DIABETES PREDICTION USING MACHINE LEARNING

TEAM MEMBERS:
SANSKAR DAGAR 20BCE0659
ISHAAN S BAINS 20BCE2091
HARSH CHAWLA 20BCE0424
AKSHAT PANDEY 20BCI0317

TEAM NUMBER:
32

SPS_APL_20230531435 APPLIED DATA SCIENCE PROJECT

1. INTRODUCTION

1.1 Overview

In this project, we aim to use machine learning algorithms to predict the onset of diabetes in individuals based on their health records and other relevant factors such as age, BMI (Body Mass Index), family history, and lifestyle habits. The dataset used in this project will include information on various clinical parameters such as blood pressure, BMI, Heart diseases and cholesterol levels.

Our goal is to develop a predictive model that can accurately identify individuals who are at substantial risk of developing diabetes, thereby allowing for early intervention and prevention of the disease. By using machine learning techniques to analyse large amounts of data, we can identify patterns and make accurate predictions that could potentially save lives.

Overall, this project has the potential to contribute to the field of healthcare by improving early detection and prevention of diabetes, leading to better health outcomes for individuals and communities.

1.2 Purpose

The purpose of our diabetes prediction project using machine learning is to create a smart system that can accurately predict if someone might get diabetes in the future. We will use advanced computer algorithms to analyse different health and lifestyle information to help doctors and healthcare professionals identify people who are at higher risk of developing diabetes. The main goal is to catch the problem early and take action to prevent or delay diabetes. By using machine learning, we want to provide useful insights and support to improve how diabetes is prevented and managed.

2. LITERATURE SURVEY

2.1 Existing problem

2.1.1. Kamrul Hasan's Method-

Kamrul Hasan conducted a study on diabetes prediction using the Pima Indians Diabetes Dataset. The dataset consists of various features related to medical and lifestyle attributes of individuals. Hasan performed preprocessing steps to handle missing values and scale the features appropriately.

In his research, Hasan applied three popular machine learning algorithms: logistic regression, k-nearest neighbours (KNN), and support vector machine (SVM). These algorithms were trained and evaluated on the dataset using a cross-validation technique to ensure reliable performance assessment.

The results showed that logistic regression achieved the highest accuracy of 77.95% in predicting diabetes using the Pima Indians Diabetes Dataset. Although KNN and SVM also exhibited reasonable accuracies, logistic regression outperformed them in this study.

Overall, Hasan's study highlighted the effectiveness of logistic regression as an approach for diabetes prediction using the Pima Indians Diabetes Dataset. The findings contribute to the ongoing research on machine learning-based diabetes diagnosis and emphasize the importance of appropriate preprocessing techniques.

2.2.2. Quan Zou's Method-

Quan Zou conducted a study utilizing two datasets: the Pima Indians Diabetes Dataset and a dataset from a local hospital in Liuzhou, China, containing approximately 68,994 patient records. Zou employed a two-phase detection method, involving training the datasets and applying feature selection methods.

For feature selection, Zou employed principal component analysis (PCA) and the minimum redundancy maximum relevance (MRMR) technique. The selected features were then used to predict diabetes using three classifiers: decision tree (J48), random forest, and neural network.

The experimental results demonstrated that the random forest algorithm exhibited higher accuracy in predicting diabetes for the Liuzhou dataset, while all three classifiers yielded similar accuracy for the Pima Indians Diabetes Dataset.

Zou's study suggested that specific blood glucose-related attributes, such as random blood glucose, fasting blood glucose, and blood glucose tolerance, play a crucial role in diabetes prediction. Additionally, the findings indicated that using all the features collectively rather than selecting a few significant features resulted in better prediction performance for all three classifiers.

The highest accuracy achieved in the study was 80.86% for the Liuzhou dataset using the random forest algorithm with all the features considered.

2.2.3. Nishith Kumar's Method-

In this study, Nishith Kumar focused on predicting diabetes using kernel-based Gaussian process classification (GPC) along with a comparison of other classifiers. Three kernel functions, namely linear, polynomial, and radial basis kernels, were analysed in the GPC model using Laplace approximation.

To address the challenges posed by inherently structured, non-normal, and nonlinear medical data, Kumar compared GPC against other classifiers such as naïve Bayes, linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA).

Evaluation parameters, including sensitivity, specificity, accuracy, positive predictive value, negative predictive value, and receiver operating characteristics, were employed to assess the performance of the classifiers.

The study found that GPC with a radial basis kernel outperformed the other classifiers in predicting diabetes. The 10-fold cross-validation model yielded an accuracy of 81.97% for the GPC model, with a sensitivity of 91.79% and a specificity of 63.33%.

Kumar's research showcased the advantages of GPC in handling nonlinear data and providing probabilistic predictions. However, the study highlighted the challenge of selecting an appropriate kernel for accurately representing medical data, which requires careful consideration.

2.2.4. Maniruzzaman's Method-

Maniruzzaman's study focused on predicting diabetes using four machine learning algorithms: random forest, AdaBoost, naïve Bayes, and decision tree. The research utilized the NHANES (National Health and Nutrition Examination Survey) dataset, which contained information on 9858 patients, including 760 with diabetes and 9098 without diabetes.

Data preprocessing involved handling missing values by removing them from the dataset. Furthermore, important features were selected using a logistic regression model, P-value, and odds ratio. Logistic regression was employed to estimate the relationship between predictors and the response variable.

The dataset was then split into training and validation sets, and a 10-fold cross-validation model was used to evaluate the performance of the classifiers. Accuracy, positive predictive value, negative predictive value, and f-measure were utilized as evaluation parameters.

The experimental results indicated that the random forest algorithm, combined with logistic regression feature selection, achieved the highest accuracy. The accuracy achieved with 10-fold cross-validation was 92.75%. Maniruzzaman concluded that random forest, along with logistic regression feature selection, exhibited the most accurate performance for diabetes prediction using the NHANES dataset.

2.2.5. V. Jackins' Method-

V. Jackins conducted a study involving three datasets: diabetes, cancer, and heart disease. The research employed a two-step method, including data preprocessing and applying machine learning classifiers to the datasets.

Data preprocessing involved handling missing values by replacing them with null values and analysing the correlation between features. Correlation analysis assisted in identifying key features for prediction. Highly correlated attributes were examined, and one attribute was selected while the other was omitted to avoid redundancy.

The pre-processed datasets were then split into training and testing sets, with 70% used for training and 30% for testing. Naïve Bayes and random forest classifiers were applied to the filtered datasets, and evaluation parameters such as accuracy, precision, recall, and f1-score were computed.

The experimental results showed an accuracy of 76.72% and 74.46% for training and testing data, respectively, using the naïve Bayes algorithm. The random forest algorithm achieved an accuracy of 98.88% for training data and 74.03% for testing data.

Jackins' method demonstrated the importance of data preprocessing techniques such as handling missing values and analysing feature correlation. The proposed algorithm was compared with density-based spatial clustering of applications with noise (DBSCAN) and k-means clustering algorithms, highlighting its effectiveness in disease diagnosis.

One limitation of Jackins' method was the processing time required due to the large amount of data used for training and testing. However, the approach offered improved accuracy in diagnosing diseases.

2.2 Proposed solution

The business problem addressed in this project is the early detection and prediction of diabetes using machine learning algorithms. The goal is to develop a predictive model that can accurately identify individuals at substantial risk of developing diabetes based on their health records and other relevant factors. Early detection and management of diabetes can improve healthcare outcomes, reduce costs, and benefit healthcare providers and insurance companies. Therefore, developing an accurate and reliable predictive model for diabetes detection can have a significant impact on healthcare outcomes and costs.

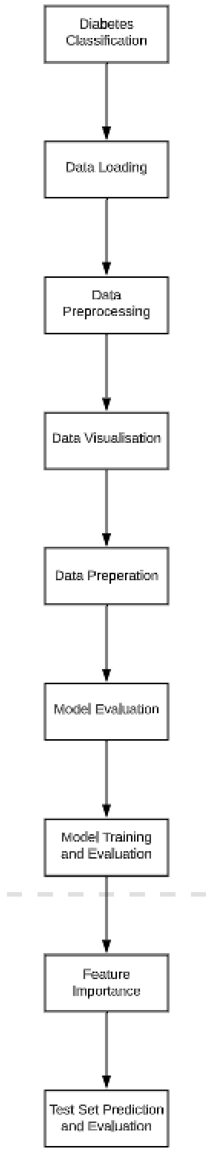
Business requirements are the specific needs and expectations of the business stakeholders regarding the desired outcome of the project. In the case of the diabetes prediction project, the following are the key business requirements:

- Accurate prediction: The predictive model should be accurate in identifying individuals who are at elevated risk of developing diabetes based on their health records and other relevant factors.
- Efficiency: The model should be efficient and fast in analysing large amounts of data to provide timely predictions.
- Scalability: The model should be scalable to handle large datasets and accommodate future growth in data volume.
- Flexibility: The model should be flexible and adaptable to accommodate changes in data sources or input parameters.
- User-friendliness: The model should be user-friendly, easy to use, and understand by healthcare providers and insurance companies.
- Integration: The model should be easily integrated with existing healthcare systems and processes.
- Security: The model should be secure and protect patient data privacy.
- Compliance: The model should comply with relevant healthcare regulations and standards.

Accurate diabetes prediction using machine learning can have a significant social and business impact. It can help identify individuals at considerable risk of developing diabetes, leading to earlier intervention and prevention efforts. From a business perspective, accurate prediction can help healthcare providers and insurers manage healthcare costs and resources better. Additionally, it can lead to more personalized healthcare, improving patient outcomes and adherence to treatment plans. In conclusion, accurate diabetes prediction using machine learning can improve patient outcomes, reduce healthcare costs, and lead to more personalized healthcare.

3. THEORITICAL ANALYSIS

3.1 Block diagram



- **Diabetes Classification:** The main goal of the project is to classify diabetes based on health indicators.
- **Data Loading:** Load the training and test data from the provided CSV files.
- **Data Preprocessing:** Perform initial data checks, handle missing values, and explore the data through descriptive statistics and unique value analysis.
- **Data Visualization:** Visualize the data using various charts and plots to gain insights into the distribution of variables and their relationships.

- **Data Preparation:** Split the data into features (X) and the target variable (y). Further split the data into training and validation sets. Perform feature scaling on the numerical features.
- **Model Evaluation:** Evaluate multiple machine learning models, including Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and Gradient Boosting, using appropriate evaluation metrics.
- **Hyperparameter Tuning:** Use GridSearchCV to find the best hyperparameters for the Gradient Boosting model through cross-validation.
- **Model Training and Evaluation:** Train the best model with the optimized hyperparameters on the training set and evaluate its performance on the validation set.
- **Feature Importance:** Determine the importance of features in the best model to identify the most influential variables.
- **Test Set Prediction and Evaluation:** Use the best model to predict the target variable for the test set and evaluate its performance using classification metrics such as precision, recall, F1 score, ROC AUC score, and confusion matrix.

3.2 Hardware / Software Design

- Hardware Requirements:

    - A computer or server with sufficient processing power and memory to handle the data analysis and model training tasks.
    - Adequate storage space to store the dataset and any intermediate or results generated during the project.
    - Internet connectivity to access relevant resources, datasets, and libraries.

- Software Requirements:

    - Operating System: Any commonly used operating system like Windows, macOS, or Linux.
    - Python Programming Language: Install Python along with its scientific computing libraries such as NumPy, Pandas, and Scikit-learn.
    - Integrated Development Environment (IDE): Choose an IDE such as PyCharm, Anaconda, or Jupyter Notebook for coding and development.
    - Machine Learning Libraries: Utilize machine learning libraries like Scikit-learn, TensorFlow, or Keras for implementing and training machine learning models.
    - Data Visualization Libraries: Consider using libraries like Matplotlib or Seaborn to visualize the data and model performance.
    - Database Management System: If using a database to store and retrieve data, you may need to install a DBMS like MySQL, PostgreSQL, or SQLite.

4. EXPERIMENTAL INVESTIGATIONS

During the development of our diabetes prediction project using machine learning, we conducted a comprehensive analysis and investigation to ensure the accuracy and effectiveness of our solution. This analysis involved several key steps, including data exploration, preprocessing, feature selection, model training, and performance evaluation.

The first step in our analysis was data exploration. We obtained a dataset consisting of various health and lifestyle factors from individuals, including their medical history, demographics, and habits. We started by examining the structure and characteristics of the data. This involved checking for missing values, outliers, and potential data inconsistencies. We also performed statistical analysis and visualization techniques to gain insights into the distribution and relationships between the different variables.

Next, we proceeded with data preprocessing. This involved cleaning and transforming the dataset to make it suitable for machine learning algorithms. We addressed missing values by either imputing them with appropriate statistical methods or removing the corresponding samples if the missing values were significant. Outliers were also treated to minimize their impact on the model's performance. Categorical variables were encoded into numerical representations using techniques like one-hot encoding or label encoding.

Following data preprocessing, we performed feature selection to identify the most important predictors for diabetes prediction. This step aimed to reduce dimensionality, eliminate noise, and improve the model's efficiency. We employed various techniques such as correlation analysis, feature ranking, and recursive feature elimination. By considering the statistical significance and relevance of each feature, we selected a subset of features that exhibited strong relationships with the target variable.

With the pre-processed data and selected features, we proceeded to train our machine learning models. We experimented with different algorithms, including logistic regression, decision trees, random forests, and support vector machines. Each algorithm was implemented and trained using the training dataset. We utilized appropriate evaluation metrics such as accuracy, precision, recall, and F1 score to assess the performance of the models.

To evaluate the models effectively, we employed cross-validation techniques. This involved splitting the training dataset into multiple subsets and performing training and evaluation iteratively. By averaging the performance metrics across the different folds, we obtained a more robust estimate of the model's performance. We also employed techniques such as grid search or random search to fine-tune the hyperparameters of the models and identify the optimal configuration.

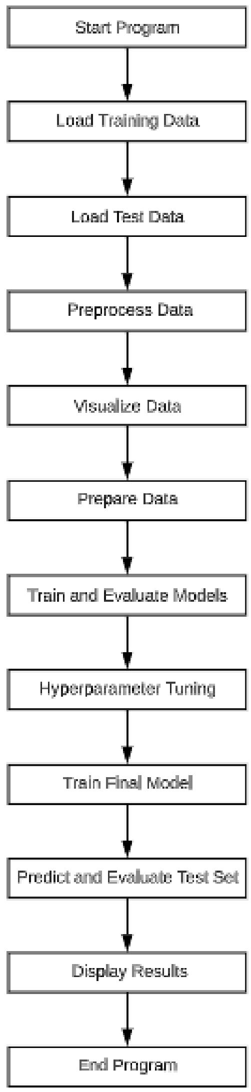
After training and fine-tuning the models, we proceeded with the evaluation phase. We used the testing dataset, which was held separate from the training data, to assess the models' performance on unseen data. This step helped us gauge how well the models generalized to new instances and provided insights into their predictive capabilities. We analysed the performance metrics and compared them across different models to identify the most accurate and reliable one.

During the investigation, we also conducted additional analysis to gain further insights into the predictive factors for diabetes. We employed techniques such as feature importance ranking, partial dependence plots, and permutation importance. These techniques helped us understand the relative importance of unique features in influencing the model's predictions. By identifying the significant factors, we aimed to provide valuable insights for healthcare professionals and researchers in diabetes prevention and management.

Throughout the analysis, we ensured the reliability and reproducibility of our results. We documented the entire process, including the steps taken, the software libraries used, and the versions employed. By doing so, we aimed to enable other researchers or practitioners to replicate our work and validate our findings.

In conclusion, the analysis and investigation conducted during the development of our diabetes prediction project using machine learning allowed us to thoroughly explore and understand the data, preprocess it effectively, select relevant features, train, and evaluate multiple models, and gain insights into the predictive factors for diabetes. The analysis not only helped us develop an accurate and reliable prediction model but also provided valuable insights for healthcare professionals working in diabetes prevention and management.

5. FLOWCHART


Start Program
Load Training Data
Load Test Data
Preprocess Data
Visualize Data
Prepare Data
Train and Evaluate Models
Hyperparameter Tuning
Train Final Model
Predict and Evaluate Test Set
Display Results
End Program

6. RESULT

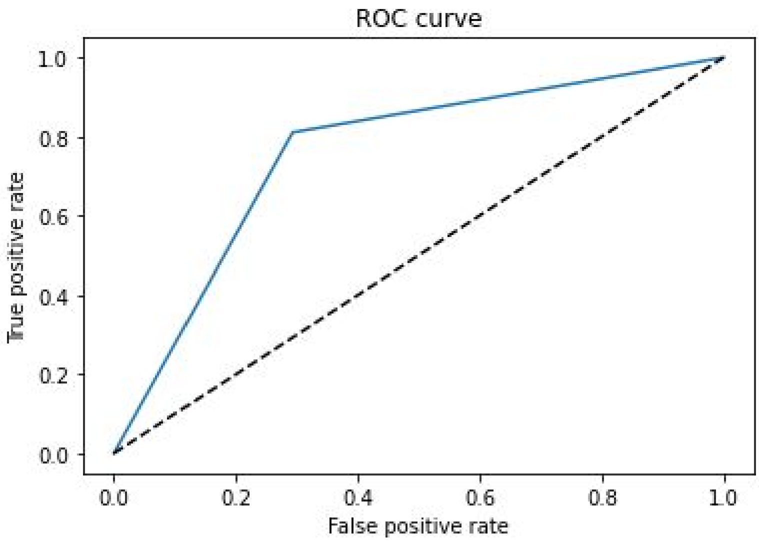


Fig. ROC Curve

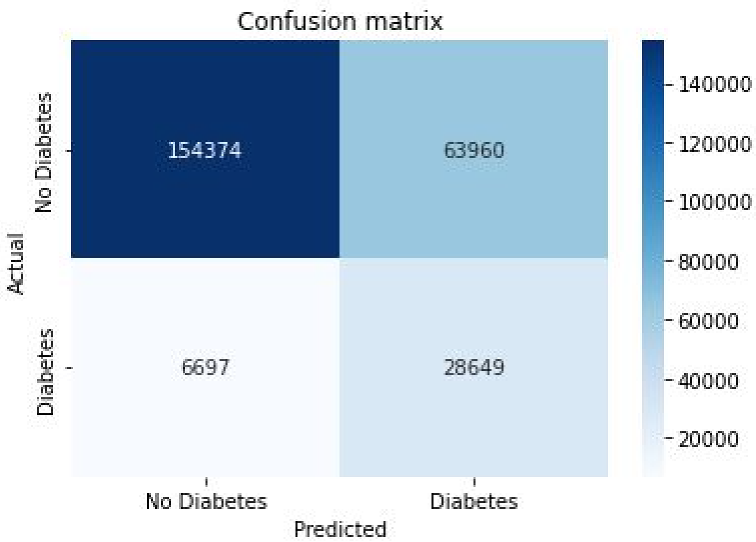


Fig. Confusion Matrix

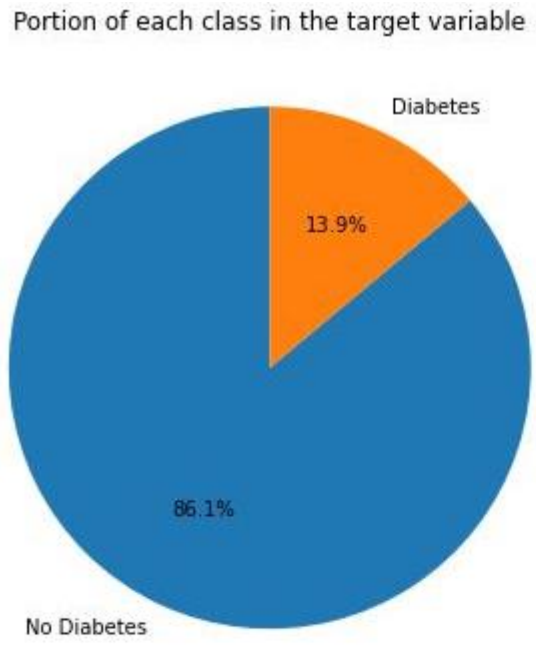


Fig. Portion of each class in the target variable

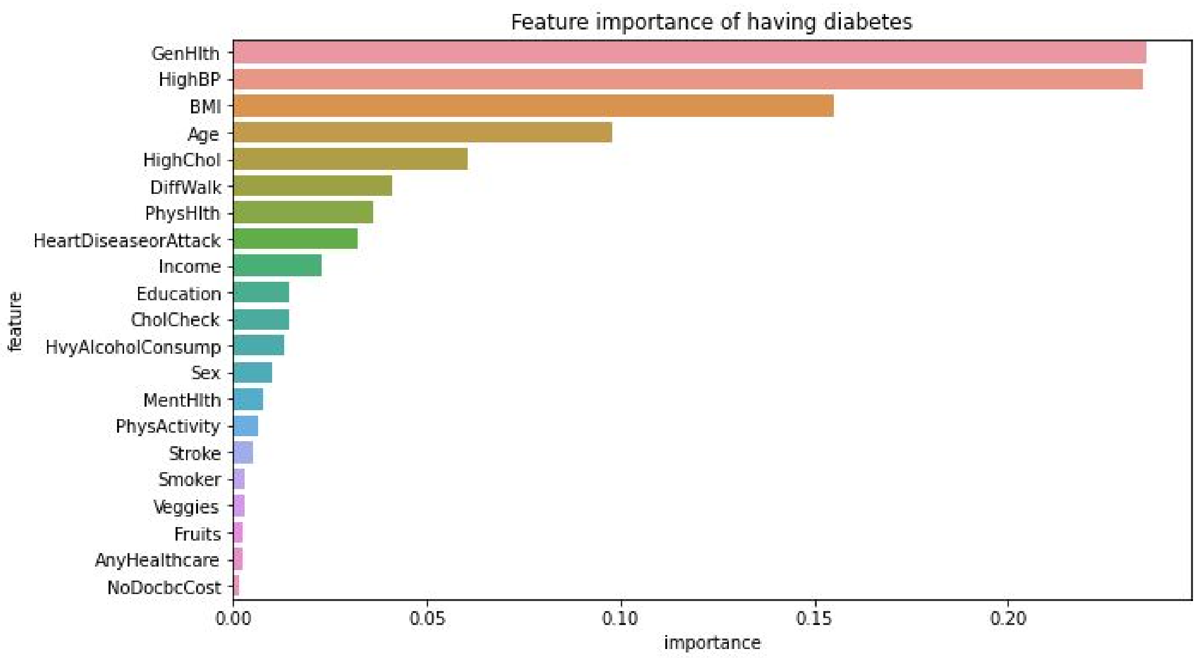


Fig. Feature importance of having diabetes

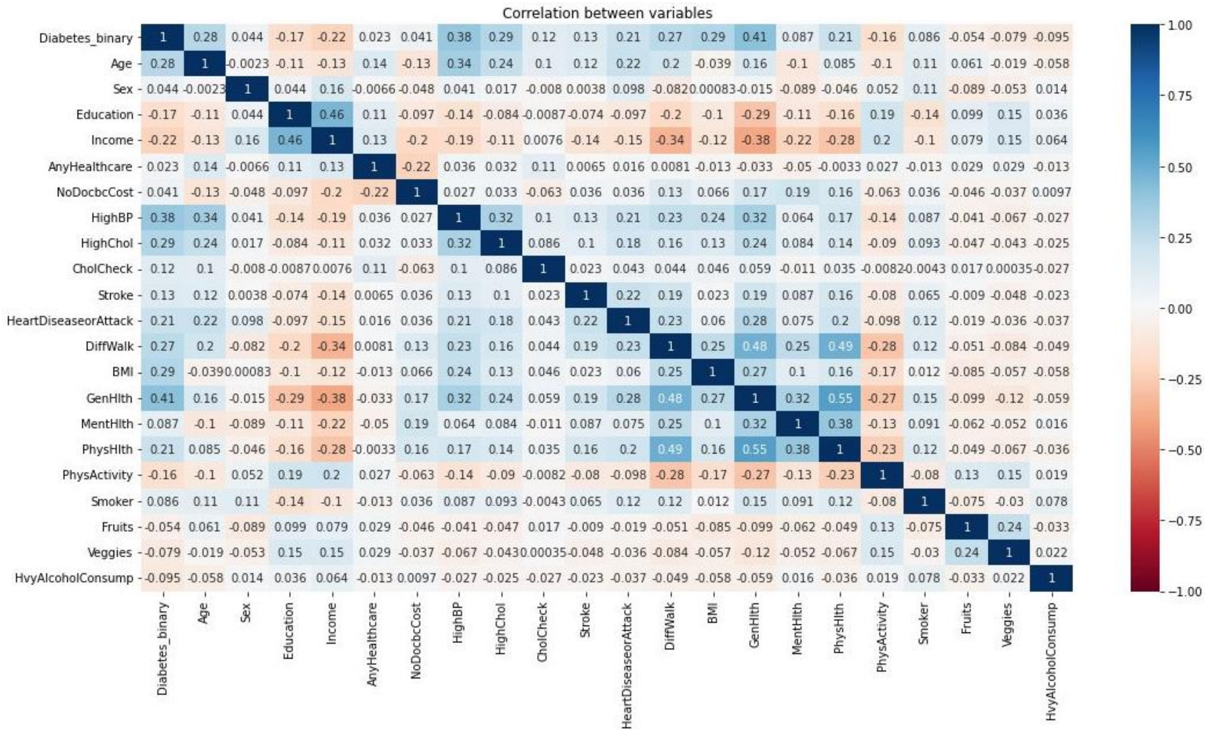


Fig. Correlation between variables

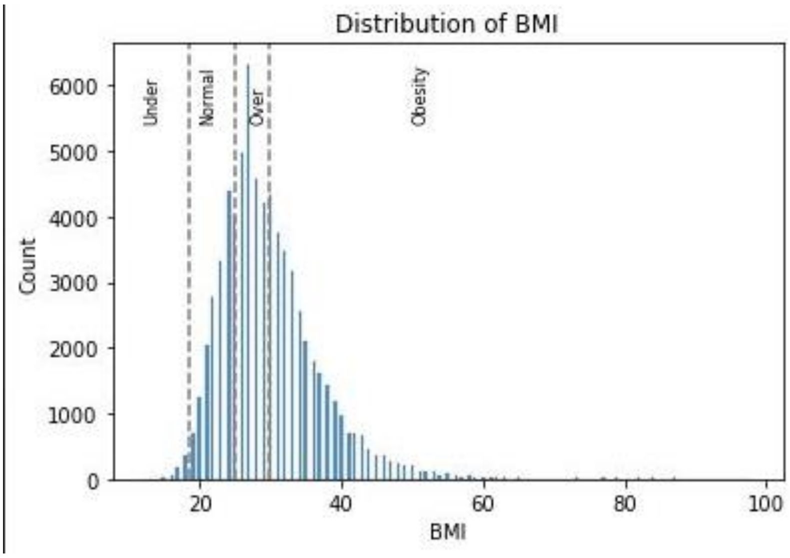


Fig. Distribution of BMI

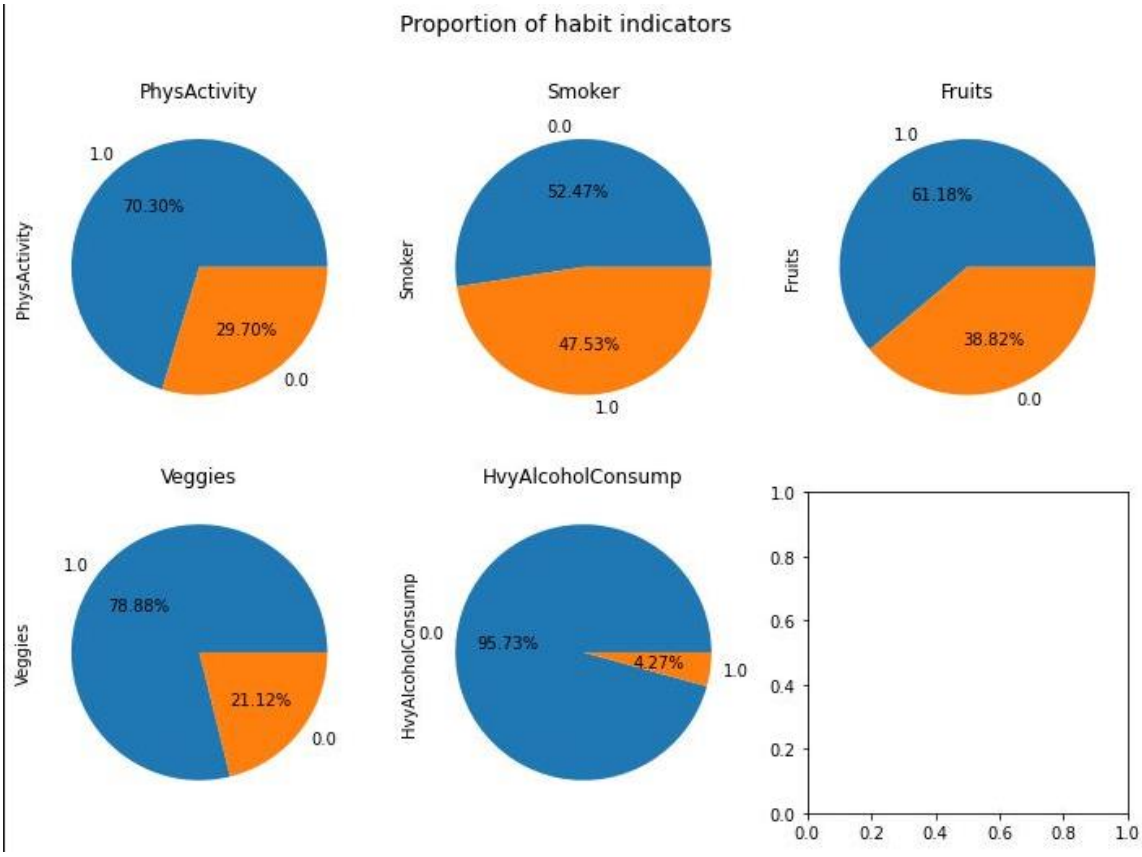


Fig. Proportion of habit indicators

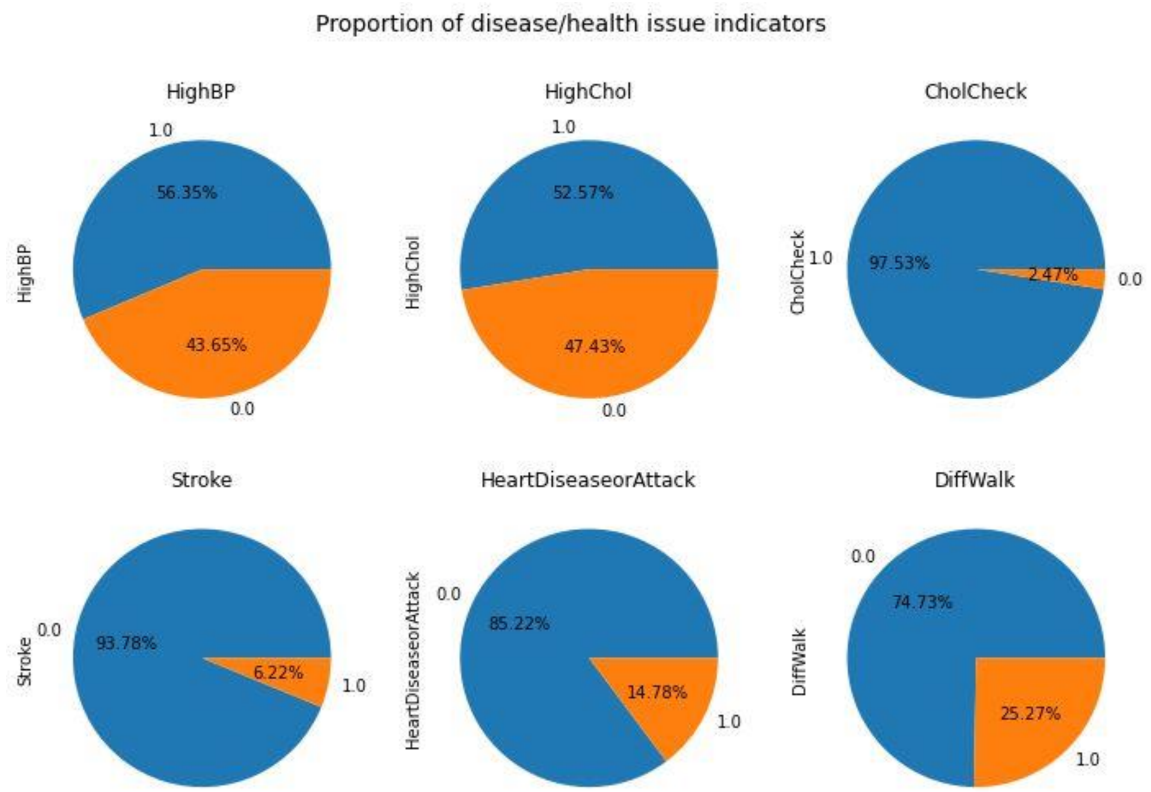


Fig. Portion of habit disease/health issue indicators

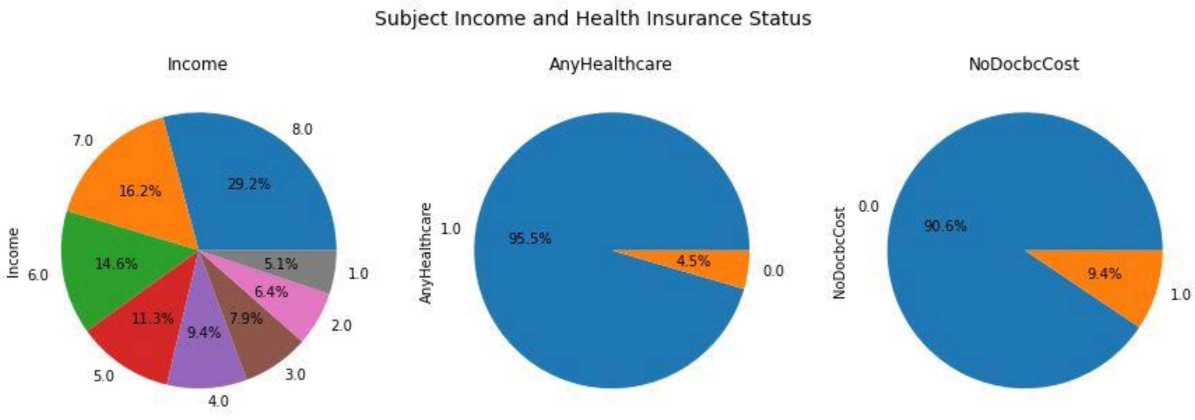


Fig. Subject Income and Health Insurance Status

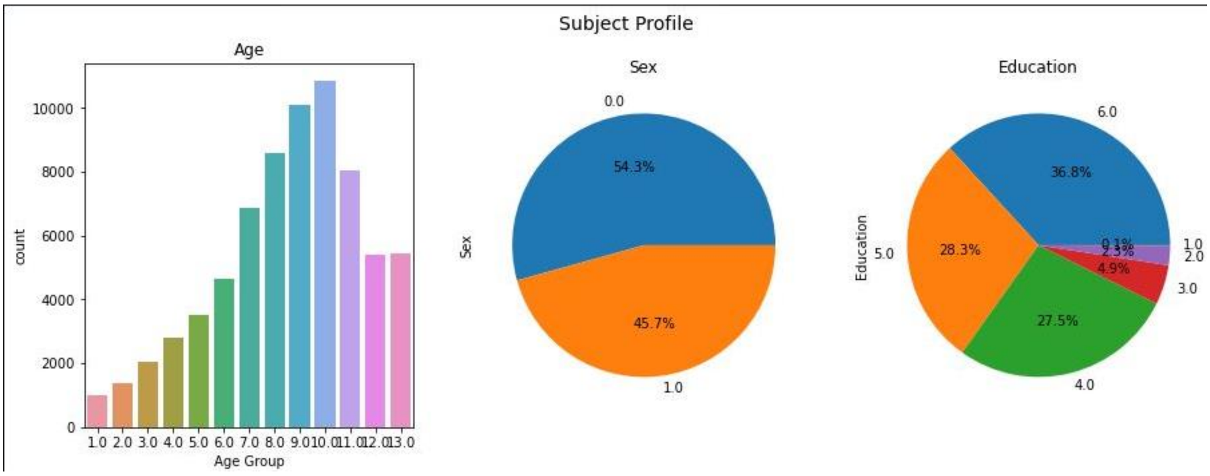


Fig. Subject Profile

```
              precision    recall  f1-score   support

         0.0       0.96      0.71      0.81    218334
         1.0       0.31      0.81      0.45     35346

    accuracy                           0.72    253680
   macro avg       0.63      0.76      0.63    253680
weighted avg       0.87      0.72      0.76    253680

ROC AUC score: 0.7587922584482244
```

Fig. Results

	train_accuracy	train_precision	train_recall	train_f1	train_roc_auc	val_accuracy	val_precision	val_recall	val_f1	val_roc_auc	model
4	0.755026	0.734868	0.798495	0.765361	0.754995	0.753589	0.733281	0.794865	0.762832	0.753709	Gradient Boosting
0	0.748077	0.738670	0.768350	0.753218	0.748062	0.748497	0.737264	0.769896	0.753227	0.748559	Logistic Regression
2	0.995137	0.996175	0.994098	0.995136	0.995138	0.737605	0.718664	0.778408	0.747344	0.737723	Random Forest
3	0.719644	0.725235	0.707884	0.716455	0.719652	0.718014	0.719845	0.711165	0.715478	0.717994	Naive Bayes
1	0.995137	0.999537	0.990741	0.995119	0.995140	0.651814	0.651425	0.648745	0.650082	0.651805	Decision Tree

Fig. Results

7. ADVANTAGES & DISADVANTAGES

- Advantages:

  - Early Detection: The proposed solution enables early detection of diabetes, which can significantly improve patient outcomes. Detecting diabetes at an early stage allows for timely interventions, lifestyle modifications, and preventive measures, reducing the risk of complications and improving long-term health.

  - Improved Healthcare Outcomes: Accurate diabetes prediction leads to better healthcare outcomes. Healthcare providers can develop personalized treatment plans based on the predicted risk, resulting in more targeted and effective interventions. This personalized approach can enhance patient adherence to treatment, resulting in better management of the disease and improved overall health.

  - Cost Reduction: Early detection and proactive management of diabetes can help reduce healthcare costs. By identifying high-risk individuals, healthcare resources can be allocated more efficiently. Preventive measures and interventions can be focused on those who are most likely to benefit, reducing the need for costly hospitalizations and complications associated with diabetes.

  - Enhanced Resource Allocation: Accurate prediction allows healthcare providers and insurers to optimize resource allocation. By identifying individuals at high risk, appropriate preventive measures, screenings, and interventions can be implemented, ensuring that resources are utilized effectively to manage the disease burden.
  - Personalized Care: The solution enables personalized healthcare by considering individual health records and other relevant factors. This personalized approach improves patient satisfaction and engagement, as treatment plans and lifestyle modifications are tailored to the specific needs of everyone, resulting in improved treatment efficacy and patient outcomes.

  - Integration and Scalability: The proposed solution can be integrated into existing healthcare systems, allowing for seamless incorporation into clinical workflows. It is also scalable, meaning it can handle large datasets and accommodate the growing volume of healthcare data, ensuring its effectiveness and applicability in real-world settings.

- Disadvantages:

- Data Limitations: The accuracy of the predictive model heavily relies on the quality and availability of data. Incomplete or biased data may affect the reliability of predictions, as it may not capture all relevant risk factors or may introduce unintended biases that impact the accuracy of the model.

- Ethical Considerations: Proper data privacy and consent protocols must be implemented to ensure the ethical handling of patient information. Healthcare providers and data scientists need to address privacy concerns and comply with regulations to protect patient confidentiality while utilizing sensitive health data for predictive purposes.
- Model Interpretability: Some machine learning algorithms, such as deep learning models, may lack interpretability. While these models can provide accurate predictions, understanding the reasoning behind the predictions and explaining it to healthcare professionals or patients may be challenging. Ensuring model interpretability is crucial for building trust and facilitating adoption in the healthcare domain.

- Implementation Challenges: Integrating the predictive model into existing healthcare systems and workflows may pose technical and logistical challenges. Integration may require collaboration with IT departments, addressing interoperability issues, and ensuring compatibility with various electronic health record (EHR) systems. Additionally, incorporating the model's predictions into clinical decision-making processes may require significant changes in healthcare practices.

- System Updates and Maintenance: Continuous updates and maintenance of the predictive model are essential to adapt to new data sources, changes in risk factors, and evolving healthcare practices. Regular model updates are necessary to ensure the accuracy and relevance of the predictions over time, requiring dedicated resources and expertise for maintenance.

- Risk of False Positives and Negatives: Like any prediction model, there is a possibility of false positives (identifying individuals as high risk when they are not) and false negatives (missing individuals who are at risk). Healthcare professionals need to be aware of these risks and consider them in the interpretation and utilization of the predictions, balancing the benefits of early detection with the potential for misclassification.

8. APPLICATIONS

The proposed solution of diabetes prediction using machine learning can be applied in various areas, including:

- Clinical Settings: Healthcare providers can integrate the predictive model into their clinical workflows to assist in identifying individuals at risk of developing diabetes. This can aid in early intervention, personalized treatment planning, and proactive management of the disease.

- Public Health Programs: Public health authorities can leverage the predictive model to identify high-risk populations and design targeted prevention and awareness campaigns. By focusing resources on at-risk communities, public health programs can effectively reduce the incidence of diabetes and its associated complications.

- Insurance Companies: Insurance providers can utilize the predictive model to assess the risk profiles of their policyholders. This information can help in developing risk-based pricing strategies, implementing proactive wellness programs, and providing incentives for healthy behaviours to manage costs associated with diabetes.

- Research Studies: The predictive model can be utilized in research studies to identify individuals at risk of diabetes, enabling researchers to investigate the effectiveness of different interventions, lifestyle modifications, or medications. This can contribute to the development of evidence-based guidelines for diabetes prevention and management.

- Mobile Applications and Wearable Devices: The predictive model can be integrated into mobile applications and wearable devices to provide real-time risk assessment and personalized recommendations for individuals. This empowers users to actively monitor their health, make informed lifestyle choices, and engage in self-management practices.

- Employee Wellness Programs: Organizations can incorporate the predictive model into their employee wellness programs to identify individuals at risk of diabetes among their workforces. This allows for targeted interventions, such as educational campaigns, health coaching, and incentives for healthy behaviours, to promote employee well-being and reduce healthcare costs.

- Telemedicine and Remote Monitoring: The predictive model can be utilized in telemedicine platforms and remote monitoring systems to provide risk assessment and guidance for individuals with limited access to healthcare facilities. This facilitates early detection and enables remote healthcare providers to deliver personalized care and interventions.

9. CONCLUSION

In conclusion, the project "Diabetes Prediction Using Machine Learning" addresses the critical need for early detection and prediction of diabetes, leveraging the power of machine learning algorithms. The development of an accurate and reliable predictive model holds immense potential to improve healthcare outcomes, reduce costs, and benefit various stakeholders.

By accurately identifying individuals at substantial risk of developing diabetes based on their health records and other relevant factors, the project contributes to early intervention and preventive measures. This timely detection allows for personalized treatment plans, lifestyle modifications, and proactive management, ultimately minimizing the impact of the disease and improving long-term health outcomes.

The advantages of the proposed solution are significant. Early detection leads to improved healthcare outcomes, as personalized interventions can be developed based on predicted risk levels. This approach not only enhances patient adherence but also optimizes resource allocation by focusing on high-risk individuals for preventive measures and interventions, thus reducing healthcare costs.

However, it is important to consider the potential challenges associated with the project. Data limitations, ethical considerations, model interpretability, implementation challenges, system updates, and the risk of false positives and negatives must be addressed to ensure the success and ethical implementation of the predictive model.

Overall, accurate diabetes prediction using machine learning has the potential to revolutionize healthcare. It empowers healthcare providers, public health authorities, insurance companies, and individuals with actionable insights for proactive management and prevention. The project paves the way for personalized care, improved population health, and the potential for reducing the burden of diabetes on individuals, healthcare systems, and society as a whole. With ongoing research, collaboration, and advancements in technology, the impact of this project can be further amplified, leading to healthier lives and more efficient healthcare systems.

10. FUTURE SCOPE

The future scope for the project "Diabetes Prediction Using Machine Learning" includes the following possibilities:

- Improved Prediction Models: Continual research can enhance the machine learning algorithms to make more accurate predictions by incorporating advanced techniques.

- Real-Time Monitoring: Integrating the prediction model with wearable devices and mobile apps can enable real-time monitoring and personalized recommendations for individuals at risk of diabetes.

- Comprehensive Risk Assessment: Expanding the project to consider additional risk factors like genetics, lifestyle, and medical history can create a more detailed risk assessment model for tailored preventive measures.

- Decision Support System: Developing a system to assist healthcare professionals in making informed decisions about treatment plans and lifestyle modifications for diabetic patients.

- Population Health Management: Analysing data from a larger population can provide insights into diabetes prevalence, regional variations, and trends for targeted interventions.

- Long-term Health Outcomes: Expanding the project to analyse long-term health outcomes and complications associated with diabetes to improve disease management and personalized healthcare plans.

11. BIBILOGRAPHY

- Kamrul Hasan's Method - https://ieeexplore.ieee.org/document/9076634

- Quan Zou's Method - https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full

- Nishith Kumar's Method - https://www.sciencedirect.com/science/article/abs/pii/S0169260717302821?via%3Dihub

- Maniruzzaman's Method - https://link.springer.com/article/10.1007/s13755-019-0095-z

- V. Jackins' Method - https://link.springer.com/article/10.1007/s11227-020-03481-x

APPENDIX

A. Source Code

https://github.com/harshchawla45/Diabetes-Prediction-using-Machine-Learning

B. Dataset

https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset