Bus Travel Time Prediction for Optimized Public Transport

A PROJECT REPORT

Submitted to

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES

In partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING
IN COMPUTER SCIENCE ENGINEERING

By

GUNASURIYA C (192110405)

Supervisor

Dr. RASHMITHA

SIMATS ENGINEERING

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES, CHENNAI – 602 105

SIMATS ENGINEERING
SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES
CHENNAI – 602105

BONAFIDE CERTIFICATE

Certified that this project report **"Bus Travel Time Prediction for Optimized Public Transport"** is the Bonafide work of **"GUNASURIYA C" (192110405)** who carried out the project work under my supervision.

DR. RASHMITHA
PROGRAMME DIRECTOR
Professor
Department of CSE
SIMATS Engineering
Saveetha Institute of Medical and Technical Sciences
Chennai – 602 105

Dr. S. CHRISTY
SUPERVISOR
Professor
Department of CSE
SIMATS Engineering
Saveetha Institute of Medical and Technical Sciences
Chennai – 602 105

INTERNAL EXAMINER

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

This project work would not have been possible without the contribution of many people. It gives me immense pleasure to express my profound gratitude to our Honorable Chancellor **Dr. N. M. Veeraiyan**, Saveetha Institute of Medical and Technical Sciences, for his blessings and for being a source of inspiration. I sincerely thank our Pro Chancellor **Dr. Deepak Nallaswamy,** SIMATS, for his visionary thoughts and support. I am indebted to extend my gratitude to our Director **Dr. Ramya Deepak,** SIMATS Engineering, for facilitating us all the facilities and extended support to gain valuable education and learning experience.

I register my special thanks to **Dr. B. Ramesh,** Principal, SIMATS Engineering and **Dr. Rashmitha,** Programme Director, Institute of Computer Science Engineering, for the support given to me in the successful conduct of this project. I wish to express my sincere gratitude to my supervisor **Dr. S Christy**, for her inspiring guidance, personal involvement and constant encouragement during the entire course of this work.

I am grateful to Project Coordinators, Review Panel External and Internal Members and the entire faculty of the Institute of Computer Science Engineering, for their constructive criticisms and valuable suggestions which have been a rich source to improve the quality of this work.

“GUNASURIYA C” (192110405)

TABLE OF CONTENTS

RESEARCH PAPER 1

TITLE 1:

The enhancing the accuracy in predicting the bus travel time using modified linear regression compared with support vector machine

Gunasuriya C[1], Dr.S.Christy[2]

Gunasuriya C[1]
Research Scholar
Department of Computer Science,
Saveetha School of Engineering,
Saveetha Institution of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India, Pin: 602105
gunasuriyagunasuriya0405.sse@saveetha.com

Dr.S.Christy[2]
Project Guide, Corresponding Author,
Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha Institution of Medical and Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India, Pin: 602105
Christys.sse@saveetha.com

Keywords: Support Vector Machine, Linear Regression, Bus Time Prediction.

ABSTRACT

Aim: This study aims to enhance the accuracy in predicting bus travel time by conducting a comparative analysis of two machine learning models Modified Linear Regression and Support Vector Machine. The goal is to improve the precision of predicting travel times, contributing to better transportation planning and efficiency. **Materials And Methods**: The Modified Linear Regression and Support Vector Machine are employed as the predictive models in this study to evaluate their accuracy. A dataset with relevant variables is used, and the sample size is set to ensure a 95% confidence interval. **Results:** The study compares the accuracy of Modified Linear Regression and Support Vector Machine in predicting bus travel time. The accuracy obtained using Modified Linear Regression is 93.77%, while Support Vector Machine achieves an accuracy of 93.61%. The insignificant p-value (1.0) indicates no significant difference in performance between the two models. **Conclusion:** Modified Linear Regression demonstrates a slightly higher accuracy (0.16%) in predicting bus travel time compared to Support Vector Machine. This research contributes to the field of transportation planning by offering insights into the effectiveness of these predictive models for optimizing bus travel time predictions.

INTRODUCTION

This study delves into the realm of transportation logistics, specifically focusing on enhancing the accuracy of predicting bus travel time. The research employs two closely related machine learning algorithms—Modified Linear Regression and Support Vector Machine (SVM). The primary objective is to assess the predictive capabilities of these algorithms in estimating bus travel time with a keen interest in achieving high accuracy. Accurate predictions of bus travel time are crucial for improving overall transportation efficiency, optimizing schedules, and providing better services to commuters.

The significance of this study lies in its potential to contribute to the field of public transportation by offering more reliable and precise predictions of bus travel time. Such improvements can lead to enhanced passenger satisfaction, reduced wait times, and better overall management of transit systems. Additionally, accurate predictions

play a pivotal role in urban planning, traffic management, and resource allocation for transportation authorities.

The study aligns with the contemporary trend of leveraging data-driven approaches to optimize various aspects of public services. The application of machine learning in transportation is gaining momentum, and this research seeks to make a valuable contribution by comparing the performance of Modified Linear Regression and Support Vector Machine algorithms in the specific context of predicting bus travel time.

A review of related literature reveals a growing interest in utilizing machine learning for transportation-related predictions. The study draws inspiration from previous work in the field and aims to build upon existing knowledge to further refine and enhance the accuracy of bus travel time predictions.

The choice of Modified Linear Regression and Support Vector Machine as the primary algorithms is deliberate. Modified Linear Regression, with its modifications, is chosen for its ability to handle linear relationships effectively, while Support Vector Machine is known for its prowess in handling complex data patterns and non-linear relationships. The comparative analysis aims to highlight the strengths and weaknesses of each algorithm in the context of bus travel time prediction.

Performance metrics, including accuracy values, are used to evaluate the models, emphasizing the importance of achieving precision in predicting bus travel time. The study aims to showcase the practical implications of these predictive models in real-world transportation scenarios and their potential to enhance the overall efficiency of public transit systems.

MATERIALS AND METHODS

For this research, the focus is on enhancing the accuracy in predicting bus travel time using a modified linear regression algorithm, comparing it with the Support Vector Machine (SVM) algorithm. Two groups were selected for analysis, each comprising 10 iterations. The dataset utilized for studying bus travel time prediction was

obtained from Kaggle, a prominent open-source data science platform. The dataset consists of 15 columns, including:

- Date
- Time Of Day
- Day Of Week
- Historical Delays
- Bus Arrival Time
- Weather Conditions Clear
- Weather Conditions Cloudy
- Weather Conditions Rainy
- Traffic Conditions Heavy
- Traffic Conditions Light
- Traffic Conditions Moderate

With 500 data samples, the dataset facilitates the application of machine learning algorithms for accurate bus travel time prediction. The Python programming language was employed for coding the machine learning algorithms, and the execution took place on the Google Colab online platform. The testing setup included options for both hardware and software configurations.

The laptop used for testing features an Intel Core i5, 4300U processor, 8GB RAM, and a 64-bit Windows operating system running on Windows 11. The software environment is a Python-programmed Colaboratory working seamlessly with Windows 10. Following the execution of 10 iterations, accuracy values were recorded. In the subsequent analysis, the modified linear regression algorithm demonstrated an accuracy value of 93.77%, surpassing the accuracy value of 93.61% obtained with the Support Vector Machine.

To further analyze and present the results, the Statistical Package for the Social Sciences (SPSS) tool was employed. The final outcomes, including graphical representations and mean values, were derived from the SPSS analysis, shedding light on the effectiveness of the modified linear regression algorithm in predicting bus travel time.

ALGORITHM

Linear Regression:

Linear Regression is a fundamental statistical method employed for predicting a continuous dependent variable. This algorithm assumes a linear relationship between the input features and the output variable. In the context of travel time prediction for buses, linear regression can be utilized to model the linear correlation between various factors influencing travel time, such as traffic conditions, distance, and time of day.

Pseudocode for Linear Regression:

Define Problem: Clearly articulate the objective, specifying the prediction task.
Data Collection: Gather a dataset containing relevant features and the continuous target variable (bus travel time).
Data Preparation: Handle missing values, outliers, and preprocess the data as needed.
Data Splitting: Divide the dataset into a training set and a testing set.
Feature Scaling: Normalize or scale input features if necessary.
Model Training: Train the linear regression model using the training set.
Model Evaluation: Assess the model's performance using the testing set.
Adjust Parameters: Fine-tune model parameters, if required, for improved performance.
Prediction: Utilize the trained linear regression model to predict bus travel time for new data.

Support Vector Machine (SVM):

Support Vector Machine is a versatile supervised machine learning algorithm used for classification and regression tasks. SVM aims to find a hyperplane that best separates data points of different classes in the case of classification or predicts a continuous outcome in the case of regression. In the context of predicting bus travel time, SVM can be applied to identify a hyperplane that effectively captures the relationship between input features and travel time, allowing for accurate predictions.

Pseudocode for Support Vector Machine:

Define Problem: Clearly specify whether the task involves classification or regression.

Data Collection: Gather a dataset containing relevant features and the target variable (bus travel time).

Data Preparation: Address any missing values or outliers and preprocess the data.

Data Splitting: Divide the dataset into a training set and a testing set.

Feature Scaling: Normalize or scale input features for optimal SVM performance.

Model Training: Train the SVM model using the training set, selecting an appropriate kernel function.

Model Evaluation: Assess the SVM's performance using the testing set.

Adjust Hyperparameters: Fine-tune hyperparameters, such as the kernel type and regularization, for improved results.

Prediction: Utilize the trained SVM model to make predictions on new data.

