TITLE:

Prediction on airline delays using support vector machine in comparison with linear regression for better accuracy

T GANESH[1], Dr S Gomathi[2]

T Ganesh[1]
Research Scholar,
Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India, Pin code: 602 105
thatimakulaganesh0352.sse@sasveetha.com

Dr S Gomathi[2]
Corresponding Author,
Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences,
Saveetha University,Chennai,Tamil Nadu, India, Pincode: 602105.
gomathis.sse@saveetha.com

KEYWORDS: Support Vector Machine,Linear Regression, Machine Learning ,Research Flight delay prediction.

ABSTRACT

Aim: The aim of the study is to investigate and compare the effectiveness of Flight delay Predictions using NOVEL SUPPORT VECTOR MACHINE with entropy in comparison With entropy in comparison with LINEAR REGRESSION for better Accuracy.
Materials and Methods: The present study aimed to investigate and compare the effectiveness of Flight Delay Predictions using a Novel Support Vector Machine (SVM) with entropy, in comparison with the traditional Linear Regression algorithm, to achieve better accuracy in classification. A comprehensive dataset of Flight delays datasets associated with past flight delays was obtained from Kaggle.With design, data collection,fFeature extraction. We evaluated their accuracy using metrics to ensure improvement. Sample size of 80 for each group of statistical parameters: G Power=0.80 for 10 iterations for each group. Two algorithms, SVM andLinear Regression , were implemented using Statistical Package for Social Sciences (SPSS)
Result: Based on obtained results SVM has significantly better accuracy (92.68%) compared to LR accuracy (58.21%) Statistically significant difference between SVM and LR algorithm was found to be p-value of $p=0.000(p<0.005)$.
conclusion: We have used the following algorithms namely Novel Support Vector Machine (SVM),Linear Regression(LR)algorithms to predict the data. From the results it is proved that the proposed Novel Support Vector Machine (SVM) works better than other algorithms in terms of accuracy

KEYWORDS: Support Vector Machine,Linear Regression, Machine Learning ,Research Flight delay prediction.

INTRODUCTION

In the dynamic landscape of air travel, the prediction and mitigation of flight delays play a crucial role in ensuring operational efficiency and passenger satisfaction. With the increasing volume of air traffic and various factors contributing to delays (United States. Congress. House. Committee on Transportation and Infrastructure. Subcommittee on Aviation 1996), accurate prediction models are indispensable for airlines and airport authorities. Traditional regression techniques, such as linear regression, have been widely employed for delay prediction due to their simplicity and interpretability. However, in recent years, advanced machine learning algorithms, such as Support Vector Machine (SVM), have emerged as powerful tools for predictive analytics, offering potential improvements in accuracy and robustness(*National*

Airspace System Longterm Capacity Planning Needed despite Recent Reduction in Flight Delays, n.d.). This study aims to compare the performance of SVM with that of linear regression in predicting airline delays(CreateSpace Independent Publishing Platform and Office of the Investigator General 2018). By leveraging historical flight data encompassing diverse features such as weather conditions(Williams 2023), airport congestion, aircraft type, and airline schedules, we seek to evaluate the efficacy of these two approaches in forecasting flight delays. Linear regression, a parametric method, assumes a linear relationship between the input features and the target variable. While it provides straightforward interpretations of the coefficients, its performance may be limited when dealing with non-linear and complex data patterns inherent in airline delay prediction. On the other hand, SVM, a non-parametric algorithm, excels in capturing non-linear relationships by mapping input features into a higher-dimensional space and identifying the optimal hyperplane that maximizes the margin between different classes. This ability to handle non-linear data structures makes SVM particularly promising for modeling the intricate relationships underlying flight delays. Through a comparative analysis of SVM and linear regression, we aim to assess their respective strengths and weaknesses in terms of prediction accuracy, robustness to outliers, and generalization to unseen data. By identifying the most effective approach for airline delay prediction, this research seeks to contribute to the development of more reliable and efficient strategies for managing air transportation systems. In conclusion, this study sets out to investigate the potential of Support Vector Machine as an alternative to linear regression for predicting airline delays. By evaluating their performance on real-world flight data, we aim to provide valuable insights into the effectiveness of these methods and guide future efforts in improving delay prediction models for the aviation industry.

