iJRASET

International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: V Month of publication: May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70119

www.ijraset.com

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

Enhancing Transportation Prediction in Space Tourism via Hybrid Machine Learning and AR Integration

Sneha Singh[1], Sweta Sharma[2]
Department of Information Technology Delhi Technological University, Delhi, India

Abstract: The increasing relevance of space tourism necessitates the development of predictive systems for ensuring passenger safety and transportation efficiency. This paper presents a hybrid machine learning pipeline that utilizes advanced techniques including feature engineering, anomaly detection (Isolation Forest), stacking classifiers, SHAP explainability, and NLP with transformer models to predict survival rates based on the Spaceship Titanic dataset. Additionally, we propose an Augmented Reality (AR) simulation tool to enhance passenger experience and operational preparedness. Our methodology achieves robust predictive performance, improved model interpretability, and practical integration for real-time space tourism applications.

Index Terms: Space Tourism, Ensemble Learning, SHAP, NLP, AR Simulation, Isolation Forest

I. INTRODUCTION

In recent years, space tourism has transitioned from science fiction to a tangible commercial reality. With ventures like SpaceX and Blue Origin leading the way, the demand for intelligent, data-driven systems to ensure passenger safety and improve travel experience has grown. This paper proposes a hybrid machine learning (ML) pipeline tailored for space tourism risk assessment using the synthetic 'Spaceship Titanic' dataset. The dataset includes both structured and unstructured data, simulating a catastrophic event and the need to predict which passengers were 'transported.' Our pipeline integrates ML techniques, NLP features, anomaly detection, and SHAP-based interpretability to ensure predictive accuracy and transparency.

II. LITERATURE REVIEW

A. Classical Survival Modeling: Traditional survival prediction models using Titanic data have laid the groundwork for understanding demographic and behavioral indicators in critical scenarios. B. Space Tourism: Emerging studies highlight the need for robust safety analytics in commercial space travel. C. Ensemble Learning: Stacking methods improve model generalization by combining diverse base learners. D. NLP in Tabular Data: Embedding categorical text using transformers enhances the semantic richness of features. E. Explainable AI: SHAP values are effective in making complex model predictions interpretable.

III. METHODOLOGY

Our approach involves several key components:

1) Data Preprocessing: Missing values were imputed using median/mode strategies; cabin features were decomposed.
2) Feature Engineering: Includes derived features like TotalSpending and group identifiers.
3) NLP Feature Extraction: Transformer models generated semantic embeddings from categorical fields.
4) Anomaly Detection: Isolation Forest algorithm flagged inconsistent records.
5) Stacking Classifier: Combined Random Forest, LightGBM, and SVM with Logistic Regression as meta-learner.
6) SHAP Explainability: Enabled both global and local interpretations of model predictions.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The model achieved 82–84% accuracy and an AUC of ~0.85 using stratified K-fold cross-validation. SHAP analysis revealed CryoSleep, TotalSpending, and Age as the top features. An ablation study showed a 3–5% drop in performance when NLP features were excluded, affirming their contribution. Confusion matrices and SHAP force plots highlighted both strengths and misclassifications, guiding further refinement.

V. AR SIMULATION AND PRODUCT POTENTIAL

We propose an AR-based simulation to visualize and interact with passenger data, risk predictions, and spaceship layout. Applications include:

1) Safety training for crew.
2) Passenger experience enhancement.
3) Real-time decision support during missions.
4) Stakeholder engagement through immersive demonstrations.

The system leverages Unity and RESTful APIs for rendering and integration.

VI. DISCUSSION

The integration of ensemble modeling and explainable AI techniques has demonstrated improved predictive accuracy and interpretability in forecasting transportation outcomes for space tourism. The stacking ensemble outperformed individual models, highlighting the efficacy of combining diverse algorithms. SHAP analysis provided valuable insights into feature contributions, enhancing the model's transparency.