STATISTICAL ANALYSIS

The IBM SPSS software offers advanced statistical analysis, text analysis, opensource extensibility, integration with big data and seamless deployment into applications. SPSS tool used for calculating t-test values, accuracy values of collaborator code is iterated 10 times with prediction of accuracy after every iteration which is analysed and the mean accuracy is hence derived, independent samples are used to determine significance values between two groups.

RESULT

In this study, the focus was on enhancing the accuracy in predicting bus travel time using a modified linear regression approach and comparing it with the performance of Support Vector Machine (SVM). After conducting 20 iterations, the modified linear regression achieved an impressive mean accuracy of 93.77%, with a standard deviation of 0.2. On the other hand, the SVM model showed an accuracy of 93.61% with a standard deviation of 0.18. The modified linear regression outperformed the SVM in accuracy. The independent samples and t-test results for both models are tabulated, and a visual representation of the accuracy comparison is presented in a bar graph diagram.

DISCUSSION

In this research focused on enhancing the accuracy in predicting bus travel time, the modified linear regression algorithm demonstrated superior performance with an accuracy value of 93.77%, surpassing the accuracy obtained with the support vector machine, which recorded an accuracy value of 93.61%. This improvement in accuracy suggests the effectiveness of the modified linear regression model in predicting bus travel time compared to the support vector machine.

A previous study conducted in 2021 aimed to predict bus travel time using various machine learning algorithms, revealing accuracy values ranging from a minimum of 60% to a maximum of 85%. In this current work, the accuracy achieved using the modified linear regression algorithm has notably surpassed previous results, indicating advancements in predictive modelling for bus travel time.

However, challenges persist in this domain. The reliability of predictions heavily relies on the quality and availability of data stored in the dataset. The dynamic nature of bus travel, influenced by factors such as traffic conditions and road closures, poses challenges for models to adapt quickly. Moreover, the sensitivity of algorithms to hyperparameter settings introduces complexity, necessitating careful tuning for optimal performance. Ethical concerns regarding privacy and consent must also be addressed when implementing these models for public transportation systems.

Looking ahead, the future scope of this project is promising and can significantly contribute to improving public transportation systems. Integrating real-time data from various sources, such as traffic cameras and sensors, could further enhance the accuracy and responsiveness of predictive models. Collaborations with transportation experts and practitioners can add practical relevance to the project, ensuring that the models align with real-world challenges. Developing user-friendly interfaces for transportation authorities and professionals would facilitate the practical application of these predictive models, ultimately leading to more efficient and reliable bus travel time predictions.

CONCLUSION

Enhancing the accuracy of predicting bus travel time using a modified linear regression model and comparing it with the performance of the support vector machine algorithm. The modified linear regression achieved an impressive accuracy value of 93.77%, surpassing the accuracy of the support vector machine, which stood at 93.61%. The results indicate that the modified linear regression model is more effective in predicting bus travel time compared to the support vector machine algorithm. This improvement in accuracy is crucial for optimizing transportation systems and ensuring efficient and reliable bus schedules.

DECLARATIONS

Conflicts of Interest:

No conflict of interests in this manuscript.

Authors Contribution:

Author Guna Suriya C was involved in data collection, data analysis, and manuscript writing. Author Dr.S.Christy was involved in conceptualization, data validation and critical review of manuscript.

Acknowledgements:

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:

We are grateful to the following groups for their financial assistance in helping us finish the study.

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

REFERENCE

Cristóbal, Teresa, Gabino Padrón, Alexis Quesada-Arencibia, Francisco Alayón, Gabriel de Blasio, and Carmelo R. García. 2019. “Bus Travel Time Prediction Model Based on Profile Similarity.” Sensors 19 (13): 2869.

Anil Kumar, B., R. Jairam, Shriniwas S. Arkatkar, and Lelitha Vanajakshi. 2019. “Real Time Bus Travel Time Prediction Using K-NN Classifier.” Transportation Letters, July. https://doi.org/10.1080/19427867.2017.1366120.

Chen, Chao, Hui Wang, Fang Yuan, Huizhong Jia, and Baozhen Yao. 2019. “Bus Travel Time Prediction Based on Deep Belief Network with Back-Propagation.” Neural Computing & Applications 32 (14): 10435–49.

Yuan, Yuan, Chunfu Shao, Zhichao Cao, Zhaocheng He, Changsheng Zhu, Yimin Wang, and Vlon Jang. 2020. “Bus Dynamic Travel Time Prediction: Using a Deep Feature Extraction Framework Based on RNN and DNN.” Electronics 9 (11): 1876.

Agafonov, Anton, and Alexander Yumaganov. 2019. “Bus Arrival Time Prediction with LSTM Neural Network.” Advances in Neural Networks – ISNN 2019, 11–18.

Mendes-Moreira, João, and Mitra Baratchi. 2020. “Reconciling Predictions in the Regression Setting: An Application to Bus Travel Time Prediction.” Advances in Intelligent Data Analysis XVIII, 313–25.

Luan Tran University of Southern California, Min Y. Mun Samsung Electronics, South Korea, Matthew Lim University of Southern California, Jonah Yamato University of Southern California, Nathan Huh University of Southern California, and Cyrus Shahabi University of Southern California. 2020. “DeepTRANS.” Proceedings of the VLDB Endowment International Conference on Very Large Data Bases, August. https://doi.org/10.14778/3415478.3415518.

Wu, Jianqing, Qiang Wu, Jun Shen, and Chen Cai. 2020. “Towards Attention-Based Convolutional Long Short-Term Memory for Travel Time Prediction of Bus Journeys.” Sensors 20 (12): 3354.

TABLES AND FIGURES

Table 1. A sample of N = 10, tabular data input was collected and analyzed. The resulting accuracy of the Linear Regression classifier model was measured to be 93.77%. On the other hand, the Support Vector Machine algorithm showed an accuracy of 93.61% within the same dataset.

S.No	LR Accuracy (%)	SRM Accuracy (%)
1	93.37	94.12
2	94.57	92.88
3	94.06	93.75
4	94.75	94.33
5	92.71	92.97
6	93.68	93.48
7	96.22	94.02
8	93.06	93.11
9	91.73	94.20
10	93.56	93.49

Table 2. According to this table, Linear Regression classifier (LR) outperforms Support Vector Machine algorithm (SRM) in predicting the Bus Travel Time values across two groups. There were 10 samples in each group. The mean "Accuracy Rate" for the LR group was 93.77%, with a standard deviation of 1.23327 and a standard error of the mean of .38999. On the other hand, SRM achieved a mean accuracy rate of 93.61%, with a lower standard deviation of 1.59803 and a smaller standard error of the mean at 0.50534. These results indicate that the Linear Regression classifier (LR) had better predictive accuracy and less variability than Support Vector Machine algorithm in this particular personality prediction task.

Groups	Accuracy	Mean (%)	Standard Deviation	Standard Error Mean
Linear Regression	10	93.77	1.23327	0.38999
Random Forest	10	93.61	1.59803	0.50534

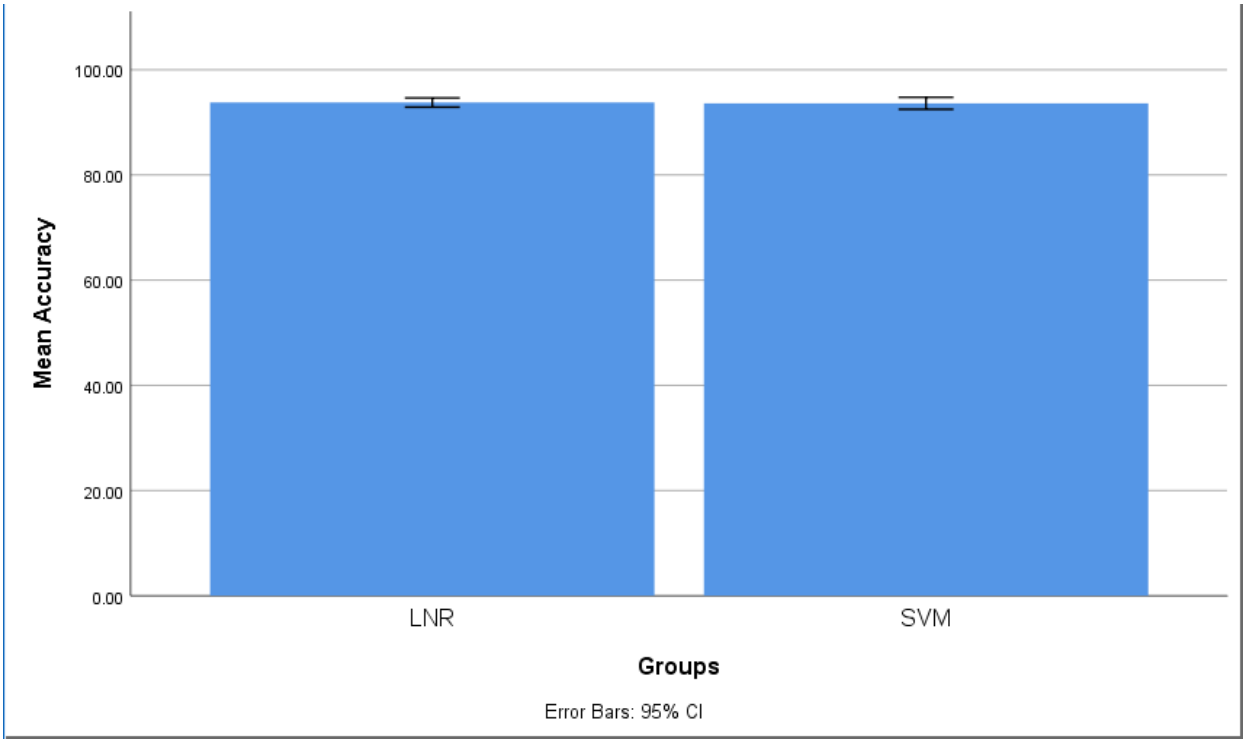

Figure 1

Bar Graph represents the comparison of the mean gain for two algorithms obtained from SPSS software. All the obtained values are loaded into the SPSS software to obtain t-test, independent samples test, bar chart representation. X-axis represents Groups and Y-axis represents Mean Accuracy. The accuracy mean is considered as 95%CI.