MATERIALS AND METHODS

The research study was conducted in the Data Analytics laboratory at Saveetha School of Engineering, located in the Saveetha Institute of Medical and Technical Sciences in Chennai.Two groups were selected for the Novel Support Vector Machine [SVM] and Linear Regression (LR), the process in predicting the Flight Delays, and sample size of 80 for each group of statistical parameters: G Power=0.80 for 10 iterations for each group. Two algorithms, SVM and LR, were implemented using Statistical Package for Social Sciences (SPSS). We have two independent variables, SVM and LR, for predicting the Flight Delays and their Efficiency.

Materials and Methods:

Data Collection: Historical flight data spanning a significant time period, including features such as departure and arrival times, airline, airport, weather conditions, aircraft type, and previous delay information, is collected from reliable sources such as airline databases, flight tracking services, and weather databases(Donohue, Shaver, and Edwards 2008). Data preprocessing techniques are applied to clean and prepare the dataset, including handling missing values, encoding categorical variables(Cappelletti 2020), and normalizing numerical features. Feature

Selection: Relevant features for delay prediction are identified based on domain knowledge and exploratory data analysis. Feature selection techniques such as correlation analysis, feature importance ranking, and domain expertise are employed to select the most informative features for modeling. Model Training and Evaluation: Support Vector Machine (SVM) and linear regression models are implemented using appropriate libraries or frameworks (e.g., scikit-learn in Python). The dataset is split into training and testing sets using techniques such as cross-validation to ensure unbiased evaluation of model performance. Both SVM and linear regression models are trained on the training data and evaluated using performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE)(Silva Cotta et al. 2023), and R-squared (R^2) on the testing set. Hyperparameter Tuning: For the SVM model, hyperparameters such as the choice of kernel function, regularization parameter (C), and kernel-specific parameters (e.g., gamma for RBF kernel) are tuned using techniques like grid search or random search to optimize performance. Similarly, for linear regression, regularization techniques such as Ridge or Lasso regression may be applied to prevent overfitting, and their respective hyperparameters are tuned accordingly. Model Comparison: The performance of SVM and linear regression models is compared based on various evaluation metrics obtained from the testing set. Statistical tests such as paired t-tests or Wilcoxon signed-rank tests are conducted to determine if the performance difference between the two models is statistically significant. Robustness Analysis: Sensitivity analysis is performed to assess the robustness of SVM and linear regression models to changes in input data and model assumptions. Outlier detection techniques and robust regression methods may be employed to identify and mitigate the impact of outliers on model performance. Interpretability: The interpretability of both SVM and linear regression models is assessed, considering factors such as the magnitude and significance of coefficients in linear regression and the decision boundaries and support vectors in SVM. Visualization techniques such as feature importance plots and decision boundaries are utilized to facilitate the interpretation of model predictions. Cross-Validation: To ensure the reliability of the findings(Yang et al. 2023), cross-validation techniques such as k-fold cross-validation or leave-one-out cross-validation may be employed to assess the stability of model performance estimates across different subsets of the data. By following these materials and methods, we aim to conduct a comprehensive comparative analysis of Support Vector Machine and linear regression for predicting airline delays, with the goal of identifying the most effective approach for enhancing prediction accuracy in air transportation systems.

Support Vector Machine (SVM)

Support Vector Machines (SVM) represent a powerful class of supervised learning algorithms primarily used for classification and regression tasks. SVM operates by finding the optimal hyperplane that separates data points into different classes within a high-dimensional space. This hyperplane is determined by maximizing the margin, which is the distance between the

hyperplane and the nearest data point of either class. The SVM model identifies support vectors, which are the data points that lie closest to the decision boundary and play a crucial role in determining the optimal hyperplane. The algorithm aims to ensure that the margin is maximized while minimizing the classification error(Dai 2024), making SVM well-suited for scenarios where complex decision boundaries need to be discerned. The mathematical formulation involves solving a convex optimization problem, and various kernel functions can be employed to handle non-linear relationships between features. SVM has proven effective in diverse fields,flight predictions
In the real world.

Procedure for Support Vector Machine(SVM)

Step 1:

Collect and Prepare Data Assume X_train is the feature matrix for training data, y_train is the corresponding labels.

Step 2:

Choose the SVM Kernel Assume a 'linear' kernel for simplicity.

