All Sciences Proceedings
http://as-proceeding.com/

2nd International Conference on Contemporary Academic Research

November 4-5, 2023 : Konya, Turkey

© 2023 Published by All Sciences Proceedings

ICCAR

https://as-proceeding.com/index.php/iccar

Boosting Predictive Power: Random Forest and Gradient Boosted Trees in Ensemble Learning

Bashar Alhajahmad*, Musa Ataş [2]

[1] *Computer Engineering/ Faculty of Engineering and Architecture, Siirt University, Turkey*
[2] *Computer Engineering/ Faculty of Engineering and Architecture, Siirt University, Turkey*

(bashar.aptech@gmail.com)

Abstract –Ensemble learning is a powerful concept in the realm of supervised machine learning, emphasizing the combination of multiple base learners or "inducers" to enhance predictive performance. This study explores the effectiveness of two ensemble algorithms, Random Forest and Gradient Boosted Trees, and their influence on predictive outcomes. The significance of ensemble methods lies in their ability to mitigate common challenges in machine learning. First, they address the issue of overfitting, which occurs when a model fits training data perfectly but fails on unseen data. Ensemble methods achieve this by averaging diverse hypotheses, reducing the risk of selecting an incorrect one and improving overall predictive performance. Second, ensemble methods provide computational advantages by avoiding local optima. Third, ensemble methods enhance representation by expanding the search space to find the best hypothesis. This extended representation facilitates more accurate modeling of complex relationships within the data. This study leverages two distinct datasets: one is hte Mushrooms and another from is the CO2 Emission by Vehicles dataset. The latter dataset, containing information on CO2 emissions from vehicles, is used for regression tasks, applying the same algorithms. The results of this study demonstrate outstanding performance from both Random Forest and Gradient Boosted Trees. In the classification task, both algorithms achieved perfect accuracy, while in the regression task, they showed remarkable explanatory power, with R-squared values of 1 for Random Forest algorithm and 0.995 for Gradient Boosted Trees. These findings emphasize the potential of ensemble learning in improving predictive accuracy and model performance.

Keywords – Random Forest, Gradient Boosted Trees, Knime, Ensemble learning, Machine Learning

I. INTRODUCTION

Ensemble learning is a versatile concept in supervised machine learning that involves the combination of multiple "inducers" or base learners to make decisions. An inducer is essentially an algorithm that takes a set of labeled examples and creates a model, such as a classifier or regressor, capable of making predictions for new, unlabeled data points. The core idea behind ensemble learning is based on the belief that by merging multiple models, the individual errors or shortcomings of a single inducer can be compensated for by the others. This collective approach tends to result in an overall prediction performance that surpasses that of any single inducer working in isolation [1].

Ensemble methods offer several compelling advantages, as highlighted by Dietterich [2] and Polikar [3]:

a) Overfitting Mitigation: When there is a limited amount of data available, a learning algorithm may overfit by fitting numerous hypotheses that work well with the training data but perform poorly on new data. Ensemble methods address this by averaging across diverse hypotheses, reducing the risk of selecting an incorrect hypothesis and improving overall predictive performance.

b) Computational Advantages: Single learners that use local search techniques may get stuck in local optima. Ensemble methods, by combining multiple learners, reduce the risk of converging to a local minimum. This broader exploration of the solution space leads to more robust and improved outcomes.

c) Enhanced Representation: The best hypothesis may extend beyond the capabilities of any individual model. By combining diverse models, ensemble methods expand the search space, resulting in a better fit within the data domain. This extended representation enables more accurate and comprehensive modeling of complex relationships in the data.

In this study, we investigated the effectiveness of two powerful ensemble algorithms: Random Forest and Gradient Boosted Trees, with a focus on their impact on predictive outcomes. Some research endeavors centered on enhancing the customer experience by simplifying the product customization process. This simplification involved a shift from a complex system that required inputting multiple parameters to create personalized products to a more user-friendly single-entry system. Behind the scenes, the platform harnessed robust machine learning (ML) algorithms, including Extreme Gradient Boosting and Random Forest ensemble learning, to efficiently map a single customer input to their desired customized product category [4].