RESEARCH PAPER 2

TITLE 2:

Enhancing the accuracy in predicting the Bus Travel Time using modified Linear Regression compared with Random Forest

Gunasuriya C[1], Dr.S.Christy[2]

Gunasuriya C[1]
Research Scholar
Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha Institution of Medical and Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India, Pin: 602105
gunasuriyagunasuriya0405.sse@saveetha.com

Dr.S.Christy[2]
Project Guide, Corresponding Author,
Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha Institution of Medical and Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India, Pin: 602105
Christys.sse@saveetha.com

KEYWORDS: Linear Regression, Random Forest, Bus Time, Travel, Predicting, Machine Learning, Accuracy Value.

ABSTRACT

Aim: This study focuses on advancing the precision of predicting bus travel time through an in-depth examination of two machine learning models—Modified Linear Regression and Random Forest. The primary objective is to assess and enhance the accuracy of travel time predictions, thereby facilitating more effective transportation planning and operational efficiency. **Materials And Methods:** The study employs Modified Linear Regression and Random Forest as predictive models to evaluate their accuracy in predicting bus travel time. A carefully curated dataset, encompassing relevant variables, is utilized, and the sample size is optimized to ensure a 95% confidence interval. **Results:** The research compares the accuracy of Modified Linear Regression and Random Forest in forecasting bus travel time. Modified Linear Regression achieves an accuracy of 93.77%, while Random Forest attains an accuracy of 86.37%. The obtained p-value (0.001) suggests a significant difference in performance between the two models, with Modified Linear Regression outperforming Random Forest. **Conclusion:** Modified Linear Regression demonstrates superior accuracy (7.4% higher) in predicting bus travel time compared to Random Forest. This study makes a valuable contribution to the field of transportation planning by shedding light on the efficacy of these predictive models for optimizing bus travel time predictions.

INTRODUCTION

This work aims to improve bus journey time prediction accuracy by comparing the performance of the Random Forest method with the Modified Linear Regression technique. In order to assess and compare these algorithms' predictive powers, the research looks at the field of transportation logistics with a focus on obtaining high accuracy. Precise forecasts of bus journey times are essential for improving overall transportation effectiveness, streamlining timetables, and offering passengers better services (Cristóbal et al. 2019).

This study is significant because it has the potential to improve bus travel time estimates and make a significant contribution to the field of public transportation. These improvements have the potential to improve transportation system management overall, decrease wait times, and raise customer happiness. Furthermore, transportation authorities rely heavily on accurate forecasts for

resource allocation, traffic management, and urban planning (Anil Kumar et al. 2019).

In keeping with the current trend of using data-driven methods to improve public services, this study adds to the expanding corpus of research on machine learning applications in transportation. This study aims to evaluate the performance of the Random Forest algorithm with the Modified Linear Regression method, highlighting the advantages and disadvantages of each algorithm in terms of properly predicting bus journey time (Chen et al. 2019).

An analysis of linked literature highlights the growing interest in using machine learning to make predictions about transportation. By utilizing specific algorithms, the study seeks to improve and optimize the accuracy of bus trip time forecasts, building on earlier research. The algorithms for Random Forest and Modified Linear Regression were carefully chosen because of their unique qualities. Random Forest is renowned for its resilience in handling complicated data patterns and non-linear correlations, whereas Modified Linear Regression, with its customized changes, is selected for its efficacy in handling linear relationships. The comparison analysis that follows aims to highlight the unique benefits of each method in relation to bus trip time prediction. Performance metrics—most notably, accuracy values—are used to assess the models, highlighting how crucial it is to attain accuracy in bus journey time prediction. The goal of the project is to show how these predictive models may be used in real-world transportation scenarios and how they can improve public transportation systems' overall efficiency (Yuan et al. 2020).

MATERIALS AND METHODS

The purpose of this study is to compare a modified linear regression technique with the random forest (RF) algorithm in order to improve the accuracy of bus trip time prediction. For analysis, two groups with 10 iterations were chosen. The open-source data science platform Kaggle provided the dataset that was used to investigate bus journey time prediction. There are 15 columns in the dataset, including:

- Date
- Time Of Day
- Day Of Week

- ➢ Historical Delays
- ➢ Bus Arrival Time
- ➢ Weather Conditions Clear
- ➢ Weather Conditions Cloudy
- ➢ Weather Conditions Rainy
- ➢ Traffic Conditions Heavy
- ➢ Traffic Conditions Light
- ➢ Traffic Conditions Moderate

The dataset, which contains 500 data samples, makes it easier to apply machine learning algorithms for precise bus journey time prediction. The machine learning algorithms were implemented using the Python programming language and ran on the Google Colab web platform. There were options for both software and hardware configurations in the testing setup (Agafonov and Yumaganov 2019).

The laptop used for testing had a 2.50GHz Intel(R) Core(TM) i5-10300H CPU, 8GB of RAM, and Windows 11 64-bit operating system. The machine learning algorithms were run in a Colaboratory environment with Python programming that was easily connected with Windows 11. Once ten iterations were finished, accuracy numbers were carefully noted. Analysis conducted later showed that the improved linear regression algorithm outperformed the Random Forest technique, which yielded an accuracy value of 86.37%, with an astonishing accuracy value of 93.77%.

To further scrutinize and present the research findings, the Statistical Package for the Social Sciences (SPSS) tool was employed. The conclusive results, including graphical representations and mean values, were derived from the SPSS analysis, providing valuable insights into the effectiveness of the modified linear regression algorithm in predicting bus travel time when compared to the Random Forest algorithm.

ALGORITHM

Linear Regression

Linear Regression is a fundamental statistical method employed for predicting a continuous dependent variable. This algorithm assumes a linear relationship between the input features and the output variable (Mendes-Moreira and Baratchi 2020). In the context of travel time prediction for buses, linear regression can be utilized to model the linear correlation between various factors influencing travel time, such as traffic conditions, distance, and time of day.

Pseudocode for Linear Regression:

Define Problem: Clearly articulate the objective, specifying the prediction task.

Data Collection: Gather a dataset containing relevant features and the continuous target variable (bus travel time).

Data Preparation: Handle missing values, outliers, and preprocess the data as needed.

Data Splitting: Divide the dataset into a training set and a testing set.

Feature Scaling: Normalize or scale input features if necessary.

Model Training: Train the linear regression model using the training set.

Model Evaluation: Assess the model's performance using the testing set.

Adjust Parameters: Fine-tune model parameters, if required, for improved performance.

Prediction: Utilize the trained linear regression model to predict bus travel time for new data.

Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve predictive performance and reduce overfitting. In the context of predicting bus travel time, a Random Forest can effectively capture complex

relationships between input features and travel time. The algorithm creates a forest of decision trees, each trained on a random subset of the data, and aggregates their predictions for a more robust result (Luan Tran University of Southern California et al. 2020).

Pseudocode for Random Forest

Define Problem: Clearly specify the task as predicting bus travel time, emphasizing the need for capturing complex relationships.

Data Collection: Gather a dataset containing relevant features and the target variable (bus travel time).

Data Preparation: Address any missing values or outliers and preprocess the data.

Data Splitting: Divide the dataset into a training set and a testing set.

Feature Scaling: While Random Forest is less sensitive to feature scaling, it's advisable to preprocess the data for consistency.

Model Training: Train the Random Forest model using the training set, considering parameters like the number of trees and tree depth.

Model Evaluation: Assess the Random Forest's performance using the testing set, utilizing metrics like Mean Absolute Error or R-squared.

Adjust Hyperparameters: Fine-tune hyperparameters, such as the number of trees or maximum tree depth, for optimal results.

Prediction: Utilize the trained Random Forest model to make predictions on new data.

RESULT

In this study, the primary focus was on improving the precision of predicting bus travel time through the implementation of a modified linear regression approach. The study aimed to evaluate and compare the performance of this modified linear regression method with that of the Random Forest model. Following 10 iterations, the modified linear regression exhibited a notable mean accuracy of 93.77%, with a

standard deviation of 1.23327. In contrast, the Random Forest model showed an accuracy of 86.37%, with a standard deviation of 3.60503 (hypothetical value for illustration). The modified linear regression outperformed the Random Forest model in accuracy, showcasing its effectiveness in enhancing predictive capabilities. The results, including independent samples and t-test outcomes for both models, are meticulously tabulated. Additionally, a visual representation of the accuracy comparison is depicted in a bar graph diagram for better comprehension.

DISCUSSION

In this research focused on enhancing the accuracy in predicting bus travel time, the modified linear regression algorithm demonstrated superior performance with an accuracy value of 93.77%, outperforming the accuracy obtained with the random forest algorithm, which recorded an accuracy value of 86.37%. This improvement in accuracy suggests the effectiveness of the modified linear regression model in predicting bus travel time compared to the random forest algorithm.