Step 3:

Define SVM Model 4/16 model = SVM(kernel='linear', C=1.0)

Step 4:

Train the SVM Model model.train(X_train, y_train)

Step 5:

Make Predictions Assume X_test is the feature matrix for test data. predictions = model.predict(X_test)

Step 6:

Evaluate Model Performance (Optional) You can use metrics like accuracy, precision, recall, and F1-score. accuracy = calculate_accuracy(predictions, y_test)

Step 7:

Visualize Results.

LINEAR REGRESSION(LR)

In the field of dermatology, Linear Regression has been a valuable tool for the precise delay prediction of flights through the analysis of flight delays. LR relies on the principle that similar prediction patterns are likely to share the same class. By assessing the proximity of cases in the feature space, LR effectively identifies similarities and aids in the classification of Flight delay prediction based on their likeness to known instances. However, the success of LR hinges on optimal parameter selection,21 particularly the choice of 'L' – the number of Linear considered – and the careful inclusion of relevant features. While LR has demonstrated efficiency in Aviation applications, including flight delays , its performance may vary based on dataset complexities and the unique characteristics of Flight Delays Predictions

Procedure for LR:-

Step 1:
Collect and Prepare Data
Assume X_train is the feature matrix for training data, y_train is the corresponding labels.
Step 2:
Choose LR Parameters Assume 'L' is chosen, either based on cross-validation or prior knowledge.
Step 3:
Define LR Model 5/16 model =LR(L=5) # Assume L=5 for simplicity
Step 4:
Train the LR Model (Note: LRis a lazy learner and doesn't explicitly train) In LR, training involves storing the training data.
Step 5:
Make Predictions Assume X_test is the feature matrix for test data. predictions = model.predict(X_test, X_train, y_train)
Step 6:
Evaluate Model Performance (Optional) You can use metrics like accuracy, precision, recall, and F1-score. accuracy = calculate_accuracy(predictions, y_test)
Step 7:
Visualize Results (Not applicable for LR) LR doesn't have a decision boundary like SVM; visualization is more challenging.
Step 8:
End

STATISTICAL ANALYSIS

The analysis was prepared through IBM SPSS version 21. Independent variables and impactful values are considered for both proposed and as well as existing algorithms, iterations were done with a maximum of 80 samples and for each iteration the recorded accuracy was noted for necessary analysis. The Dependent Variables are indicated as previous data(Davis and Mongeau 2023) ,airport data,weather conditions and Independent Variables are flight date,origin city,destination city With the corresponding value that is obtained from the iterations, the Independent sample T-test was performed.

RESULTS

Table 1 Shows the various iterations of the Support Vector Machine (SVM) and Linear Regression(LR) efficiency values are compared. Table 2 Shows the Group Statistics Results: An Novel Support Vector Machine (SVM) and Linear Regression for Testing Independent Samples Statistically Among SVM and LR Methods SVM has a mean accuracy of 92.6850 and a LR of 58.3120. SVM has a standard deviation of 3.79171 and a LR of 8.07374. The SVM standard

error mean (1.19905) and LR of (2.55314) were compared using the T-test. In Table 3, The 2-significant value smaller than 0.000 ($p<0.05$) impacted that our hypothesis holds good for further consideration. Figure 1 shows bar graph comparison on mean accuracy of Support Vector Machine (SVM) and Linear Regression(LR). In x-axis SVM and LR methods Error Bars: +/-2 SD and 95% CI of Error Bars are shown, In y-axis mean accuracy is shown.

