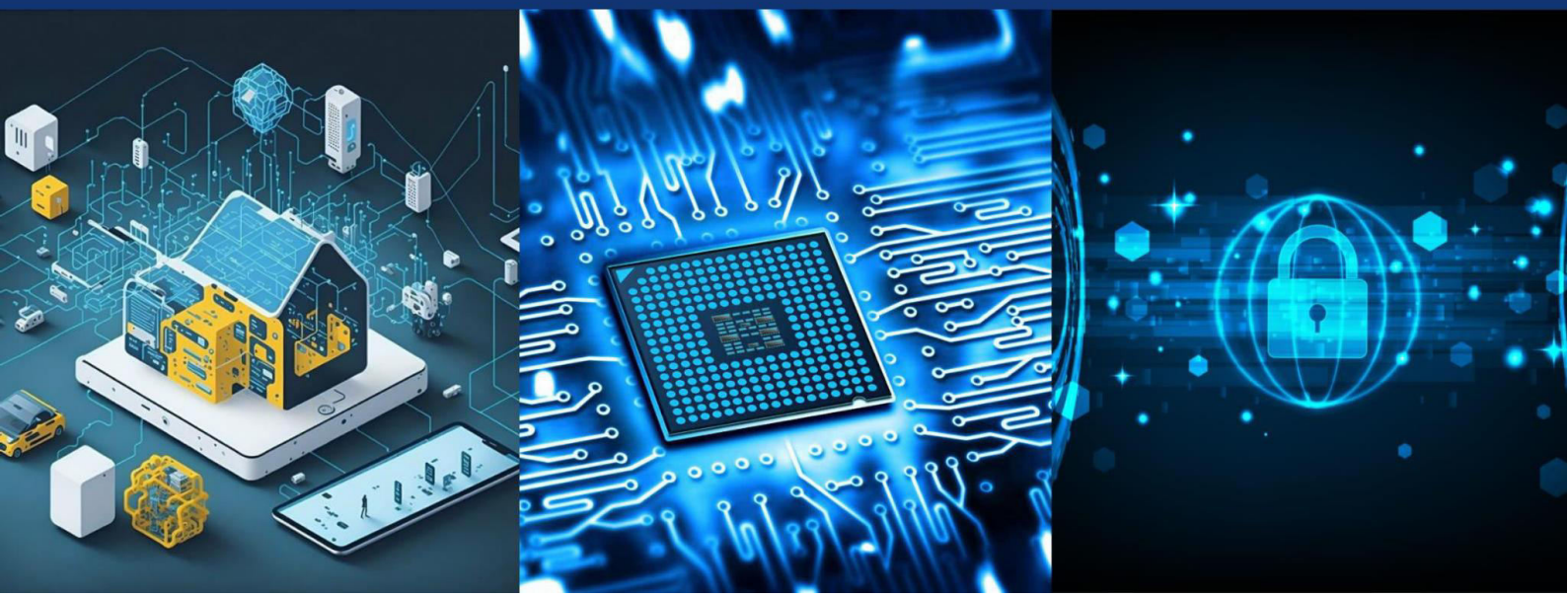
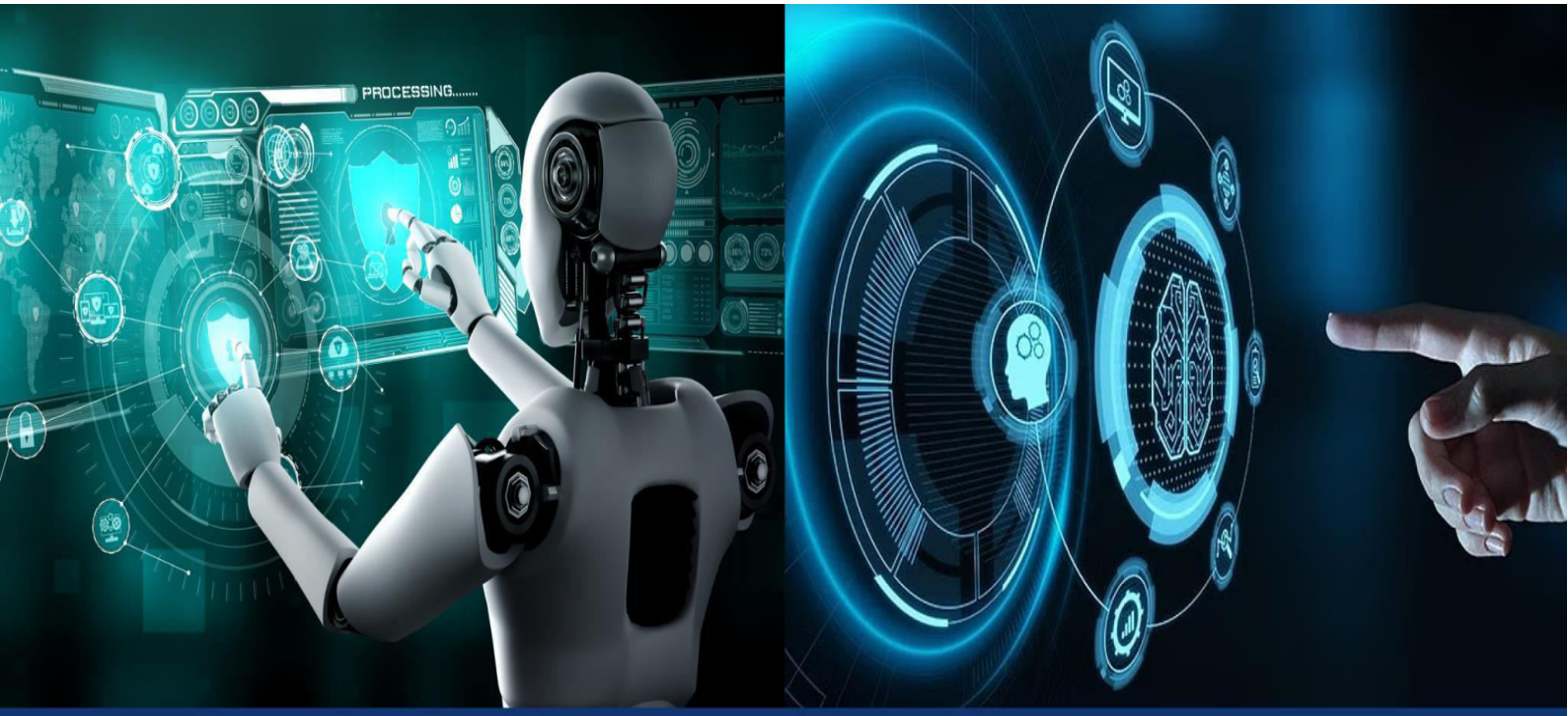
ISSN(O): 2320-9801
ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771 **Volume 14, Issue 4, April 2026**

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404117

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Comparative Study of Supervised Learning Models for Robust and Scalable Predictive Analytics

N. Prasad[1], M. Sheeba Gracy [2], M.Sai Chandu[3], P.Koti Ganesh [4], R. Vooha Devi[5]

Assistant Professor, Department of Information Technology, Sir C R Reddy College of Engineering, Eluru, India[1]

Final Year Students, Department of Information Technology, Eluru, India[2-5]

ABSTRACT: Machine learning models are widely used for predictive decision-making in real-world systems; however, selecting the most suitable model remains a challenging task due to differences in performance and scalability. This project focuses on developing a comparative framework that evaluates multiple supervised learning models to identify the most effective approach for predictive analytics. The proposed system uses Logistic Regression, Support Vector Machine (SVM), and Random Forest to analyze a healthcare diabetes dataset and generate predictions. The models are tested under real-world- like conditions and evaluated using metrics such as accuracy, precision, recall, and F1-score. The system improves reliability by comparing model performance and selecting the most robust approach. This helps in building efficient and scalable predictive systems for practical applications.

KEYWORDS: Model Comparison, Evaluating Performance , Robustness and Scalability .

I. INTRODUCTION

This paper addresses Machine Learning (ML) as one of the most important technologies in modern computing. It enables systems to analyze large volumes of data and make predictions with minimal human intervention. ML is widely used in fields such as healthcare, finance, transportation, and business analytics, where it helps identify patterns, predict outcomes, and improve decision-making processes. However, selecting the most suitable model remains a challenge, as different algorithms perform differently depending on the dataset and problem complexity. In real-world applications, relying on a single model without proper comparison can lead to inaccurate predictions and poor generalization.

Different algorithms behave differently depending on factors such as dataset size, feature distribution, and data quality. In many real-world applications, a single model is used without proper evaluation, which can lead to inaccurate predictions and poor performance when deployed in practical environments. Models that perform well under controlled conditions may fail when exposed to noisy or incomplete data. To address this issue, it is essential to adopt a comparative approach that evaluates multiple models under consistent conditions. This project focuses on developing a structured framework that compares three supervised learning models—Logistic Regression, Support Vector Machine (SVM), and Random Forest—for predictive analytics. The system includes key stages such as data preprocessing, feature scaling, model training, and performance evaluation to ensure fairness and consistency in comparison. The models are evaluated using multiple performance metrics, including accuracy, precision, recall, and F1-score, which provide a balanced assessment of their effectiveness. This helps in understanding not only which model performs better but also why it performs better under specific conditions. The models are evaluated using multiple performance metrics, including accuracy, precision, recall, and F1-score, which provide a balanced assessment of their effectiveness. This helps in understanding not only which model performs better but also why it performs better under specific conditions.

The objective of this work is to improve predictive decision-making by identifying a model that offers better robustness and scalability. By selecting the most reliable model through systematic comparison, the proposed framework supports the development of efficient and practical machine

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

learningsolutionsforrealworldapplications.Theproposedframeworkemphasizes the importance of reproducibility and consistency in performance model evaluation .

By maintaining a uniform preprocessing and training pipeline, the system ensures that all models are tested under the same conditions, reducing bias in performance comparison. This structured approach allows for more reliable conclusions and helps in selecting models that can generalize well to unseen data.

II. LITERATURE REVIEW

Machine learning has become an essential tool for making predictions and supporting decision-making in real-world areas such as healthcare, finance, manufacturing, transportation, and business analytics. Earlier models such as linear regression, logistic regression, and decision trees were widely used because they were simple and easy to understand, allowing users to clearly see how input features influenced the output.
As the need for better prediction accuracy increased, more advanced models such as Support Vector Machines,

Random Forests, and ensemble techniques were developed. These models are capable of handling complex patterns and large datasets more effectively, but they also introduce challenges in terms of model selection and performance comparison. Different models often produce different results on the same dataset, making it difficult to identify the most suitable approach.

To address this issue, researchers have focused on comparative analysis of machine learning models. Comparative studies evaluate multiple algorithms under the same conditions to ensure fair assessment and reliable conclusions. These studies typically involve data preprocessing, feature scaling, model training, and evaluation using standard performance metrics. Such an approach helps in understanding the strengths and limitations of each model.

Various evaluation metrics such as accuracy, precision, recall, and F1-score are commonly used to measure model performance. Accuracy provides an overall correctness measure, while precision and recall give deeper insights into classification performance, especially in cases where class distribution is uneven.
Therefore, a structured framework for comparing multiple supervised learning models is necessary. By evaluating models under consistent conditions and using multiple performance metrics, it becomes possible to identify the most robust and scalable model for predictive analytics. This approach ensures better decision-making and supports the development of efficient real-world machine learning systems.