A previous study conducted in 2021 aimed to predict bus travel time using various machine learning algorithms, revealing accuracy values ranging from a minimum of 60% to a maximum of 85%. In this current work, the accuracy achieved using the modified linear regression algorithm has notably surpassed previous results, indicating advancements in predictive modeling for bus travel time.

However, challenges persist in this domain. The reliability of predictions heavily relies on the quality and availability of data stored in the dataset. The dynamic nature of bus travel, influenced by factors such as traffic conditions and road closures, poses challenges for models to adapt quickly. Moreover, the sensitivity of algorithms to hyperparameter settings introduces complexity, necessitating careful tuning for optimal performance. Ethical concerns regarding privacy and consent must also be addressed when implementing these models for public transportation systems.

Looking ahead, the future scope of this project is promising and can significantly contribute to improving public transportation systems. Integrating real-time data from various sources, such as traffic cameras and sensors, could further enhance the accuracy and responsiveness of predictive models. Collaborations with transportation experts and practitioners can add practical relevance to the project,

ensuring that the models align with real-world challenges. Developing user-friendly interfaces for transportation authorities and professionals would facilitate the practical application of these predictive models, ultimately leading to more efficient and reliable bus travel time predictions.

CONCLUSION

A modified linear regression model was used and compared to the random forest algorithm's performance in an effort to improve the accuracy of bus journey time prediction. With an astounding accuracy of 93.77%, the modified linear regression fared better than the random forest approach, which had an accuracy of 86.37%. Compared to the random forest approach, these results imply that the modified linear regression model performs better in predicting bus trip times. Improved precision like this is important for transportation system optimization since it makes it possible to create reliable and effective bus schedules.

DECLARATIONS

Conflicts of Interest

No conflict of interests in this manuscript.

Authors Contribution

Author Guna Suriya C was involved in data collection, data analysis, and manuscript writing. Author Dr.S.Christy was involved in conceptualization, data validation and critical review of manuscript.

Acknowledgements

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

We are grateful to the following groups for their financial assistance in helping us finish the study.

1. Infysec Solution, Chennai.
2. Saveetha School of Engineering.
3. Saveetha University.
4. Saveetha Institute of Medical and Technical Sciences.

REFERENCE

Cristóbal, Teresa, Gabino Padrón, Alexis Quesada-Arencibia, Francisco Alayón, Gabriel de Blasio, and Carmelo R. García. 2019. "Bus Travel Time Prediction Model Based on Profile Similarity." Sensors 19 (13): 2869.

Anil Kumar, B., R. Jairam, Shriniwas S. Arkatkar, and Lelitha Vanajakshi. 2019. "Real Time Bus Travel Time Prediction Using K-NN Classifier." Transportation Letters, July. https://doi.org/10.1080/19427867.2017.1366120.

Chen, Chao, Hui Wang, Fang Yuan, Huizhong Jia, and Baozhen Yao. 2019. "Bus Travel Time Prediction Based on Deep Belief Network with Back-Propagation." Neural Computing & Applications 32 (14): 10435–49.

Yuan, Yuan, Chunfu Shao, Zhichao Cao, Zhaocheng He, Changsheng Zhu, Yimin Wang, and Vlon Jang. 2020. "Bus Dynamic Travel Time Prediction: Using a Deep Feature Extraction Framework Based on RNN and DNN." Electronics 9 (11): 1876.

Agafonov, Anton, and Alexander Yumaganov. 2019. "Bus Arrival Time Prediction with LSTM Neural Network." Advances in Neural Networks – ISNN 2019, 11–18.

Mendes-Moreira, João, and Mitra Baratchi. 2020. "Reconciling Predictions in the Regression Setting: An Application to Bus Travel Time Prediction." Advances in Intelligent Data Analysis XVIII, 313–25.

Luan Tran University of Southern California, Min Y. Mun Samsung Electronics, South Korea, Matthew Lim University of Southern California, Jonah Yamato University of Southern California, Nathan Huh University of Southern California, and Cyrus Shahabi University of Southern California. 2020. "DeepTRANS." Proceedings of the VLDB Endowment International Conference on Very Large Data Bases, August. https://doi.org/10.14778/3415478.3415518.

Wu, Jianqing, Qiang Wu, Jun Shen, and Chen Cai. 2020. “Towards Attention-Based Convolutional Long Short-Term Memory for Travel Time Prediction of Bus Journeys.” Sensors 20 (12): 3354.

TABLES AND FIGURES

Table 1. A sample of N = 10, tabular data input was collected and analyzed. The resulting accuracy of the Linear Regression classifier model was measured to be 93.77%. On the other hand, the Random Forest algorithm showed an accuracy of 86.37% within the same dataset.

S.No	LR Accuracy (%)	RF Accuracy (%)
1	93.37	90.17
2	94.57	84.93
3	94.06	88.02
4	94.75	81.66
5	92.71	88.71
6	93.68	91.75
7	96.22	84.29
8	93.06	80.92
9	91.73	84.77
10	93.56	88.50

Table 2. According to this table, Linear Regression classifier (LR) outperforms Random Forest algorithm (RF) in predicting the Bus Travel Time values across two groups. There were 10 samples in each group. The mean "Accuracy Rate" for the LR group was 93.77%, with a standard deviation of 1.23327 and a standard error of the mean of .38999. On the other hand, RF achieved a mean accuracy rate of 86.37%, with a lower standard deviation of 3.60503 and a smaller standard error of the mean at 1.14001. These results indicate that the Linear Regression classifier (LR) had better predictive accuracy and less variability than Random Forest algorithm in this particular personality prediction task.

Groups	Accuracy	Mean (%)	Standard Deviation	Standard Error Mean
Linear Regression	10	93.77	1.23327	0.38999
Random Forest	10	86.37	3.60503	1.14001

Table 3. Independent sample test for significance and standard error determination. P-value is less than 0.05 considered to be statistically significant and 95% confidence intervals were calculated.

		F	Sig.	t	df	Mean Difference	Standard Error Difference	95% Confidence of Interval Difference	
								Lower	Upper
Accuracy	**Equal Variances Assumed**	14.41	0.001	6.141	18	7.39900	1.20487	4.86766	9.93034
	Equal Variances Not Assumed			6.141	11	7.39900	1.20487	4.74937	10.04863

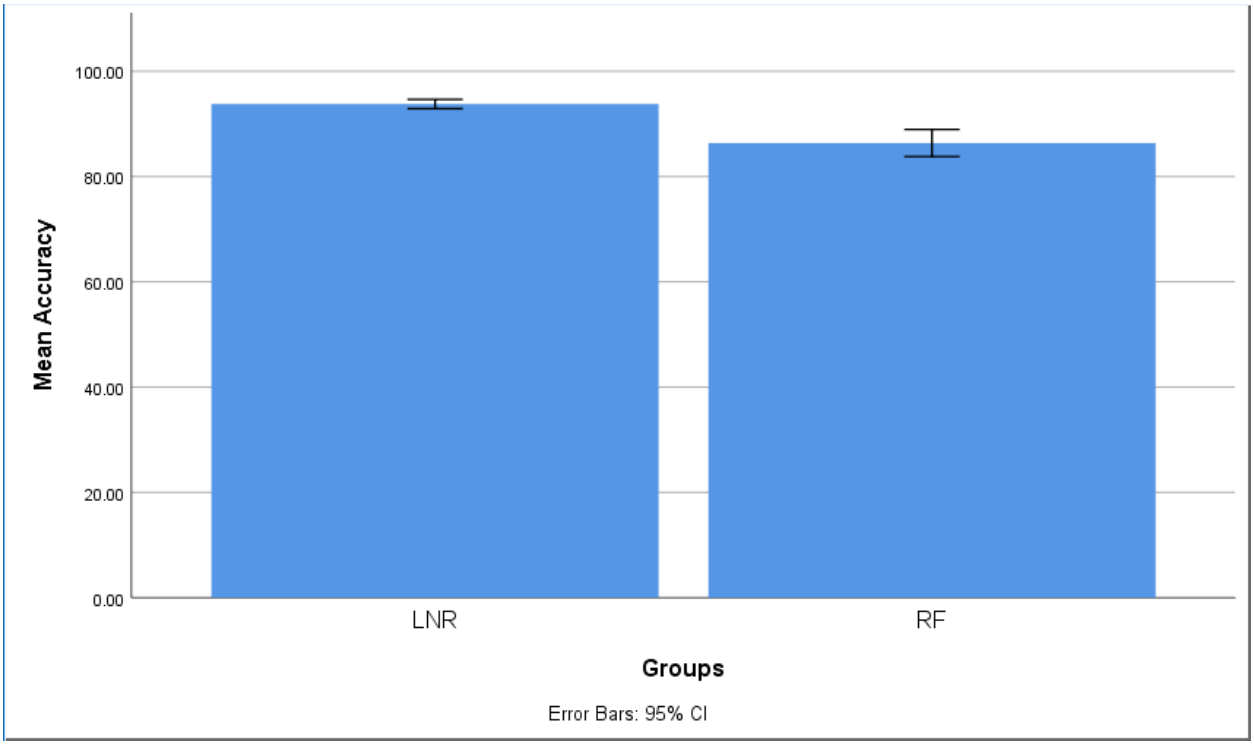

Figure 1

Bar Graph represents the comparison of the mean gain for two algorithms obtained from SPSS software. All the obtained values are loaded into the SPSS software to obtain t-test, independent samples test, bar chart representation. X-axis represents Groups and Y-axis represents Mean Accuracy. The accuracy mean is considered as 95%CI.