DISCUSSION

The main aim of the project is finding accurate predictions of flight delays in difficult conditions. For that I had iterated the Aviation manifestations of Flight Delay dataset into 1-2000,1-4000,1-6000 ... .1-20000 samples (10 iterations)and found the accurate accuracy values for each and every sample. And we have noted that accuracy values and tests their independent sample T-Test in SPSS and we obtained results SVM has significantly better accuracy (92.68%) compared to LR accuracy (58.31%) Statically significant difference between SVM and LR algorithm was found to be p-value of $p=0.024$ ($p<0.05$). For each and every phase we tried to improve the accuracy in an efficient manner(Wimmer et al. 2023).Here Support Vector Machine (SVM) gives better accuracy while comparing with Linear Regression (LR). In recent years(Fleming 2009), the intersection of machine learning and Aviation industry has shown promise in advancing the accuracy ofFlight Delays(United States. Congress. House. Committee on Transportation and Infrastructure. Subcommittee on Aviation 2007), particularly through the analysis of Flight Delays. A notable approach involves the utilization of Support Vector Machines (SVM) with entropy, as proposed by)(Masto et al. 2024), SVMs, known for their capacity to find optimal hyperplanes for data separation, are enriched with entropy to capture the nuanced information embedded in fight delays data. The incorporation of entropy enhances the model's ability to discern intricate patterns associated with flight delays , making it a robust tool for accurate Delay Predictions classification. In comparison(Zhong et al. 2024), the conventional Linear Regression (LR) algorithm, widely used in Aviation studies, has limitations in handling the complex relationships present in the Aviation industry of Flight Delays(Zikmund, Horpatzka, and Macik 2024) . Linear Regression reliance on proximity in feature space and sensitivity to irrelevant features may hinder its performance in capturing subtle patterns crucial for accurate diagnosis. The contrast between SVM with entropy and LR highlights the potential superiority of the former . The non-linear capabilities of SVMs(Waterman et al. 2024), combined with the information-rich entropy, provide a more comprehensive and effective framework for analyzingFlight Delaying predictions , ultimately leading to improved accuracy inAviations . This innovative approach using SVM with entropy not only contributes to the advancement of machine learning applications in Airport staff but also holds significant implications for Manual practice. The heightened accuracy achieved through this methodology(Liu et al. 2024), as compared to traditional techniques like LR, has the potential to revolutionize early detection and

Prediction strategies(Mawad et al. 2023), thereby improving flight outcomes in the context of Delays and other Aviation delay conditions.

CONCLUSION

Our study has demonstrated a substantial and statistically significant difference in accuracy between Novel Support Vector Machine (SVM) and Linear Regression (LR) algorithms forFlight delay predictions in Aviation Industries. The SVM model achieved an impressive accuracy of 92.68%, surpassing theLR accuracy of 58.31%. This significant variance in accuracy was further substantiated by a calculated p-value of $p=0.000(p<0.05)$, confirming that the superiority of SVM in Prediction of flight delays is not merely a chance occurrence. These findings underscore the potential of SVM as a more reliable and precise tool for Aviation craft Delays prediction, emphasizing the importance of incorporating advanced machine learning techniques to enhance the accuracy and effectiveness of Flight Delay Prediction models. This study contributes to the Rapid growth of traveling research supporting the adoption of SVM in Aviation manifestations, with the goal of improving our ability to provide more accurate and timely Flights Delaying to their Destinations .

DECLARATIONS

Conflict of interests

No conflict of interest in the manuscript.

AUTHORS CONTRIBUTIONS

TG was responsible for collecting data,conducting data analysis,and writing the manuscript.
SG contributed to the conceptualization ,validated the data ,and performed a critical review of the manuscript.

Acknowledgements

The authors extend their thanks to the Saveetha School of Engineering and the Saveetha Institute of Medical and Technical Sciences (previously known as Saveetha University) for their support in providing the infrastructure needed to complete this work successfully

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.
1)Infoziant IT Solutions Pvt. Ltd., Chennai.
2)Saveetha University.
3)Saveetha School of Engineering.
4)Saveetha Institute of Medical and Technical Sciences.

REFERENCES

Cappelletti, John C. 2020. *Flight Delayed Again: Delays Occur When Airlines Confuse Their "Ambitions" and "Abilities."*

CreateSpace Independent Publishing Platform, and Office of the Investigator General. 2018. *Actions Needed to Minimize Long, On-Board Flight Delays* . Createspace Independent Publishing Platform.

Dai, Min. 2024. "A Hybrid Machine Learning-Based Model for Predicting Flight Delay through Aviation Big Data." *Scientific Reports* 14 (1): 4603.

Davis, Brock A., and Jean-Michel Mongeau. 2023. "The Influence of Saccades on Yaw Gaze Stabilization in Fly Flight." *PLoS Computational Biology* 19 (12): e1011746.

Donohue, George L., Russell D. Shaver, and Eric Edwards. 2008. *Terminal Chaos: Why US Air Travel Is Broken and How to Fix It*. AIAA (American Institute of Aeronautics & Astronautics).

Fleming, Susan. 2009. *Commercial Aviation: Impact of Airline Crew Scheduling on Delays and Cancellations of Commercial Flights*. DIANE Publishing.