Authors	Year	Title
Zhang et al.	2026	Machine Learning Evaluation Framework
Kumar & Lee	2025	Scalable Machine Learning Models
Singh et al.	2024	Ensemble Learning (Random Forest)
Wang & Patel	2023	Hybrid Machine Learning Model

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404117

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

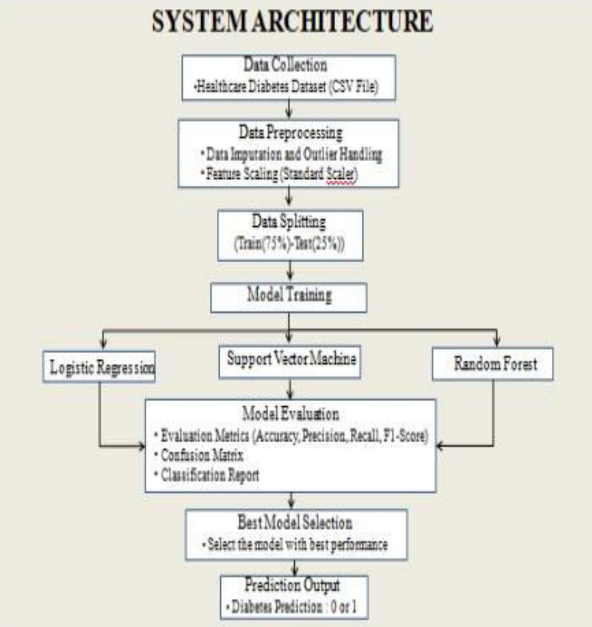
III. PROBLEM STATEMENT

Machine learning models are widely used for making predictions in real-world applications, but selecting the most suitable model remains a major challenge. Different models often produce different results depending on the dataset, its quality, and the complexity of the problem. In many cases, a single model is chosen without proper comparison, which can lead to inaccurate predictions and poor performance in practical scenarios. This becomes more critical when dealing with real-world data that may contain noise, missing values, or varying patterns. .This challenge highlights the growing need for machine learning models that are not only accurate but also explainable. This challenge highlights the need for a structured framework that can compare multiple machine learning models under the same conditions. Such an approach helps in identifying the most effective model based on performance metrics and ensures better decision- making in predictive analytics systems.

IV. SYSTEM ARCHITECTURE

The system architecture is designed as a clear, step-by-step workflow that makes the entire predictive analytics process easy to understand and implement. It begins with data collection, where the healthcare diabetes dataset is loaded and prepared for analysis. This is followed by data inspection and preprocessing, where missing values are handled, duplicate entries are removed, and the dataset is cleaned to ensure quality. In the preprocessing stage, feature scaling is applied so that all input values are standardized and suitable for model training.

Once the data is prepared, the system moves to the model training phase, where multiple supervised learning models—Logistic Regression, Support Vector Machine (SVM), and Random Forest—are trained using the same dataset. After training, each model is evaluated using performance metrics such as accuracy, precision, recall, and F1-score to measure their effectiveness. The system then proceeds to the comparison stage, where the results of all models are analyzed to identify the most reliable and efficient one. Based on this analysis, the best-performing model is selected.



This structured workflow ensures consistency,improves reliability, and helps in selecting a robust and scalable model for real-world predictive analytics applications.

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404117

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. EXISTING SYSTEM

In many real-world applications, existing systems for predictive decision-making rely on traditional machine learning models or basic analytical approaches. These systems are designed to work with structured data, where past information is used to identify patterns and generate predictions. They are commonly used in domains such as healthcare for disease prediction, finance for credit scoring and fraud detection, and industrial environments for monitoring system performance and detecting faults.

