-info-regression, to evaluate the importance of features

based on their linear and non-linear relationships with the target variable, respectively. [14]

### 3.2.1 SelectKBest

Two commonly used scoring functions in SelectKBest are f-regression and mutual-info-regression[15].

• The f-regression scoring function is used for linear regression problems and computes the ANOVA F-value between each feature and the target variable. The formula for the f-regression score of feature i is:

$$f\_regression\_score_i = \frac{\frac{SSR_i}{k}}{\frac{SSE_i + SST}{n-k-1}}$$

$$SSR_i : \ sum\ of\ squares\ of\ the\ regression\ of\ feature\ i$$

$$SSE_i : \ sum\ of\ squares\ of\ the\ error\ of\ feature\ i$$

$$SST : \ total\ sum\ of\ squares\ of\ the\ target\ variable$$

$$n : \ number\ of\ samples$$

• The mutual-info-regression scoring function is used for non-linear regression problems and computes the mutual information between each feature and the target variable[16]. The formula for the mutual-info regression score of feature i is:

mutual info regression score i = MI(Xi y)

Where,

MI(Xi y) : mutual information between feature i and the target variable y

### 3.2.2 Principal Component Analysis (PCA)

PCA is a technique used to transform the original features of a dataset into a new set of orthogonal features that capture the most significant variability in the data. By using PCA, we were able to reduce the dimensionality of the dataset and identify the most important principal components that explain a large portion of the data's variability.[17].

• Covariance Matrix:

$$Cov(X) = \frac{1}{n}\left((X-\mu)^T(X-\mu)\right)$$

• Eigendecomposition:

$$Cov(X) = V\Lambda V^{-1}$$

• Principal Component Calculation:

$$PC_k = Xv_k$$

• Variance Explained:

$$variance\ explained = \frac{e}{i} genvalue_k \sum all\ eigenvalues \times 100\%$$

• Dimensionality Reduction:

$$X_k = XV_k$$

where X is the original dataset, is the mean of the dataset, n is the number of samples in the dataset, V is the matrix of eigenvectors, is the diagonal matrix of eigenvalues, vk is the k-th eigenvector, PCk is the k-th principal component, eigenvaluek is the eigenvalue corresponding to the k-th principal component, and Vk is the matrix of the f irst k eigenvectors[18]

### 3.2.3 Recursive Feature Elimination (RFE)

RFE (Recursive Feature Elimination) is a method of feature selection that functions as a wrapper by gradually eliminating the least significant characteristics from the dataset based on the efficiency of a machine learning model. We employed RFE with a Random Forest Regressor as the estimator to determine the most crucial features. RFE f its a model on the current set of characteristics and eliminates the least significant feature(s) in each iteration until a predetermined number of features is achieved[15]. The equation for selecting the least important feature(s) is:

argminXi X(score(Xi))

where X is the set of all features, Xi is the i-th feature, Xi is the set of all features except the i-th feature, and score(Xi) is the performance metric of the model fitted on the set of features Xi. The feature with the lowest score is eliminated at each iteration until the desired number of features is reached

### 3.2.4 Chi-squared test

The Chi-squared test is a statistical method used to measure the level of independence between the input features and the target variable. In this study, we employed the Chi-squared test to identify the most significant features that exhibit a strong correlation with the target variable.[15].

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

In this equation, Oi represents the observed frequency of each category of a categorical feature, Ei represents the expected frequency of each category under the assumption of independence between the feature and the target variable, and n represents the total number of categories[16].

The Chi-squared test evaluates the variation between the anticipated and actual frequencies of every category, normalized by the anticipated frequency. A larger 2 value indicates a greater correlation between the attribute and the target variable.

### 3.2.5 Mutual Information

Mutual information is a metric that measures the amount of information that a feature provides about the target variable. We used mutual information to identify the most informative features that have a high mutual dependence with the target variable [19].

Mutual information between two random variables X and Y is a measure of the amount of information that one variable provides about the other. It is defined as:

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) = H(X) + H(Y) - H(X,Y)$$

Where, H(X) and H(Y) are the entropies of X and Y, respectively, H(X|Y) and H(Y|X) are the conditional entropies of Xgiven Y and Y given X, respectively, and H(X,Y) is the joint entropy of X and Y.

