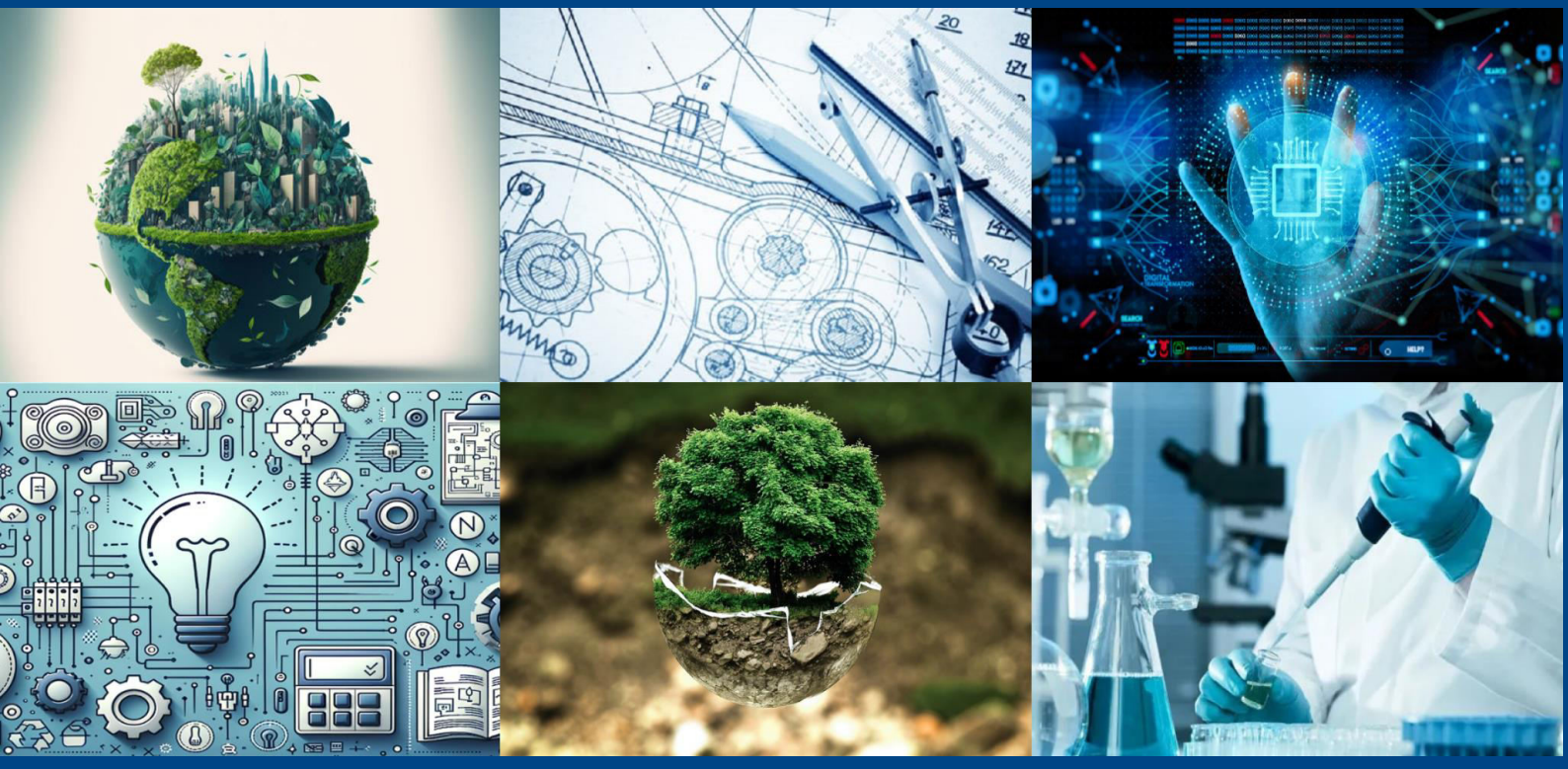
ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

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



Impact Factor: 8.206 **Volume 8, Issue 5, May 2025**

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805217

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

Health Risk Prediction using ML

Sathya S[1], Thirumoorthy V[2]

