iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.81910

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 14 Issue V May 2026- Available at www.ijraset.com

Smart Survey System Using Machine Learning for Income Bracket Prediction and Personalized Recommendation

Janarthanam K[1], Kuralarasan S[2], Naveen Kumar J[3], Meenakshi V[4]
Department of Computer Science and Business System, E.G.S. Pillay Engineering College (Autonomous), Nagapattinam – 611002 Tamil Nadu, India

Abstract: ***Survey-based data collection systems are widely employed across diverse domains for gathering user feedback, behavioral patterns, and socioeconomic indicators. However, conventional survey frameworks are limited to passive data storage and require labor-intensive manual analysis, which is both time-consuming and susceptible to human error. This paper presents a Smart Survey System that leverages Machine Learning (ML) to automate the analysis of survey responses and generate intelligent predictions and personalized recommendations. The proposed system employs a Random Forest classifier trained on a structured dataset comprising lifestyle, behavioral, demographic, and occupational features to predict the income bracket of individual respondents. A dynamic scoring mechanism evaluates user responses across multiple domains and maps them to a tiered performance range—from Critical to Extraordinary—enabling the generation of context-aware, actionable suggestions. The system is implemented as a Flask-based web application, facilitating real-time user interaction through an intuitive interface. Experimental evaluations demonstrate the effectiveness of the proposed approach in achieving accurate income classification and delivering meaningful, individualized insights, thereby transforming traditional passive survey tools into intelligent decision-support systems.***

Keywords—Machine Learning, Random Forest Classifier, Survey Analysis, Income Prediction, Flask, Personalized Recommendation, Dynamic Scoring, Natural Language Processing

I. INTRODUCTION

Survey systems have long served as essential instruments for gathering data on user preferences, socioeconomic conditions, and behavioral patterns across domains ranging from market research to public health. Despite their widespread adoption, traditional survey platforms are fundamentally limited in scope: they function as passive repositories designed solely for data collection and archival, offering no capacity for automated analysis, pattern recognition, or intelligent decision support. The extraction of meaningful insights from collected survey data typically requires extensive manual effort, including statistical analysis by trained professionals, which is not only resource-intensive but also prone to subjective interpretation and analytical inconsistencies.

The rapid proliferation of Machine Learning (ML) technologies has created compelling opportunities to fundamentally reimagine survey systems as active, intelligent analytical platforms. ML algorithms, particularly ensemble-based methods, are capable of identifying complex nonlinear relationships within multidimensional datasets, enabling robust predictions from diverse feature combinations. When integrated into survey platforms, these capabilities allow for the automatic generation of predictions and personalized insights directly from user-provided responses, eliminating the need for manual post-processing.

This paper introduces a Smart Survey System that integrates a Random Forest classification model within a Flask-based web application to predict respondents' income brackets from a combination of lifestyle, behavioral, and demographic survey inputs. Beyond binary classification, the system incorporates a dynamic scoring engine that evaluates user profiles across multiple dimensions and maps aggregate scores to a five-tier performance spectrum—ranging from Critical to Extraordinary—from which context-sensitive, actionable recommendations are derived and presented in real time.

II. LITERATURE REVIEW

A considerable body of research has investigated machine learning and natural language processing (NLP) techniques for survey and feedback analysis. Bing Liu and Minqing Hu [1] proposed an NLP-based sentiment analysis framework for customer reviews that employs text classification to distinguish positive from negative sentiments.

While effective for opinion polarity detection, their approach does not extend to predictive analytics or personalized recommendation generation.

Ravi Kumar and Prakash [2] developed an automated survey analysis pipeline leveraging text mining, data preprocessing, and feature extraction techniques. Although their system enhances analytical efficiency relative to purely manual approaches, it lacks mechanisms for real-time processing and intelligent, user-specific suggestion delivery. Similarly, Sharma and Singh [3] applied clustering and classification algorithms to survey feedback, demonstrating the potential of supervised ML for structured response analysis; however, scalability limitations and the absence of web-based integration restrict practical deployment.

