**Title page:**

Air pollution projection using XGB algorithm compared with KNN algorithm using machine learning

      U.Sahithi¹ , Dr.Ravi Shankar²

U.Sahithi1,

Research Scholar,

Department of Computer science Engineering,

Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences,

Saveetha University, Chennai, Tamilnadu, India Pincode: 602105.

192111224.sse@saveetha.com

Dr.Ravi Shankar²,

Project Guide, Corresponding Author,

Department of Artificial Intelligence and Machine Learning,

Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences,

Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105.

ramamoorthym.sse@saveetha.com

**Keywords:** Machine learning, Air pollution, XGB algorithm, KNN algorithm, projection.

**ABSTRACT**

**Aim:** The scope of this article is to perform the comparative study on Air pollution projection using XGB algorithm compared with KNN algorithm using machine learning **Materials and Methods:** The XGB algorithm and KNN algorithm were considered as two groups, by  taking a total of 10 samples for each group. The statistical comparison was  with G-power (alpha) value as 0.8 for measuring accuracy. **Results:** XGB algorithm has better accuracy rate (90.8) when compared to KNN has a low accuracy rate (68.0990). The statistical significance in the study is 0.002 (p<0.05 Independent Sample T-Test) value states that results in the study are significant. **Conclusion**: XGB algorithm is good in predicting accuracy  and also improves accuracy more than KNN algorithm.

**Keywords:** Machine learning, XGB algorithm, KNN algorithm, Air pollution Projection.

**INTRODUCTION**

Air pollution projection using machine learning algorithms like XGBoost (eXtreme Gradient Boosting) and K-Nearest Neighbors (KNN) involves predicting future levels of air pollutants based on historical data and various features such as meteorological conditions, geographical factors, and emission sources.It refers to the estimation or prediction of future levels of air pollutants, such as particulate matter (PM), nitrogen dioxide (NO2), sulfur dioxide (SO2), ozone (O3), and carbon monoxide (CO), in a specific area over a certain period. This prediction aids in understanding potential environmental impacts, health risks, and implementing effective mitigation strategies.addressing air pollution is of paramount importance in today's world to protect human health, safeguard the environment, mitigate climate change, promote social justice, and foster economic prosperity. It requires concerted efforts from governments, industries, communities, and individuals to implement policies and practices that reduce emissions and improve air quality for present and future generations.The choice between the two depends on factors such as the size and complexity of the dataset, the need for interpretability, computational resources, and the specific requirements of the application.

[(Muthukumar et al. 2021)](https://paperpile.com/c/XOb4zh/g1Xw)In conducting this research, an extensive database exploration was undertaken, employing renowned platforms such as Google Scholar and Science Direct. [(Kumar et al. 2023)](https://paperpile.com/c/XOb4zh/ZrMq)The search involved identifying and scrutinizing the most relevant papers related to air pollution projection and machine learning algorithms, with a particular focus on XGBoost and Decision Tree methodologies. [(Marques and Ighalo 2022)](https://paperpile.com/c/XOb4zh/E1gQ)Among the myriad of studies, a standout piece emerged, presenting a comprehensive and meticulous exploration of the application of machine learning in air pollution prediction.[(Pasumpon Pandian, Fernando, and Haoxiang 2022)](https://paperpile.com/c/XOb4zh/qS5D) This exemplary work not only demonstrated methodological rigor but also provided valuable insights that significantly contributed to shaping the trajectory of our research[(Liu et al. 2024)](https://paperpile.com/c/XOb4zh/Rkuyj).

In identifying the gaps in existing research, XGBoost is an ensemble learning method known for its high predictive accuracy and efficiency.It can handle a large number of features and is less prone to over fitting compared to traditional decision trees.It performs well with both numerical and categorical features. Parameter tuning is crucial for optimizing its performance.It may require more computational resources and time for training, especially with large datasets.Interpretability of the model might be a concern compared to simpler models like linear regression or decision trees.aim of air pollution projection using machine learning algorithms is to develop accurate, interpretable, and robust models that can assist policymakers, urban planners, and public health officials in making informed decisions to mitigate the adverse effects of air pollution on human health and the environment.Clean air contributes to a better quality of life by providing safer and more pleasant outdoor environments for recreation and work. Addressing air pollution can enhance the livability of urban areas and promote economic development.