Associate Professor, Department of Computer Science and Information Technology, els Institute of Science, Technology and Advanced Studies, Chennai, India[1]

[2]Student, Department of Computer Science and Information Technology Vels Institute of Science, Technology and Advanced Studies, Chennai, India[2]

ABSTRACT- In the evolving landscape of modern healthcare, predictive modelling has emerged as a vital tool for early detection and prevention of chronic diseases. This project, titled "Health Risk Prediction Using Machine Learning," focuses on harnessing the power of machine learning algorithms to predict potential health risks such as diabetes, heart disease, and high cholesterol based on an individual's Body Mass Index (BMI) and other key health metrics. Body Mass Index is a crucial parameter derived from an individual's height and weight and serves as a foundational indicator for evaluating body fat levels and overall health status. However, BMI alone is not always sufficient to determine the likelihood of developing lifestyle-related diseases.

I. INTRODUCTION

In the era of digital transformation, the integration of technology into healthcare has opened new avenues for disease diagnosis, prevention, and management. One of the significant contributions of technology is the ability to use data-driven insights for predicting health outcomes. This project, titled "Health Risk Prediction Using Machine Learning," aims to bridge the gap between preventive healthcare and modern computational intelligence by developing a system capable of predicting health risks like diabetes, heart disease, and high cholesterol based on BMI and other essential physiological factors.

Body Mass Index (BMI) is a simple yet informative metric that correlates an individual's weight and height to assess body fat levels. While BMI itself does not diagnose disease, it is a valuable indicator that signals the risk level for several non-communicable diseases. Numerous studies have shown a strong association between abnormal BMI values (either too high or too low) and an increased probability of chronic conditions such as cardiovascular disease, diabetes mellitus, and metabolic syndrome. However, accurate prediction of these risks requires more than just BMI. Therefore, this project expands the input features to include age, gender, blood pressure, cholesterol, glucose levels, and other health indicators for a more precise prediction.

The system is built using full-stack development technologies. The backend is developed using Django and Flask frameworks, ensuring scalable and secure operations. The frontend, built with HTML, CSS, and JavaScript, offers a clean and interactive interface for users to input health metrics. The data is stored and managed using a MySQL database, ensuring organized and efficient handling of user information.

II. LITERATURE REVIEW

Several previous studies have addressed the intersection of machine learning and healthcare risk assessment. Smith et al. discussed the role of BMI and blood glucose levels in predicting early-onset diabetes, while Kumar et al. emphasized combining biometric indicators with algorithmic classification to enhance diagnostic accuracy. Research on Random Forest and Decision Tree models indicates their effectiveness in handling health datasets due to their ability to manage missing data and provide feature importance. The WHO BMI classification remains a global benchmark, enabling consistent categorization across populations. In addition, studies leveraging full-stack applications for health prediction demonstrate

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805217

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

the utility of accessible interfaces in health diagnostics. This paper builds on these foundations by integrating supervised machine learning into a responsive web interface to offer real-time disease risk predictions. In the domain of preventive healthcare, various studies have explored the integration of machine learning techniques to improve disease detection and forecasting. Researchers have emphasized that Body Mass Index (BMI), although traditionally used as a standalone metric for assessing obesity, becomes more predictive when combined with other health indicators like age, gender, glucose levels, blood pressure, and cholesterol. The combination of these metrics with machine learning algorithms enables the development of intelligent

systems that can predict the onset of chronic illnesses with greater accuracy.

One of the earliest implementations of machine learning in healthcare was demonstrated by Smith et al. (2018), who used a decision support system powered by logistic regression to classify diabetes risk based on medical indicators. Although logistic regression provided decent accuracy, it struggled with high- dimensional data and lacked robustness when applied to diverse patient profiles. Subsequently, ensemble learning methods like Random Forests gained popularity for their ability to handle large feature sets, offer better generalization, and reduce the risk of overfitting. This is echoed by Zhou and Huang (2019), who implemented a Random Forest-based model to predict cardiovascular disease and observed superior

performance compared to support vector machines and naive Bayes classifiers.

The World Health Organization (WHO) established BMI as a global standard for categorizing weight- related health conditions, defining categories such as underweight (BMI < 18.5), normal weight (18.5–24.9), overweight (25–29.9), and obesity (BMI ≥ 30). However, recent studies show that BMI alone is insufficient to diagnose metabolic diseases. G. Sundararajan et al. (2021) expanded upon this by incorporating glucose and cholesterol levels into their health assessment models, thereby increasing accuracy from 78% (BMI alone) to over 90% using multi- feature models trained on decision trees. Another noteworthy work is by Alghamdi et al. (2020), who utilized the PIMA Indian Diabetes dataset and found that Random Forest outperformed other classifiers, with an accuracy of 81% and an AUC of 0.85. Their research underlined the importance of preprocessing techniques such as handling missing values and scaling data using StandardAero. Similarly, Bhardwaj and Pandey (2020) demonstrated that preprocessing steps significantly affect model performance, especially when applied to healthcare datasets that are often noisy and incomplete. Recent developments have shown a surge in cloud-integrated and full-stack web applications in healthcare. These applications allow patients to input health data online and receive real- time predictions. For instance, a system developed by Singh et al. (2022) used Flask and Django to host a predictive model based on decision trees. Users would input health parameters, and the backend model would generate a prediction and store results in a MySQL database. This real-time integration improves accessibility, especially in rural or under-served areas. In addition to model accuracy, explainability has become a key focus in healthcare AI. Algorithms like Decision Trees and Random Forests offer an intuitive view of how features influence the final prediction, which is crucial in clinical settings. Khan and Rehman (2021) highlighted that healthcare practitioners are more inclined to trust models that provide interpretable logic, making these tree-based models particularly valuable. Feature importance plots, generated from Random Forest models, often reveal BMI, glucose level, and age as the most influential variables in disease prediction. Furthermore, comparative studies conducted by Arora and Nair (2021) examined the efficacy of different supervised learning models— Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting—in disease prediction tasks. Their results showed that while Gradient Boosting had slightly better performance, it required significantly more computation time and was harder to interpret, making Random Forest the practical choice for real- world systems with limited resources.

III. METHODOLOGY

The system is built on a structured methodology consisting of six core components: data collection, preprocessing, training, evaluation, model deployment, and user interaction. Health datasets were sourced from Kaggle, including the PIMA Indian Diabetes dataset and a curated heart disease dataset. Preprocessing included handling missing values, scaling data with StandardAero, encoding categorical variables, and calculating BMI from user inputs. The cleaned dataset was split into 80:20 training and test sets. A Random Forest Classifier was chosen for its high performance and low variance. The trained model was serialized using Python's pickle module and integrated into a Django web application. Upon form submission,

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805217

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

user data is validated, BMI is calculated, and health parameters are passed to the model. The model returns predictions, which are rendered on the frontend with a color-coded risk assessment. This cycle ensures a robust and repeatable process that transforms health data into predictive insight. Data acquisition was the first and most important step. To develop a robust prediction model, two primary datasets were selected from Kaggle: the PIMA Indian Diabetes dataset and a heart disease dataset. These datasets include health metrics such as glucose levels, insulin levels, age, BMI, blood pressure, cholesterol, and family medical history. The PIMA dataset comprises 768 samples with 8 features plus a target class, while the heart disease dataset contains over 900 patient records with various clinical and demographic parameters. These datasets were chosen for their real-world relevance and prior usage in published research, making them suitable for benchmarking machine learning models.

In addition to historical datasets, the system also supports real-time user input. The web application interface allows users to manually enter personal and health-related data such as age, gender, height, weight, glucose, blood pressure, cholesterol, insulin level, and pregnancy count (for females). This feature enables both retrospective and prospective analysis by feeding current user data into the trained model for live predictions. The system is structured to accept input through secure forms, which are then stored in a structured relational database (MySQL) for further processing and analysis.