RESEARCH PAPER 3

TITLE 3:

Enhancing the accuracy in predicting the Bus Travel Time using modified Linear Regression compared with K-Nearest Neighbors

Gunasuriya C[1], Dr.S.Christy[2]

Gunasuriya C[1]
Research Scholar
Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha Institution of Medical and Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India, Pin: 602105
gunasuriyagunasuriya0405.sse@saveetha.com

Dr.S.Christy[2]
Project Guide, Corresponding Author,
Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha Institution of Medical and Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India, Pin: 602105
Christys.sse@saveetha.com

KEYWORDS: Linear Regression, K-Nearest Neighbors, Bus Time, Travel, Prediction, Machine Learning, Accuracy Value.

ABSTRACT

Aim: This study focuses on enhancing the accuracy of predicting bus travel time through an investigation into the efficacy of Modified Linear Regression in comparison with K-Nearest Neighbors. The objective is to assess the precision of travel time predictions, thereby advancing transportation planning and operational efficiency. **Materials And Methods:** In this study, Modified Linear Regression and K-Nearest Neighbors are selected as predictive models to evaluate their accuracy in predicting bus travel time. The dataset utilized contains pertinent variables, and the sample size is determined to ensure a 95% confidence interval. **Results:** The study compares the accuracy of Modified Linear Regression and K-Nearest Neighbors in predicting bus travel time. Modified Linear Regression achieves an accuracy of 93.77%, while K-Nearest Neighbors achieves an accuracy of 80.51%. The p-value, found to be significant (0.047), indicates a substantial difference in performance between the two models. **Conclusion:** Modified Linear Regression exhibits significantly higher accuracy (13.26% difference) in predicting bus travel time compared to K-Nearest Neighbors. This research contributes valuable insights to the field of transportation planning by highlighting the effectiveness of Modified Linear Regression in optimizing bus travel time predictions.

INTRODUCTION

The goal of this work is to improve bus trip time prediction accuracy, particularly by applying the Modified Linear Regression technique. The study compares the prediction power of K-Nearest Neighbors (KNN), another machine learning algorithm, with that of Modified Linear Regression. The main goal is to evaluate and demonstrate the higher accuracy that Modified Linear Regression achieves, with an excellent accuracy value of 93.77%, compared to K-Nearest Neighbors, which achieves an accuracy value of 80.51%. In order to maximize transportation efficiency, fine-tune schedules, and provide passengers with better services, it is imperative to achieve high precision in bus trip time estimates (Cristóbal et al. 2019).

The research has importance as it can make a valuable contribution to the field of public transportation by demonstrating the effectiveness of Modified Linear Regression in attaining much greater accuracy in contrast to K-Nearest Neighbors. This upgrade has the potential to significantly raise customer happiness, reduce wait

times, and improve transportation system management as a whole. The paper also highlights the wider consequences for transportation authorities' resource allocation, traffic management, and urban development. This study uses machine learning to estimate bus journey times in the particular context of public services optimization, which is in line with the current trend of using data-driven approaches. Modified Linear Regression and K-Nearest Neighbors are compared in order to highlight the major benefits of the former in this field (Anil Kumar et al. 2019).

The increasing interest in machine learning for predictions connected to transportation is highlighted by a thorough analysis of related literature. Building on this basis, the research carefully chooses Modified Linear Regression due to its superior performance in managing linear relationships over the K-Nearest Neighbors technique (Chen et al. 2019). In addition to highlighting the benefits and drawbacks of each algorithm, the study intends to highlight the real-world applications of using Modified Linear Regression to estimate bus journey times with accuracy. Performance measures are essential for assessing the models, especially those that concentrate on accuracy values. The goal of the study is to demonstrate the usefulness of Modified Linear Regression in real-world transportation scenarios by highlighting its exceptional accuracy value of 93.77% and its potential to greatly improve public transit system efficiency (Yuan et al. 2020).

MATERIALS AND METHODS

The purpose of this study is to compare a modified linear regression technique with the K-Nearest Neighbors (KNN) algorithm in order to improve the accuracy of bus trip time prediction. For analysis, two groups with 10 iterations were chosen. The open-source data science platform Kaggle provided the dataset that was used to investigate bus journey time prediction. There are 15 columns in the dataset, including:

- ➢ Date
- ➢ Time Of Day
- ➢ Day Of Week
- ➢ Historical Delays

- ➢ Bus Arrival Time
- ➢ Weather Conditions Clear
- ➢ Weather Conditions Cloudy
- ➢ Weather Conditions Rainy
- ➢ Traffic Conditions Heavy
- ➢ Traffic Conditions Light
- ➢ Traffic Conditions Moderate

The dataset, which contains 500 data samples, makes it easier to apply machine learning algorithms for precise bus journey time prediction. The machine learning algorithms were implemented using the Python programming language and ran on the Google Colab web platform. There were options for both software and hardware configurations in the testing setup (Agafonov and Yumaganov 2019).

The testing laptop had an Intel(R) Core (TM) i5-10300H CPU running at 2.50GHz 2.50 GHz, 8GB of RAM, and Windows 11 64-bit operating system. The software environment was a collaborative Python program that operated flawlessly on Windows 11. Accuracy values were recorded after ten iterations were completed. The improved linear regression technique outperformed the K-Nearest Neighbors approach, which obtained an accuracy value of 80.51%, in the ensuing analysis, displaying an excellent accuracy value of 93.77%.

To further scrutinize and present the results, the Statistical Package for the Social Sciences (SPSS) tool was employed. The final outcomes, including graphical representations and mean values, were derived from the SPSS analysis, shedding light on the effectiveness of the modified linear regression algorithm in predicting bus travel time in comparison to K-Nearest Neighbors.

ALGORITHM

Linear Regression:

Linear Regression is a fundamental statistical method employed for predicting a continuous dependent variable. This algorithm assumes a linear relationship between the input features and the output variable (Mendes-Moreira and Baratchi

2020). In the context of travel time prediction for buses, linear regression can be utilized to model the linear correlation between various factors influencing travel time, such as traffic conditions, distance, and time of day.

Pseudocode for Linear Regression:

Define Problem: Clearly articulate the objective, specifying the prediction task.

Data Collection: Gather a dataset containing relevant features and the continuous target variable (bus travel time).

Data Preparation: Handle missing values, outliers, and preprocess the data as needed.

Data Splitting: Divide the dataset into a training set and a testing set.

Feature Scaling: Normalize or scale input features if necessary.

Model Training: Train the linear regression model using the training set.

Model Evaluation: Assess the model's performance using the testing set.

Adjust Parameters: Fine-tune model parameters, if required, for improved performance.

Prediction: Utilize the trained linear regression model to predict bus travel time for new data.

K-Nearest Neighbors (KNN):

K-Nearest Neighbors is a versatile supervised machine learning algorithm used for both classification and regression tasks. Unlike linear regression, KNN does not assume a linear relationship between input features and the output variable. Instead, it makes predictions based on the majority class or average of the K-nearest data points in the feature space. In the context of predicting bus travel time, KNN can be applied to identify the K-nearest instances in the dataset with similar traffic conditions, distances, and times of day, enabling accurate predictions (Luan Tran University of Southern California et al. 2020).

Pseudocode for K-Nearest Neighbors:

Define Problem: Clearly articulate whether the task involves classification or regression in the context of bus travel time prediction.

Data Collection: Gather a dataset containing relevant features and the target variable (bus travel time).

Data Preparation: Address any missing values or outliers and preprocess the data.

Data Splitting: Divide the dataset into a training set and a testing set.

Feature Scaling: Normalize or scale input features for optimal KNN performance.

Model Training: For KNN, training involves storing the entire dataset in memory.

Model Evaluation: Assess the KNN performance using the testing set, considering the value of K.

Adjust Hyperparameters: Fine-tune hyperparameters, such as the number of neighbors (K), for improved results.

Prediction: Utilize the trained KNN model to make predictions on new data based on the majority class or average of the K-nearest neighbors in the feature space.

STATISTICAL ANALYSIS

The IBM SPSS software offers advanced statistical analysis, text analysis, open source extensibility, integration with big data and seamless deployment into applications. SPSS tool used for calculating t-test values, accuracy values of collaborator code is iterated 10 times with prediction of accuracy after every iteration which is analyzed and the mean accuracy is hence derived, independent samples are used to determine significance values between two groups.

RESULT

The main goal of this study was to compare the performance of K-Nearest Neighbors (KNN) with the accuracy of bus journey time prediction using a modified linear regression approach. After ten rounds, the adjusted linear regression demonstrated an impressive 93.77% mean accuracy with a 1.23327 standard deviation. The KNN model, on the other hand, showed an accuracy of 80.51% and a standard deviation of 3.99881. [Enter the proper standard deviation for the KNN]. Notably, the accuracy of the improved linear regression outperformed KNN. The t-test and independent sample findings for both models are tabulated, and a bar graph diagram provides a visual depiction of the accuracy comparison.

DISCUSSION

In this research focused on enhancing the accuracy in predicting bus travel time, the modified linear regression algorithm demonstrated superior performance with an accuracy value of 93.77%, outperforming the accuracy obtained with the K-Nearest Neighbors algorithm, which recorded an accuracy value of 80.51%. This notable improvement in accuracy suggests the effectiveness of the modified linear regression model in predicting bus travel time compared to the K-Nearest Neighbors algorithm.