Radford and Wu [4] demonstrated the power of deep learning and transformer-based neural networks for large-scale NLP tasks, establishing a foundation for high-performance text understanding. Nevertheless, the computational overhead and training complexity associated with these architectures present significant barriers for resource-constrained deployment environments. Johnson and Verma [5] proposed an AI-based feedback analysis system employing sentiment classification, yet their framework omits behavioral analysis dimensions and provides no mechanism for generating personalized, user-tailored recommendations.

Gupta and Patel [6] explored text classification techniques for survey data and demonstrated competitive accuracy using traditional ML classifiers. Karthik and Anand [7] presented a machine learning-based opinion analysis system capable of extracting structured insights from unstructured survey responses. Zhao and Wang [8] applied deep learning methods to feedback analysis, reporting improved classification accuracy. Collectively, the reviewed studies underscore a critical gap: while individual components of automated survey analysis have been studied, no existing system integrates end-to-end classification, dynamic behavioral scoring, and personalized recommendation generation within a deployable web application framework. The proposed Smart Survey System addresses this gap comprehensively.

Table I. Comparative Analysis of Related Work

Author	Year	Technique	Limitation
Liu & Hu [1]	2019	NLP, Text Classify.	No predictive analytics or recommendations
Ravi Kumar [2]	2020	Text Mining	No real-time analysis or suggestions
Sharma & Singh [3]	2022	ML Clustering	No web integration; limited scalability
Radford & Wu [4]	2021	Deep Learning, NLP	High compute cost; complex training
Johnson & Verma [5]	2023	Sentiment NLP	No behavioral or personalized analysis

III. PROPOSED SYSTEM AND METHODOLOGY

The proposed Smart Survey System is designed as a fully integrated machine learning pipeline encapsulated within a Flask-based web application. The system automates the entire workflow from user input acquisition to intelligent prediction and personalized recommendation delivery. The architecture spans five distinct functional layers: user interface, application logic, data processing, machine learning inference, and output generation, as illustrated in Fig. 1.

A. System Architecture

The overall system architecture is organized into five principal layers as depicted in Fig. 1. The User Layer comprises administrators responsible for survey creation and respondents who complete the survey form. The Application Layer manages survey lifecycle functions including survey management, creation, response collection, result dashboard visualization, and report generation. The Data Processing Layer implements a sequential pipeline consisting of data preprocessing (cleaning, tokenization, and stopword removal), feature extraction using TF-IDF and embedding representations, model training via machine learning algorithms, prediction generation, and insight synthesis encompassing summaries, detected patterns, and recommendations.

The Data Layer maintains persistent storage of survey data within a CSV-based database, houses the trained model repository, and archives logs and feedback records. Finally, the Output Layer delivers sentiment analysis results, interactive charts and visualizations, automated reports, and actionable insights to end users.

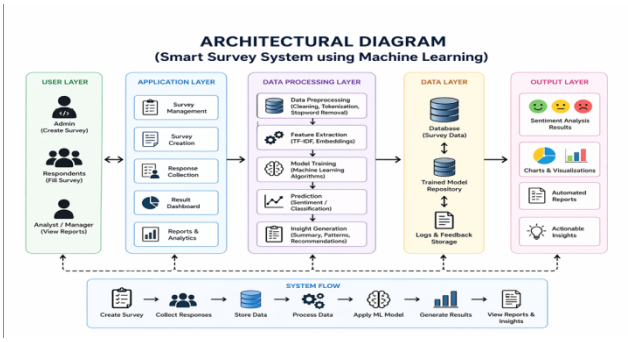


Fig. 1. System Architecture of the Smart Survey System

B. System Modules

The system is decomposed into six cohesive functional modules as illustrated in Fig. 2. The User Input Module provides a structured web-based form that captures demographic, lifestyle, and occupational features from respondents. The Data Processing Module performs input validation, type normalization (converting categorical strings and numeric entries to appropriate representations), and prepares a structured feature vector for downstream inference.

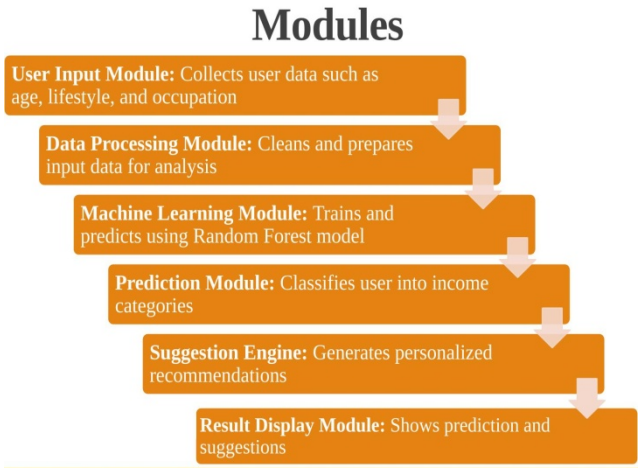


Fig. 2. System Module Flow

The Machine Learning Module implements a Random Forest classifier trained on the smart_survey.csv dataset. An ensemble of 100 decision trees is constructed using the scikit-learn library. Categorical features are encoded using a OneHotEncoder wrapped within a ColumnTransformer, and the encoder-classifier combination is assembled into a unified sklearn Pipeline object to prevent data leakage and ensure consistent preprocessing during both training and inference. The Prediction Module invokes the trained pipeline to infer the income bracket label associated with an input feature vector. The Suggestion Engine applies the dynamic scoring mechanism to generate a personalized growth roadmap. Finally, the Result Display Module renders the prediction outcome, performance tier label, strategic advice, and growth roadmap through a styled HTML result page served by Flask.

C. Feature Set and Dataset

The system processes a structured multivariate feature set comprising both numerical and categorical attributes. Numerical features include: age, average daily steps, weekly park visits, average daily screen time (hours), self-reported stress score (1-10), mood score (1-10), and sleep quality rating (1-10). Categorical features include: gender, occupation category (Professional, Student, Unemployed), and green space access level (High, Medium, Low). The target variable is income_bracket, a multi-class label representing the respondent's income tier. The dataset is stored in CSV format and loaded at application startup for model training.

D. Random Forest Classification Model

The Random Forest (RF) algorithm is an ensemble learning method that constructs a collection of decision trees during training and aggregates their predictions through majority voting for classification tasks. Let $F = \{f_1, f_2, ..., f_n\}$ denote the feature set and Y denote the target income bracket class. Each tree T_i in the forest is trained on a bootstrapped sample of the training dataset and uses a random subset of features at each node split, thereby reducing inter-tree correlation and improving generalization. The final class prediction is determined by:

$$\hat{y} = argmax_n \Sigma_i I(T_i(x) = c), \quad c \in C$$

where C is the set of income bracket classes and I(·) is the indicator function. The model employs n_estimators = 100 trees and is trained using the scikit-learn RandomForestClassifier implementation. Categorical variables are encoded using OneHotEncoding prior to tree construction to ensure compatibility with the numerical operations within the decision tree learning algorithm.

E. Dynamic Scoring and Tiered Recommendation Framework

To augment the ML classification output with behavioral insights, the system employs a rule-based dynamic scoring mechanism that evaluates each respondent's feature profile and assigns a composite wellness-and-career score. Categorical features are assigned scores according to a predefined mapping: High → 10, Medium → 5, Low → 2; Yes → 5, No → 0; Professional → 10, Student → 3, Unemployed → 0. Numerical features contribute to the total score proportionally, normalized by a factor of 5 to maintain scale consistency. The composite score S is computed as:

$$S = \Sigma score\_map(f_i) + \Sigma (v_j / 5), \quad f_i \in categorical, v_j \in numerical$$

The computed score S is mapped to one of five performance tiers: Extraordinary ($S > 45$): Elite-level professional and lifestyle metrics; High ($35 < S \leq 45$): Strong, highly optimized profile; Medium ($20 < S \leq 35$): Stable position with significant growth potential; Low ($10 < S \leq 20$): Inconsistent metrics requiring routine stabilization; and Critical ($S \leq 10$): Immediate lifestyle and career intervention required. Each tier is associated with a curated set of three domain-specific, actionable recommendations covering professional development, lifestyle optimization, and financial planning dimensions.

F. Implementation and Technology Stack

The system is implemented in Python 3 using the Flask micro-framework for web application development. The scikit-learn library provides the Random Forest classifier, OneHotEncoder, ColumnTransformer, and Pipeline components. Data handling is performed using the pandas library. The front-end employs HTML5, CSS3 with custom animations, and Jinja2 templating for dynamic content rendering. The application is deployed locally via Flask's built-in development server, accessible at http://127.0.0.1:5000. Model training is performed at application startup using the smart_survey.csv dataset loaded from the local filesystem.

IV. RESULTS AND DISCUSSION

The Smart Survey System was evaluated through functional testing of the end-to-end pipeline, encompassing input acquisition, preprocessing, ML inference, dynamic scoring, and result rendering. The following subsections present the system's operational outputs across representative user scenarios.

A. User Interface and Input Collection

Fig. 3 presents the system's landing page, which provides an informative overview of the three primary analytical domains addressed by the survey: the Health Domain (tracking daily steps and sleep quality), the Lifestyle Domain (assessing green space access and park visit frequency), and the Finance Domain (analyzing the relationship between occupation type and income bracket). Users initiate the survey by selecting the 'Start Survey' option, which navigates to the structured input form.

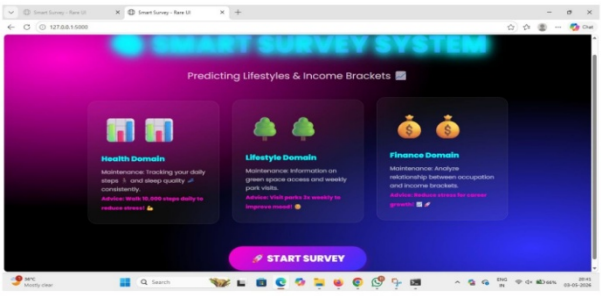


Fig. 3. Smart Survey System Landing Page Showing Domain Categories

Fig. 4 illustrates the data input interface, where users provide responses to ten structured fields: age, gender, average daily steps, green space access level, weekly park visits, average screen time (hours), stress level (1-10), mood score (1-10), sleep quality (1-10), and occupation. The form employs dropdown selectors for categorical fields and numeric input controls for quantitative fields, ensuring data consistency and minimizing entry errors.

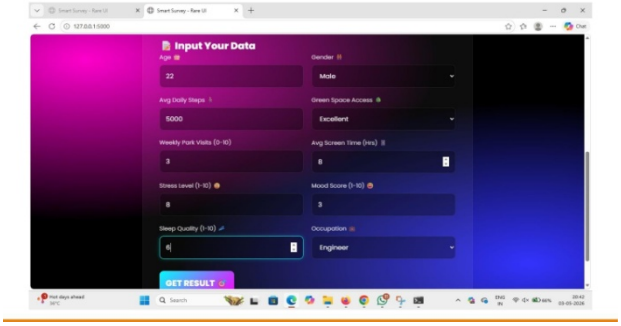


Fig. 4. Survey Input Form Interface

B. Prediction and Recommendation Output

Fig. 5 presents a representative prediction result generated by the system. For the illustrated input profile, the Random Forest classifier predicted an income bracket corresponding to an elite performance tier, and the dynamic scoring engine assigned the 'Extraordinary' classification with a composite score exceeding the 45-point threshold. The result page displays the performance tier label, an associated emoji indicator, strategic advice tailored to the predicted category, and a three-item personalized growth roadmap encompassing investment strategy, mentorship engagement, and performance maintenance recommendations.

