iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** II **Month of publication:** February 2024

DOI: https://doi.org/10.22214/ijraset.2024.58316

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue II Feb 2024- Available at www.ijraset.com

Forecasting Agricultural Crop Profitability and Yield in India

Astha Tripathi[1], Mahesh G[2], Avneet Singh Charabra[3], Divyam Bahl[4]
Dept of CSE, JSS Academy of Technical Education, Noida, Uttar Pradesh, India

Abstract: ***"India's heavy reliance on agriculture underscores the importance of accurately estimating agricultural production, considering the interplay of organic, economic, and seasonal variables, particularly amid a growing population. Predicting crop yields is crucial for farmers' planning, covering storage and marketing strategies. However, this task is complex and requires foresight. Data mining techniques emerge as a potent tool, leveraging extensive datasets to extract invaluable insights. By employing methods like Random Forest, this research offers a swift yet comprehensive examination of crop yield forecasts for specific regions. Predictive analyses like these act as a crucial asset, enabling stakeholders to make well-informed decisions grounded in expected patterns and trends."***
Keywords: Crop Profitability, Prediction, Agriculture, Machine Learning

I. INTRODUCTION

The agricultural sector stands as a cornerstone in India's economic development and sustenance of food security, supporting the livelihoods of over 50% of its population. However, this crucial sector grapples with multifaceted challenges, including erratic crop yields, market price fluctuations, and restricted access to essential resources and technological advancements. Addressing these challenges demands the implementation of effective tools and strategies aimed at bolstering the efficiency and profitability of India's agriculture. One such initiative, the 'Forecasting Agricultural Crop Profitability and Yield in India,' aims to fill this gap by harnessing machine learning methodologies to predict crop yield and profitability. By meticulously gathering and analyzing data on various factors influencing crop productivity and profitability—ranging from weather conditions and soil quality to market dynamics—and employing machine learning algorithms for predictive insights, this project aims to empower farmers to make informed decisions regarding crop selection, optimal planting schedules, and resource distribution.

Moreover, the predictive model formulated through this endeavor holds promise not only for individual farmers but also for agricultural entities and governmental bodies. Its potential utility spans efficient resource planning, thereby bolstering food security and sustainability endeavors across the nation.

Key facets of this project involve a comprehensive collation and analysis of diverse datasets encompassing agricultural statistics, weather patterns, and market trends. The predictive model, rooted in machine learning, not only adeptly learns from new data but also extrapolates accurate predictions based on historical patterns. An intuitive user interface is envisaged, enabling farmers to seamlessly input data and receive informed predictions about crop yield and profitability.

Ultimately, this project aspires to furnish a valuable resource that empowers farmers and agricultural organizations, optimizing their operational resources and enhancing efficiency and profitability. It also endeavors to contribute tangibly to India's agricultural landscape by fortifying food security, promoting sustainability, and advocating for environmentally conscious practices.

II. RELATED WORK

Sl.No	Author	Year	Technique Used	Result
11	Ersin Elbasi, Elda Cina	2023	Naive Bayes Classifier, Random Forest	Our research emphasizes the importance of integrating machine learning and IoT sensors in agriculture for optimizing crop production, waste reduction, and enhancing food security.

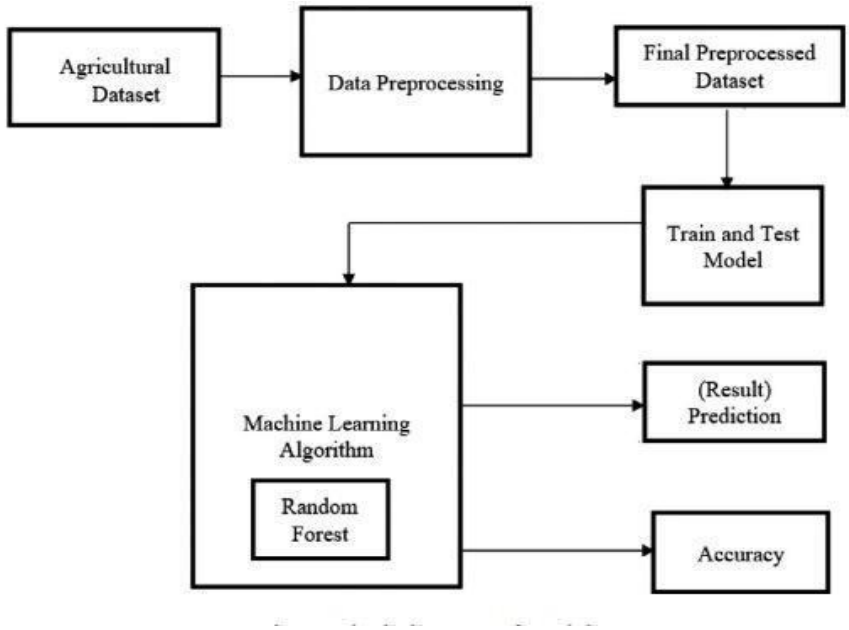
©IJRASET: All Rights are Reserved

9.	Jagadish Timsina ,Sudarshan Dutta, Krishna Prasad Devkota ,Somsubhra Chakraborty , Ram Krishna Neupane ,Sudarshan Bista,, Lal Prasad Amgain, Kaushik Majumdar	2022	Nutrient Expert (NE) for fertilizer recommendations, government recommendation (GR), and farmers' practices,RAndom Forest	Nutrient Expert recommendations outperform government and farmers' practices, yielding higher crops and profits with lower greenhouse gas emissions, showcasing its potential for sustainable cereal production in the region.
3.	Sonal Agarwa and Sandhya Tarar	2021	SVM, LSTM, RNN	Utilizing SVM, LSTM, and RNN, the model achieves a 97% accuracy in predicting optimal crops, benefiting farmers in diverse conditions.
10.	Janmejay Pant, R.P. Pant ,Manoj Kumar Singh , Devesh Pratap Singh , Himanshu Pant	2021	Gradient Boosting Regressor, Random Forest Regressor, SVM, Decision Tree Regressor	This research on crop yield prediction in India employs machine learning, highlighting the Decision Tree Regressor's 96% accuracy, with a focus on potatoes and suggesting potential improvements through the inclusion of more relevant features.
1.	Ms Kavita, Pratistha Mathur	2020	Decision Tree, Linear Regression, Lasso regression, and Ridge Regression	Random Forest outperforms other models in crop yield prediction, indicating a need for larger datasets and further exploration of deep learning models.
2.	Shruthi Gowda, Sangeetha Reddy	2020	Random Forest, Polynomial Regression, Decision Tree algorithms	Machine Learning (Random Forest, Polynomial Regression, Decision Tree) showed improved crop yield prediction, with Random Forest as the standout model.
4.	Saqid A Mulla, S.A. Qadri	2020	Decision Tree algorithm, EDA, datasets, rainfall data, WPI, web application.	The model predicts crop yield and prices, aiding farmers in making informed decisions for sustainable agriculture.
6.	Thomas van Klompenburg, Ayalew Kassahun, Cagatay Catal	2020	Systematic Literature Review(SLR)	The study highlights the diverse approaches to crop yield prediction with machine learning, showcasing variability in features and models using CNN, LSTM, and DNN.
7.	Potnuru Sai Nishant ,Pinapa Sai Venkat , Bollu Lakshmi Avinash, B. abber	2020	Kernel Ridge, Lasso, and ENet algorithms	Enhanced crop yield prediction has been attained through the utilization of advanced regression techniques like Kernel Ridge, Lasso, and ENet, coupled with improved accuracy via Stacked Regression, demonstrating promising

				practical applications in agriculture.
8.	Pallavi Kamath , Pallavi Patil, Shrilatha S, Sushma, Sowmya	2020	Random forest Algorithm, Seaborn library for data visualization, and Weather APIs for fetching weather data.	The study successfully utilized data mining with the Random Forest algorithm, achieving a notable 98% accuracy in predicting crop yield, demonstrating its effectiveness for enhancing agricultural planning and decision-making.
12	Nischitha K	2020	Using SVM	GUI system predicts crops, provides nutrients, seed, yield, market data, empowering informed decisions, advancing agriculture.
5.	S.Bhanumathi, M.Vineeth and N.Rohit	2019	Data Mining, Random Forest, Backpropagation, NumPy, Pandas, Matplotlib, Scikit-learn,TensorFlow	The study effectively used Random Forest and Backpropagation algorithms for crop yield prediction, favoring Random Forest due to lower error rates, and envisions future development of a user-friendly web app.