In a separate research paper, the authors provided a detailed account of their approach to building machine learning models, with a particular emphasis on the use of gradient boosting and random forest models for predicting real GDP growth. This study was dedicated to examining the real GDP growth of Japan and involved generating forecasts spanning the years from 2001 to 2018 [5].

In another research article, the researchers delved into practical applications and conducted a comprehensive analysis of the performance of various machine learning techniques. These techniques included Random Forests, Gradient Boosted Trees, and diverse ensembles of these methods, all within the domain of statistical arbitrage. For model training, the authors utilized historical data encompassing lagged returns of all stocks in the S&P 500, paying careful attention to addressing survivor bias. This research spanned a timeframe from 1992 to 2015, during which they generated daily one-day-ahead trading signals based on the probability forecasts of individual stocks outperforming the broader market [6].

II. Materials and Method

Describe in detail the materials and methods used when conducting the study. The citations you make from different sources must be given and referenced in references.

In this research, the authors harnessed two distinct datasets to conduct our analysis. The first dataset, sourced from The Audubon Society Field Guide to North American Mushrooms, was made available by Jeff Schlimmer at the University of California, Irvine (UCI). It can be accessed at https://archive.ics.uci.edu/dataset/73/mushroom. This dataset encompasses 8,124 instances and 22 distinct features. Notably, it provides detailed descriptions of hypothetical samples corresponding to 23 different species of gilled mushrooms within the Agaricus and Lepiota Family. Each species in this dataset is categorized as either definitely edible, definitely poisonous, or of unknown edibility with a non-recommendation. Our application of this dataset primarily focused on classification tasks, where we employed both Random Forest and Gradient Boosted Trees algorithms.

The second dataset, drawn from Kaggle and titled "CO2 Emission by Vehicles," can be accessed at https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles?resource=download. This dataset contains information that sheds light on how CO2 emissions from vehicles vary in relation to various vehicle characteristics. It comprises a total of 7,385 rows and 12 columns. In our study, we engaged this dataset for regression tasks, utilizing Random Forest and Gradient Boosted Trees algorithms. It's worth noting that, to streamline our analysis, we excluded three columns from this dataset as they were deemed to have no significant impact on the target class. These columns are identified as "Make," "Model," and "Vehicle Class."

For the implementation of our analysis, we opted for the versatile Knime software, a free and opensource tool developed in Java. Knime is well-equipped to handle both structured and unstructured data and offers robust data visualization capabilities. As detailed in reference [7], this software played a pivotal role in our data analysis pipeline.

To provide a glimpse into the contents of these datasets, Table 1 offers a snapshot of the Mushrooms dataset's contents, presented using the Table View component. Similarly, Table 2 provides a snapshot of the CO2 Emission dataset's contents.

Table 1: Table View of the Mushroom dataset.

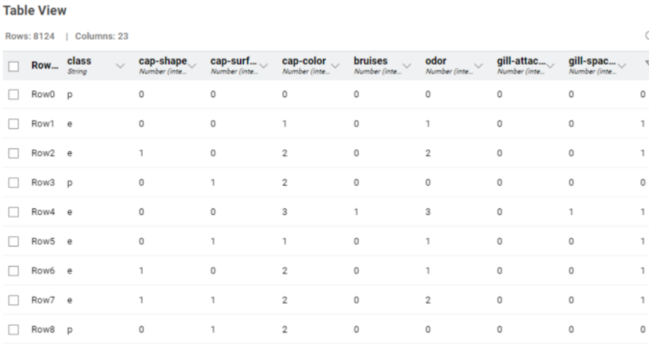


Table View

Rows: 8124 | Columns: 23

☐	Row...	class String	cap-shape Number (inte...	cap-surf... Number (inte...	cap-color Number (inte...	bruises Number (inte...	odor Number (inte...	gill-attac... Number (inte...	gill-spac... Number (inte...	
☐	Row0	p	0	0	0	0	0	0	0	0
☐	Row1	e	0	0	1	0	1	0	0	1
☐	Row2	e	1	0	2	0	2	0	0	1
☐	Row3	p	0	1	2	0	0	0	0	0
☐	Row4	e	0	0	3	1	3	0	1	1
☐	Row5	e	0	1	1	0	1	0	0	1
☐	Row6	e	1	0	2	0	1	0	0	1
☐	Row7	e	1	1	2	0	2	0	0	1
☐	Row8	p	0	1	2	0	0	0	0	0

Table 2: Table View of the CO2 Emission by Vehicles dataset.

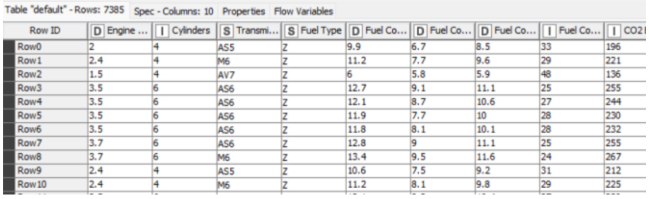


Table "default" - Rows: 7385 | Spec - Columns: 10 | Properties | Flow Variables

Row ID	D Engine ...	I Cylinders	S Transmi...	S Fuel Type	D Fuel Co...	D Fuel Co...	D Fuel Co...	I Fuel Co...	I CO2
Row0	2	4	AS5	Z	9.9	6.7	8.5	33	196
Row1	2.4	4	M6	Z	11.2	7.7	9.6	29	221
Row2	1.5	4	AV7	Z	6	5.8	5.9	48	136
Row3	3.5	6	AS6	Z	12.7	9.1	11.1	25	255
Row4	3.5	6	AS6	Z	12.1	8.7	10.6	27	244
Row5	3.5	6	AS6	Z	11.9	7.7	10	28	230
Row6	3.5	6	AS6	Z	11.8	8.1	10.1	28	232
Row7	3.7	6	AS6	Z	12.8	9	11.1	25	255
Row8	3.7	6	M6	Z	13.4	9.5	11.6	24	267
Row9	2.4	4	AS5	Z	10.6	7.5	9.2	31	212
Row10	2.4	4	M6	Z	11.2	8.1	9.8	29	225

In the scope of our study, we crafted a total of four distinct workflows using the Knime environment. The initial two workflows, visualized in Figure 1 and Figure 2, were devised to address the regression problem using the CO2 Emission dataset. Specifically, we applied both the Random Forest and Gradient Boosted Tree algorithms to this dataset in these workflows.

Conversely, the remaining two workflows, showcased in Figure 3 and Figure 4, were tailored to tackle the classification task using the Mushrooms dataset. These workflows were designed to employ the Random Forest and Gradient Boosted Tree algorithms in order to effectively address the classification challenges presented by the Mushrooms dataset.

To gauge the effectiveness of our classification models, we leverage the Scorer component. This tool empowers us to examine the confusion matrix and obtain essential statistical metrics, which, in turn, offer valuable insights into the accuracy and efficiency of our models. In a parallel manner, for assessing the performance of our regression models, we rely on the Numeric Scorer component. This instrumental feature enables us to access crucial statistics like the R-squared value, furnishing us with a comprehensive view of the model's performance in the realm of regression analysis.

To enhance our assessment of model performance, we integrated the ROC Curve component into our analysis. This valuable component produces ROC-AUC charts, which provide a visual means of gauging and understanding the performance of our models.

