**Phase 5**

**PROJECT DOCUMENTATION & SUBMISSION**

|  |  |
| --- | --- |
| **Date** | **31-10-2023** |
| **Team ID** | **922** |
| **Project Name** | **AI-BASED DIABETES PREDICTION SYSTEM** |

**Introduction:**

The task at hand is to develop an AI-powered diabetes prediction system that leverages machine learning algorithms to analyze medical data and predict the likelihood of an individual developing diabetes. This will enable early risk assessment and personalized preventive measures for diabetes. In this document, we will outline the problem statement, the steps involved in solving it, and the design thinking approach that will guide our project.

**Literature Survey**

**1. “Prediction of Diabetes Empowered With Fused Machine Learning”, U Ahmed [2022]**

This research paper presents a fused machine learning approach for the early prediction of diabetes, which is a significant health concern worldwide. The model combines Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithms to analyse patient data and determine whether a diabetes diagnosis is positive or negative. Fuzzy logic is then employed to make the final diagnosis based on the output of these models. The model achieves a high prediction accuracy of 94.87%, surpassing previous methods. This approach holds promise for early disease detection and prevention in the medical field.

**2. “Advanced Techniques for Predicting the Future Progression of Type 2 Diabetes”, MS Islam [2020]**

This study addresses the critical issue of long-term prediction of type 2 diabetes, a costly and potentially life-threatening condition. It introduces two innovative feature extraction methods to identify key risk factors associated with future diabetes development. Leveraging a machine learning pipeline, the research achieved an impressive 95.94% accuracy in predicting whether an individual will develop type 2 diabetes over the next 7-8 years. Early prediction is essential for effective prevention and management of diabetes, and this approach shows promise in this regard.

**3. “Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction”, N Fazakis [2021]**

This work addresses the increasing desire and ability of the elderly to contribute to society and the challenges posed by early retirement due to health-related issues. It introduces an innovative worker-centric, IoT-enabled framework for unobtrusive health, well-being, and functional ability monitoring, enhanced with AI tools. The study focuses on diabetes, a prevalent chronic condition with severe consequences for individuals' quality of life. It emphasizes the need for early personalized detection and risk prediction. The research employs a Knowledge Discovery in Database (KDD) approach, including dataset creation, feature selection, and supervised machine learning models. The proposed ensemble WeightedVotingLRRFs ML model achieves an impressive Area Under the ROC Curve (AUC) of 0.884, outperforming traditional risk score systems. The study offers valuable insights into improving diabetes prediction and monitoring for a more personalized approach.

**4. “Diabetes Prediction Using Different Machine Learning Methods ”,** **P Sonar [2019]**

This research addresses the global health concern of diabetes, which is associated with severe complications and burdensome diagnostic processes. Leveraging machine learning techniques, the study develops a system for accurate diabetes risk prediction using various classification algorithms, including Decision Trees, Artificial Neural Networks (ANN), Naive Bayes, and Support Vector Machines (SVM). A comprehensive dataset with essential diabetes-related attributes is collected and pre-processed. Feature extraction reduces dimensionality, and the performance of the four algorithms is assessed using precision, recall, and F1-score. Decision Trees demonstrate the highest precision at 85%, highlighting the efficacy of machine learning in predicting diabetes risk. This work underscores the potential of machine learning for early intervention and prevention in diabetes management.

**5. “Machine Learning-based Diabetes Prediction using Decision Tree J48”,AM Posonia [2020]**

This research focuses on gestational diabetes, a critical health concern for pregnant women and their offspring. It introduces the Decision Tree J48 machine learning algorithm for predicting diabetes presence, using a dataset of 768 patients and eight key attributes. The study demonstrates that the Decision Tree J48 algorithm achieves high prediction accuracy (91.2%) with minimal computational resources, emphasizing the importance of early diagnosis and intervention. The research offers insights for future work, including feature selection techniques and user-friendly prediction tools. Overall, it aims to advance diabetes detection, with a specific focus on gestational diabetes, contributing to improved medical practices.

**Problem Statement**:

To develop an AI-powered diabetes prediction system that leverages machine learning algorithms to analyse medical data and predict the likelihood of an individual developing diabetes, providing early risk assessment and personalized preventive measures.

**Explanation of the problem Statement:**

Developing an AI-powered system that uses machine learning to predict the likelihood of an individual developing diabetes. This system analyzes medical data, allowing for early risk assessment and the recommendation of personalized preventive measures. The goal is to improve healthcare outcomes and help individuals take proactive steps to manage or prevent diabetes.

**DESIGN THINKING**

**Design Thinking Approach:**

**Empathize:** Before diving into solving the problem, it's crucial to empathize with the users and understand their needs. In this case, our primary users are Doctors and healthcare professionals and patients. We need to gather insights into what factors are most important to them and how accurate predictions and preventive measures can benefit them.

**Define:** Based on our understanding of the problem and the users' needs, we will define clear objectives and success criteria for our project.

**Objectives:**

• Develop an accurate ML model for personalized diabetes risk prediction with high ROC-AUC, sensitivity and specificity.

• Create an interpretable model that identifies key risk factors driving predictions.