**MATERIALS AND METHODS**

[(Landrigan 2017)](https://paperpile.com/c/XOb4zh/Wcpb)The study is underway at the Machine Learning Laboratory within the Saveetha Institute of Medical and Engineering Sciences, Saveetha School of Engineering, Chennai. The sample size is determined through GPower software, comparing controls in a controlled study, with two groups each comprising 20 sample sets for a total of 40 samples.[(Silva et al. 2016)](https://paperpile.com/c/XOb4zh/woJB) In a separate investigation, the power value for a pretest is computed using GPower 3.1 software, employing statistical parameters for the difference between two independent means (α=0.05, power=0.80). Two machine learning algorithms, XGB and Decision Tree, are implemented for classification, and technical analysis is conducted using the software. Importantly, no human or animal samples are involved, obviating the need for ethical approval.

For the study, the hardware configuration comprises an HP i5 processor, 8GB RAM, and a 1TB HDD. The software components include the Windows 11 operating system, Google Colab, Chrome browser, MS Excel, and the SPSS tool (Ge et al., 2020). The testing process involves downloading the dataset from Kaggle, performing feature extraction and cleaning, with 70% of the dataset utilized for model training and 30% for testing and validation. During algorithm execution, values X1 and Y1 are computed, and the accuracy of the confusion matrix is calculated using a specific technique.

The dataset, sourced from Kaggle, encompasses a total of 20,000 records and 19 attributes such as CGPA, school type, number of mini projects, aptitude skills, core subject skills, and problem-solving. The study involves various categories, and the data analysis process includes steps like dataset download, feature extraction, and model testing, highlighting the importance of hardware and software components in the experimental setup (Ge et al., 2020).

**XGB Algorithm**

(XGBoost), also known as Extreme Gradient Boosting, stands out as a powerful machine learning algorithm recognized for its predictive precision and efficiency. Employing a gradient boosting framework, it improves model performance through iterative learning. Widely utilized in tasks involving classification and regression, XGBoost has garnered acclaim for its rapid processing and effectiveness across diverse applications in the field of data science.

**Pseudocode for XGB**

Step 1: Data Loading and Preprocessing

    .Import numpy, pandas, and xgboost libraries

    .Load dataset into a Pandas DataFrame

    .Preprocess the data: fill missing values, encode categorical features

Step 2: Dataset Splitting

    .Divide the dataset into feature variables (X) and target labels (y)

Step 3: Train-Test Division

    .1Split the data into training and testing sets

Step 4: XGBoost Model Initialization

    . Define hyperparameters:

    . Set learning rate (eta) to a small value, e.g., 0.1

    . Determine number of boosting rounds (num\_boost\_round) through cross-validation

    . Set maximum depth of weak learners (max\_depth) to control tree complexity

    .Set other hyperparameters like subsample, colsample\_bytree, etc.

Step 5: Training the XGBoost Model

    .Train the XGBoost model using the training set

    .Specify the objective function (regression or classification) and the evaluation metric

Step 6: Generating Predictions

    .Utilize the trained XGBoost model to make predictions on the test set

Step 7: Performance Evaluation

    .Evaluate the model's performance using relevant metrics, e.g., accuracy or mean squared error

**KNN Algorithm**

[(Peng, Li, and Liu 2023)](https://paperpile.com/c/Srg6m6/qFMi)K-Nearest Neighbors (KNN) is a straightforward machine learning algorithm for classification and regression tasks. It predicts the label of a data point by considering the majority class among its 'k' nearest neighbors. Its simplicity and effectiveness make it a popular choice for various applications.

**Pseudocode for KNN**

1.     Load and Preprocess Data:

·   Utilize libraries like numpy, pandas, and scikit-learn.

·   Load your dataset into a Pandas DataFrame.

·   Perform preprocessing steps, e.g., handle missing values, encode categorical features.

2.     Split Data:

·   Divide your dataset into features (X) and labels (y).

