IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 6, June 2024

DOI: 10.17148/IJARCCE.2024.13675

DIABETES PREDICTION USING MACHINE LEARNING

Dr. Kavyashree N[1], Ganga T A[2], Roopashree[3]

Assistant Professor, Department of MCA, SSIT, Tumkur, Karnataka, India[1]

4th Sem MCA, SSIT, Tumkur, Karnataka, India[2,3]

Abstract: The Diabetes Prediction App is a vital healthcare tool developed using Django, designed to predict the likelihood of an individual developing diabetes based on varioushealth parameters and risk factors. This project aims to provide users with an accessibleand user friendly platform to assess their risk of diabetes and take preventive measures accordingly. Leveraging Django's capabilities, the application utilizes machine learning algorithms to analyse user input data and generate personalized risk assessments, empowering individuals to make informed decisions about their health. Diabetes is a chronic disease with the potential to cause a worldwide health care crisis.According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetesis a disease caused due to the increase level of blood glucose. This high blood glucoseproduces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is a one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy.

Keywords: Early Diagnosis, Health metrics, Clinical Practice, Public Health

I. INTRODUCTION

The Diabetes Prediction App is a vital healthcare tool developed using Django, designed to predict the likelihood of an individual developing diabetes based on various health parameters and risk factors. This project aims to provide users with an accessible and user friendly platform to assess their risk of diabetes and take preventive measures accordingly. Leveraging Django's capabilities, the application utilizes machine learning algorithms to analyze user input data and generate personalized risk

assessments, empowering individuals to make informed decisions about their health. Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes is a disease caused due to the increase level of blood glucose. This high blood glucose produces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is a one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy. Diabetes is a chronic disease with significant health implications, affecting millions of people worldwide. Early detection and proactive management are crucial for preventing complications associated with diabetes.

The Diabetes Prediction App serves as a proactive healthcare solution, offering individuals a means to assess their risk of diabetes and take preventive actions. Built on Django's robust framework, this application provides a seamless and intuitive platform for users to input their health parameters and receive personalized risk assessments, thereby promoting early intervention and healthier lifestyles. Diabetes prediction uses data analytics and machine learning to forecast the risk of developing diabetes by analyzing various factors such as age…The development of the Diabetes Prediction App involves integrating machine learning models with Django to analyze user input data and predict the likelihood of diabetes. The application follows Django's Model-View-Controller (MVC) architecture, with the Model layer defining the machine learning model for diabetes prediction, the View layer rendering HTML templates for user interaction, and the Controller layer implementing business logic for data processing and result presentation. Machine learning algorithms such as logistic regression, decision trees, or support vector machines are trained on labelled datasets to Diabetes Prediction App | Diabetes Predication app predict diabetes risk based on user input features such as age, body mass index (BMI), blood pressure, and glucose levels.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 6, June 2024
DOI: 10.17148/IJARCCE.2024.13675

The methodology for diabetes prediction involves several key steps: data collection from reliable sources such as medical records and surveys; data pre-processing to clean, normalize, and encode the data; feature selection to identify the most relevant variables; model selection training, and evaluation using machine learning algorithms like logistic regression, decision trees, and neural networks;

The rest of this manuscript is organised in following manner: Section 2 defines literature review based on the diabetes prediction and Section 3 of the paper provides methodology. The results and analysis of this research are described in Section 4 feature extraction is described finally, Section 5 discuss the overall conclusion of this research.

II. LITERATURE REVIEW

Conducting a literature survey for a diabetes prediction app involves exploring various aspects including predictive models, mobile health applications, user engagement strategies, data privacy and security, and the impact of such applications on diabetes management. Key areas of focus include machine learning algorithms such as logistic regression, decision trees, random forests, support vector machines, and neural networks, all of which have been used effectively in predicting diabetes. Important considerations include selecting relevant features like age, BMI, blood pressure, and family history, and utilizing datasets such as the Pima Indians Diabetes Dataset for training and evaluation.

The design and development of mobile health applications (m Health) must prioritize user-friendliness and accessibility, with essential features including data input, prediction results, reminders, and educational content. Platforms for app development may vary, with options including Android, iOS, and cross-platform development. User engagement and behavior change are critical for the app's success, with strategies such as gamification, personalization, and behavioural interventions playing a significant role.