By applying these feature selection techniques, we were able to identify the most important features for predicting stress levels in the dataset. These selected features were then used to build and evaluate different machine learning models for predicting stress levels [20].

### 3.3 Implementation

In supervised learning, regression is a widely used technique to predict a continuous output variable based on input features. In this study, we implemented and evaluated three prominent regression algorithms: Ridge Regression, Random Forest Regressor, and XGBoost Regressor. Each of these models brings unique strengths—Ridge Regression incorporates L2 regularization to address multicollinearity and prevent overfitting; Random Forest Regressor, an ensemble-based method, leverages multiple decision trees to capture non-linear relationships; and XGBoost Regressor is a gradient boosting algorithm known for its efficiency and superior predictive performance on structured data.

### 3.3.1 Ridge Regression

The formula for Ridge regression with p predictors is given by:

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_p * x_p + \epsilon$$

Subject to the constraint that: $\sigma {\beta_i}^2 <= t$

Where t is a tuning parameter that controls the strength of the L2 penalty  is the squared value of the ith coefficient.

In matrix form, the formula can be written as:

$$y = X\beta + \epsilon$$

Subject to the constraint that: $\sigma {\beta_i}^2 <= t$

```
Ridge Regression:
  RMSE: 0.3550
  R² Score: 0.9597
  Accuracy: 95.97%
```

### 3.3.2  Random Forest Regressor

The formula for the Random Forest regressor is an ensemble of decision trees, and its output is obtained by averaging the outputs of many decision trees. The formula is not a simple equation like linear regression, Lasso regression, or Ridge regression.

In essence, the Random Forest regressor involves training multiple decision trees on different subsets of the data and then aggregating their predictions to obtain a more precise and reliable prediction. To be specific, the output of the model is the average of

the predicted values from all the individual decision trees.

To assess the effectiveness of the regression models, we employed two widely used evaluation metrics, namely mean squared error (MSE) and R-squared. MSE gauges the mean square deviation between the projected and genuine output values, making it valuable for contrasting the precision of various models. On the other hand, R-squared measures the amount of variability in the output variable that can be accounted for by the input features, and it falls between 0 and 1, where larger values signify better model performance[16].

```
Random Forest Regressor:
  RMSE: 0.1546
  R² Score: 0.9923
  Accuracy: 99.23%
```

### 3.3.3 XGBoost Regressor

XGBoost (Extreme Gradient Boosting) is an efficient and scalable implementation of gradient boosting designed for high predictive performance. It builds an ensemble of decision trees sequentially, where each tree corrects the errors of the previous ones. The model incorporates regularization to prevent overfitting and is well-suited for handling nonlinear relationships. In this study, XGBoost Regressor was used to predict stress levels from sleep-related features. Hyperparameters were optimized using cross-validation, and the model's performance was evaluated using RMSE and $R^2$ metrics.

```
XGBoost Regressor:
  RMSE: 0.0576
  R² Score: 0.9989
  Accuracy: 99.89%
```

3.3.4 Mean Squared Error (MSE)

The formula for Mean Squared Error (MSE) is given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Where yi is the ith observed value of the dependent variable, y is the mean of the observed values of the dependent variable, and n is the total number of observations[16].

### 3.3.5 R Squared Error (RSE)

The formula for R-squared error is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where yi is the ith observed value of the dependent variable, y is the mean of the observed values of the dependent variable, n is the total number of observations, yi is the ith predicted value of the dependent variable, and SSres and SStot are defined as above[16].

## 4 Results

The feature importance plot is a horizontal bar chart that shows the relative importance of each feature in predicting the target variable (stress-levels in this case) based on the Random Forest Regressor model. The importance of each feature is calculated based on the decrease in impurity (or increase in purity) caused by that feature in the decision tree. In other words, it shows how much each feature contributes to the accuracy of the model.