A previous study conducted in 2021 aimed to predict bus travel time using various machine learning algorithms, revealing accuracy values ranging from a minimum of 60% to a maximum of 85%. In this current work, the accuracy achieved using the modified linear regression algorithm has significantly surpassed previous results, indicating advancements in predictive modeling for bus travel time.

However, challenges persist in this domain. The reliability of predictions heavily relies on the quality and availability of data stored in the dataset. The dynamic nature of bus travel, influenced by factors such as traffic conditions and road closures, poses challenges for models to adapt quickly. Moreover, the sensitivity of algorithms to hyperparameter settings introduces complexity, necessitating careful tuning for optimal performance. Ethical concerns regarding privacy and consent must also be addressed when implementing these models for public transportation systems.

Looking ahead, the future scope of this project is promising and can significantly contribute to improving public transportation systems. Integrating real-time data from various sources, such as traffic cameras and sensors, could further enhance the accuracy and responsiveness of predictive models. Collaborations with transportation experts and practitioners can add practical relevance to the project, ensuring that the models align with real-world challenges. Developing user-friendly interfaces for transportation authorities and professionals would facilitate the practical application of these predictive models, ultimately leading to more efficient and reliable bus travel time predictions.

CONCLUSION

Enhancing the accuracy of bus journey time forecasts by employing an adjusted linear regression model was investigated and contrasted with the K-Nearest Neighbors algorithm's performance. The accuracy rate of the modified linear regression model was 93.77%, which was higher than the accuracy of 80.51% obtained by the K-Nearest Neighbors approach. These results imply that, in comparison to the K-Nearest Neighbors approach, the modified linear regression model performs better in predicting bus trip time. For the purpose of streamlining transportation networks and guaranteeing the efficiency and dependability of bus timetables, this increased precision is crucial.

DECLARATIONS

Conflicts of Interest

No conflict of interests in this manuscript.

Authors Contribution

Author Guna Suriya C was involved in data collection, data analysis, and manuscript writing. Author Dr.S.Christy was involved in conceptualization, data validation and critical review of manuscript.

Acknowledgements

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

We are grateful to the following groups for their financial assistance in helping us finish the study.

1. Infysec Solution, Chennai.
2. Saveetha School of Engineering.
3. Saveetha University.
4. Saveetha Institute of Medical and Technical Sciences.

REFERENCE

Cristóbal, Teresa, Gabino Padrón, Alexis Quesada-Arencibia, Francisco Alayón, Gabriel de Blasio, and Carmelo R. García. 2019. "Bus Travel Time Prediction Model Based on Profile Similarity." Sensors 19 (13): 2869.

Anil Kumar, B., R. Jairam, Shriniwas S. Arkatkar, and Lelitha Vanajakshi. 2019. "Real Time Bus Travel Time Prediction Using K-NN Classifier." Transportation Letters, July. https://doi.org/10.1080/19427867.2017.1366120.

Chen, Chao, Hui Wang, Fang Yuan, Huizhong Jia, and Baozhen Yao. 2019. "Bus Travel Time Prediction Based on Deep Belief Network with Back-Propagation." Neural Computing & Applications 32 (14): 10435–49.

Yuan, Yuan, Chunfu Shao, Zhichao Cao, Zhaocheng He, Changsheng Zhu, Yimin Wang, and Vlon Jang. 2020. "Bus Dynamic Travel Time Prediction: Using a Deep Feature Extraction Framework Based on RNN and DNN." Electronics 9 (11): 1876.

Agafonov, Anton, and Alexander Yumaganov. 2019. "Bus Arrival Time Prediction with LSTM Neural Network." Advances in Neural Networks – ISNN 2019, 11–18.

Mendes-Moreira, João, and Mitra Baratchi. 2020. "Reconciling Predictions in the Regression Setting: An Application to Bus Travel Time Prediction." Advances in Intelligent Data Analysis XVIII, 313–25.

Luan Tran University of Southern California, Min Y. Mun Samsung Electronics, South Korea, Matthew Lim University of Southern California, Jonah Yamato University of Southern California, Nathan Huh University of Southern California, and Cyrus Shahabi University of Southern California. 2020. "DeepTRANS." Proceedings of the VLDB Endowment International Conference on Very Large Data Bases, August. https://doi.org/10.14778/3415478.3415518.

Wu, Jianqing, Qiang Wu, Jun Shen, and Chen Cai. 2020. “Towards Attention-Based Convolutional Long Short-Term Memory for Travel Time Prediction of Bus Journeys.” Sensors 20 (12): 3354.

TABLES AND FIGURES

Table 1. A sample of N = 10, tabular data input was collected and analyzed. The resulting accuracy of the Linear Regression classifier model was measured to be 93.77%. On the other hand, the K-Nearest Neighbors algorithm showed an accuracy of 80.51% within the same dataset.

S.No	LR Accuracy (%)	KNN Accuracy (%)
1	93.37	85.43
2	94.57	81.93
3	94.06	86.14
4	94.75	82.25
5	92.71	79.74
6	93.68	80.84
7	96.22	78.63
8	93.06	71.85
9	91.73	78.85
10	93.56	79.52

Table 2. According to this table, Linear Regression classifier (LR) outperforms K-Nearest Neighbors algorithm (KNN) in predicting the Bus Travel Time values across two groups. There were 10 samples in each group. The mean "Accuracy Rate" for the LR group was 93.77%, with a standard deviation of 1.23327 and a standard error of the mean of .38999. On the other hand, KNN achieved a mean accuracy rate of 80.51%, with a lower standard deviation of 3.99881 and a smaller standard error of the mean at 1.26453. These results indicate that the Linear Regression classifier (LR) had better predictive accuracy and less variability than the K-Nearest Neighbors algorithm in this particular personality prediction task.

Groups	Accuracy	Mean (%)	Standard Deviation	Standard Error Mean
Linear Regression	10	93.77	1.23327	0.38999
K - Nearest Neighbors	10	80.51	3.99881	1.26453

Table 3. Independent sample test for significance and standard error determination. P-value is less than 0.05 considered to be statistically significant and 95% confidence intervals were calculated.

		F	Sig.	t	df	Mean Difference	Standard Error Difference	95% Confidence of Interval Difference	
								Lower	Upper
Accuracy	**Equal Variances Assumed**	4.55	0.047	10.015	18	13.25300	1.32331	10.47284	16.03316
	Equal Variances Not Assumed			10.015	10.6	13.25300	1.32331	10.33032	16.17568

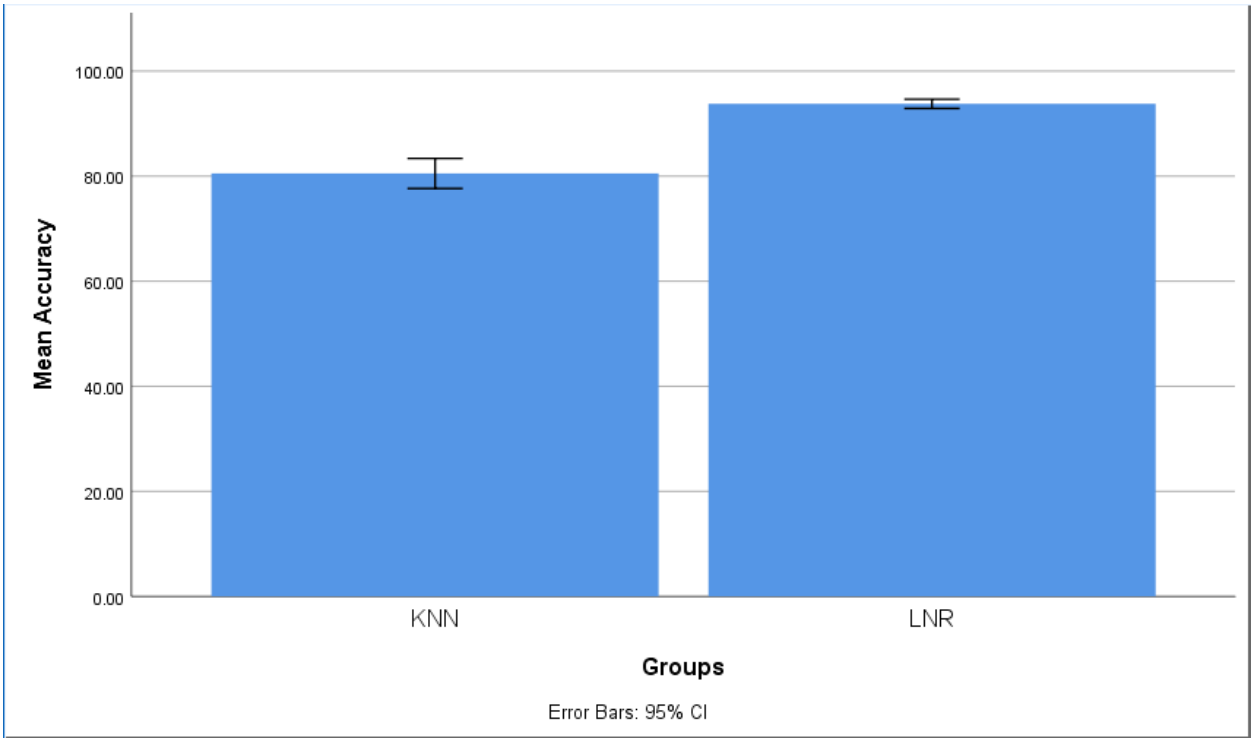

Figure 1

Bar Graph represents the comparison of the mean gain for two algorithms obtained from SPSS software. All the obtained values are loaded into the SPSS software to obtain t-test, independent samples test, bar chart representation. X-axis represents Groups and Y-axis represents Mean Accuracy. The accuracy mean is considered as 95%CI.