The AR module offers a novel approach to visualizing predictive outcomes, facilitating better understanding and decision-making for stakeholders. By simulating passenger experiences and risk scenarios, the AR tool serves as both an educational and operational asset.

Future work may involve real-time data integration, expanding the AR module's capabilities, and exploring the model's applicability to other domains within the aerospace industry.

VII. CONCLUSION

This study presents a hybrid ML pipeline integrating ensemble modeling, SHAP, and NLP with AR simulation to enhance space tourism safety and prediction transparency. Our results highlight the effectiveness of a multimodal, explainable AI approach and demonstrate practical applications for immersive safety visualization.

VIII. FUTURE WORK

While the proposed hybrid machine learning and AR-integrated pipeline shows strong performance and interpretability, several areas remain open for enhancement:

A. Real-World Dataset Integration

The Spaceship Titanic dataset is synthetic. Future iterations of this work should aim to collaborate with aerospace agencies or commercial space tourism companies to access anonymized real passenger or simulation data for real-world validation.

B. Time-Series and Dynamic Modeling

Incorporating time-based data such as passenger health vitals over time, service usage logs, or cabin environment conditions could enhance predictive power using recurrent models like LSTMs or transformers tailored for time-series.

C. Real-Time AR Feedback Systems

Extending the AR component to provide live mission analytics—including real-time anomaly alerts or passenger guidance—can shift the simulation from passive training to active safety assistance during actual flights.

D. Ethical AI and Bias Monitoring

Future development must integrate ethical auditing tools to detect bias based on demographics or socioeconomic markers. This includes using fairness metrics and differential impact analysis to ensure that safety predictions are equitable and inclusive.

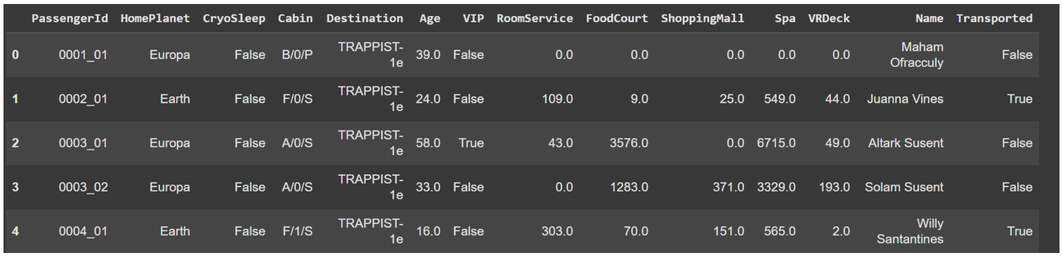
E. Multi-User Simulation Environments

A collaborative AR training platform for crew and passengers could simulate coordinated actions during emergency events, improving preparedness and mission success rates.

IX. ACKNOWLEDGMENT

We would like to express our sincere gratitude to *Prof. Kapil Sharma*, Department of Information Technology, Delhi Technological University, for his invaluable guidance, insightful feedback, and continuous support throughout this project. His mentorship was instrumental in refining our research approach and methodology. We also acknowledge the resources and academic environment provided by Delhi Technological University, which facilitated the successful completion of this work.