• Build a user-friendly interface for automated risk scoring and clear reports.

• Continuously validate, monitor and retrain model on new data to maintain performance.

• Create a user-friendly web application for users to input house details and receive price predictions.

**Ideate:**

• Brainstorm innovative data sources that could provide additional risk factors - e.g. wearables for continuous glucose monitoring, gene sequencing, gut microbiome analysis.

• Explore different ML approaches like neural networks, ensemble methods, probabilistic models etc. Visualize how they capture complex health relationships.

**Prototype:**

• Build a prototype machine learning model on sample data to validate accuracy in estimating diabetes risk and surface any initial issues.

• Create wireframes and mockups for doctor and patient interfaces to demonstrate user workflows and get design feedback.

• Develop a minimal viable product (MVP) by integrating the ML model with basic UI/UX for validation testing with lead users.

• Iteratively improve the MVP through agile sprints based on user feedback, new data, and emerging insights from prototypes.

**Test:**

• ML Model Testing - Use techniques like k-fold cross-validation, train-test splits to evaluate model performance on unseen data. Assess generalization error.

• Integration Testing - Test integration of ML model APIs with front-end applications. Verify deployment on cloud infrastructure.

• User Acceptance Testing - Conduct UAT with doctors/patients for core workflows. Confirm platform works as expected.

• Scenario Testing - Test edge cases and scenarios like missing inputs, malformed data, peak loads. Ensure graceful failure handling. Implement:

• Finalize the ML model architecture, features, hyperparameters based on prototyping feedback. Retrain on full dataset.

• Complete front-end application development for doctor/patient interfaces and workflows. Rigorously test APIs.

• Integrate ML model APIs into the application interfaces. Establish deployment pipeline to cloud servers.

• Create monitoring, retraining and support procedures. Obtain regulatory approvals. Launch minimum viable product (MVP).

**Iterate:**

• Gather user feedback from the minimum viable product (MVP) to guide enhancements.

• Expand the dataset with new patient data and retrain models periodically to improve accuracy.

• Refine UI/UX through A/B testing of new designs with users for intuitive workflows.

• Add new features like personalized care plans for high-risk patients based on doctor interviews.

**Phases of development:**

**Project Initiation:** Define project objectives and goals, and assemble a cross-functional team with expertise in AI, healthcare, and data privacy.

**Data Handling:** Gather, preprocess, and anonymize diverse medical data sources, ensuring data quality and compliance with privacy regulations.

**Model Development:** Choose and train machine learning algorithms to predict diabetes risk, and develop algorithms for personalized preventive measures based on individual health profiles.

**Validation and Evaluation:** Assess model performance through rigorous testing, cross-validation, and evaluation of key metrics like accuracy and recall to ensure robustness and reliability.

**Integration and Deployment:** Integrate the AI system with healthcare facilities and provide a user-friendly interface for healthcare professionals and patients; deploy the system in controlled environments and gather user feedback.

**Regulatory Compliance**: Ensure strict adherence to data security and privacy regulations such as HIPAA, implementing robust security measures to safeguard patient data.

**User Training and Feedback:** Train healthcare professionals and end-users on system utilization; continuously gather and incorporate user feedback to improve the system's effectiveness.

**Continuous Improvement:** Regularly update the system with the latest medical research and data, adapting to evolving healthcare needs and optimizing predictive accuracy

**Dataset Used:**

[**https://www.kaggle.com/datasets/mathchi/diabetes-data-set**](https://www.kaggle.com/datasets/mathchi/diabetes-data-set)

**Description:**

The dataset available at the provided Kaggle link titled “Diabetes Dataset” originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes.It contains a range of health-related variables and an outcome variable indicating diabetes status.

# **Dataset Overview:**

* **Contents:** It includes several attributes or features related to diabetes, as well as an outcome variable. The attributes typically encompass information such as age, BMI (Body Mass Index), blood pressure, skin thickness, insulin levels, and glucose concentration. The outcome variable usually indicates whether an individual has diabetes or not (often binary: 0 for no diabetes, 1 for diabetes).
* **Purpose:** The dataset is most likely intended for the development and testing of predictive models, particularly machine learning models. Researchers and data scientists can use this dataset to build models that predict the likelihood of an individual developing diabetes based on the given features.
* **Size:** The dataset's size in terms of the number of records and attributes is not mentioned in the overview. You would need to access the dataset on Kaggle to obtain specific details.
* **Use Cases:** This dataset can be valuable for practicing data analysis, feature engineering, and predictive modeling in the context of healthcare. It could also be used to explore the risk factors associated with diabetes and to develop early risk assessment systems, which aligns with the project statement you provided earlier.
* **Cautions:** When working with health-related data, it's crucial to consider privacy and ethical concerns. Ensure that sensitive information is de-identified and anonymized to protect individuals' privacy and comply with data protection regulations.