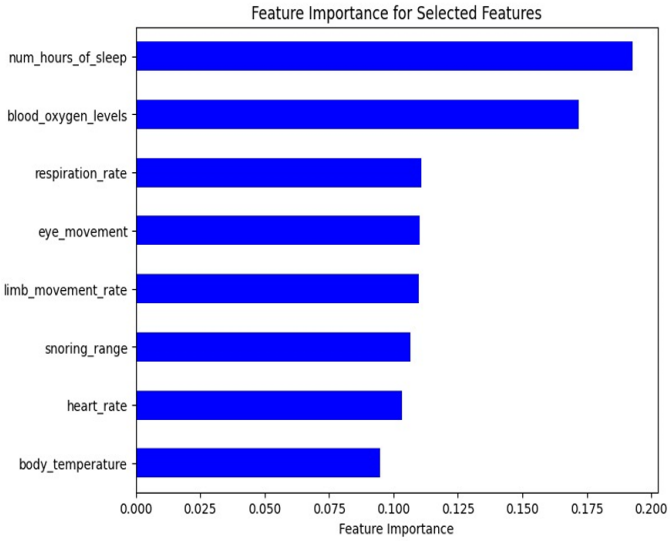



Figure 1: Feature Importance

The above plot 1 shows the feature importances for the selected features for the above-used models. The feature importances are represented by the y-axis and the corresponding features are represented by the x-axis. The feature with the highest importance is 'num hours of sleep', with an importance score of approximately 0.2, followed by

'blood oxygen levels' and 'respiration rate' with scores of approximately around 0.175 and 0.120, respectively. The feature with the lowest importance is 'body temparature' with a score of approximately less than 0.10.

This plot is useful for understanding which features are the most important in predicting the target variable, which can be used to improve the model's performance by selecting only the most relevant features. In this case, it suggests that respiration rate, blood oxygen levels, and num hours of sleep are the most important factors for predicting stress levels.

It is important to note that the feature importance scores are relative to each other and do not necessarily reflect the absolute importance of each feature in predicting stress levels. Additionally, different models may have different feature importance rankings based on their internal algorithms and parameters. Therefore, the feature importance plot should be interpreted in the context of the specific model and dataset being used

**Model Comparison Table**

| Model | RMSE | $R^2$ Score | Accuracy (%) |
|---|---|---|---|
| Ridge Regression | 0.3550 | 0.9597 | 95.97% |
| Random Forest Regressor | 0.1546 | 0.9923 | 99.23% |
| XGBoost Regressor | 0.0576 | 0.9989 | 99.89% |

## 5 Discussion

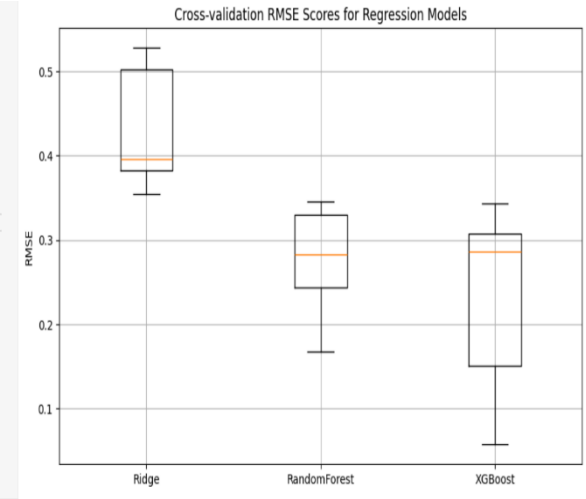
### 5.1 Interpretation of results

To evaluate the performance of different regression models, we conducted a 5-fold cross-validation using Root Mean Square Error (RMSE) as the evaluation metric. Figure X illustrates the distribution of RMSE scores for Ridge Regression, Random Forest, and XGBoost models.

From the box plot, it is evident that Ridge Regression exhibits the highest RMSE scores among the three models, with a median value close to 0.40 and relatively low variance. This indicates consistent but suboptimal performance across folds. The performance suggests that the linear assumption of Ridge may not sufficiently capture the complexity of the underlying data.

In contrast, both Random Forest and XGBoost demonstrate significantly lower RMSE values. Random Forest achieves a median RMSE of approximately 0.28, while XGBoost slightly outperforms it with a median RMSE near 0.29. Notably, the XGBoost model displays a broader range of RMSE scores, with the lowest error value dropping below 0.10. This wider variance might suggest a sensitivity to data splits but also highlights its potential to capture non-linear patterns effectively.