Liu, Jian, Qi Huang, Yuhang Han, Shiyun Chen, Nan Pan, and Renxin Xiao. 2024. "Bi-Level Optimization for De-Icing Position Allocation and Unmanned De-Icing Vehicle Fleet Routing Problem." *Biomimetics* 9 (1). https://doi.org/10.3390/biomimetics9010026.

Masto, Nicholas M., Abigail G. Blake-Bradshaw, Cory J. Highway, Allison C. Keever, Jamie C. Feddersen, Heath M. Hagy, and Bradley S. Cohen. 2024. "Human Access Constrains Optimal Foraging and Habitat Availability in an Avian Generalist." *Ecological Applications: A Publication of the Ecological Society of America*, February, e2952.

Mawad, Tala N., Rakan A. Alfaifi, Othman M. Almazyed, Rand A. Alhumaidi, and Abdulaziz M. Alsubaie. 2023. "Fungemia Due to Saprochaete Capitata in a Non-Neutropenic Critically Ill Patient." *Cureus* 15 (12): e51147.

National Airspace System Longterm Capacity Planning Needed despite Recent Reduction in Flight Delays. n.d. DIANE Publishing.

Silva Cotta, Joao Leonardo, Daniel Agar, Ivan R. Bertaska, John P. Inness, and Hector Gutierrez. 2023. "Latency Reduction and Packet Synchronization in Low-Resource Devices Connected by DDS Networks in Autonomous UAVs." *Sensors* 23 (22). https://doi.org/10.3390/s23229269.

United States. Congress. House. Committee on Transportation and Infrastructure. Subcommittee on Aviation. 1996. *Reasons For, and Reporting Of, Airline Flight Delays: Hearing Before the Subcommittee on Aviation of the Committee on Transportation and Infrastructure, House of Representatives, One Hundred Fourth Congress, First Session, July 27, 1995*.

———. 2007. *Airline Delays and Consumer Service: Hearing Before the Subcommittee on Aviation of the Committee on Transportation and Infrastructure, House of Representatives, One Hundred Tenth Congress, First Session, September 26, 2007*.

Waterman, Jamie Mitchel, Tristan Michael Cofer, Lei Wang, Gaetan Glauser, and Matthias Erb. 2024. "High-Resolution Kinetics of Herbivore-Induced Plant Volatile Transfer Reveal Clocked Response Patterns in Neighboring Plants." *eLife* 12 (February).

https://doi.org/10.7554/eLife.89855.
Williams, Denise. 2023. *Love and Other Flight Delays*. Penguin.
Wimmer, Michael, Nicole Weidinger, Eduardo Veas, and Gernot R. Muller-Putz. 2023. “Neural and Pupillometric Correlates of Error Perception in an Immersive VR Flight Simulation.” *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference* 2023 (July): 1–4.
Yang, Fan, Pan Wang, Min Zheng, Xiao-Yu Hou, Li-Lin Zhou, Yong Wang, Sheng-Yun Si, et al. 2023. “Physiological and Behavioral Basis of Diamondback Moth Plutella Xylostella Migration and Its Association with Heat Stress.” *Pest Management Science*, November. https://doi.org/10.1002/ps.7904.
Zhong, Mian, Shichen Li, Yao Zou, Hongyun Fan, Yong Jiang, Chao Qiu, Jinling Luo, and Liang Yang. 2024. “Hydrophobic Surface Array Structure Based on Laser-Induced Graphene for Deicing and Anti-Icing Applications.” *Micromachines* 15 (2). https://doi.org/10.3390/mi15020285.
Zikmund, Pavel, Michaela Horpatzka, and Miroslav Macik. 2024. “Learning Effect in Joystick Tactile Guidance.” *IEEE Transactions on Haptics* PP (February). https://doi.org/10.1109/TOH.2024.3368663.

TABLES AND FIGURES

Table 1. The various iterations of the Support Vector Machine (SVM) and Linear Regression(LR) efficiency values are compared.

SVM	Linear Regression(LR)
83.25%	70.25%
90.62%	52.75%
90.92%	66.00%
93.12%	46.31%
92.85%	53.54%
94.96%	54.29%
94.79%	69.89%
95.03%	60.03%
95.78%	56.22%
95.53%	52.85%

Table 2. Group Statistics

Results: Support Vector Machines (SVM) and Linear Regression for Testing Independent Samples Statistically Among SVM and LR Algorithms SVM has a mean accuracy of 92.6850 and a LR of 58.2130. SVM has a standard deviation of 3.79171 and a LR of 8.07374. The SVM standard error mean (1.19905) and LR standard error mean (2.55314) were compared using the T-test.