III. PROPOSED METHODOLOGY

Process Flow Diagram



A. Data Collection

The initial phase involves gathering data on diverse factors influencing crop yield and profitability, encompassing weather patterns, soil attributes, and market rates. This data is sourced from a variety of outlets, including governmental bodies, research institutions, and online repositories.

1) Dataset1

datainput - DataFrame

Index	Crop	State	ultivation (`/Hecta	Cultivation (`/Hec	Production (`/Qui	d (Quintal/ Hecta	Support price
0	ARHAR	Uttar Pradesh	9794.05	23076.7	1941.55	9.83	6000
1	ARHAR	Karnataka	10593.1	16528.7	2172.46	7.47	6000
2	ARHAR	Gujarat	13468.8	19551.9	1898.3	9.59	6000
3	ARHAR	Andhra Pradesh	17051.7	24171.7	3670.54	6.42	6000
4	ARHAR	Maharashtra	17130.5	25270.3	2775.8	8.72	6000
5	COTTON	Maharashtra	23711.4	33116.8	2539.47	12.69	5515
6	COTTON	Punjab	29047.1	50828.8	2003.76	24.39	5515
7	COTTON	Andhra Pradesh	29140.8	44756.7	2509.99	17.83	5515
8	COTTON	Gujarat	29616.1	42070.4	2179.26	19.05	5515
9	COTTON	Haryana	29919	44018.2	2127.35	19.9	5515

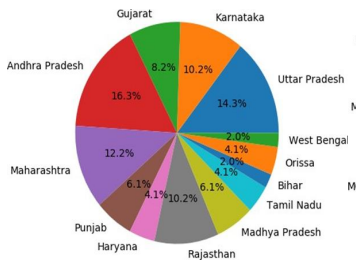
We consolidated data from diverse origins [1], featuring columns such as:
Crop types
States
Cultivation costs (A2+FL) per hectare
Cultivation costs (C2) per hectare
Production costs (C2) per quintal
Yield obtained
For each entry, profit was computed utilizing the formula:
Profit = (Yield * Support Price) - (C1 + C2 + (Yield * Cp))

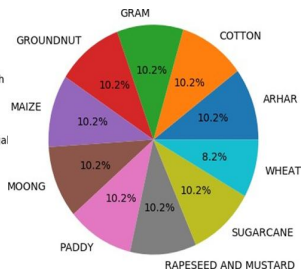
where:
C1 represents cultivation costs (A2+FL) per hectare
C2 signifies cultivation costs (C2) per hectare
Cp denotes production costs (C2) per quintal
The govt. fixes support prices [2] per Quintal for various commodities, for example various Kharif and Rabi crops.If the yield produced will result in profit based on support prices declared by the government, class 1 was allotted; else it was classified as class 0.

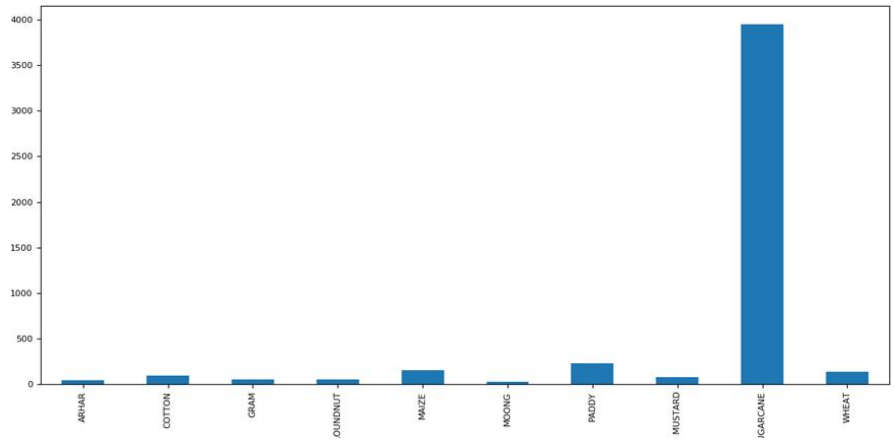
2) Dataset2
In the second dataset, the columns include State_Name, District_Name, Crop_Year, Season, Crop, Area, and Production. The objective is to forecast crop production using regression techniques.

datainput - DataFrame

Index	State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
0	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254	2000
1	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2	1
2	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102	321
3	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	176	641
4	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Cashewnut	720	165
5	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Coconut	18168	6.51e+07
6	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Dry ginger	36	100
7	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Sugarcane	1	2
8	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Sweet potato	5	15
9	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Tapioca	40	169
10	Andaman and Nicobar Islands	NICOBARS	2001	Kharif	Arecanut	1254	2061
11	Andaman and Nicobar Islands	NICOBARS	2001	Kharif	Other Kharif pulses	2	1
12	Andaman and Nicobar Islands	NICOBARS	2001	Kharif	Rice	83	300
13	Andaman and Nicobar Islands	NICOBARS	2001	Whole Year	Cashewnut	719	192







B. Data Preprocessing

Once data is gathered, it undergoes processing to make it usable. This includes cleaning to eliminate missing or incorrect values and standardizing to ensure all variables share the same scale. In our dataset, after incorporating the support price column and profit, labeled as 0 and 1 respectively, preprocessing techniques were employed to address missing values. The crops and state columns underwent label encoding, followed by the application of a one-hot encoder to avoid ranking implications.

datainput - DataFrame

Index	Crop	State	Cost of Cultivation (`/Hectare) A2+FL	Cost of Cultivation (`/Hectare) C2	Cost of Production (`/Quintal) C2	Yield (Quintal/ Hectare)	Support price
0	ARHAR	Uttar Pradesh	9794.05	23076.7	1941.55	9.83	6000
1	ARHAR	Karnataka	10593.1	16528.7	2172.46	7.47	6000
2	ARHAR	Gujarat	13468.8	19551.9	1898.3	9.59	6000

A snapshot of dataframe

C. Model Development

After the data is pre-processed, machine learning algorithms such as linear regression and decision trees are used to analyze the data and make predictions. The algorithms are trained on a subset of the data, and their performance is tested on a separate validation set to ensure that the models are accurate and reliable.

In our system, we conduct tests on multiple algorithms, comparing and selecting the best one based on the analysis of the classification report. It must calculate the accuracy for both the training and testing datasets, determine specificity, False Positive rate, precision, and recall, and subsequently compare these metrics across various algorithms using Python code.

A. The Involvement steps in the Process Include

1) Identify and define the problem.
2) Prepare the necessary data.
3) Evaluate different algorithms.
4) Enhance and optimize the results.
5) Make predictions based on the refined model.

Classification algorithms will be utilized on dataset 1, whereas regression will be employed for predicting production in dataset 2.

B. Applied Algorithms Include

1) Classification

- Decision Tree
- Logistic Regression
- K Nearest Neighbor
- Random Forest Classifier

2) Clustering

3) Regression
- Decision Tree
- Random forest

D. Model Evaluation

The developed models are evaluated using various performance metrics such as mean absolute error, root mean squared error, and accuracy. The models are also compared with other existing models to determine their effectiveness.

First, let's briefly comprehend certain performance evaluation metrics:

1) General Definitions

a) True Positive (TP) represents the count of instances where the system accurately identifies a condition in the presence of its actual occurrence.

b) True Negative (TN) signifies the count of instances where the system correctly does not identify a condition that is indeed absent.

When assessing precision, which quantifies the ratio of true positive observations to the total predicted positive observations, a lower false positive rate indicates higher precision. In the present study, a noteworthy precision score of 0.788 has been achieved.

Regarding Recall, it measures the proportion of correctly predicted positive observations relative to the total observations in the actual "yes" class. For instance, in a model predicting defaulters, Recall (Sensitivity) is calculated as TP / (TP + FN), offering insight into how effectively the model captures actual positive cases among all instances in the "yes" class.

2) F1 Score

The F1 score involves computing the weighted average of Precision and Recall through a calculated process, factoring in false positives and false negatives. Understanding accuracy intuitively can be challenging, but F1 proves more valuable, especially with uneven class distribution. Accuracy is preferred when the costs of false positives and false negatives are similar. To delve into precision and recall, it's essential that the costs of false positives and false negatives vary significantly.

Formula for General F-Measure:

$$F\text{-Measure}=2TP/(FP+FN+2TP)$$

Formula for F1-Score:

The F1-Score is calculated as

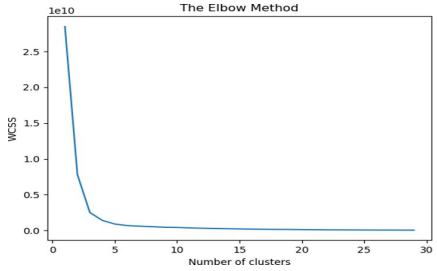
$$2\times(Recall\times Precision)/(Recall+Precision)$$

3) Precision

How reliable is the model's positive outcome prediction? Precision, calculated as TP/(TP+FP), becomes crucial, especially in scenarios where the cost of false positives is substantial. Consider a context like skin cancer detection. A low-precision model could lead to numerous patients being incorrectly diagnosed with melanoma, causing additional tests and psychological stress. High false positives might lead to a desensitization to alarms, as they become frequent and often inaccurate.

E. Clustering

Following the application of clustering, we generated an elbow graph to ascertain the optimal number of clusters. A common method for determining this optimal number involves calculating the Within-Cluster-Sum-of-Squares (WCSS), which quantifies the sum of squared distances between each data point within all clusters and their respective centroids. The objective is to minimize this sum.



IV. RESULT & DISCUSSION

Among classification algorithms, Logistic Regression emerged as the most effective in forecasting the profitability of a specific crop, considering factors such as the state, cultivation costs (C1, C2), production costs (Cp), and government support prices for the 2020-21 period. However, the performance of the second dataset was not as satisfactory. To enhance model accuracy, supplementary variables such as rainfall and temperature should be incorporated.

Algorithm	Precision		Recall		F1 Score		Accuracy
	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1	
Logistic Regression	1.0	0.89	0.86	1.0	0.92	0.94	0.93
DecisionTree	0.78	1.0	1.0	0.75	0.88	0.86	0.87
Randomfor-est	0.86	0.75	0.75	0.86	0.80	0.80	0.80
K nearest	0.50	0.86	0.75	0.67	0.60	0.75	0.69

Algorithm	R2 score	Mean absolute error
Decision Tree	0.84	167163.3086041714
Random Forest	0.91	155503.99436675265

The developed models are continuously monitored and updated as needed to ensure their accuracy and effectiveness. Any changes or updates to the model are carefully tested to ensure that they do not negatively impact its performance.

V. CONCLUSION

In conclusion, the Profitability and Yield Prediction project utilizing Python and machine learning present a promising avenue for enhancing agricultural efficiency in India. The project focuses on predicting crop yields, considering factors like weather conditions, soil quality, and market prices. By employing a range of machine learning algorithms like Logistic Regression and Decision Trees, the project endeavors to equip farmers with the necessary insights for making informed choices regarding crop selection and allocation of resources. Continuous monitoring and collaboration with stakeholders ensure the accuracy and effectiveness of the predictive models. The potential impact extends beyond individual farmers to benefit agricultural entities and governmental bodies, fostering food security and sustainability. Overall, the research demonstrates the potential of technology-driven solutions in optimizing agricultural practices for the development and prosperity of the country.