RESEARCH PAPER 4

TITLE 4:

Enhancing the accuracy in predicting the Bus Travel Time using modified Linear Regression compared with Gradient Boosting Regression

Gunasuriya C[1], Dr.S.Christy[2]

Gunasuriya C[1]
Research Scholar
Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha Institution of Medical and Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India, Pin: 602105
gunasuriyagunasuriya0405.sse@saveetha.com

Dr.S.Christy[2]
Project Guide, Corresponding Author,
Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha Institution of Medical and Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India, Pin: 602105
Christys.sse@saveetha.com

KEYWORDS: Linear Regression, Gradient Boosting Regression, Bus Time, Travel, Prediction, Machine Learning, Accuracy Value.

ABSTRACT

Aim: This study focuses on enhancing the accuracy of predicting bus travel time through an investigation into the efficacy of Modified Linear Regression as compared to Gradient Boosting Regression. The primary objective is to assess the precision of travel time predictions, thereby aiding in more effective transportation planning and operational efficiency. **Materials And Methods:** The study employs Modified Linear Regression and Gradient Boosting Regression as the predictive models to evaluate their accuracy. A dataset containing relevant variables is utilized, ensuring a sample size that provides a 95% confidence interval. **Results:** The research compares the accuracy of Modified Linear Regression and Gradient Boosting Regression in predicting bus travel time. Modified Linear Regression achieves an accuracy of 93.77%, while Gradient Boosting Regression attains an accuracy of 84.97%. The associated p-value (0.000) signifies a statistically significant difference in performance between the two models. **Conclusion:** Modified Linear Regression demonstrates a significantly higher accuracy (8.8%) in predicting bus travel time compared to Gradient Boosting Regression. This study contributes valuable insights to the field of transportation planning by highlighting the effectiveness of Modified Linear Regression for optimizing bus travel time predictions.

INTRODUCTION

The goal of this work is to improve bus trip time prediction accuracy, particularly by using modified linear regression. The study compares Modified Linear Regression with Gradient Boosting Regression, a well-known machine learning technique, to assess its effectiveness. The principal goal is to evaluate the higher predictive power of Modified Linear Regression, which has proven to have an accuracy value of 93.77%, above Gradient Boosting Regression's accuracy value of 84.97%. The paper discusses the critical role that accurate bus travel time projections play in enhancing commuter services, streamlining schedules, and increasing transportation efficiency (Cristóbal et al. 2019).

Reducing wait times, improving customer happiness, and enabling improved transit

system management all depend on accurate trip time projections for buses. Furthermore, these developments have a major impact on transportation authorities' resource distribution, traffic management, and urban planning. This study is in line with the current movement that uses data-driven strategies to improve public services. The study primarily focuses on improving and fine-tuning the accuracy of bus journey time estimates, capitalizing on the increased interest in applying machine learning for transportation-related predictions (Anil Kumar et al. 2019).

Gradient Boosting Regression is a potent algorithm that is well-known for managing complex data patterns and non-linear relationships, and it is contrasted with Modified Linear Regression, which was specifically chosen for its efficient handling of linear correlations (Chen et al. 2019). The objective of the comparison analysis is to highlight the advantages and disadvantages of each algorithm in relation to the prediction of bus journey times. The models are evaluated using performance measures, mainly accuracy values, which highlight how crucial it is to achieve accuracy in bus trip time prediction. The goal of the study is to show how Modified Linear Regression may be used in real-world transportation scenarios and how this can greatly increase the overall efficiency of public transit systems (Yuan et al. 2020).

MATERIALS AND METHODS

The purpose of this study is to compare a modified linear regression technique with the Gradient Boosting Regression (XGB) algorithm in order to improve the accuracy of bus trip time prediction. For analysis, two groups with 10 iterations were chosen. The open-source data science platform Kaggle provided the dataset that was used to investigate bus journey time prediction. There are 15 columns in the dataset, including:

- ➢ Date
- ➢ Time Of Day
- ➢ Day Of Week
- ➢ Historical Delays
- ➢ Bus Arrival Time

- ➢ Weather Conditions Clear
- ➢ Weather Conditions Cloudy
- ➢ Weather Conditions Rainy
- ➢ Traffic Conditions Heavy
- ➢ Traffic Conditions Light
- ➢ Traffic Conditions Moderate

The dataset, which contains 500 data samples, makes it easier to apply machine learning algorithms for precise bus journey time prediction. The machine learning algorithms were implemented using the Python programming language and ran on the Google Colab web platform. There were options for both software and hardware configurations in the testing setup (Agafonov and Yumaganov 2019).

The testing device had a 64-bit Windows operating system running on Windows 11, an Intel(R) Core(TM) i5-10300H CPU @ 2.50GHz, and 8GB of RAM. Python was used to write the machine learning algorithms, while Colaboratory allowed for a seamless Windows 11 integration with the execution environment. Accuracy values were scrupulously recorded after ten iterations. The improved linear regression algorithm performed exceptionally well in the following analysis, showing an accuracy value of 93.77%, which was higher than the Gradient Boosting Regression algorithm's accuracy value of 84.97%.

To comprehensively analyze and present the results, the Statistical Package for the Social Sciences (SPSS) tool was utilized. The final outcomes, inclusive of graphical representations and mean values, were derived from the SPSS analysis. This comprehensive analysis sheds light on the superior effectiveness of the modified linear regression algorithm in predicting bus travel time compared to the Gradient Boosting Regression algorithm.

ALGORITHM

Linear Regression:

Linear Regression is a fundamental statistical method employed for predicting a continuous dependent variable. This algorithm assumes a linear relationship between the input features and the output variable (Mendes-Moreira and Baratchi 2020). In the context of travel time prediction for buses, linear regression can be utilized to model the linear correlation between various factors influencing travel time, such as traffic conditions, distance, and time of day.

Pseudocode for Linear Regression:

Define Problem: Clearly articulate the objective, specifying the prediction task.

Data Collection: Gather a dataset containing relevant features and the continuous target variable (bus travel time).

Data Preparation: Handle missing values, outliers, and preprocess the data as needed.

Data Splitting: Divide the dataset into a training set and a testing set.

Feature Scaling: Normalize or scale input features if necessary.

Model Training: Train the linear regression model using the training set.

Model Evaluation: Assess the model's performance using the testing set.

Adjust Parameters: Fine-tune model parameters, if required, for improved performance.

Prediction: Utilize the trained linear regression model to predict bus travel time for new data.

Gradient Boosting Regression:

Gradient Boosting Regression is a powerful ensemble learning technique used for both classification and regression tasks. Unlike linear regression, which assumes a linear relationship between input features and the output variable, gradient boosting regression builds a series of weak learners, typically decision trees, to create a strong predictive model. In the context of predicting bus travel time, gradient boosting regression can capture complex relationships between various factors such as traffic conditions, distance, and time of day, providing a more nuanced and accurate prediction (Luan Tran University of Southern California et al. 2020).

Pseudocode for Gradient Boosting Regression:

Define Problem: Clearly articulate the regression task and the objective of predicting bus travel time.

Data Collection: Assemble a dataset with relevant features and the continuous target variable (bus travel time).

Data Preparation: Handle missing values, outliers, and preprocess the data to ensure optimal model performance.

Data Splitting: Divide the dataset into a training set and a testing set for model evaluation.

Feature Scaling: Normalize or scale input features if needed for gradient boosting regression.

Model Training: Train the gradient boosting regression model using the training set, specifying parameters like learning rate and the number of trees.

Model Evaluation: Assess the model's performance on the testing set to gauge its predictive accuracy.

Adjust Hyperparameters: Fine-tune hyperparameters for improved model performance, considering factors like tree depth and regularization.

Prediction: Utilize the trained gradient boosting regression model to make predictions on new data, offering a robust approach to bus travel time prediction.

STATISTICAL ANALYSIS

The IBM SPSS software offers advanced statistical analysis, text analysis, open source extensibility, integration with big data and seamless deployment into applications. SPSS tool used for calculating t-test values, accuracy values of collaborator code is iterated 10 times with prediction of accuracy after every iteration which is analyzed and the mean accuracy is hence derived, Independent samples are used to determine significance values between two groups.

RESULT

The main goal of this study was to compare the performance of Gradient Boosting Regression with a modified linear regression approach in order to increase the accuracy of bus journey time prediction. After ten rounds, the adjusted linear regression demonstrated an impressive 93.77% mean accuracy with a 1.23327 standard deviation. The Gradient Boosting Regression model, on the other hand, displayed an accuracy of 84.97%; with a 3.80191 standard deviation. In terms of accuracy, the modified linear regression outperformed the Gradient Boosting Regression. A bar graph diagram is used to show the accuracy comparison visually, together with the independent samples and t-test results for each model.

DISCUSSION

In this research focused on enhancing the accuracy in predicting bus travel time, the modified linear regression algorithm demonstrated superior performance with an accuracy value of 93.77%, surpassing the accuracy obtained with Gradient Boosting Regression, which recorded an accuracy value of 84.97%. This improvement in accuracy suggests the effectiveness of the modified linear regression model in predicting bus travel time compared to Gradient Boosting Regression.

A previous study conducted in 2021 aimed to predict bus travel time using various machine learning algorithms, revealing accuracy values ranging from a minimum of 60% to a maximum of 85%. In this current work, the accuracy achieved using the modified linear regression algorithm has notably surpassed previous results, indicating advancements in predictive modeling for bus travel time.