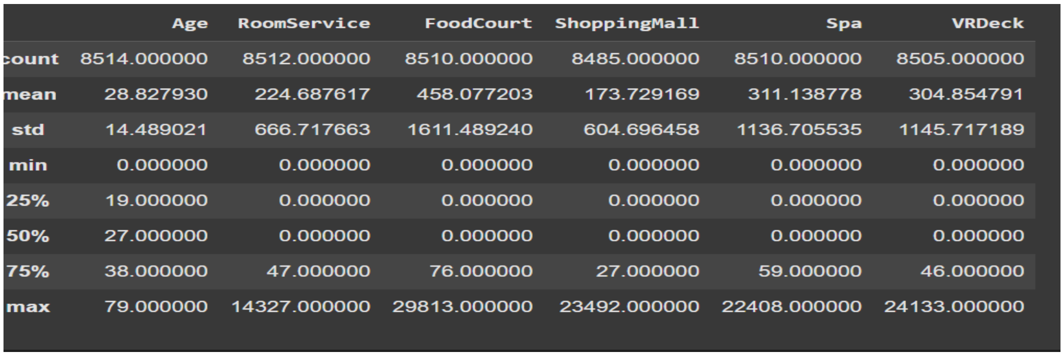
X. APPENDIX: FIGURES AND VISUALIZATIONS



	PassengerId	HomePlanet	CryoSleep	Cabin	Destination	Age	VIP	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck	Name	Transported
0	0001_01	Europa	False	B/0/P	TRAPPIST-1e	39.0	False	0.0	0.0	0.0	0.0	0.0	Maham Ofracculy	False
1	0002_01	Earth	False	F/0/S	TRAPPIST-1e	24.0	False	109.0	9.0	25.0	549.0	44.0	Juanna Vines	True
2	0003_01	Europa	False	A/0/S	TRAPPIST-1e	58.0	True	43.0	3576.0	0.0	6715.0	49.0	Altark Susent	False
3	0003_02	Europa	False	A/0/S	TRAPPIST-1e	33.0	False	0.0	1283.0	371.0	3329.0	193.0	Solam Susent	False
4	0004_01	Earth	False	F/1/S	TRAPPIST-1e	16.0	False	303.0	70.0	151.0	565.0	2.0	Willy Santantines	True

Figure 1. Sample of Passenger Data

This figure shows a snapshot of the dataset containing features such as PassengerId, HomePlanet, CryoSleep, and spending behavior. These attributes form the basis for survival prediction modeling.



	Age	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck
count	8514.000000	8512.000000	8510.000000	8485.000000	8510.000000	8505.000000
mean	28.827930	224.687617	458.077203	173.729169	311.138778	304.854791
std	14.489021	666.717663	1611.489240	604.696458	1136.705535	1145.717189
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	19.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	27.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	38.000000	47.000000	76.000000	27.000000	59.000000	46.000000
max	79.000000	14327.000000	29813.000000	23492.000000	22408.000000	24133.000000

Figure 2. Summary Statistics of Numerical Features

A statistical summary of key numerical variables like Age, RoomService, FoodCourt, and TotalSpending. These summaries inform imputation strategies and feature scaling.

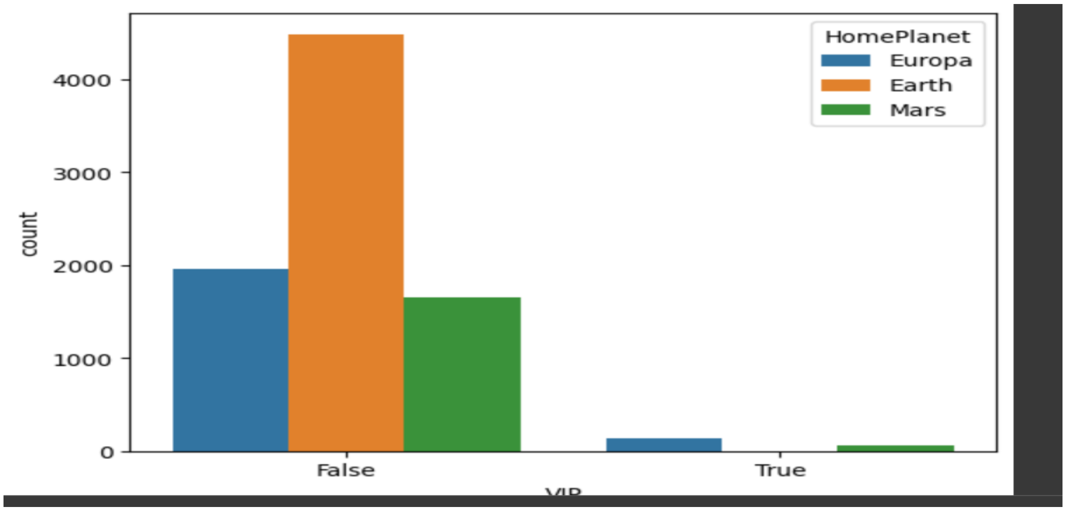


Figure 3. Distribution of VIP Status by HomePlanet

A bar chart showing the relationship between VIP passengers and their home planets, indicating socioeconomic or demographic clusters.

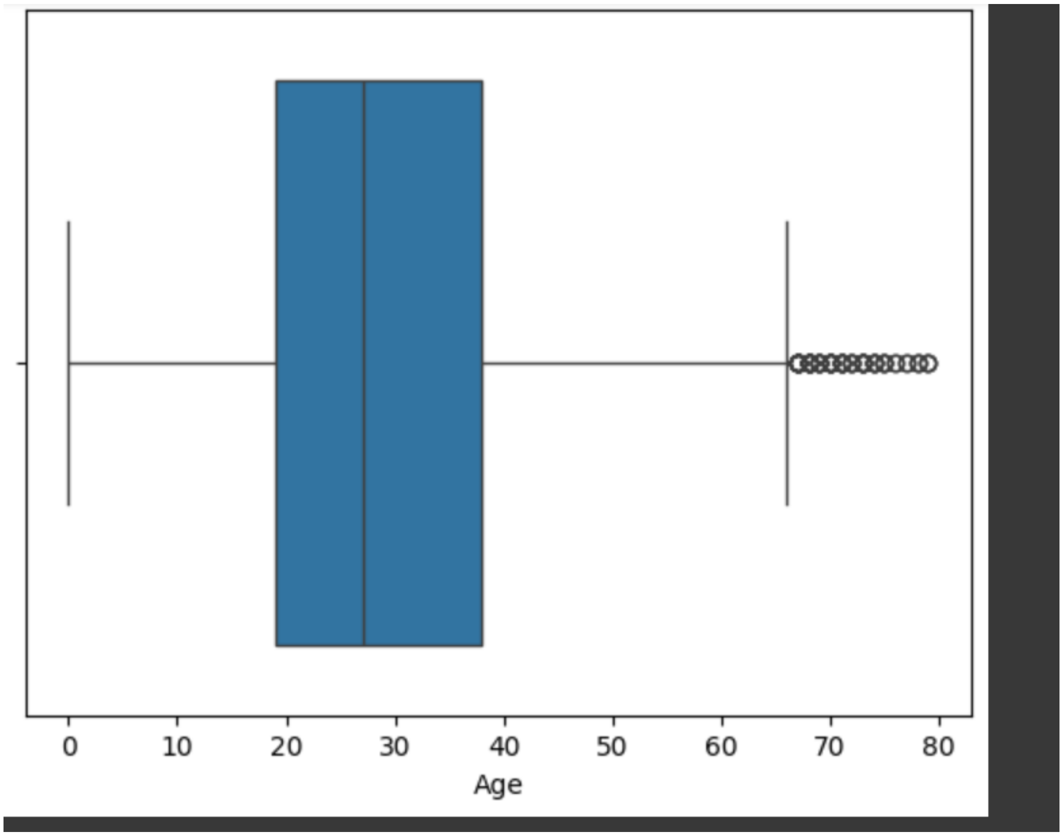


Figure 4. Boxplot of Passenger Age

This visualization helps identify the age distribution across passengers, highlighting outliers and skewness, and informing binning strategies.

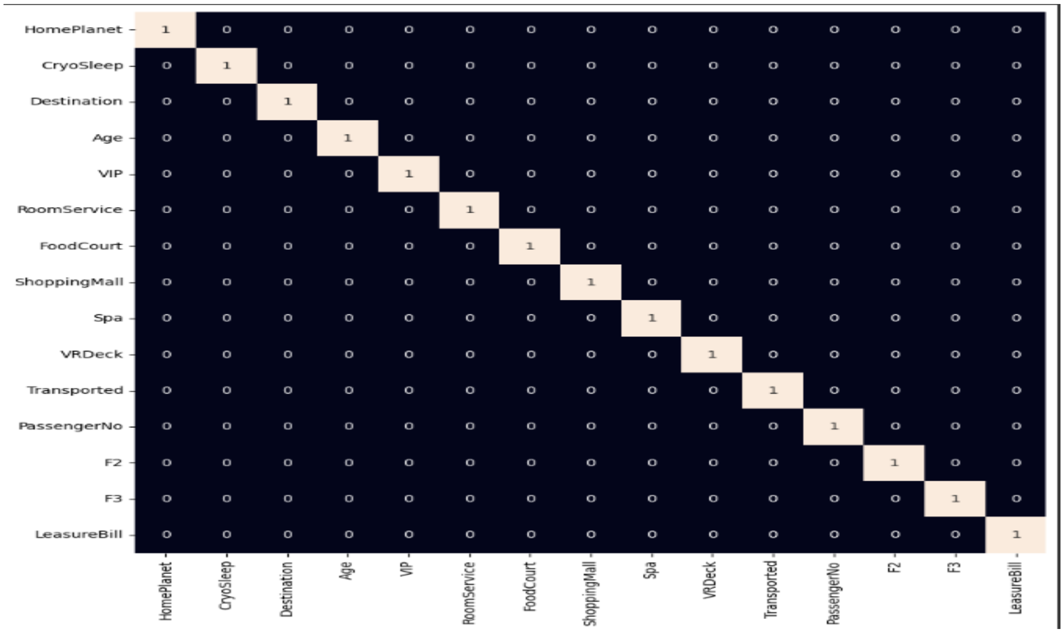


Figure 5. Correlation Matrix of Dataset Features

A heatmap showing Pearson correlation among numeric features. It reveals multicollinearity, feature interactions, and redundancy.

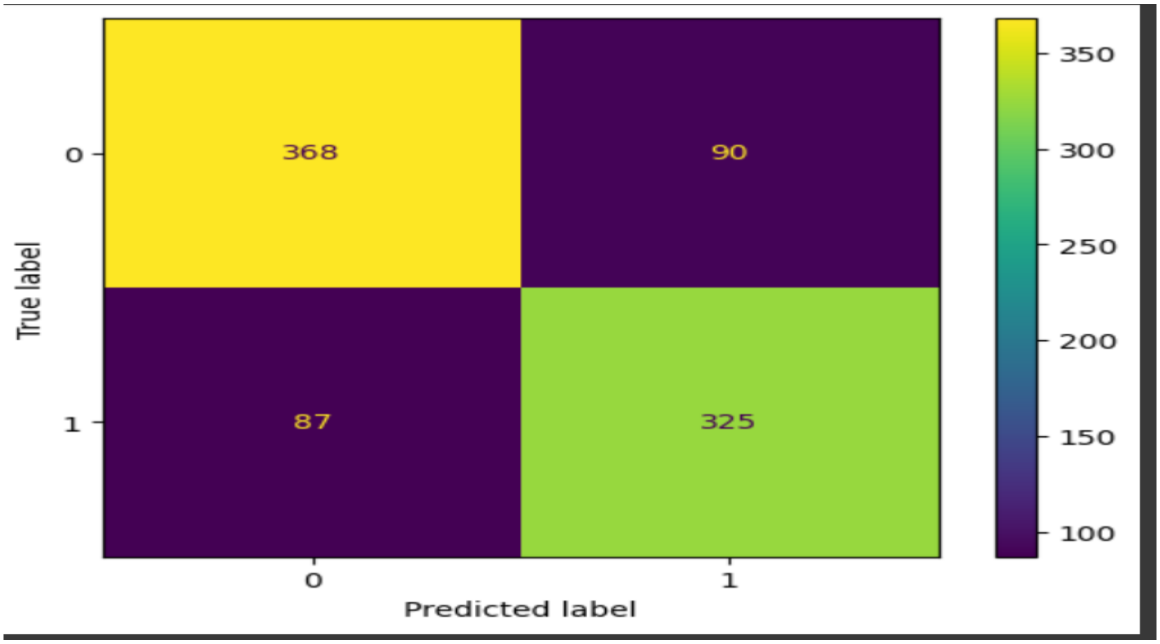


Figure 6. Confusion Matrix for Stacked Model

This confusion matrix summarizes the model's classification performance, showing true positives (325), true negatives (368), false positives (90), and false negatives (87). It highlights good separation between the 'Transported' and 'Not Transported' classes, with a slight imbalance in false predictions.

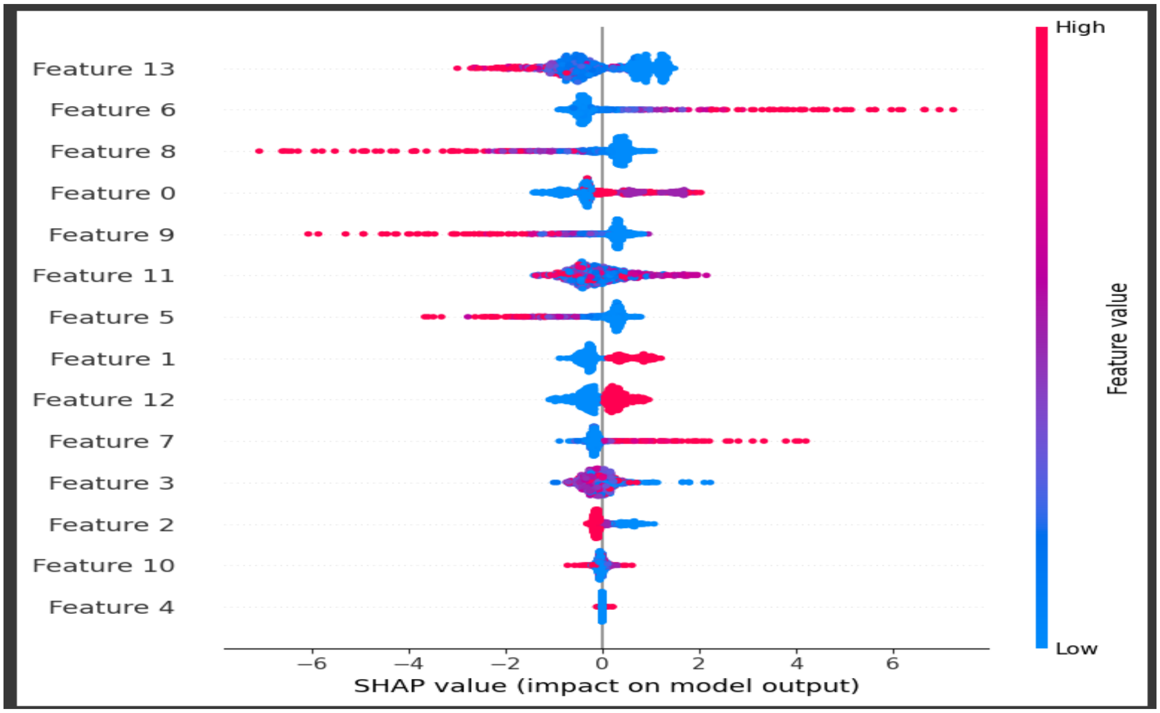
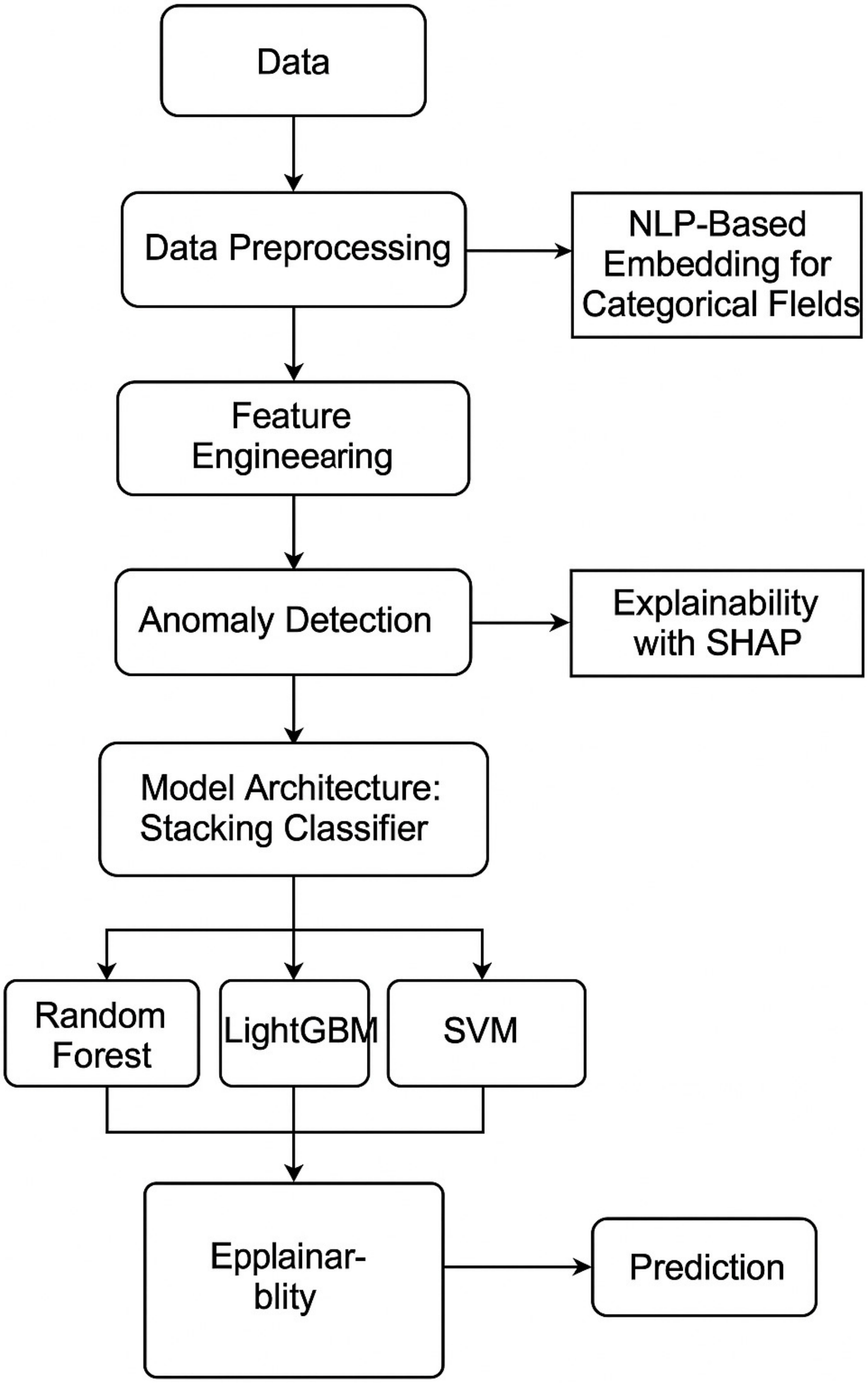


Figure 7. SHAP Summary Plot for Feature Importance

This SHAP summary plot visualizes the impact of each feature on the model's predictions. Features are ranked by importance, and color gradients represent feature values. The spread along the X-axis indicates the extent to which each feature influences the model output.



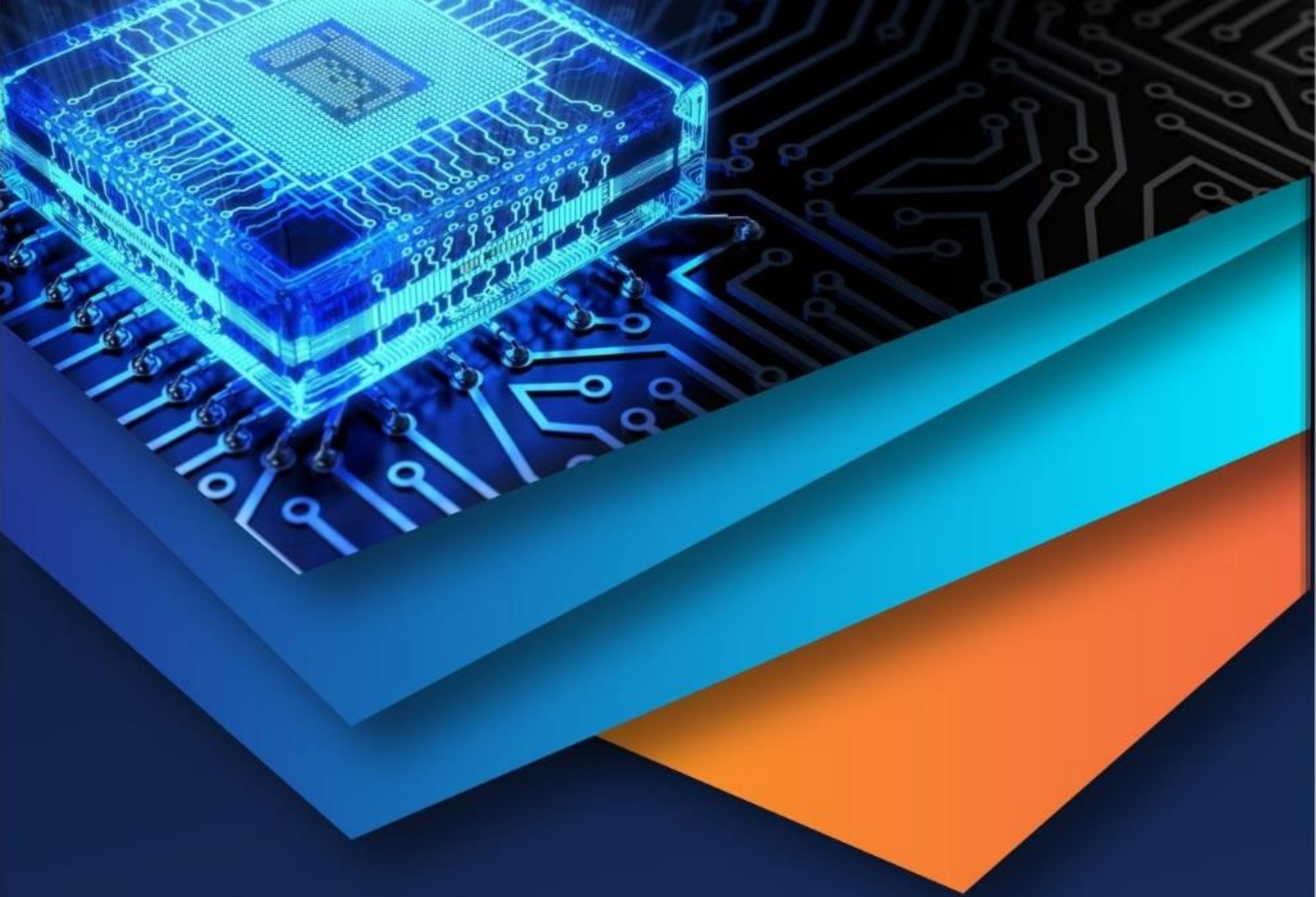
visual diagram of the machine learning pipeline

REFERENCES

[1] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794.

[2] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” Adv. Neural Inf. Process. Syst., vol. 30, pp. 4765–4774, 2017.

[3] A. Vaswani et al., “Attention is all you need,” in Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.



10.22214/IJRASET

INDEX COPERNICUS
45.98

ISRA JIF
IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 ✆ (24*7 Support on Whatsapp)