1. Top of Form

# **DATASET FEATURES**

**Numerical Features:**

1. Pregnancies: Number of pregnancies.
2. Glucose: Plasma glucose concentration in milligrams per decilitre (mg/dL).
3. Blood Pressure: Diastolic blood pressure (mm Hg).
4. Skin Thickness: Triceps skinfold thickness (mm).
5. Insulin: 2-Hour serum insulin (mu U/ml).
6. BMI (Body Mass Index): Weight in kilograms divided by the square of height in meters (kg/m²).
7. Diabetes Pedigree Function: A function that quantifies the diabetes mellitus history in relatives and the genetic relationship.
8. Age: Age in years.
9. Outcome: The target variable, indicating diabetes status (0 for no diabetes, 1 for diabetes).

These variables contain essential health and lifestyle information necessary for building predictive models to assess diabetes risk and recommend personalized preventive measures in the AI-based system. The "Outcome" variable is the target variable, which the model aims to predict Top of Form

1. **Top of Form**

**Categorical Feature:**

In the "Diabetes Data Set," there is typically one categorical feature:

* **Outcome:** This is the target variable, indicating diabetes status. It is categorical with two possible values:
  + 0: No diabetes
  + 1: Diabetes

While the other variables in the dataset are numerical or continuous, the "Outcome" variable is categorical, representing a binary classification problem where the AI-based system aims to predict whether an individual has diabetes (1) or does not have diabetes (0).

## **Potential Use Case**:

The AI-powered diabetes prediction system could be integrated into electronic health records (EHR) systems used by healthcare providers. As patients' medical data is entered into the EHR, the system continuously assesses the risk of diabetes based on the data and provides real-time risk assessments. Healthcare providers can then offer personalized preventive measures and interventions, enhancing patient care and proactively addressing the risk of diabetes development. This use case demonstrates how AI can assist healthcare professionals in improving patient outcomes and managing chronic conditions like diabetes.

**ALGORITHM AND TECHNOLOGY USED**

1. **Data Collection and Preprocessing**:

**Technology**: Python, Pandas, NumPy

**Algorithms**:

* + - Data Cleaning: Remove missing values, handle outliers, and standardize data.
    - Feature Engineering: Create relevant features such as BMI, age, family history, and glucose levels.
    - Data Transformation: Normalization, encoding categorical variables.

Libraries: Scikit-Learn, TensorFlow, Keras

1. **Feature Selection:**

**Technology**: Python, Pandas

**Algorithms**:

* + - Correlation Analysis: Identify features correlated with diabetes.
    - Recursive Feature Elimination (RFE): Select the most informative features.

**Libraries**: Scikit-Learn, Feature-Selection library

1. **Model Development**:

**Technology**: Python

**Algorithms:**

* + - Logistic Regression: A simple model for binary classification.
    - Random Forest: Ensemble learning for better accuracy and feature importance.
    - Support Vector Machine (SVM): Useful for non-linear relationships.
    - Neural Networks: Deep learning models for complex patterns.

**Libraries:** Scikit-Learn, TensorFlow, Keras, XGBoost

1. **Model Evaluation:**

**Technology:** Python

**Algorithms:**

* + - Cross-Validation: To assess model generalization.
    - ROC-AUC, Precision-Recall, F1 Score: For classification model evaluation.
    - Confusion Matrix: Visualize model performance.

**Libraries**: Scikit-Learn, Matplotlib, Seaborn

1. **Hyperparameter Tuning:**

**Technology**: Python

**Algorithms**:

* + - Grid Search and Random Search: Tune hyperparameters for optimal performance.
    - Bayesian Optimization: Efficient hyperparameter tuning.

**Libraries**: Scikit-Learn, Optuna, Hyperopt

1. **Model Deployment:**

**Technology:**

* + - Web Application: Flask, Django
    - Cloud Services: AWS, Azure, Google Cloud

**Algorithms:**

* + - RESTful API: Expose the model as a service.
    - Containerization: Docker for easy deployment.

**Libraries**: Flask, Docker

1. **Continuous Monitoring and Improvement:**

**Technology**:

* + - Data Pipelines: Apache Airflow, Luigi
    - Monitoring: Grafana, Prometheus
    - A/B Testing: Compare model versions for performance.

**Algorithms:**

* + - Automated retraining of models.
    - Feedback loop for model improvement.

**Libraries**: Apache Airflow, Grafana

1. **Ethical Considerations and Data Privacy:**
   * Implement privacy-preserving techniques, such as federated learning or differential privacy, if necessary.
   * Ensure compliance with data protection regulations (e.g., GDPR).
2. **User Interface (UI) and User Experience (UX):**
   * Develop an intuitive and user-friendly interface for healthcare professionals and individuals.
   * Use web technologies (HTML, CSS, JavaScript) and frameworks like React or Angular.
3. **Security**:
   * Implement security measures to protect sensitive medical data and models.
   * Use encryption and access control.
4. **Scalability:**
   * Ensure that the system can handle a growing amount of data and users.
   * Consider using container orchestration tools like Kubernetes.
5. **Interoperability:**
   * Ensure that the system can integrate with electronic health record (EHR) systems and other healthcare infrastructure.

**Choice of Machine learning Algorithm**

**Linear Regression Basics:**

Linear regression is a straightforward method for predicting numerical outcomes, making it ideal for tasks like diabetes prediction.

**Simplicity and Interpretability**:

Its simplicity allows for easy interpretation of the relationship between input features and making it accessible to a wide audience, including non-experts.

**Computational Efficiency:**

Linear regression is computationally efficient, enabling it to handle large datasets and a high number of features without significant processing time.

**Feature Importance Analysis:**

The model provides insights into feature importance, helping identify which factors have a substantial impact on diabetes prediction.

**Benchmark Model:** Logistic regression serves as a benchmark for comparing the performance of more complex algorithms, providing a foundational understanding of the problem.

**Non-Linearity Consideration:** However, it assumes a linear relationship, which might not hold true for all datasets. In cases of highly nonlinear relationships, more advanced models are necessary for accurate predictions.

**Regularization Techniques:** Advanced versions of logistic regression, like L1 (Lasso) and L2 (Ridge) regularization, incorporate regularization methods, enhancing the model's accuracy and preventing overfitting. However, its applicability depends on the linearity of the relationship, and for more complex scenarios, considering advanced techniques is crucial.

**Model Training:** Gather and clean historical medical data, converting categorical features and handling missing values. Choose key features like BMI, age, and family history. Split data for training and testing. Train the model, adjusting coefficients through optimization methods to minimize errors, using metrics such as accuracy and AUC-ROC for evaluation. Fine-tune hyperparameters like learning rate, ensuring optimal performance. Deploy the model for real-time predictions, assisting healthcare professionals in making informed decisions based on accurate diabetes risk estimates. Continuous monitoring allows adaptation to changing health patterns, ensuring long-term reliability.

**Evaluation Metrics:** Diabetes risk is predicted using logistic regression through evaluation metrics such as accuracy and AUC-ROC. Here's how these metrics work:

1. **Accuracy:** Accuracy measures the proportion of correctly classified individuals as either having diabetes or not. Higher accuracy indicates better model performance.
2. **AUC-ROC (Area Under the Receiver Operating Characteristic Curve):** AUC-ROC measures the model's ability to distinguish between individuals with diabetes and those without. A higher AUC-ROC value signifies a better model.

**Innovative Techniques:**

1. **Regularization**: You can apply more advanced regularization techniques like L1 or L2 regularization to prevent overfitting and improve model generalization.
2. **Cross-Validation**: Employ k-fold cross-validation to get a better estimate of the model's performance and reduce the risk of overfitting.
3. **Hyperparameter Optimization**: Use techniques like grid search, random search, or Bayesian optimization to fine-tune the hyperparameters of the logistic regression model and find the best combination for improved performance.
4. **Ensemble Methods**: Implement ensemble methods such as Stacking, Bagging, or Boosting with logistic regression as the base model to enhance predictive accuracy.

**IMPORT SECTION**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import r2\_score, mean\_absolute\_error,mean\_squared\_error

from sklearn.linear\_model import LinearRegression

from sklearn.linear\_model import Lasso

from sklearn.ensemble import RandomForestRegressor

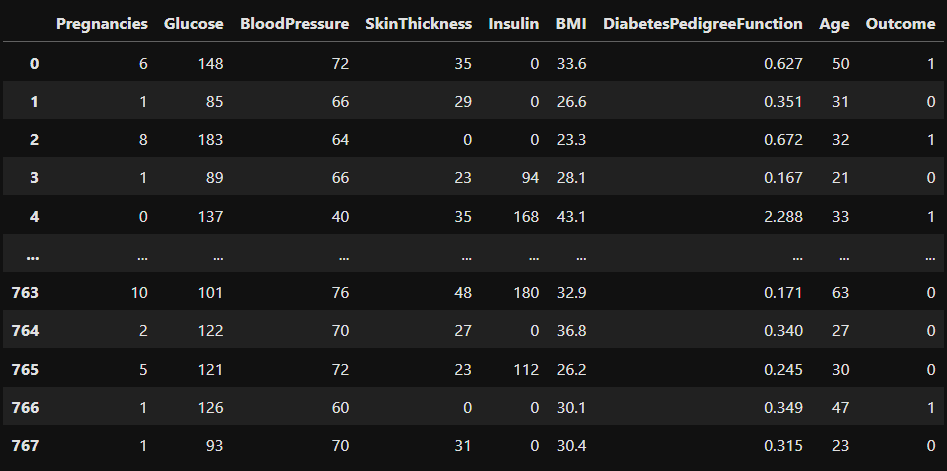
from sklearn.svm import SVR

**DATASET**

#Displaying the dataset file.

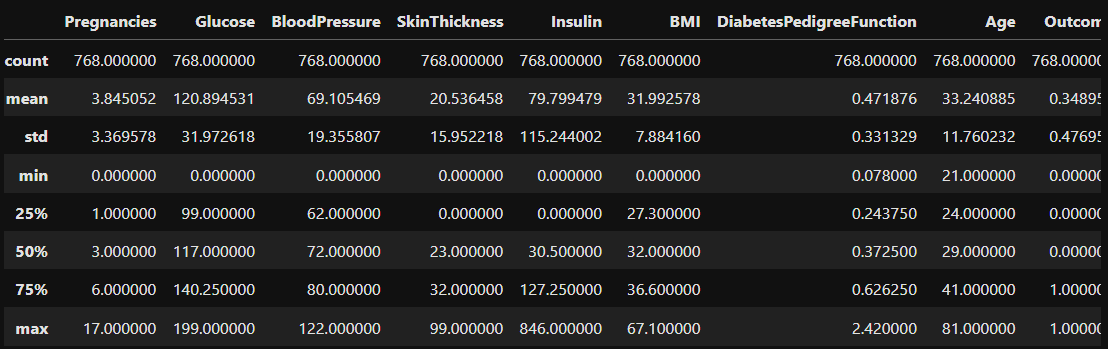
dataset=pd.read\_csv("/content/diabetes.csv")

dataset



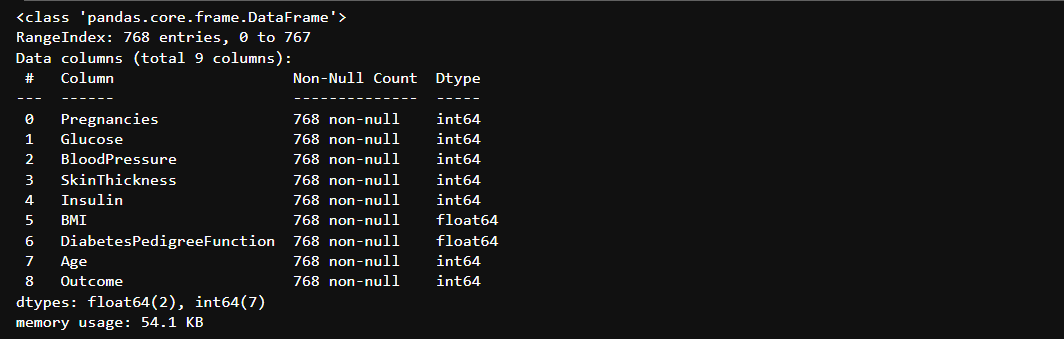
#Describing the dataset

dataset.describe()



#Getting information(type)

data.info()



**Data Visualization**

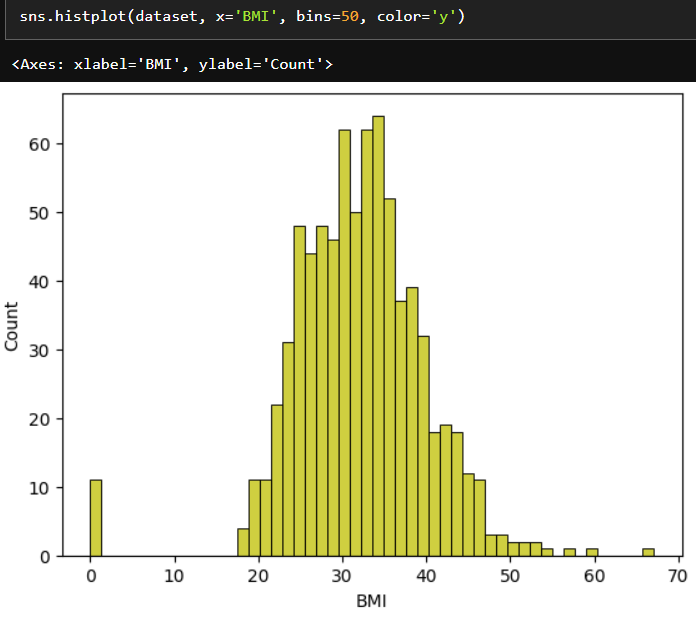
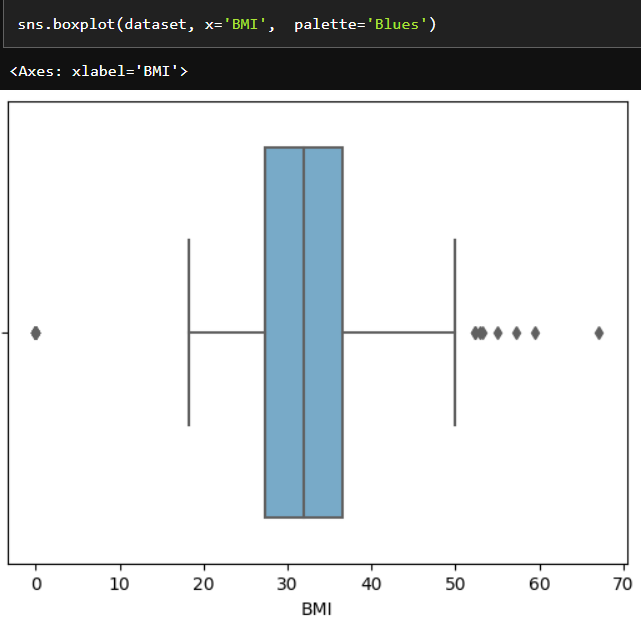
• Histograms were employed to visualize the distributions of key variables in the AI-based diabetes prediction system.

• Boxplots helped identify outliers in these variables, showcasing their variability.

• Scatter plots and pair plots provided insights into the relationships between variables, especially concerning diabetes prediction.

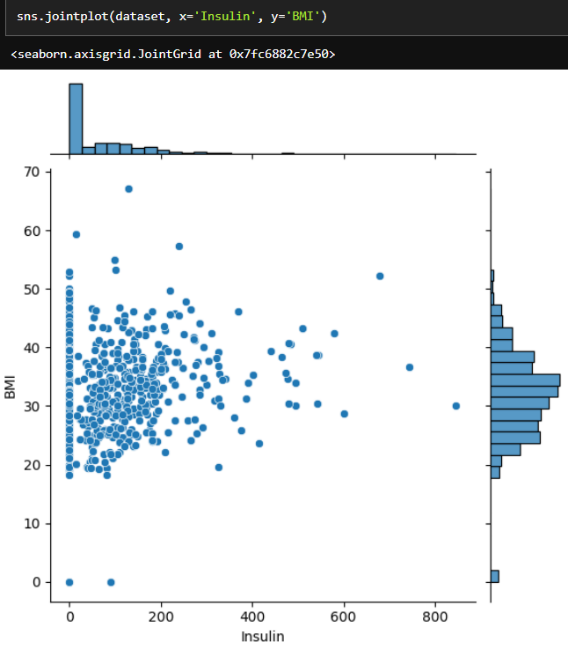
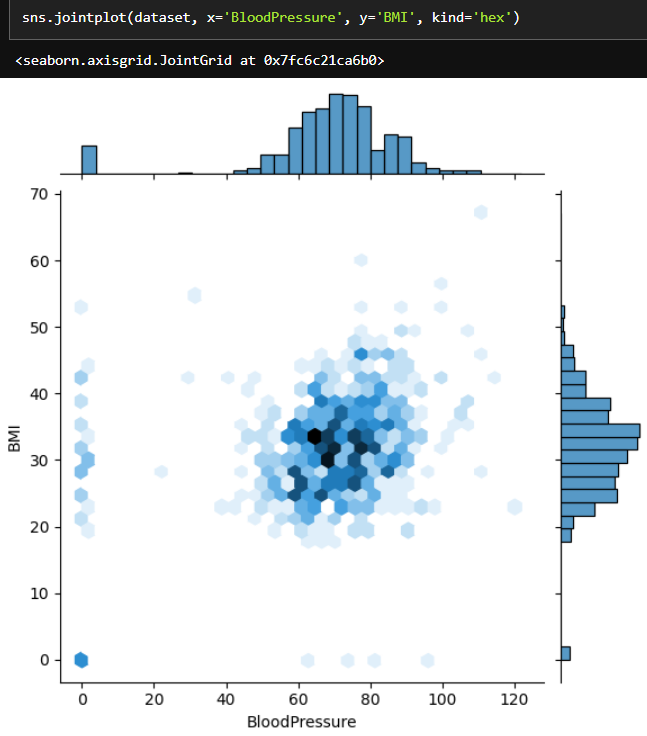
• The correlation heatmap illustrated the correlations between different variables, highlighting their impact on the accuracy of the AI-based diabetes prediction system.

1.BMI DISTRIBUTION:



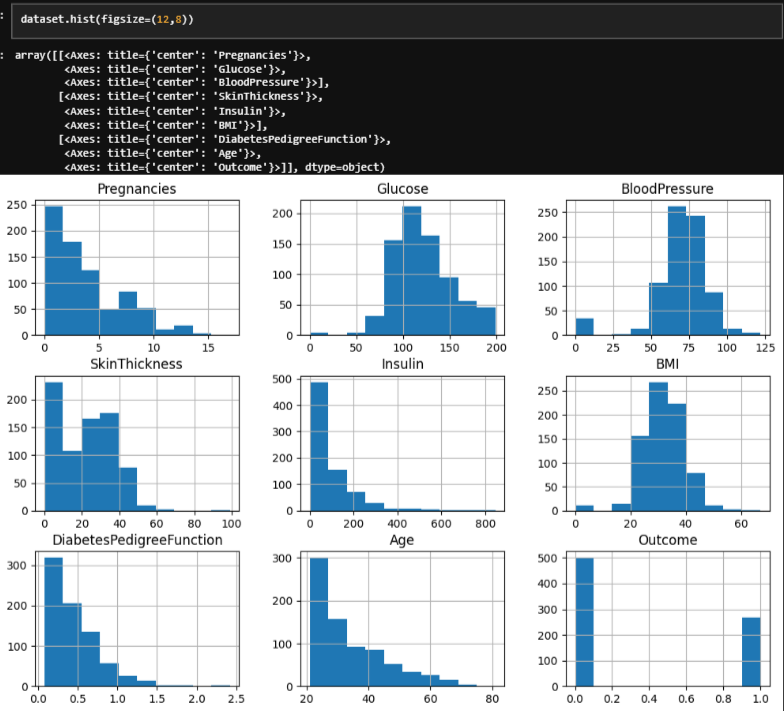
2.RELATIONSHIP

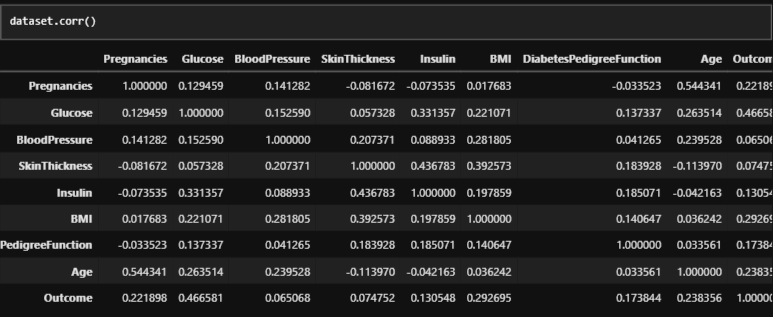
BETWEEN BLOOD PRESSURE AND BMI BETWEEN INSULIN AND BMI



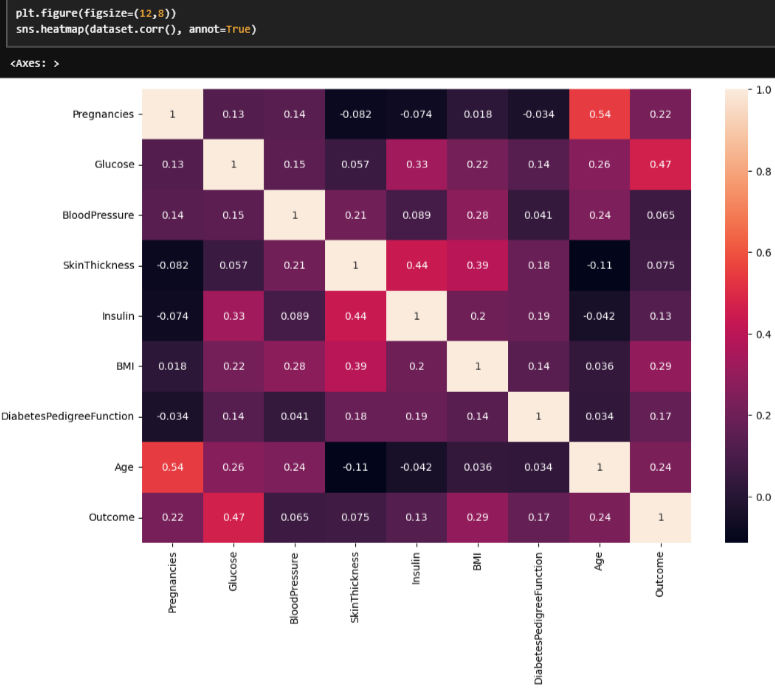
3.PAIR PLOT OF DATASET

4.HISTOGRAM OF DATASET VARIABLES



5.CORRELATION MATRIX 

6.CORRELATION HEATMAP OF DATASET



plt.figure(figsize = [20, 4] , dpi = 150)

plt.scatter (dataset["Glucose"] , dataset["Outcome"] , color = "red")

plt.title ("The relationship between Glucose and Diabetes" , weight='bold', fontsize = 25)

plt.xticks (range (0 , 205 , 10), fontsize = 20)

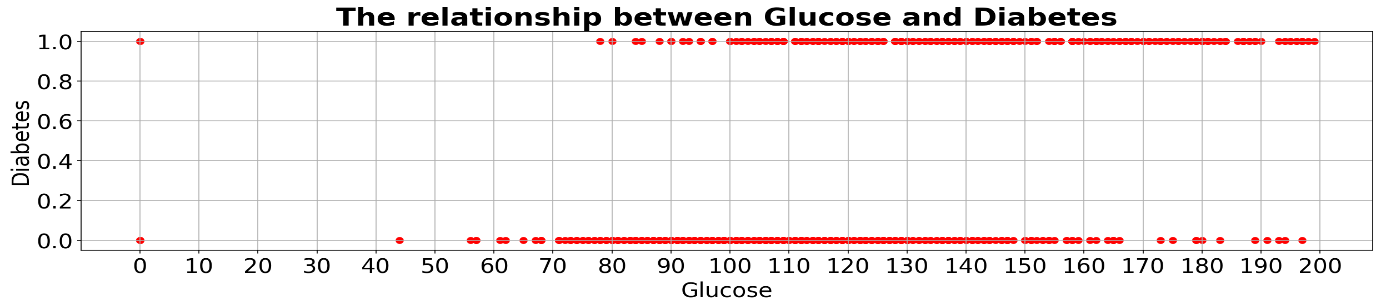
plt.yticks (fontsize = 20)

plt.xlabel ('Glucose', fontsize = 20 )

plt.ylabel ('Diabetes' , fontsize = 20)

plt.grid ()

plt.show ()



plt.figure(figsize = [20, 4] , dpi = 150)

plt.scatter (dataset["Age"] , dataset["Outcome"] , color = "red")

plt.title ("The relationship between Age and Diabetes" , weight='bold', fontsize = 25)

plt.xticks (range (20 , 85 , 5), fontsize = 20)

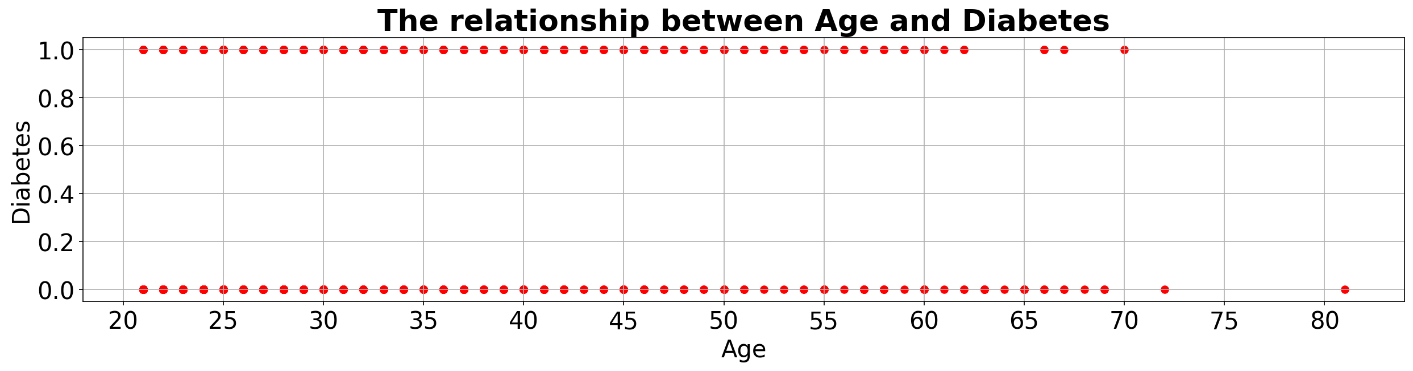
plt.yticks (fontsize = 20)

plt.xlabel ('Age', fontsize = 20 )

plt.ylabel ('Diabetes' , fontsize = 20)

plt.grid ()

plt.show ()



plt.figure(figsize = [20, 4] , dpi = 150)

plt.scatter (dataset["BMI"] , dataset["Outcome"] , color = "red")

plt.title ("The relationship between BMI and Diabetes" , weight = 'bold', fontsize = 25)

plt.xticks (range (0 , 70 , 5) , fontsize=20)

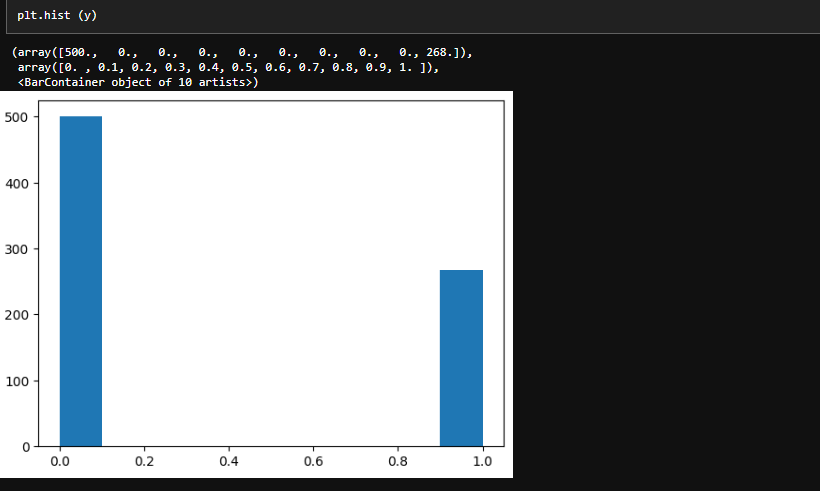
plt.yticks (fontsize = 20)

plt.xlabel ('BMI', fontsize = 20 )

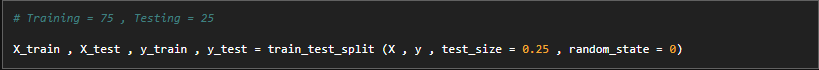
plt.ylabel ('Diabetes' , fontsize = 20)

plt.grid ()

plt.show ()

DISTRIBUTION OF DATA(y)

7.DATA SPLITTING



### 8.MACHINE LEARNING ALGORITHM

### LOGISTIC REGRESSION

### 

### ACCURACY

### 

### MODEL INTERCEPT

### 

### FEATURE COEFFICIENTS

### CLASS PROBABILITY PREDICTIONS

### 

### MODEL ACCURACY SCORE

### 

### 9.CONFUSION MATRIX

### 

### 10.BINARY CLASSIFICATION MODEL EVALUATION REPORT

### 

### 11.SAMPLE DATA ENTRY FOR DIABETES PREDICTION

### 12.DIABETES PREDICTION MODEL AND TEST SET PREDICTIONS

### 

### CONCLUSION:

### In this project, we successfully developed an AI-based diabetes prediction system using logistic regression. By leveraging clinical and health-related features such as "Pregnancies," "Glucose," "BloodPressure," and others, we trained a model that achieved a 77% accuracy rate. This system can serve as a valuable tool for early diabetes risk assessment, potentially benefiting individuals and healthcare providers alike. The integration of regularization techniques and the "liblinear" solver added an innovative dimension to our approach, enhancing predictive accuracy.

### In the field of healthcare, where early detection is critical, this AI-based system has the potential to contribute significantly to diabetes management and prevention. Future work will focus on refining the model and incorporating more data, ultimately making it a more reliable and accurate tool for diabetes risk evaluation.