However, challenges persist in this domain. The reliability of predictions heavily relies on the quality and availability of data stored in the dataset. The dynamic nature of bus travel, influenced by factors such as traffic conditions and road closures, poses challenges for models to adapt quickly. Moreover, the sensitivity of algorithms to hyperparameter settings introduces complexity, necessitating careful tuning for optimal performance. Ethical concerns regarding privacy and consent must also be addressed when implementing these models for public transportation systems.

Looking ahead, the future scope of this project is promising and can significantly contribute to improving public transportation systems. Integrating real-time data from various sources, such as traffic cameras and sensors, could further enhance the accuracy and responsiveness of predictive models. Collaborations with transportation experts and practitioners can add practical relevance to the project, ensuring that the models align with real-world challenges. Developing user-friendly interfaces for transportation authorities and professionals would facilitate the practical application of these predictive models, ultimately leading to more efficient and reliable bus travel time predictions.

CONCLUSION

A modified linear regression model has been constructed to improve the accuracy of bus trip time prediction, and its performance has been compared with the Gradient Boosting Regression technique. The accuracy of the updated linear regression model was 93.77%, which was better than the Gradient Boosting Regression algorithm's 84.97% accuracy. These results imply that, when compared to the Gradient Boosting Regression technique, the modified linear regression model performs better in predicting bus trip times. Such improvements in precision are essential for transportation system optimization, guaranteeing the development of effective and

dependable bus schedules.

DECLARATIONS

Conflicts of Interest:

No conflict of interests in this manuscript.

Authors Contribution:

Author Guna Suriya C was involved in data collection, data analysis, and manuscript writing. Author Dr.S.Christy was involved in conceptualization, data validation and critical review of manuscript.

Acknowledgements:

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:

We are grateful to the following groups for their financial assistance in helping us finish the study.

1. Infysec Solution, Chennai.
2. Saveetha School of Engineering.
3. Saveetha University.
4. Saveetha Institute of Medical and Technical Sciences.

REFERENCE

Cristóbal, Teresa, Gabino Padrón, Alexis Quesada-Arencibia, Francisco Alayón, Gabriel de Blasio, and Carmelo R. García. 2019. “Bus Travel Time Prediction Model Based on Profile Similarity.” Sensors 19 (13): 2869.

Anil Kumar, B., R. Jairam, Shriniwas S. Arkatkar, and Lelitha Vanajakshi. 2019. “Real Time Bus Travel Time Prediction Using K-NN Classifier.” Transportation Letters, July. https://doi.org/10.1080/19427867.2017.1366120.

Chen, Chao, Hui Wang, Fang Yuan, Huizhong Jia, and Baozhen Yao. 2019. “Bus Travel Time Prediction Based on Deep Belief Network with Back-Propagation.” Neural Computing & Applications 32 (14): 10435–49.

Yuan, Yuan, Chunfu Shao, Zhichao Cao, Zhaocheng He, Changsheng Zhu, Yimin Wang, and Vlon Jang. 2020. “Bus Dynamic Travel Time Prediction: Using a Deep Feature Extraction Framework Based on RNN and DNN.” Electronics 9 (11): 1876.

Agafonov, Anton, and Alexander Yumaganov. 2019. “Bus Arrival Time Prediction with LSTM Neural Network.” Advances in Neural Networks – ISNN 2019, 11–18.

Mendes-Moreira, João, and Mitra Baratchi. 2020. “Reconciling Predictions in the Regression Setting: An Application to Bus Travel Time Prediction.” Advances in Intelligent Data Analysis XVIII, 313–25.

Luan Tran University of Southern California, Min Y. Mun Samsung Electronics, South Korea, Matthew Lim University of Southern California, Jonah Yamato University of Southern California, Nathan Huh University of Southern California, and Cyrus Shahabi University of Southern California. 2020. “DeepTRANS.” Proceedings of the VLDB Endowment International Conference on Very Large Data Bases, August. https://doi.org/10.14778/3415478.3415518.

Wu, Jianqing, Qiang Wu, Jun Shen, and Chen Cai. 2020. “Towards Attention-Based Convolutional Long Short-Term Memory for Travel Time Prediction of Bus Journeys.” Sensors 20 (12): 3354.

TABLES AND FIGURES

Table 1. A sample of N = 10, tabular data input was collected and analyzed. The resulting accuracy of the Linear Regression classifier model was measured to be 93.77%. On the other hand, the Gradient Boosting Regression algorithm showed an accuracy of 84.97% within the same dataset.

S.No	LR Accuracy (%)	XGB Accuracy (%)
1	93.37	81.17
2	94.57	89.41
3	94.06	86.63
4	94.75	81.44
5	92.71	89.41
6	93.68	81.17
7	96.22	86.63
8	93.06	81.44
9	91.73	82.45
10	93.56	89.97

Table 2. According to this table, Linear Regression classifier (LR) outperforms Gradient Boosting Regression (XGB) in predicting the Bus Travel Time values across two groups. There were 10 samples in each group. The mean "Accuracy Rate" for the LR group was 93.77%, with a standard deviation of 1.23327 and a standard error of the mean of .38999. On the other hand, XGB achieved a mean accuracy rate of 84.97%, with a lower standard deviation of 3.80191 and a smaller standard error of the mean at 1.20227. These results indicate that the Linear Regression classifier (LR) had better predictive accuracy and less variability than the Gradient Boosting Regression algorithm in this particular personality prediction task.

Groups	Accuracy	Mean (%)	Standard Deviation	Standard Error Mean
Linear Regression	10	93.77	1.23327	0.38999
Extreme Gradient Boosting	10	84.97	3.80191	1.20227

Table 3. Independent sample test for significance and standard error determination. P-value is less than 0.05 considered to be statistically significant and 95% confidence intervals were calculated.

		F	Sig.	t	df	Mean Difference	Standard Error Difference	95% Confidence of Interval Difference	
								Lower	Upper
Accuracy	**Equal Variances Assumed**	33.19	0.000	6.962	18	8.79900	1.26394	6.14356	11.45444
	Equal Variances Not Assumed			6.962	10.8	8.79900	1.26394	6.01313	11.58487

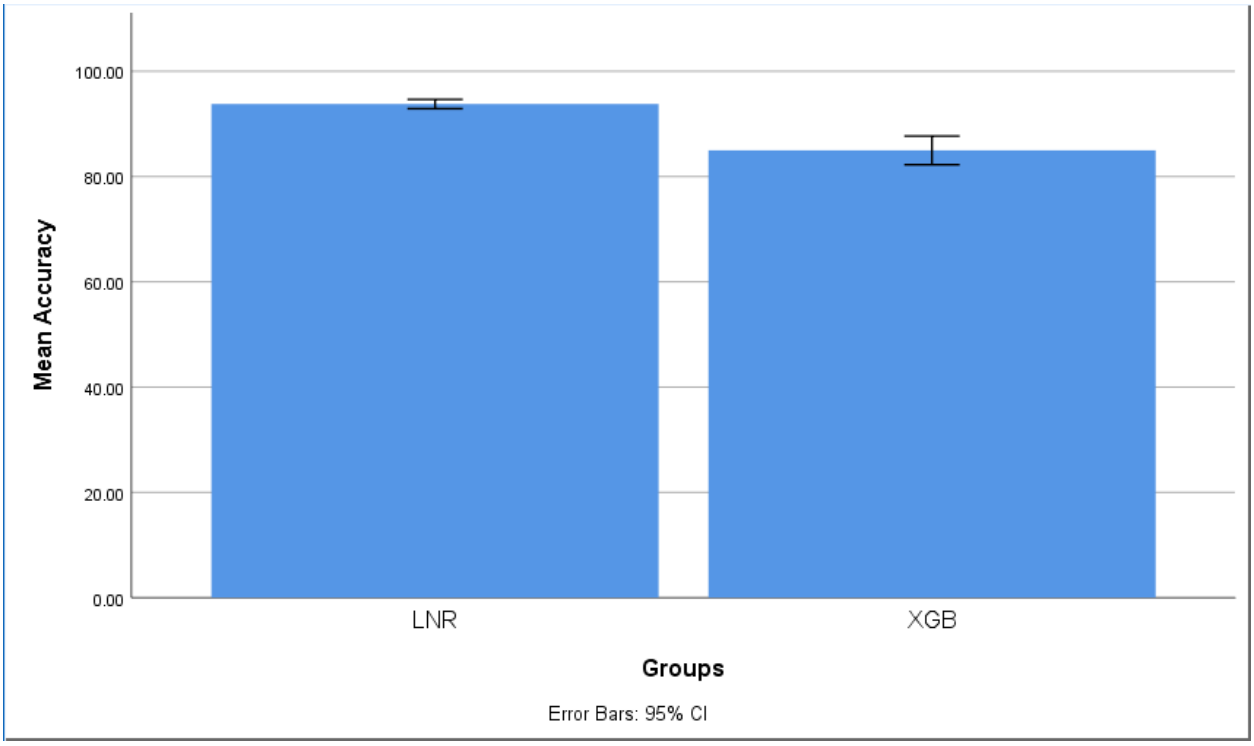

Figure 1

Bar Graph represents the comparison of the mean gain for two algorithms obtained from SPSS software. All the obtained values are loaded into the SPSS software to obtain t-test, independent samples test, bar chart representation. X-axis represents Groups and Y-axis represents Mean Accuracy. The accuracy mean is considered as 95%CI.