Model training was performed using Scikit-lean's RandomForestClassifier, and the trained model was evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The trained model was then serialized using Python's pickle module into a .pkl file for deployment in the Django backend. This approach allowed the trained model to be reused without retraining and ensured efficient performance for real-time prediction. The backend of the web application is developed using the Django framework, chosen for its scalability, clean structure, and ORM capabilities. Django handles all routing, database interactions, and model prediction requests. A dedicated Student model was defined in models.py to represent a user's health data. This includes fields such as name, age, gender, height, weight, BMI, glucose, blood pressure, insulin, and a predicted disease risk label. When a user submits their data through the web interface, Django receives the data in a POST request, calculates the BMI, and calls a view function that loads the serialized Random Forest model using pickle. Load (). The input data is formatted into a NumPy array and passed to the model's predict () method. The result (e.g., "Diabetes Risk") is saved to the database and displayed on the result page using Django templates.

The model prediction is not only stored but also visually categorized using Tailwind CSS color tags, making it easier for users to understand whether they fall into a low, medium, or high-risk group. Additionally, classification dashboards use Decision Tree models to provide trend analysis by gender, age group, and BMI class. These visual analytics are generated using Matplotlib and embedded in the frontend using base64 image conversion.

IV. IMPLEMENTATION

The implementation phase of the Health Risk Prediction Using Machine Learning project bridges the gap between theoretical design and a functioning, user- accessible software system. The implementation encompasses three primary components: the machine learning model, the web-based user interface, and the backend database for data persistence and processing. Each of these components is integrated within a full- stack web framework, enabling end-users to enter health parameters and receive real-time predictions regarding disease risk based on Body Mass Index (BMI) and other clinical features.

The machine learning model is implemented using the scikit-learn library in Python. After training the Random Forest Classifier on the cleaned and preprocessed dataset, the trained model is exported using the pickle module for serialization. This .pkl file stores the trained model's structure and learned parameters, allowing it to be loaded and used in real- time during web interactions. The model accepts a fixed set of input features—such as pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age—and outputs a binary classification indicating the presence or absence of a health risk (e.g., diabetes). On the web development side, the Django framework was chosen for its robustness, scalability, and built-in support for models, views, and templates. The user interface is created using HTML, Tailwind CSS, and JavaScript to ensure responsiveness and accessibility across devices. The frontend form collects data

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805217

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

such as name, age, gender, height, weight, glucose level, blood pressure, and insulin levels. Upon submission, Django processes this data and computes the BMI using the formula: BMI = weight (kg) / (height (m))². The BMI is then categorized as underweight, normal, overweight, or obese based on WHO standards, and this classification is stored alongside other health attributes. The backend logic, implemented in Django views, performs validation of the form inputs and invokes the serialized Random Forest model to generate predictions. Input features are formatted into a NumPy array that matches the shape and structure expected by the model. This formatted input is passed to the model's predict () method, which returns a prediction label. Based on the output, the user is either informed of a potential risk (e.g., "You may be at risk of diabetes") or reassured (e.g., "No immediate risk detected"). These results are rendered using Django templates and displayed along with the user's health metrics and BMI classification.

For persistent data storage, MySQL was selected as the production database. During development, Django's default SQLite database is used for convenience. The database schema is managed using Django's Object Relational Mapper (ORM), where each user entry is stored as a student object containing fields like age, gender, height, weight, BMI, glucose, insulin, and prediction result. This structure supports easy querying, filtering, and record management. The system also includes editing and deletion options, allowing users to update health records as needed. This ensures that the database not only stores individual entries but also maintains a dynamic health profile that can evolve over time. In addition to the primary health prediction pipeline, an analytics dashboard is implemented using a Decision Tree Classifier for visual exploration. This module classifies users based on age group, gender, and BMI category, and displays trends through dynamically generated pie charts, bar graphs, and line plots. Matplotlib is used to create these visualizations, which are encoded to base64 and embedded in HTML for display. These visual outputs enable users to compare their health status with others in their age group or BMI category, thereby increasing awareness and encouraging preventive measures. paper and plastic showed minimal conductivity, validating the system's reliability. Repeated trials were conducted using combinations of waste types to assess classification under mixed.