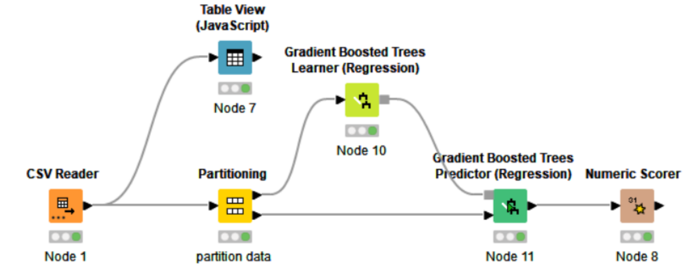


Fig. 1 Gradient Boosted Trees (Regression) workflow

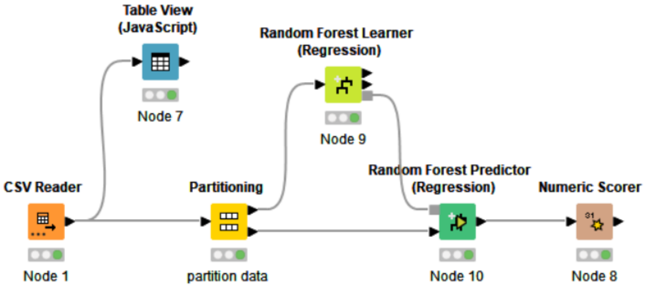


Fig. 2 Random Forest (Regression) workflow

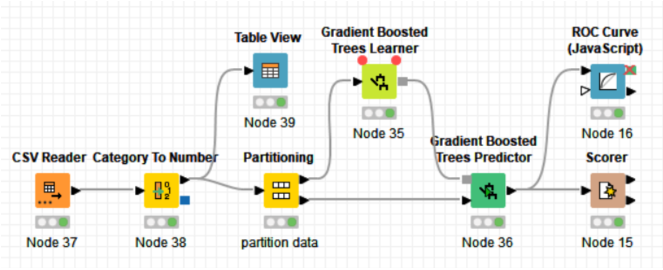


Fig. 3 Gradient Boosted Trees (Classification) workflow

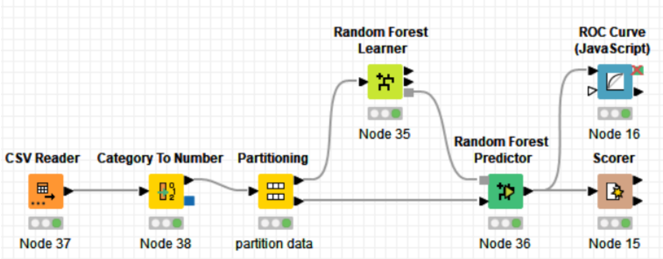


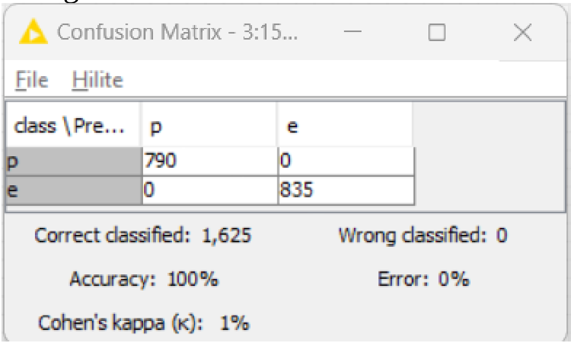
Fig. 4 Random Forest (Classification) workflow

III. RESULTS

Figures 5 and 6 provide visual representations of the confusion matrices, complete with accuracy scores, resulting from the application of the Random Forest and Gradient Boosted Trees algorithms to the mushroom dataset. The confusion matrix serves as a pivotal tool for calculating essential metrics such as Recall, Precision, Specificity, Accuracy, and, notably, for generating AUC-ROC curves, which offer an insightful view of model performance.

In contrast, Figures 9 and 10 present ROC Curve charts, offering a graphical depiction of the algorithm outcomes from both Random Forest and Gradient Boosted Trees. These charts provide a comprehensive visualization of the models' discrimination capabilities and performance in classification tasks.

Lastly, Figures 7 and 8 in the analysis illustrate the R-squared values following the application of the same algorithms to the Mushroom dataset, shedding light on the effectiveness of the models in explaining the variance observed in the data.



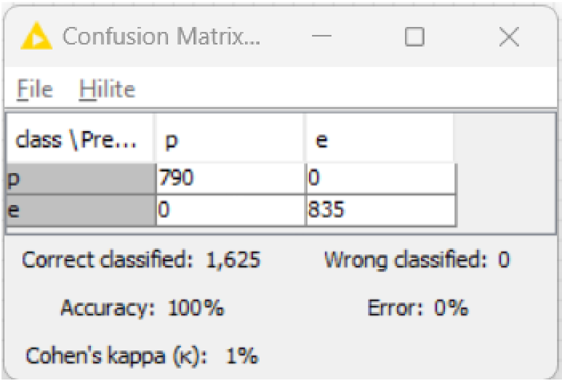
class \ Pre...	p	e
p	790	0
e	0	835

Correct classified: 1,625 Wrong classified: 0

Accuracy: 100% Error: 0%

Cohen's kappa (κ): 1%

Fig. 5 Confusion Matrix, Accuracy and Error rate for Random Forest Classification