VI. FUTURE SCOPE

There are numerous opportunities for advancing research and development in predicting the profitability and yield of agricultural crops in India using Python. These include enhancing the accuracy and reliability of predictive models by collecting and analyzing extensive data and exploring more sophisticated machine learning algorithms. Additionally, there is potential to expand the model's applicability by including additional crops and regions, providing a more comprehensive understanding of crop yield and profitability. Moreover, investigating advanced machine learning techniques such as deep learning can further improve predictive models. Efforts can also focus on enhancing the user experience and accessibility of the model, potentially through developing user-friendly interfaces or integrating with existing agricultural tools and platforms. Overall, these initiatives show great promise for advancing the agricultural sector and contributing to India's overall development and prosperity.

REFERENCES

[1] Kavita, Ms, and Pratistha Mathur. (2020) “Crop Yield Estimation in India Using Machine Learning.” In 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), 220–224. doi:10.1109/ICCCA49541.2020.9250915.

[2] Fan, Wu, Chen Chong, Guo Xiaoling, Yu Hua, and Wang Juyun. (2015) “Prediction of Crop Yield Using Big Data.” In 2015 8th International Symposium on Computational Intelligence and Design (ISCID), 1:255–260. doi:10.1109/ISCID.2015.191.

[3] Kamath, Pallavi, Pallavi Patil, Shrilatha S, Sushma, and Sowmya S. (2021) “Crop Yield Forecasting Using Data Mining.” Global Transitions Proceedings, International Conference on Computing System and its Applications (ICCSA- 2021), 2 (2): 402–407. doi:10.1016/j.gltp.2021.08.008.

[4] Wigh, Daniel S., Jonathan M. Goodman, and Alexei A. Lapkin. “A Review of Molecular Representation in the Age of Machine Learning.” WIREs Computational Molecular Science n/a (n/a): e1603. doi:10.1002/wcms.1603.

[5] Kavita, and Pratistha Mathur. (2021) “Satellite-Based Crop Yield Prediction Using Machine Learning Algorithm.” In 2021 Asian Conference on Innovation in Technology (ASIANCON), 1–5. doi:10.1109/ASIANCON51346.2021.9544562.

[6] Bali, Nishu, and Anshu Singla. (2022) “Emerging Trends in Machine Learning to Predict Crop Yield and Study Its Influential Factors: A Survey.” Archives of Computational Methods in Engineering 29 (1): 95–112. doi:10.1007/s11831-021-09569-8.

[7] van Klompenburg, Thomas, Ayalew Kassahun, and Cagatay Catal. (2020) "Crop Yield Prediction Using Machine Learning: A Systematic Literature Review." Computers and Electronics in Agriculture 177 (October): 105709. doi:10.1016/j.compag.2020.105709.

[8] Shen, Dinggang, Guorong Wu, and Heung-Il Suk. (2017) "Deep Learning in Medical Image Analysis." Annual Review of Biomedical Engineering 19 (1): 221–248. doi:10.1146/annurev-bioeng-071516-044442.

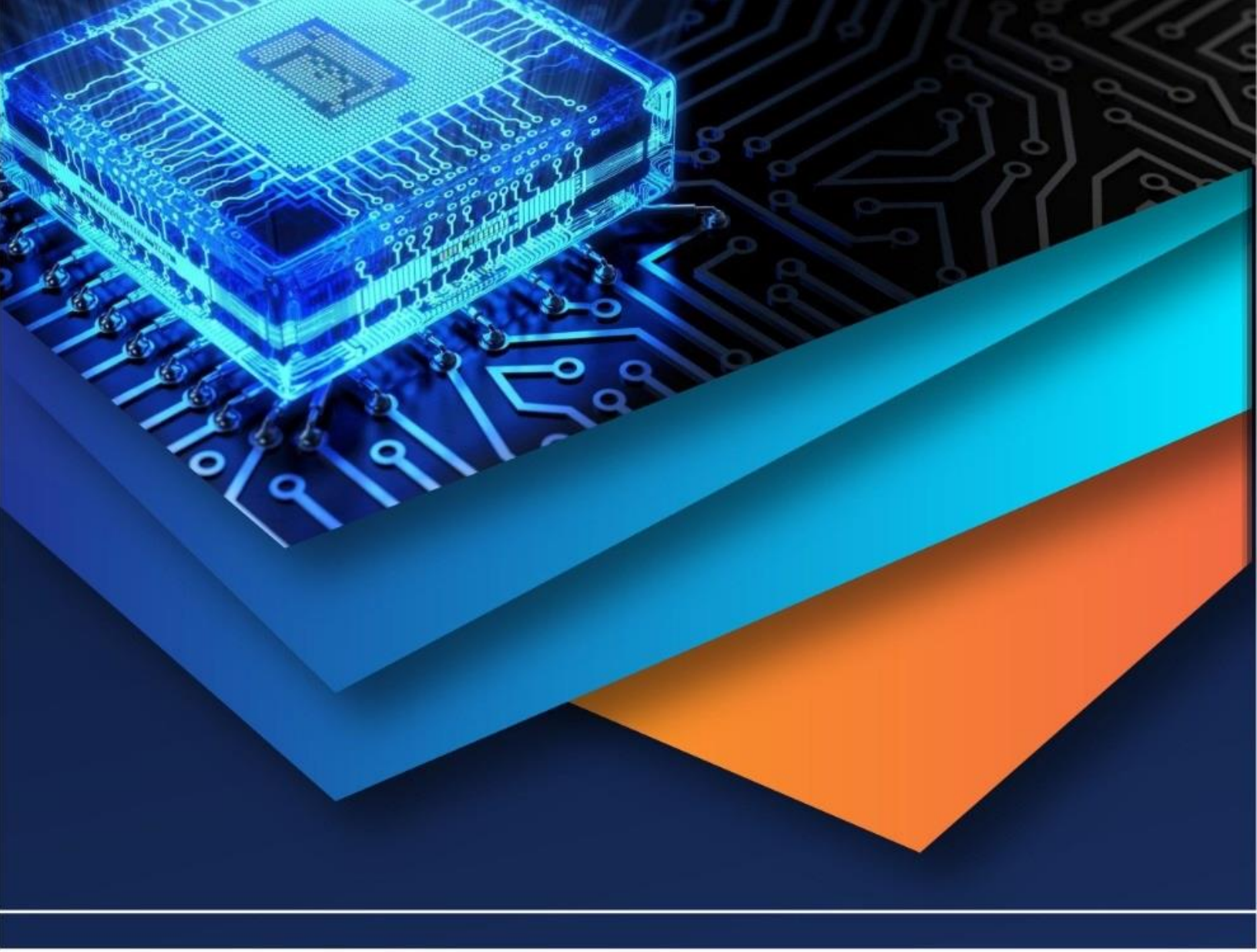
[9] Minaee, Shervin, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. (2021) "Deep Learning--Based Text Classification: A Comprehensive Review." ACM Computing Surveys 54 (3): 1–40. doi:10.1145/3439726.

[10] Belgiu, Mariana, and Lucian Drăguţ. (2016) "Random Forest in Remote Sensing: A Review of Applications and Future Directions." ISPRS Journal of Photogrammetry and Remote Sensing 114 (April): 24–31. doi:10.1016/j.isprsjprs.2016.01.011.

[11] Suthaharan, Shan. (2016) "Support Vector Machine." In Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, edited by Shan Suthaharan, 207–235. Integrated Series in Information Systems. Boston, MA: Springer US. doi:10.1007/978-1-4899-7641-3_9.

[12] Hochreiter, Sepp, A. Steven Younger, and Peter R. Conwell. (2001) "Learning to Learn Using Gradient Descent." In Artificial Neural Networks — ICANN 2001, edited by Georg Dorffner, Horst Bischof, and Kurt Hornik, 87–94. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. doi:10.1007/3-540-44668- 0_13.

[13] Sherstinsky, Alex. (2020) "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network." Physica D: Nonlinear Phenomena 404 (March): 132306. doi:10.1016/j.physd.2019.132306. [15]Yu, Yong, Xiaosheng Si, Changh.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)