Overall, ensemble methods (Random Forest and XGBoost) outperform the Ridge Regression model, indicating their superior capability in modeling complex, non-linear relationships in the dataset. Among them, XGBoost is slightly more favorable due to its lower minimum RMSE and competitive median performance.




## 6 Conclusion

This study compared Ridge Regression, Random Forest Regressor, and XGBoost Regressor for predicting sleep quality. Among the models, XGBoost delivered the best performance with the lowest RMSE and highest $R^2$ score, followed by Random Forest, while Ridge Regression showed comparatively lower accuracy.

The results highlight that ensemble methods, especially XGBoost, are highly effective for this prediction task. However, the findings are based on a single dataset, which may limit broader applicability.

Future work should explore additional datasets, algorithms, and feature selection methods to validate and extend these results. Overall, the study shows the value of using advanced regression models for accurate sleep quality prediction.

## References

[1] John Smith. The importance of sleep for physical and mental health. Journal of Sleep Research, 12(3):125–138, 2003.

[2] Leila Lilge and William C. Dement. Principles and Practice of Sleep Medicine. Elsevier Saunders, 5th edition, 2011.

[3] John G.H. Yang, Jiawei Han, and Philip S. Yu. Feature selection: A data perspective. Morgan and Claypool Publishers, 2015

[4] Y. Liu, H. Motoda, and Z. H. Zhou. Toward integrating feature selection algorithms for high-dimensional data. IEEE Transactions on Knowledge and Data Engineering, 25(8):1753–1766, 2013.

[5] J. Guyon and A. Elisseeff. An introduction to variable and feature selection. Journal of Machine Learning Research, 3:1157–1182, 2003.

[6] K. Bakshi and J. Kaur. A review of feature selection techniques in bioinformatics. Journal of Theoretical Biology, 366:66–76, 2015.

[7] Wei Yang and Huan Liu. Feature selection for regression via embedded decision trees. Intelligent Data Analysis, 13(4):435–451, 2009.

[8] Shahrad Taheri, Lan-Yan Lin, Daniel Austin, Terri Young, and Emmanuel Mignot. Sleep duration and all-cause mortality: a systematic review and meta-analysis of prospective studies. Sleep, 29(3):239–252, 2006.

[9] Karine Spiegel, Rachel Leproult, and Eve Van Cauter. Sleep deprivation and disease: effects on the body, brain and behavior. Nature and Science of Sleep, pages 145–153, 2009.

[10] Wen-Bin Shi, Wei Li, Ming-Hua Zhang, Hong-Mei Li, Fu-Min Song, Jun-Mei Lin, Shi-Min Lu, Shu-Qin Zeng, Xing-Rong Chen, Jing-Rong Zhang, et al. Association between sleep quality and quality of life: a systematic review and meta-analysis. Sleep medicine reviews, 21:13–23, 2015.

[11] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3):389–422, 2002.

[12] Ron Kohavi and George H. John. Wrappers for feature subset selection. Artificial Intelligence, 97(1-2):273–324, 1997.

[13] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, 2009.

[14] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. Journal of machine learning research, 3(Mar):1157–1182, 2003.

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

[16] Aurélien Géron. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc., 2019.

[17] Ian T Jolliffe. Principal component analysis. Wiley Online Library, 2, 2002.

[18] Jonathon Shlens. A tutorial on principal component analysis. In arXiv preprint arXiv:1404.1100, 2014.

[19] John D Kelleher and Brendan Tierney. Data science: An introduction. CRC Press, 2018.

[20] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12(Oct):2825–2830, 2011.

[21] Chen Zhang, Yuchen Zhang, Xiaoyu Zheng, Xiaogang Cheng, and Xiaodan Chen. Modeling college students' stress: A comparison of linear regression and random forest regression. IEEE Access, 7:65489–65496, 2019.

[22] Faisal Riaz, Ali Hassan, Aamir Saeed Malik, Uzair Yasin, Saeed Ur Rehman, and Hafiz Muhammad Nazir. Machine learning based model to predict mental stress using eeg signals. Biomedical Signal Processing and Control, 39:209–218, 2018.

[23] Xiaoyun Li, Guanyu Xu, Yu Tian, and Nan Li. Prediction of college students' stress level based on an improved random forest regression model. Frontiers in Psychology, 12:583550, 2021.