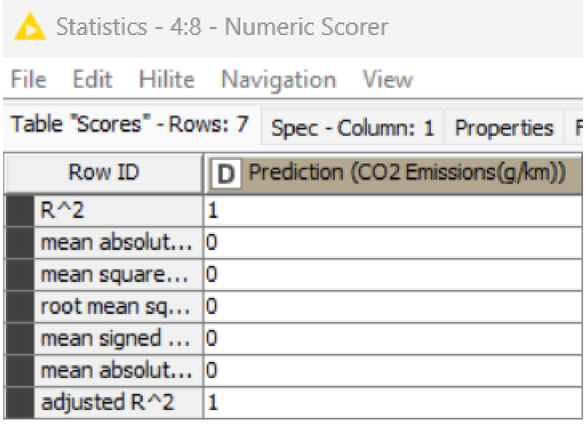
class \ Pre...	p	e
p	790	0
e	0	835

Correct classified: 1,625 Wrong classified: 0

Accuracy: 100% Error: 0%

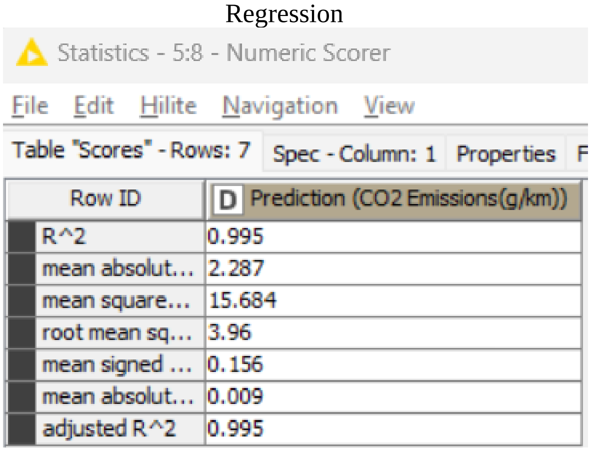
Cohen's kappa (κ): 1%

Fig. 6 Confusion Matrix, Accuracy and Error rate for Gradient Boosted Trees Classification


Statistics - 4:8 - Numeric Scorer
File Edit Hilite Navigation View
Table "Scores" - Rows: 7
Spec - Column: 1
Properties
F

Row ID	D Prediction (CO2 Emissions(g/km))
R^2	1
mean absolut...	0
mean square...	0
root mean sq...	0
mean signed ...	0
mean absolut...	0
adjusted R^2	1

Fig. 7 Numeric Scorer Statistics for Random Forest Regression


Statistics - 5:8 - Numeric Scorer
File Edit Hilite Navigation View
Table "Scores" - Rows: 7
Spec - Column: 1
Properties
F

Row ID	D Prediction (CO2 Emissions(g/km))
R^2	0.995
mean absolut...	2.287
mean square...	15.684
root mean sq...	3.96
mean signed ...	0.156
mean absolut...	0.009
adjusted R^2	0.995