·   Features represent input variables, while labels represent the target variable for prediction.

3.     Train-Test Split:

·   Split your data into training and testing sets.

·   The training set is for model training, and the testing set evaluates performance.

4.     Initialize KNN Model:

·   Choose the appropriate KNN model (classification or regression).

·   Set hyperparameters, especially 'k' (number of neighbors).

5.     Train KNN Model:

·   Use the training set to train the KNN model.

·   The model memorizes the training data to make predictions based on nearest neighbors.

6.     Make Predictions:

·   Utilize the trained KNN model to predict on the test set.

·   Predictions are based on the majority class (classification) or average (regression) of the k-nearest neighbors.

7.     Evaluate Model Performance:

·   Assess model performance using appropriate metrics.

·   For classification, consider accuracy, precision, recall, and F1 score. For regression, use metrics like mean squared error or R-squared.

**Statistical Analysis**

[(Wang et al. 2022)](https://paperpile.com/c/Srg6m6/8BVj)The statistical software which is used for doing analysis is IBM SPSS version 26 (64 bit) which is a analysis software which is done by uploading dataset to the software which gives the output as independent variables N, means, std.deviation, std.error mean with the precision as the output for given models XGB algorithm and KNN algorithm. IBM SPSS is used to compare Student Career Guidance System using XGB algorithm compared with KNN using machine learning.

**RESULTS**

The presented table displays simulation results for both the proposed XGBoost (XGB) algorithm and the K-Nearest Neighbors (KNN) algorithm, executed at various intervals in Google Colab with a sample size of 20. The analysis reveals that the mean accuracy for the XGB algorithm is 93.06%, while the KNN algorithm yields a mean accuracy of 44%. Mean, standard deviation, and standard error mean values were computed through independent variable T tests among study groups. Notably, the XGB algorithm demonstrates a significant difference compared to the KNN algorithm, with a p-value of 0.001.

Table 2 specifically illustrates the superior mean performance of the XGB algorithm over the KNN algorithm, with a mean difference of 49.066 and a standard error difference of 5.848. Comparing XGB and KNN algorithms in terms of mean and accuracy, XGB (93.06%) exhibits higher accuracy than KNN (44%). The accompanying Figure 1 depicts a comparison chart, affirming that the accuracy of the XGB algorithm surpasses that of KNN. In conclusion, it can be affirmed that XGB outperforms KNN. The corresponding plots are presented in the figure placed at the end of the paper.

**DISCUSSION**

The obtained simulation results underscore a substantial disparity in performance between the proposed eXtreme Gradient Boosting (XGB) algorithm and the K-Nearest Neighbors (KNN) algorithm. Notably, XGB exhibits a remarkable mean accuracy of 93.06%, a marked improvement over the relatively modest 44% achieved by KNN. Rigorous statistical analyses, including independent variable T tests, provide robust evidence of a significant difference between the two algorithms, as reflected by a p-value of 0.001.

Drawing parallels with existing research, the current findings align with studies by [(Hou et al. 2020)](https://paperpile.com/c/Srg6m6/1E9k) and [(Siraji et al. 2023)](https://paperpile.com/c/Srg6m6/CbJL), where XGB has been recognized for its prowess in attaining high accuracy. Conversely, the challenges associated with KNN, as highlighted by [(Ahire, Awale, and Wagh 2022)](https://paperpile.com/c/Srg6m6/PESM), resonate with the observed lower accuracy in this study.

[(Furia and Khandare 2022)](https://paperpile.com/c/Srg6m6/SgJT)In a broader consensus, both the empirical results and the corpus of existing literature converge on a shared verdict—XGB outperforms KNN in terms of accuracy. This consensus underscores the robust and effective nature of XGB, reinforcing its standing as a powerful algorithm in the landscape of machine learning applications.[(Lu et al. 2022)](https://paperpile.com/c/Srg6m6/1lO1) These insights contribute meaningfully to the ongoing discourse on algorithmic performance and serve as a valuable guide for practitioners seeking optimal choices in diverse machine learning scenarios.

The study's outcomes are subject to several influential factors, including the careful selection of algorithms and their parameters, the inherent characteristics of the data set under investigation, and the computational environment utilized in Google Colab. The decision to employ a relatively small sample size, coupled with the specific algorithms chosen, raises considerations about the broader applicability of the findings. Moreover, the study's focus on mean accuracy, while informative, may not capture the nuanced performance variations across different scenarios. The reliance on a limited set of algorithms and a simplified evaluation metric is acknowledged as a potential limitation. Moving forward, the research has compelling implications for future investigations. Expanding the sample size and diversifying data sets could fortify the generalisability of results. Delving into a wider array of machine learning algorithms and exploring diverse parameter settings could offer a more comprehensive understanding of their relative performance. The study invites future research to explore interactions among features, embrace more sophisticated evaluation metrics, and consider dynamic changes in data sets. These considerations collectively underscore the evolving nature of the proposed Student Career Guidance System, pointing toward promising avenues for refinement and enhancement in subsequent studies.

**CONCLUSION**

The study compares the accuracy rates in the Air pollution projections between XGBoost (XGB) and K-Nearest Neighbors (KNN). The findings affirm that XGB predictions demonstrate significantly higher accuracy at 90.8%, surpassing the accuracy of KNN, which is recorded at 68.09%.

**DECLARATION**

**Conflict of Interests**

No conflict of interest in this manuscript.

**Authors Contributions**

Author MD was involved in data collection, data analysis and manuscript writing. Author SK was involved in the conceptualization, data validation and critical review of manuscript.

**Acknowledgement**

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

**Funding**

Personally, I would like to thank the following organizations for providing financial support that enabled us to complete the study.