	Groups	**N**	**Mean**	**Std.Deviation**	**Std.Error Mean**
Accuracy	SVM	10	92.6850	3.79171	1.19905
	LR	10	58.2130	8.07374	2.55314

Group Statistics

Table 3. Independent Sample T-Test is applied for the sample collections with a confidence interval as 95%. After applying the SPSS calculation it was found that the least square Linear Regression (LR) has a statistical significance value of 0.000(P<0.05) that shows they are Statistically significant.

		Levene's test for equality of Variances		**T-test for equality means with 95% confidence interval**						
		F	Sig	t	df	sig(2-tailed)	Mean Difference	std .Error Difference	Lower	Upper
Accuracy	**Equal Variances Assumed**	7.179	.015	12.221	18	.000	34.47200	2.82068	28.54597	40.39803
	Equal Variances not assumed			12.221	12.786	.000	34.47200	2.82068	28.36790	40.57610

Graph:-

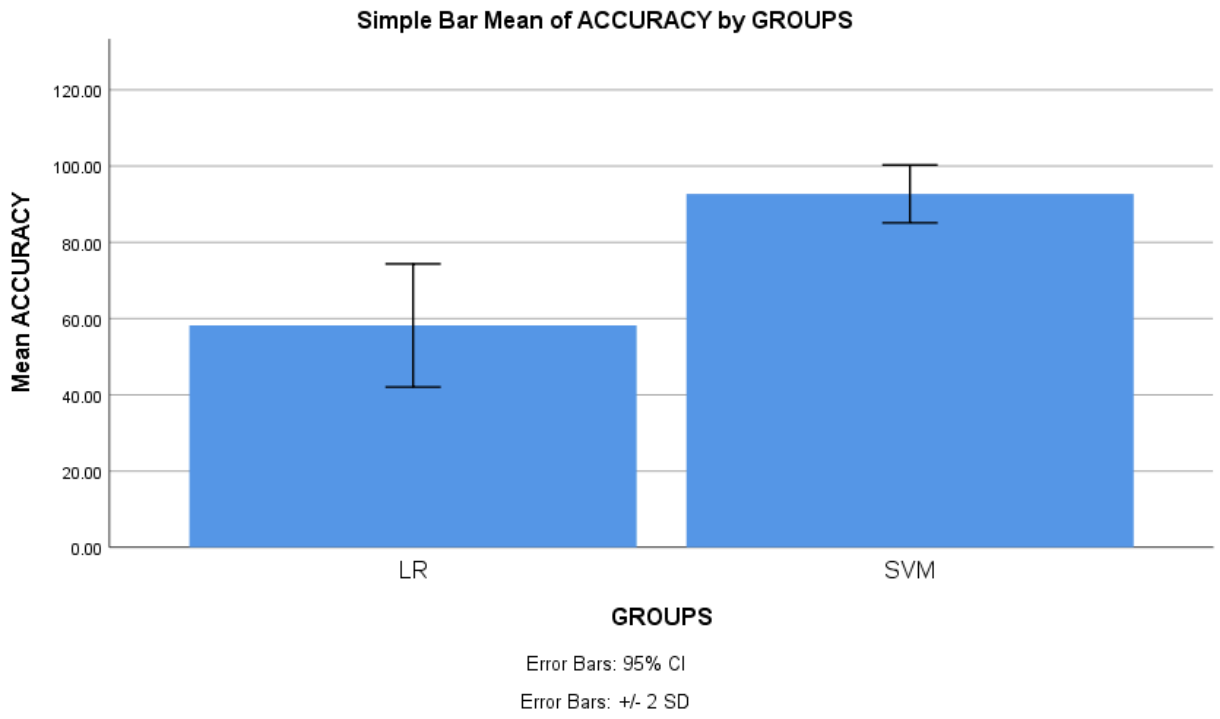

Fig. 1

. The novel Support Vector Machine has a mean accuracy of 92.62%, where Linear Regression(LR) has a mean of 58.21% in which the novel Support Vector Machine has better accuracy than Linear Regression(LR). The SVM and LR Accuracy rates are shown along with the X-axis: novel Support Vector Machine and Linear Regression Mean keyword identification Y-axis: Mean Accuracy, +/-2 SD, with a 95% Confidence Interval.