Fig. 8 Numeric Scorer Statistics for Gradient Boosted Trees Regression

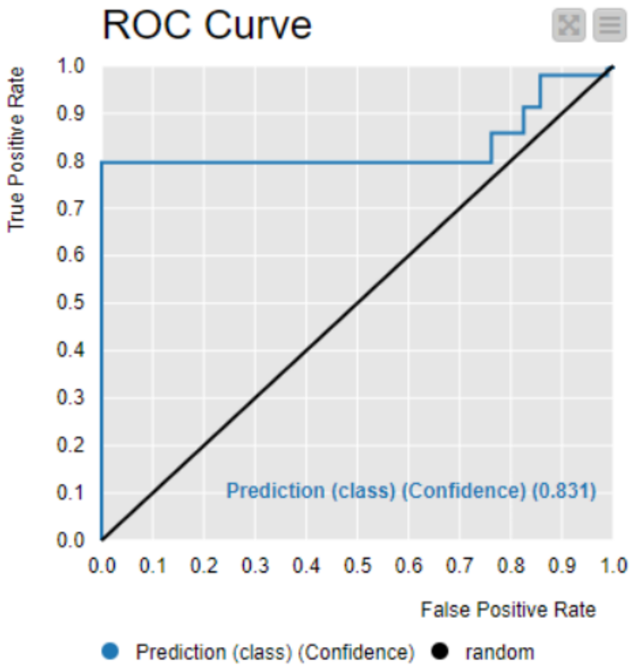


Fig. 9 ROC Curve chart for Gradient Boosted Trees Model

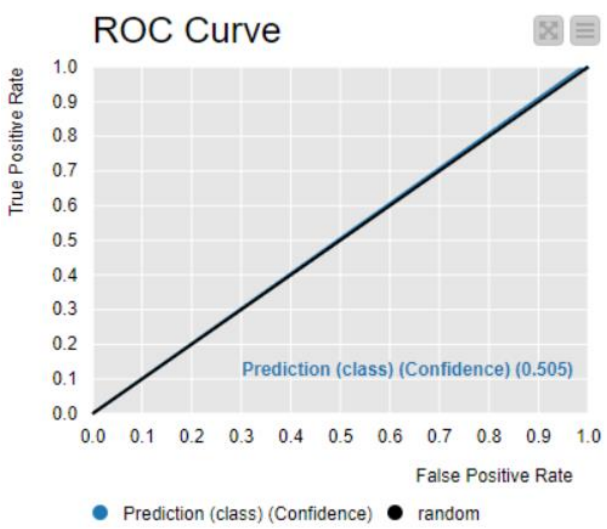


Fig. 10 ROC Curve chart for Random Forest Model

IV. DISCUSSION

Upon a careful examination of the confusion matrices depicted in Figures 5 and 6, it becomes strikingly apparent that both the Random Forest and Gradient Boosted Trees algorithms delivered outstanding performance in the classification task. Remarkably, both algorithms achieved a flawless 100% accuracy, resulting in a remarkable 0% error rate in our findings.

In contrast, when we turn our attention to Figures 7 and 8, we observe that the R-squared value equalled 1 when we applied the Random Forest algorithm, signifying a perfect fit of the model to the data. Moreover, when we employed the Gradient Boosted Trees algorithm, the R-squared value registered at an impressive 0.995, indicating an exceptionally strong explanatory power of the model in relation to the dataset.

The coefficient of determination, often denoted as R-squared, serves as a valuable metric that provides insights into the quality of a model's fit. In the context of regression analysis, R-squared offers a statistical assessment of how effectively the regression line approximates the actual data.

When R-squared equals 1, it signifies that the independent variable X fully accounts for all the variations observed in the dependent variable Y. This represents the optimal scenario we aim for, where the model perfectly captures the relationship between the variables.

Conversely, when R-squared equals 0, it indicates that none of the variations in the dependent variable Y can be attributed to the independent variable X. In this scenario, the model fails to explain or predict any of the variations observed in the data [8].

V. CONCLUSION

In conclusion, this study delved into the world of ensemble learning, specifically focusing on the application of two powerful algorithms, Random Forest and Gradient Boosted Trees. We investigated their impact on predictive outcomes, with a particular emphasis on classification and regression tasks.

The results of our analysis showcased exceptional performance from both Random Forest and Gradient Boosted Trees in the classification task, achieving a remarkable 100% accuracy and a 0% error rate. Moreover, the regression models using these algorithms demonstrated outstanding explanatory power, with R-squared values of 1 and 0.995, indicating a nearly perfect fit to the data.

In summary, the findings of this study emphasize the significance of ensemble learning and the efficacy of the algorithms examined in improving model performance and accuracy, which can be of great value in real-world applications across diverse industries and domains.

REFERENCES

[1] Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1249.

[2] Dietterich, T. G. (2002). Ensemble learning. In The handbook of brain theory and neural networks (Vol. 2, pp. 110–125). Cambridge, MA: MIT Press.

[3] Polikar, R. (2006). Ensemble based systems in decision making. IEEE Circuits and Systems Magazine, 6(3), 21–45.

[4] Kahiomba Kiangala, S., & Wang, Z. (2021). An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. Machine Learning with Applications, 4.

[5] Yoon, J. Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach. Comput Econ 57, 247–265 (2021).

[6] Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research, 259(2), 689-702.

[7] Kalpana Rangra and Dr. K. L. Bansal, "Comparative Study of Data Mining Tools", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 6, 2014.

[8] Coefficient of Determination, R-squared. [Online]. Available: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html.