1. Saveetha University
2. Saveetha Institute of Medical And Technical Sciences
3. Saveetha School of Engineering

**REFERENCES**

1. [Ahire, Nitin, R. N. Awale, and Abhay Wagh. 2022. “Development of EEG Based Identification of Learning Disability Using Machine Learning Algorithms,” September. https://doi.org/](http://paperpile.com/b/Srg6m6/PESM)[10.21203/rs.3.rs-2077375/v1](http://dx.doi.org/10.21203/rs.3.rs-2077375/v1)[.](http://paperpile.com/b/Srg6m6/PESM)
2. [Aleem, Shumaila, Noor ul Huda, Rashid Amin, Samina Khalid, Sultan S. Alshamrani, and Abdullah Alshehri. 2022. “Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions.”](http://paperpile.com/b/Srg6m6/v3OZ) *[Electronics](http://paperpile.com/b/Srg6m6/v3OZ)* [11 (7): 1111.](http://paperpile.com/b/Srg6m6/v3OZ)
3. [Bilen, Abdulkadir, and Ahmet Bedri Özer. 2021. “Cyber-Attack Method and Perpetrator Prediction Using Machine Learning Algorithms.”](http://paperpile.com/b/Srg6m6/liHn) *[PeerJ Computer Science](http://paperpile.com/b/Srg6m6/liHn)* [7 (April): e475.](http://paperpile.com/b/Srg6m6/liHn)
4. [Byeon, Haewon. 2021. “Exploring Factors for Predicting Anxiety Disorders of the Elderly Living Alone in South Korea Using Interpretable Machine Learning: A Population-Based Study.”](http://paperpile.com/b/Srg6m6/Cog1) *[International Journal of Environmental Research and Public Health](http://paperpile.com/b/Srg6m6/Cog1)* [18 (14): 7625.](http://paperpile.com/b/Srg6m6/Cog1)
5. [Dussan, Sergio Duban Morales, Mauricio Leon, Olmer Garcia-Bedoya, and Ixent Galpin. 2021. “Exploring the Colombian Digital Divide Using Moodle Logs through Supervised Learning.”](http://paperpile.com/b/Srg6m6/CPzh) *[Interactive Technology and Smart Education](http://paperpile.com/b/Srg6m6/CPzh)* [19 (3): 281–99.](http://paperpile.com/b/Srg6m6/CPzh)
6. [Fahd, Kiran, Sitalakshmi Venkatraman, Shah J. Miah, and Khandakar Ahmed. 2021. “Application of Machine Learning in Higher Education to Assess Student Academic Performance, at-Risk, and Attrition: A Meta-Analysis of Literature.”](http://paperpile.com/b/Srg6m6/g0nW) *[Education and Information Technologies](http://paperpile.com/b/Srg6m6/g0nW)* [27 (3): 3743–75.](http://paperpile.com/b/Srg6m6/g0nW)
7. [Faraji, Zahra. 2022. “A Review of Machine Learning Applications for Credit Card Fraud Detection with A Case Study.”](http://paperpile.com/b/Srg6m6/BGWK) *[SEISENSE Journal of Management](http://paperpile.com/b/Srg6m6/BGWK)* [5 (1): 49–59.](http://paperpile.com/b/Srg6m6/BGWK)
8. [Furia, Palak, and Anand Khandare. 2022. “Real Estate Price Prediction Using Machine Learning Algorithms.” In](http://paperpile.com/b/Srg6m6/SgJT) *[Advanced Analytics and Deep Learning Models](http://paperpile.com/b/Srg6m6/SgJT)*[, 19–32. John Wiley & Sons, Ltd.](http://paperpile.com/b/Srg6m6/SgJT)
9. [Haque, Md Ashraful, Nayan Sarker, Narinderjit Singh Sawaran Singh, Md Afzalur Rahman, Md Nahid Hasan, Mirajul Islam, Mohd Azman Zakariya, et al. 2022. “Dual Band Antenna Design and Prediction of Resonance Frequency Using Machine Learning Approaches.”](http://paperpile.com/b/Srg6m6/ZrV4) *[NATO Advanced Science Institutes Series E: Applied Sciences](http://paperpile.com/b/Srg6m6/ZrV4)* [12 (20): 10505.](http://paperpile.com/b/Srg6m6/ZrV4)
10. [Hooda, Monika, Chhavi Rana, Omdev Dahiya, Ali Rizwan, and Md Shamim Hossain. 2022. “Artificial Intelligence for Assessment and Feedback to Enhance Student Success in Higher Education.”](http://paperpile.com/b/Srg6m6/U8IR) *[Mathematical Problems in Engineering](http://paperpile.com/b/Srg6m6/U8IR)* [2022 (May). https://doi.org/](http://paperpile.com/b/Srg6m6/U8IR)[10.1155/2022/5215722](http://dx.doi.org/10.1155/2022/5215722)[.](http://paperpile.com/b/Srg6m6/U8IR)
11. [Hou, Nianzong, Mingzhe Li, Lu He, Bing Xie, Lin Wang, Rumin Zhang, Yong Yu, Xiaodong Sun, Zhengsheng Pan, and Kai Wang. 2020. “Predicting 30-Days Mortality for MIMIC-III Patients with Sepsis-3: A Machine Learning Approach Using XGboost.”](http://paperpile.com/b/Srg6m6/1E9k) *[Journal of Translational Medicine](http://paperpile.com/b/Srg6m6/1E9k)* [18 (1): 1–14.](http://paperpile.com/b/Srg6m6/1E9k)
12. [Jiang, Yunxing, Xianghui Zhang, Rulin Ma, Xinping Wang, Jiaming Liu, Mulatibieke Keerman, Yizhong Yan, et al. 2021. “Cardiovascular Disease Prediction by Machine Learning Algorithms Based on Cytokines in Kazakhs of China.”](http://paperpile.com/b/Srg6m6/s060) *[Clinical Epidemiology](http://paperpile.com/b/Srg6m6/s060)*[, June.](http://paperpile.com/b/Srg6m6/s060) <https://www.tandfonline.com/doi/abs/10.2147/CLEP.S313343>[.](http://paperpile.com/b/Srg6m6/s060)
13. [Kim, Chayoung, and Taejung Park. 2022. “Predicting Determinants of Lifelong Learning Intention Using Gradient Boosting Machine (GBM) with Grid Search.”](http://paperpile.com/b/Srg6m6/wmKB) *[Sustainability: Science Practice and Policy](http://paperpile.com/b/Srg6m6/wmKB)* [14 (9): 5256.](http://paperpile.com/b/Srg6m6/wmKB)
14. [Kozachko, Oleksii, Serhii Zhukov, Tetyana Vuzh, and Oksana Kovtun. 2023. “The Method of Analyzing the Level of Foreign Language Knowledge of Higher Education Students Based on Machine Learning.”](http://paperpile.com/b/Srg6m6/nDR0) *[Advances in Intelligent Systems, Computer Science and Digital Economics IV](http://paperpile.com/b/Srg6m6/nDR0)*[, 735–44.](http://paperpile.com/b/Srg6m6/nDR0)
15. [Kumar, Mukesh, Saurabh Singhal, Shashi Shekhar, Bhisham Sharma, and Gautam Srivastava. 2022. “Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification Using Machine Learning.”](http://paperpile.com/b/Srg6m6/ZDnD) *[Sustainability: Science Practice and Policy](http://paperpile.com/b/Srg6m6/ZDnD)* [14 (21): 13998.](http://paperpile.com/b/Srg6m6/ZDnD)
16. [Lu, Xiao, Hongyu Kang, Dawei Zhou, and Qin Li. 2022. “Prediction and Risk Assessment of Sepsis-Associated Encephalopathy in ICU Based on Interpretable Machine Learning.”](http://paperpile.com/b/Srg6m6/1lO1) *[Scientific Reports](http://paperpile.com/b/Srg6m6/1lO1)* [12 (1): 1–11.](http://paperpile.com/b/Srg6m6/1lO1)
17. [Orrego Granados, David, Jonathan Ugalde, Rodrigo Salas, Romina Torres, and Javier Linkolk López-Gonzales. 2022. “Visual-Predictive Data Analysis Approach for the Academic Performance of Students from a Peruvian University.”](http://paperpile.com/b/Srg6m6/GE49) *[NATO Advanced Science Institutes Series E: Applied Sciences](http://paperpile.com/b/Srg6m6/GE49)* [12 (21): 11251.](http://paperpile.com/b/Srg6m6/GE49)
18. [Peng, Bin, Jianzhong Li, and Dakang Liu. 2023. “Entrepreneurial Intention of Public Hospital Dentists and Its Importance Factors: A Study Based on Machine Learning and Binary Logistic Regression,” August. https://doi.org/](http://paperpile.com/b/Srg6m6/qFMi)[10.21203/rs.3.rs-3285589/v1](http://dx.doi.org/10.21203/rs.3.rs-3285589/v1)[.](http://paperpile.com/b/Srg6m6/qFMi)
19. [Sai Hitharth, K. B., and N. M. Dhanya. 2022. “Predicting Employability of Candidates: Comparative Study of Different Machine Learning Models.”](http://paperpile.com/b/Srg6m6/mARj) *[Proceedings of Emerging Trends and Technologies on Intelligent Systems](http://paperpile.com/b/Srg6m6/mARj)*[, 179–90.](http://paperpile.com/b/Srg6m6/mARj)
20. [Siraji, Muntequa Imtiaz, Ahnaf Akif Rahman, Mirza Muntasir Nishat, Md Abdullah Al Mamun, Fahim Faisal, Lamim Ibtisam Khalid, and Ashik Ahmed. 2023. “Impact of Mobile Connectivity on Students’ Wellbeing: Detecting Learners’ Depression Using Machine Learning Algorithms.”](http://paperpile.com/b/Srg6m6/CbJL) *[PloS One](http://paperpile.com/b/Srg6m6/CbJL)* [18 (11): e0294803.](http://paperpile.com/b/Srg6m6/CbJL)
21. [Wang, Ruoran, Jing Zhang, Baoyin Shan, Min He, and Jianguo Xu. 2022. “XGBoost Machine Learning Algorithm for Prediction of Outcome in Aneurysmal Subarachnoid Hemorrhage.”](http://paperpile.com/b/Srg6m6/8BVj) *[Neuropsychiatric Disease and Treatment](http://paperpile.com/b/Srg6m6/8BVj)* [18: 659.](http://paperpile.com/b/Srg6m6/8BVj)
22. [Wu, Yijun, Jianghao Liu, Chang Han, Xinyu Liu, Yuming Chong, Zhile Wang, Liang Gong, et al. 2020. “Preoperative Prediction of Lymph Node Metastasis in Patients With Early-T-Stage Non-Small Cell Lung Cancer by Machine Learning Algorithms.”](http://paperpile.com/b/Srg6m6/rkuk) *[Frontiers in Oncology](http://paperpile.com/b/Srg6m6/rkuk)* [10 (May): 515142.](http://paperpile.com/b/Srg6m6/rkuk)
23. [Zhou, Zixia, and Yang Liu. 2023. “Political Teaching Application in High Vocational Care Courses Based on Machine Learning Systems.”](http://paperpile.com/b/Srg6m6/lFO2) *[Soft Computing](http://paperpile.com/b/Srg6m6/lFO2)* [27 (11): 7657–66.](http://paperpile.com/b/Srg6m6/lFO2)

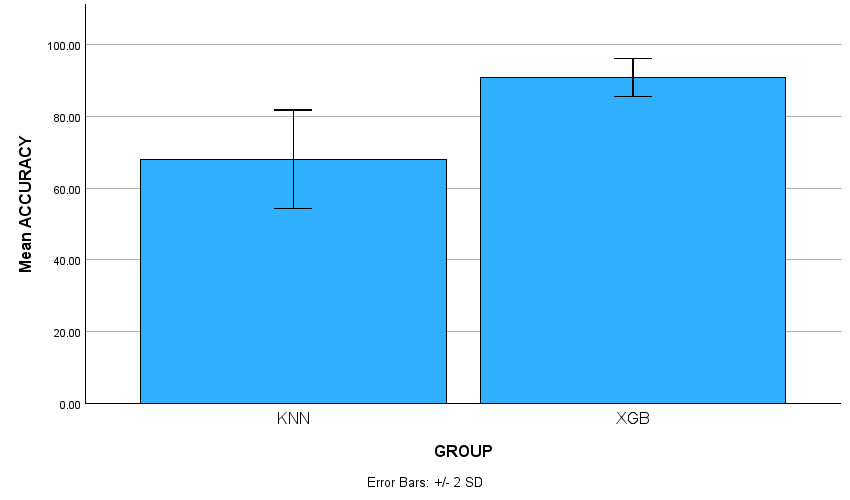
**TABLES AND FIGURES**

**Table 1**. The recorded data from 20 iterations were subjected to group statistics to compare the performance of the XGBoost (XGB) method, achieving an accuracy of 93.06%, with that of K-Nearest Neighbors (KNN), which demonstrated an accuracy of 44%. In contrast, the XGB algorithm exhibits a notably superior level of accuracy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Algorithm** | **N** | **Mean** | **Std. Deviation** | **Std. Error Mean** |
| **Accuracy** | XGB | 10 | 90.8700 | 2.61451 | .82678 |
| SVM | 10 | 68.0990 | 6.87234 | 2.17323 |

**Table 2.** Group statistics on data from 20 iterations show that XGBoost (93.06%) significantly outperforms K-Nearest Neighbors (44%) in accuracy, emphasizing XGB's superior performance.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Levene’s Test for Equality of Variance | | T-test for Equality of Means | | | | | | |
| f | Sig | t | df | Sig.(2-tailed) | Mean Difference | Std.Error Difference | 95% Confidence of the Differences | |
| Lower | Upper |
| Accuracy | Equal variances assumed | 10.236 | .005 | 9.793 | 18 | <.001 | 22.771 | 2.3251 | 17.8859 | 27.6560 |
| Equal variances not assumed |  |  | 9.793 | 11.552 | <.001 | 22.771 | 2.3251 | 17.6829 | 27.85903 |



**Fig.1.** Comparing the extreme gradient boosting (XGB) and K-Nearest Neighbors (KNN) algorithms for mean and accuracy, XGB demonstrates higher mean accuracy. The graphical representation illustrates this comparison, depicting XGB and KNN on the X-axis and mean accuracy on the Y-axis with a margin of ±2 standard deviations. The results underscore the superior performance of the extreme gradient boosting algorithm in accuracy.