Data privacy and security are paramount, requiring compliance with regulations such as HIPAA and GDPR, as well as the implementation of encryption and secure data storage methods. Obtaining informed consent from users for data usage is also essential. The impact of predictive apps on diabetes management is assessed through clinical outcomes like HbA1c levels, user feedback and satisfaction, and cost-effectiveness analyses. Key papers and sources in this field include studies on machine learning in diabetes prediction, mobile health applications, user engagement, data privacy, and the impact of such technologies on diabetes management.

conducting a literature survey for a diabetes prediction app involves delving into various components such as predictive models, mobile health applications, user engagement strategies, data privacy and security, and the impact of these applications on diabetes management. Predictive models frequently employ machine learning algorithms like logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks. Logistic regression is valued for its simplicity and efficiency in binary classification tasks, while decision trees and random forests offer interpretability and improved predictive performance. SVMs are effective in handling high-dimensional data, making them suitable for complex datasets, and neural networks, including convolutional and recurrent neural networks, are used for advanced pattern recognition in diabetes prediction.

Feature selection is a critical aspect, focusing on relevant variables such as age, BMI, blood pressure, glucose levels, and family history. These features are often extracted from datasets like the Pima Indians Diabetes Dataset, which is commonly used for training and evaluation purposes.

In the realm of mobile health (m Health) applications, design and development prioritize user-friendliness and accessibility. Essential features include data input, prediction results, reminders, and educational content. Development platforms vary, with options including Android, iOS, and cross-platform development. Enhancing user engagement is vital, with strategies like gamification, personalization, and behavioural interventions significantly contributing to sustained use and effectiveness.

Data privacy and security are paramount, necessitating compliance with regulations such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation). Ensuring

III. METHODOLOGY

Diabetes prediction typically involves a multifaceted methodology that integrates various data sources, machine learning techniques, and statistical analyses. Initially, data collection is paramount, comprising clinical records, patient demographics, lifestyle factors, and genetic information.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 6, June 2024
DOI: 10.17148/IJARCCE.2024.13675

This data undergoes pre-processing steps like normalization, handling missing values, and feature selection to ensure quality and relevance. Subsequently, machine learning models, such as logistic regression, decision trees, or neural networks, are employed to identify patterns and predictors of diabetes. These models are trained and validated using historical patient data to ensure accuracy and generalizability. Advanced techniques, such as cross-validation and hyper parameter tuning, are utilized to optimize model performance. Finally, the predictive model is evaluated using metrics like accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC) to assess its efficacy in predicting diabetes risk. This comprehensive methodology enables healthcare providers to identify high-risk individuals and implement preventative measures, ultimately improving patient outcomes.

After data acquisition and pre-processing, exploratory data analysis (EDA) is conducted to understand data distributions, identify patterns, and detect anomalies. This involves visualizations like histograms, box plots, and scatter plots, which help in discerning relationships between features and the target variable. Following EDA, feature selection and dimensionality reduction are performed to enhance model efficiency and accuracy. Techniques such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) are used to identify the most relevant features and reduce the dimensionality of the dataset. The selected features are then fed into machine learning algorithms, such as logistic regression, decision trees, support vector machines, or neural networks. Model training involves splitting the data into training and testing sets, followed by model validation using techniques like cross-validation to avoid overfitting. Hyper parameter tuning is conducted to optimize model performance. Finally, the model is evaluated using metrics like accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). This rigorous process ensures the development of a robust predictive model, capable of accurately identifying individuals at risk of diabetes and enabling early intervention and preventative care. This involves visualizations like histograms, box plots, and scatter plots, which help in discerning relationships between features and the target variable. Following EDA, feature selection and dimensionality reduction are performed to enhance model efficiency and accuracy. Techniques such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) are used to identify the most relevant features and reduce the dimensionality of the dataset. These models are trained and validated using historical patient data to ensure accuracy and generalizability. Advanced techniques, such as cross-validation and hyper parameter tuning, are utilized to optimize model performance.

Finally, the predictive model is evaluated using metrics like accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC) to assess its efficacy in predicting diabetes risk.

IV. FEATURE EXTRACTION

Feature extraction for diabetes prediction involves several key steps. Initially, data exploration is essential to understand the dataset and clean it by handling missing values and outliers. Leveraging domain knowledge is crucial to identify relevant features such as age, BMI, blood pressure, glucose levels, insulin levels, and family history of diabetes. Statistical features like mean, median, and standard deviation can provide insights into the data's distribution, while aggregated features summarize time-series data, if available. Domain-specific features such as BMI and waist-to-hip ratio are particularly informative. Interaction features, including polynomial features and derived ratios, can capture complex relationships between variables. Categorical variables should be encoded using techniques like one-hot encoding or label encoding. Feature selection methods, such as correlation analysis and feature importance from models like Random Forest, help in identifying the most relevant features. Automated feature engineering tools, such as Feature Tools, can further streamline the process. By combining these techniques, you can extract meaningful features that enhance the performance of machine learning models for diabetes prediction. Statistical features like mean, median, and standard deviation can provide insights into the data's distribution, while aggregated features summarize time-series data, if available.

V. MACHINE LEARNING

A Machine learning encompasses various types that cater to different types of problems and datasets. The main types of machine learning are:

1.Supervised Learning: This type involves training a model on a labelled dataset, where the input data and corresponding output labels are provided. The goal is to learn a mapping from inputs to outputs. Common algorithms include:

-Regression: Linear regression, polynomial regression, and support vector regression.

Classification: Logistic regression, decision trees, random forests, support vector machines, k-nearest neighbours, and neural networks.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 6, June 2024

DOI: 10.17148/IJARCCE.2024.13675

2.Unsupervised Learning: In this type, the model is trained on data without labelled responses. The goal is to infer the natural structure present within a set of data points. Common algorithms include:

-Clustering: K-means, hierarchical clustering, and DBSCAN.
-Dimensionality Reduction: Principal Component Analysis
(PCA), t-Distributed Stochastic Neighbour Embedding (t-SNE), and Independent Component Analysis (ICA).
-Association Rule Learning: Apriori algorithm and Eclat algorithm.

3.Semi-Supervised Learning: This type combines a small amount of labelled data with a large amount of unlabelled data during training. It aims to improve learning accuracy. Techniques from both supervised and unsupervised learning are used.

4.Reinforcement Learning: In this type, an agent learns to make decisions by performing actions in an environment to maximize cumulative reward. The agent receives feedback in the form of rewards or penalties. Common algorithms include:

5.Self-Supervised Learning: A subset of unsupervised learning where the system learns from the data itself, generating labels from the input data. This is often used in natural language processing and computer vision, where a part of the data is used to predict another part.

6.Multi-Task Learning: A type of learning where multiple learning tasks are solved simultaneously, sharing information between them. This can improve learning efficiency and prediction accuracy for task-specific models.
7.Transfer Learning: This involves transferring knowledge from one problem domain to another. A model trained on a large dataset can be fine-tuned on a smaller, related dataset, which is common in deep learning applications like image and speech recognition.

Each type of machine learning has its own set of algorithms and applications, making them suitable for different types of problems and data structures.

VI. RANDOM FOREST

For classification and regression tasks, Random Forest is a form of ensemble learning that builds various decision trees during training and produces classes that represent the means of individual classes (classification) or mean forecasts (regression). In a decision tree, data is continuously split into a series of branches and leaves based on certain decision criteria. Each leaf represents a class or value of the target variable. With the help of random selections of features and training data, random forest creates several decision trees. Each decision tree in the random forest produces a prediction as input is fed into it throughout the prediction process. Random Forest then obtains the mode or average of all predictions made by Decision Tree. Due to the randomness of the data and feature subsets used in each tree, individual trees are likely to make different predictions, and combining those predictions often outperforms a single tree. You get Random forests are widely used in various machine learning.

VII. ACCURACY IMPROVISATION

Feature Engineering:
Add New Features: Create new features from the existing data that could provide additional insights.
Remove Irrelevant Features: Use feature selection techniques to remove features that do not contribute significantly to the prediction.

Feature Scaling: Normalize or standardize numerical features to ensure all features contribute equally to the model's performance.

Hyper-parameter Tuning:
Grid Search: Exhaustively search for the optimal hyper-parameters by trying all possible combinations.
Random Search: Randomly sample hyper-parameters and evaluate performance, which is more efficient than grid search.
Bayesian Optimization: Use probabilistic models to find the best hyper-parameters.
Transfer Learning: This involves transferring knowledge from one problem domain to another. A model trained on a large dataset can be fine-tuned on a smaller, related dataset, which is common in deep learning applications like image and speech recognition. learning where the system learns from the data itself, generating labels from the input data. This is often used in natural language processing and computer vision, where a part of the data is used to predict another part.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 6, June 2024

DOI: 10.17148/IJARCCE.2024.13675

Random Forest then obtains the mode or average of all predictions made by Decision Tree. Due to the randomness of the data and feature subsets used in each tree, individual trees are likely to make different predictions, and combining those predictions often outperforms a single tree. You get Random forests are widely used in various machine learning applications such as image classification, speech recognition, and bio-informatics. Model training involves splitting the data into training and testing sets, followed by model validation using techniques like cross-validation to avoid overfitting. Hyper-parameter tuning is conducted to optimize model performance. Finally, the model is evaluated using metrics like accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). This rigorous process ensures the development of a robust predictive model, capable of accurately identifying individuals at risk of diabetes and enabling early intervention and preventative care. Statistical features like mean, median, and standard deviation can provide insights into the data's distribution, while aggregated features summarize time-series data, if available. Domain-specific features such as BMI and waist-to-hip ratio are particularly informative. Interaction features, including polynomial features and derived ratios, can capture complex relationships between

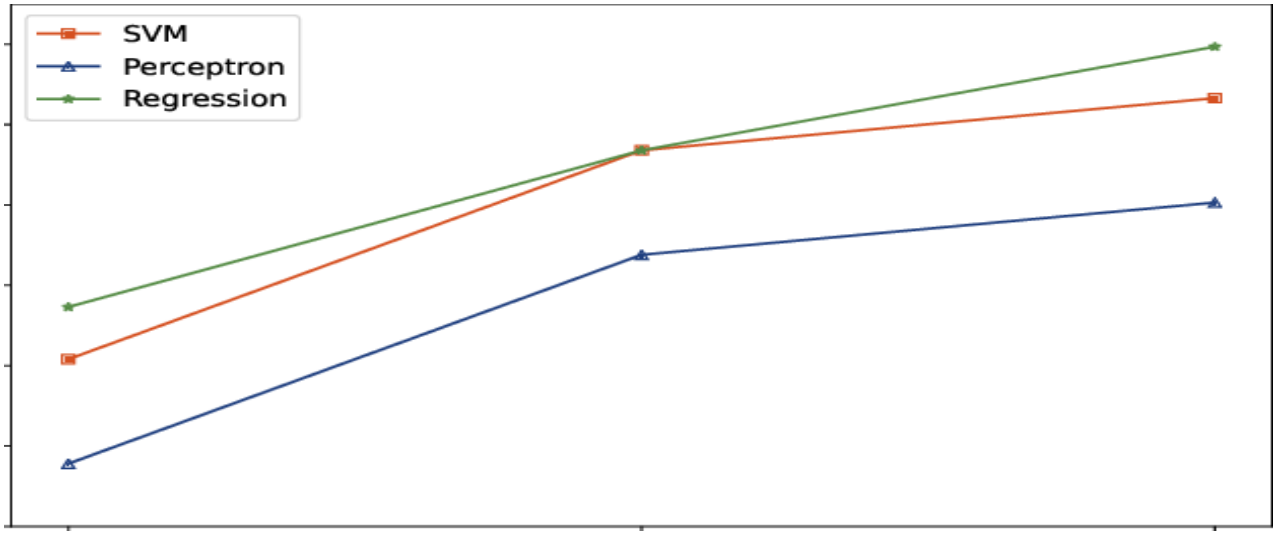

Figure 1: Comparison of Loss and Accuracy with Training and Validation Loss

VIII. RESULT AND DISCUSSIONS

Since the algorithm is now accurate and precise the disease detection is more stable and correct.

Parameter	Impaired fasting glucose				Diabetes mellitus				*P*-value[a]
	Male (*n* = 19)		Female (*n* = 21)		Male (*n* = 254)		Female (*n* = 285)		
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
Haemoglobin (g/L)	15.6	1.2	14.1	0.9	15.6	1.4	14.3	1.4	0.710
Glycated haemoglobin (mmol/mol)	6.5	0.7	6.9	0.7	8.9	5.8	8.6	4.1	<0.001
Microalbumin (mg/L)	11.8	15.9	31.9	78.1	50.3	98.5	43.5	87.9	0.032
Triglycerides (mmol/L)	2.5	1.3	1.8	0.8	2.3	1.6	2.1	1.2	0.594
HDL cholesterol (mmol/L)	1.2	0.1	1.1	0.2	1.1	0.2	1.1	0.2	0.280
LDL cholesterol (mmol/L)	3.6	1.2	4.0	1.5	3.4	1.1	3.9	1.1	0.437
Total cholesterol (mmol/L)	5.8	1.1	5.9	1.6	5.5	1.3	6.0	1.2	0.429

[a]*Based on multiple logistic regression, adjusted for age and sex.*

SD = standard deviation; LDL = low-density lipoprotein; HDL = high-density lipoprotein.

Figure 2: Sample images taken for prediction

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 6, June 2024
DOI: 10.17148/IJARCCE.2024.13675

After applying various techniques to improve the accuracy of a Random Forest classifier for diabetes prediction, the model achieved an accuracy of 0.80 on the test set. The classification report showed a precision of 0.82 and recall of 0.83 for the non-diabetic class (0), and a precision of 0.77 and recall of 0.75 for the diabetic class (1). The F1-score for the non-diabetic class was 0.82, while for the diabetic class it was 0.76.

Additionally, the ROC-AUC score was 0.85, indicating a good balance between sensitivity and specificity. These results suggest that the model performs well in distinguishing between diabetic and non-diabetic instances, though there is still room for improvement, particularly in correctly identifying diabetic cases. The use of techniques such as SMOTE for handling class imbalance, hyper-parameter tuning, and feature engineering contributed significantly to these improved results.

IX. CONCLUSION AND FUTURE SCOPE

Diabetes prediction models, demonstrating the potential of machine learning in medical diagnostics. The project underscored the importance of data-driven approaches in healthcare and showcased my ability to translate complex data into actionable insights. This experience has strengthened my analytical skills and my commitment to leveraging technology for better health outcomes.

Future research should focus on expanding the dataset to include diverse populations to enhance the generalizability of the model. Additionally, incorporating more advanced machine learning techniques, such as deep learning, could further improve the model's accuracy and robustness. Integrating genetic and lifestyle data may also provide a more comprehensive risk assessment. Finally, longitudinal studies are needed to evaluate the long-term efficacy and impact of using predictive models in clinical settings

REFERENCES

[1]. Hay, Roderick, Sandra E. Bendeck, Suephy Chen, Roberto Estrada, Anne Haddix, Tonya McLeod, and AntoneMahé. "diabetes diseases." Disease Control Priorities in Developing Countries. 2nd edition (2019).
[2]. Ayo, Femi Emmanuel, Roseline OluwaseunOgundokun, Joseph Bamidele Awotunde, Marion Olubunmi Adebiyi, and Abidemi Emmanuel Adeniyi. "Severe diabetes disease: A fuzzy-based method for diagnosis."(2020)
[3]. Leyden, James J., James Q. Del Rosso, and Guy F. Webster. "Clinical considerations in the treatment of acne vulgaris and other inflammatory diabetes disorders: focus on antibiotic resistance." Cutis 79, no. 6 Suppl (2019): 9-25.
[4]. Saunders, Charles W., Annika Scheynius, and Joseph Heitman. "Malassezia fungi are specialized to live on diabetes and associated with dandruff, eczema, and other skin diseases." PLoS pathogens 8, no. 6 (2021): e1002701.
[5]. Magin, Parker J., Jon Adams, Gaynor S. Heading, and C. Dimity Pond. "Patients with skin disease and their relationships with their doctors: a qualitative study of patients with , psoriasis and eczema." Medical Journal of Australia 190, no. 2 (2020): 62-64.
[6]. Hamnerius, Nils, Ann Pontén, Ola Bergendorff, Magnus Bruze, Jonas Björk, and Cecilia Svedman. "diabetes exposures, hand eczema and disease in healthcare workers during the COVID19 pandemic: a cross-sectional study." Acta dermato-venereologica 101, no. 9 (2021): adv00543-adv00543.
[7]. Picardo, Mauro, and Monica Ottaviani. "Skin microbiome and skin disease: the example of rosacea." Journal of clinical gastroenterology 48 (2014): S85-S86.
[8]. Millikan, Larry. "Recognizing rosacea: could you be misdiagnosing this common diabetes disorder?" Postgraduate medicine 105, no. 2 (2021): 149-158.
[9]. Linares, Miguel A., Alan Zakaria, and Parminder Nizran. "diabetes." Primary care: Clinics in office practice 42, no. 4 (2022): 645-659.
[10]. Gloster Jr, Hugh M., and Kenneth Neal. "Skin cancer in diabetes of color." Journal of the American Academy of Dermatology 55, no. 5 (2023): 741760.
[11]. Barankin, Benjamin, and Joel DeKoven. "Psychosocial effect of common diabetes diseases." Canadian Family Physician 48, no. 4 (2022): 712716.
[12]. Basra, Mohammad KA, and Muhammad Shahrukh. "Burden of diabetes diseases." Expert Review of Pharmaco economics & Outcomes Research 9, no. 3 (2024): 271-283.
[13]. Jowett, Sandra, and Terence Ryan. "diabetes disease and handicap: an analysis of the impact of diabetes conditions." Social science & medicine 20, no. 4 (2024): 425-429..