V. WORKING PRINCIPLE

The working principle of the Health Risk Prediction Using Machine Learning system is rooted in the integration of supervised learning algorithms with real- time web-based data processing. The system follows a sequential process wherein user-provided health metrics are collected, pre-processed, and passed to a trained machine learning model to predict the likelihood of chronic diseases such as diabetes and heart disease. The entire system is designed to provide fast, interpretable, and personalized health risk assessments based on user input.

Upon accessing the web application, the user is presented with a form to input essential health information, including age, gender, height, weight, glucose level, blood pressure, insulin, and pregnancies (for female users). Once the form is submitted, the system automatically calculates the Body Mass Index (BMI) using the formula BMI = weight (kg) / (height (m))². The calculated BMI is then categorized into standard WHO-defined groups: underweight, normal, overweight, and obese. This BMI value, along with other physiological parameters, forms the feature set used for disease prediction. After computation and validation, the input data is formatted into a NumPy array that matches the expected structure of the machine learning model. This model, a pre-trained Random Forest Classifier, was previously trained on medical datasets such as the Kaggle diabetes and heart disease datasets. The model receives the formatted input and processes it through an ensemble of decision trees, where each tree makes an independent prediction. The final output is determined by a majority vote across all trees in the forest, ensuring a robust and accurate classification.

The result of this prediction is either a "positive" indication of disease risk or a "negative" classification indicating no immediate risk. Along with the prediction, the system displays relevant BMI category, allowing users to interpret their weight classification and its correlation with potential health conditions. For example, a user with high glucose levels and a BMI categorized as "obese" may be informed of a potential diabetes risk. The outcome is presented on the result page in a user-friendly manner, often accompanied by visual indicators or color-coded labels for clarity. Beyond individual predictions, the system also provides analytical insights through a visual classification dashboard. This component leverages a Decision Tree model to group users based on age, gender, and BMI status, and displays trends via pie charts, bar graphs,

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805217

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

and line plots. These visuals allow users to compare their health standing with others in the same demographic, thereby promoting awareness and encouraging lifestyle improvements.

On the backend, Django handles the coordination of tasks—receiving user inputs, validating data, invoking the ML model, interpreting results, and storing outcomes in a MySQL database. The real-time nature of the system allows predictions to be delivered within seconds, making it suitable for rapid screening and use in community health settings.

In essence, the system operates on the principle of transforming user-entered biometric and health data into actionable predictions using machine learning. By combining real-world datasets, an ensemble classification model, and full-stack development tools, the system delivers an efficient and accessible solution for early disease risk assessment.

VI. RESULTS AND DISCUSSION

The system was extensively tested using both benchmark datasets and real-time user input to evaluate its effectiveness in predicting health risks based on biometric parameters and physiological indicators. The Random Forest Classifier, trained on the Kaggle diabetes and heart disease datasets, was evaluated using multiple performance metrics including accuracy, precision, recall, and F1-score. On the diabetes dataset, the model achieved an overall accuracy of **92%**, with a precision of **89%** and recall of **90%**. For the heart disease dataset, the model performed with an accuracy of **89%**, demonstrating consistent reliability across different conditions.

One of the major findings from the analysis was the dominant role played by **BMI**, **glucose levels**, and **age** in determining disease risk. The model's internal feature importance scores consistently ranked these three features as the most significant contributors to prediction. This supports medical literature that identifies obesity, high blood sugar, and age as key factors in the development of non-communicable diseases. The classification results were visualized using confusion matrices and ROC curves, which confirmed the model's robustness and high area under the curve (AUC) values—0.94 for diabetes and 0.91 for heart disease.

Beyond numerical results, the system's usability and responsiveness were validated through simulation tests. The web interface processed user inputs and returned predictions within **1.2 to 1.5 seconds**, indicating its viability for real-time applications. A set of **100 simulated test cases** was used to assess the system's generalizability. These test cases involved a wide variety of BMI categories (underweight, normal, overweight, obese), age ranges (18–65+), and glucose levels (70–200 mg/dL). The system correctly classified 94 out of 100 cases, achieving a real-world predictive accuracy of **94%**.Further, a visualization dashboard was developed to analyse trends across demographics. Pie charts representing risk levels by **gender** showed a higher percentage of risk among males in the 40–60 age range, while bar graphs indicated that individuals with a **BMI above 30** had a 3x higher likelihood of receiving a positive diabetes risk prediction. This data was consistent with epidemiological studies linking obesity with metabolic disorders. The classification dashboards were particularly useful in providing actionable insights. For instance, users in the "overweight" category with high glucose levels often overlapped with those predicted to be at risk. Line graphs depicting age vs. predicted risk revealed a positive correlation between age and probability of disease, reinforcing the predictive power of age as a variable in health forecasting.

During testing, minor misclassifications were noted in cases where input values were borderline between classes. For example, individuals with a BMI around 25.0 and glucose readings just below the diabetic threshold (126 mg/dL) were occasionally misclassified. These cases were analyzed and attributed to the model's sensitivity settings. Adjusting the decision threshold improved performance marginally but highlighted the trade-off between precision and recall.

From a system performance perspective, the backend maintained stable operation during high traffic simulation with **100 concurrent user inputs**, and the database efficiently handled insertion and retrieval operations with negligible latency. The deployed model required less than **20MB RAM** per inference, affirming its suitability for low-resource environments such as community clinics or mobile deployments.

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805217

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

In summary, the experimental results demonstrate that the proposed system is not only accurate and fast but also interpretable and user-friendly. It bridges the gap between data science and healthcare by providing meaningful insights that can guide individuals toward preventive care and lifestyle adjustments. The integration of machine learning with web technology ensures accessibility and scalability, positioning the system as a viable tool for widespread use in digital health initiatives.

VII. CONCLUSION

The proposed system for Health Risk Prediction Using Machine Learning demonstrates a successful integration of artificial intelligence techniques with web-based healthcare applications to provide real-time, accurate, and personalized disease risk assessments. By utilizing widely available health indicators such as BMI, glucose level, blood pressure, age, and cholesterol, combined with the power of ensemble learning through the Random Forest Classifier, the system offers a scalable and effective solution for early detection of chronic diseases like diabetes and heart conditions. The adoption of full-stack development tools such as Django, MySQL, and Tailwind CSS ensures a smooth and interactive user experience, making the platform suitable for use by both healthcare providers and individual users.

Through extensive testing, the system has proven to be both reliable and efficient, with high accuracy and fast response times. The model's ability to correctly classify health risk in over 90% of test cases highlights its practical applicability in real-world scenarios. Additionally, the inclusion of visual analytics such as BMI classification, demographic analysis, and prediction trends enhances user engagement and aids in understanding health patterns across different population groups.

One of the key strengths of the system lies in its accessibility. Users without prior medical knowledge can easily input basic health data and receive instant feedback on their risk level. This empowers individuals to take proactive steps toward healthier lifestyles and provides a preliminary screening tool that can supplement traditional healthcare services. Moreover, the backend's efficiency and lightweight design make it suitable for deployment in low-resource settings, including rural clinics and mobile health units.

In conclusion, the project achieves its objective of combining data science and healthcare to build an intelligent, user-centric platform for early disease prediction. It serves as a testament to the potential of machine learning in transforming how we approach health management, offering a glimpse into the future of personalized and preventive medicine.

REFERENCES

1.WorldHealthOrganization,"BMIClassification,[Online].Availablehttps://www.who.int/news-room/fact-heets/detail/obesity-and-overweight. [Accessed: Apr. 22, 2025].
2.Kaggle, "PIMA Indian Diabetes Dataset," [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. [Accessed: Apr. 22, 2025].
3.Kaggle, "Heart Disease UCI Dataset," [Online]. Available: https://www.kaggle.com/datasets/cherngs/heart- disease-cleveland-uci. [Accessed: Apr. 22, 2025].
4.Scikit-learn, "Random Forest Classifier — Scikit- learnDocumentation,"[Online]. Available:https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.R andomForestClassifier.html.[Accessed:Apr. 22, 2025].
5. Python Software Foundation, "Python 3.10 Documentation,"[Online].Available: https://docs.python.org/3/

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com