Commonly used models include Logistic Regression, Decision Trees, Support Vector Machines (SVM), and Random Forest. Each of these models has its own advantages and is selected based on the problem requirements and data characteristics. These systems are generally effective in handling classification tasks and provide reasonable accuracy when applied to suitable datasets. Most existing systems depend on a single model without performing a proper comparison with other algorithms. This approach can lead to suboptimal performance, as different models behave differently depending on data characteristics such as size, distribution, and noise. A model that performs well in one scenario may fail in another, making it difficult to ensure consistent results. Another limitation of existing systems is the lack of focus on robustness and scalability. This limits the ability to make informed decisions about model selection.

VI. PROPOSED SYSTEM

The proposed system is a machine learning framework designed to provide accurate predictions by comparing multiple supervised learning models and selecting the most suitable one. Unlike traditional systems that rely on a single model, this approach focuses on evaluating different algorithms to improve performance, reliability, and scalability. The overall architecture follows a structured pipeline that includes data collection, preprocessing, feature processing, model training, evaluation, comparison, and prediction. This organized design makes the system easy to understand, maintain, and apply in real-world applications. The process begins with collecting the dataset from a reliable source, which in this case is a healthcare diabetes dataset. The data is then prepared through preprocessing steps such as handling missing values, removing duplicates, and cleaning inconsistencies. Feature scaling is applied to normalize the data so that all input variables contribute equally during model training. This ensures better performance and stability across the different types of models.

Once the data is prepared, the system moves to the model training phase, where multiple supervised learning models—Logistic Regression, Support Vector Machine (SVM), and Random Forest—are trained using the same dataset. Using multiple models allows the system to capture different patterns in the data and ensures a fair comparison. Each model is evaluated using accuracy, precision, recall, and F1-score, and the best-performing model is selected based on overall performance. The final model is then used to generate predictions on new data, making the system reliable and efficient for predictive analytics tasks.

VII. RESULT

- The project titled **"Comparative Study of Supervised Learning Models for Robust and Scalable Predictive Analytics"** focuses on evaluating multiple machine learning models to identify the most effective one for prediction tasks.
- The system collects and preprocesses the dataset before using it to train different supervised learning models for predictive analysis.
- The trained model is evaluated using performance metrics such as:
  - Accuracy
  - Classification report
  - Confusion matrix

These evaluation methods help demonstrate the system's ability to analyse data effectively and produce reliable predictions.

A key feature of the project is the comparison of multiple models instead of relying on a single approach.

The system implements Logistic Regression, Support Vector Machine (SVM), and Random Forest to capture different data patterns.

The results show that each model performs differently based on the dataset and evaluation metrics.

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404117

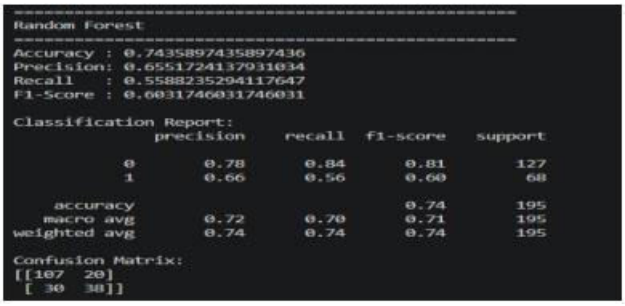
www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

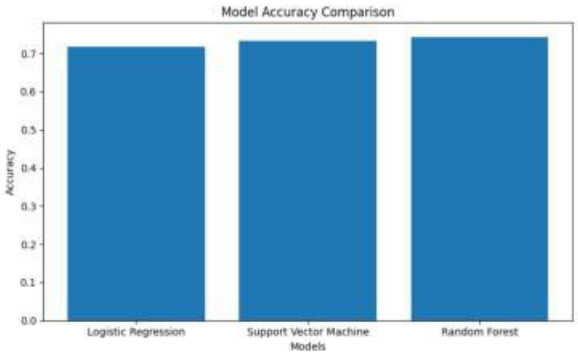
International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Among the models, Random Forest provides better overall performance and stability compared to others.

Output screens:





Overall, the results highlight that the proposed system effectively compares multiple machine learning models and identifies the most suitable one, achieving a good balance between prediction accuracy and computational efficiency, making it well-suited for real-world applications such as healthcare prediction, financial analysis, and decision-support systems.

VIII. CONCLUSION

This paper presented the "Comparative Study of Supervised Learning Models for Robust and Scalable Predictive Analytics," which focuses on developing a system that improves prediction accuracy by comparing multiple machine learning models. In this approach, data is first collected and preprocessed to ensure it is clean and suitable for training. The system then applies different supervised learning models, including Logistic Regression, Support Vector Machine (SVM), and Random Forest, for predictive analysis.

Once the models are trained, they are evaluated using standard performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix. These metrics help in understanding how well each model performs and whether it can produce reliable predictions on new data. A key aspect of this project is the comparison of multiple models under the same conditions, which allows for identifying the most effective and consistent approach. This comparative analysis improves decision-making by selecting a modelthat provides better accuracy and scalability making the system suitable for real-world applications.

IX. FUTURE WORK

The proposed machine learning framework can be further enhanced in several important ways to improve its performance, scalability, and real-world applicability. One key improvement is to extend the system to handle more complex and diverse datasets, including large-scale and real- time data. This would allow the framework to be applied in broader domains such as healthcare analytics, financial forecasting, and intelligent business systems. Another important area for future work is the inclusion of additional advanced machine learning and

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

deep learning models to improve prediction performance. Integrating techniques such as hyperparameter tuning and optimization methods can further enhance model accuracy and efficiency. This would help in identifying more robust models for different types of datasets and problem scenarios. The system can also be improved by enabling real-time prediction and deployment in cloud- based environments. This would support faster processing, better scalability, and continuous model updates, making the system more practical for real-world applications. Additionally, incorporating automated model selection techniques can further simplify the process of choosing the best model for predictive analytics tasks.

REFERENCES

[1] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[2] C. Cortes and V. Vapnik, "Support Vector Networks," Machine Learning, vol. 20, pp. 273–297, 1995.
[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. New York: Springer, 2009.
[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[5] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do We Need Hundreds of Classifiers to Solve Real-World Classification Problems?" Journal of Machine Learning Research, vol. 15, pp. 3133–3181, 2014.
[6] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. Sebastopol, CA: O'Reilly Media, 2019.
[7] C. M. Bishop, Pattern Recognition and Machine Learning. New York: Springer, 2006.
[8] Scikit-learn Developers, Scikit-learn Documentation. [Online]. Available: https://scikit-learn.org [9] NumPy Developers, NumPy Documentation. [Online]. Available: https://numpy.org
[10] Pandas Developers, Pandas Documentation. [Online]. Available: https://pandas.pydata.org
[11] Matplotlib Developers, Matplotlib Documentation. [Online]. Available: https://matplotlib.org [12] Seaborn Developers, Seaborn Documentation. [Online]. Available: https://seaborn.pydata.org [13] IBM Cloud Education, "Supervised Learning." [Online]. Available: https://www.ibm.com/topics/supervised-learning
[14] Analytics Vidhya, "Supervised Machine Learning Algorithms." [Online]. Available: https://www.analyticsvidhya.com
[15] Kaggle Learn, "Introduction to Machine Learning." [Online]. Available: https://www.kaggle.com/learn/intro-to-machine-learning
[16] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA: Morgan Kaufmann, 2011.
[17] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016. [18] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with Applications in R. New York: Springer, 2013.
[19] K. P. Murphy, Machine Learning: A Probabilistic Perspective. Cambridge, MA: MIT Press, 2012. Sir C R Reddy College of Engineering Department of IT COMPARATIVE STUDY OF SUPERVISED LEARNING MODELS FOR ROBUST AND SCALABLE PREDICTIVE ANALYTICS 49
[20] S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed. Birmingham, UK: Packt Publishing, 2019.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details