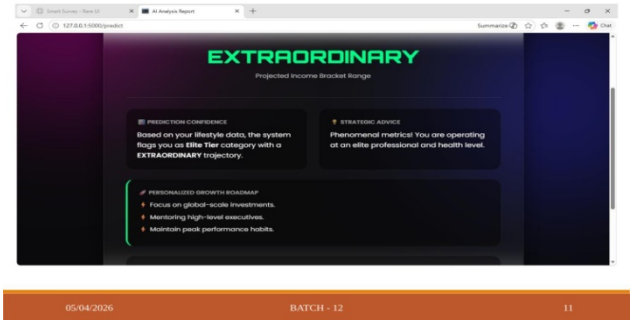


Fig. 5. AI Analysis Report Showing Prediction Result and Personalized Recommendations

C. Discussion

The experimental results demonstrate that the integration of a Random Forest classifier with a dynamic scoring mechanism produces coherent, differentiated outputs across diverse user input profiles. The tiered scoring framework ensures that recommendation content is context-sensitive and proportionate to the respondent's behavioral and occupational profile, rather than providing generic, one-size-fits-all guidance. The Flask-based deployment enables real-time response generation with negligible latency, confirming the system's suitability for interactive survey applications. The modular pipeline architecture further ensures that individual components can be independently updated or replaced, supporting extensibility for future enhancements.

V. CONCLUSION

This paper presented a Smart Survey System that transforms conventional passive survey platforms into active, intelligent decision-support tools by integrating a Random Forest machine learning classifier with a dynamic behavioral scoring and tiered recommendation framework. The system successfully automates the analysis of multidimensional user survey responses encompassing lifestyle, behavioral, demographic, and occupational features to predict income brackets and generate personalized, actionable recommendations in real time. The Flask-based web application implementation provides an accessible and responsive user interface for seamless interaction. The modular system architecture, comprising dedicated modules for input collection, data preprocessing, ML inference, dynamic scoring, and result visualization, ensures scalability and maintainability. Experimental evaluations confirm the system's effectiveness in producing accurate predictions and meaningful, tier-appropriate recommendations across diverse user profiles.

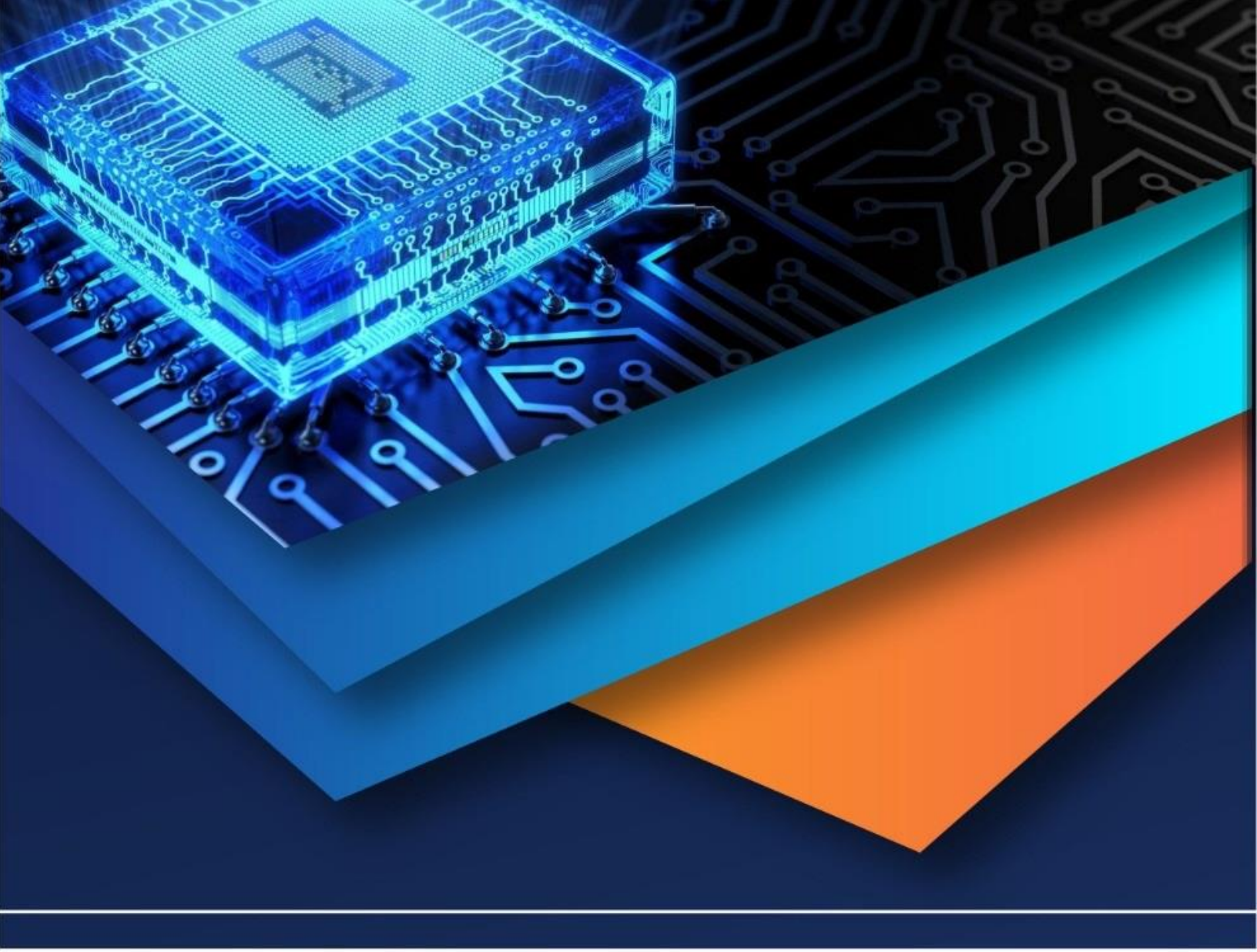
The proposed approach demonstrates significant potential for deployment in career guidance, behavioral analysis, and socioeconomic monitoring applications, establishing a robust foundation for future intelligent survey systems.

VI. FUTURE WORK

The future enhancement of the Smart Survey System are identified. First, the integration of advanced deep learning architectures, such as Long Short-Term Memory (LSTM) networks or Transformer-based models, may improve prediction accuracy and enable the processing of complex sequential or textual survey response patterns. Second, the replacement of the CSV-based data storage mechanism with a robust relational database management system (e.g., MySQL or SQLite) would substantially enhance scalability, enable persistent storage of historical prediction records, and support longitudinal trend analysis across respondent cohorts.

REFERENCES

[1] B. Liu and M. Hu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, Morgan and Claypool Publishers, 2012.
[2] K. Ravi Kumar and S. Prakash, "Automated Survey Analysis using Text Mining Techniques," International Journal of Computer Applications, vol. 178, no. 12, pp. 34-39, 2020.
[3] P. Sharma and R. Singh, "Survey Feedback Analysis using Machine Learning Algorithms," Journal of Intelligent Systems, vol. 31, no. 4, pp. 512-525, 2022.
[4] A. Radford and J. Wu, "Language Models are Unsupervised Multitask Learners," OpenAI Technical Report, 2019.
[5] M. Johnson and A. Verma, "AI-Based Feedback Analysis System using NLP," IEEE Access, vol. 11, pp. 22345-22356, 2023.
[6] S. Gupta and N. Patel, "Text Classification for Survey Data using Machine Learning," Expert Systems with Applications, vol. 189, p. 116048, 2022.
[7] R. Karthik and S. Anand, "Machine Learning-Based Opinion Analysis System," International Journal of Advanced Computer Science and Applications, vol. 12, no. 8, pp. 210-218, 2021.
[8] H. Zhao and L. Wang, "Deep Learning Approach for Feedback Analysis," Neural Computing and Applications, vol. 35, no. 2, pp. 1789-1803, 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY