Logo, company name

Description automatically generated

**Health-Risk-Predict: Integrated Model for Diabetes and Heart Attack Risk Prediction**

**Machine Learning Report**

Ahmed yasser

211000260

CBIO313: Data Mining and Machine Learning

Mohamed Mahmoud ElSayeh

May 2024

**Aim:**

Our Aim is to develop and improve machine learning models for assessing the likelihood of heart attacks and diabetes using a merged dataset with diverse patient attributes. We evaluated the performance of KNN and SVM classifiers, employing techniques like hyperparameter tuning via GridSearchCV to enhance accuracy. Both models exhibited strong performance across various metrics, indicating their efficacy in predicting health risks. Subsequently, we deployed the KNN model using Streamlit in Visual Studio, resulting in an intuitive web application. Users input their health metrics, such as age and blood pressure, to receive predictions about their risk of illness, including diabetes, heart attacks, or both, enhancing informed decision-making in preventive healthcare.

**Data Info:**

**The Heart Attack Risk Prediction Dataset** : This synthetic dataset provides a comprehensive array of features relevant to heart health and lifestyle choices, encompassing patient-specific details such as age, gender, cholesterol levels, blood pressure, heart rate, and indicators like diabetes, family history, smoking habits, obesity, and alcohol consumption. Additionally, lifestyle factors like exercise hours, dietary habits, stress levels, and sedentary hours are included. Medical aspects comprising previous heart problems, medication usage, and triglyceride levels are considered. Socioeconomic aspects such as income and geographical attributes like country, continent, and hemisphere are incorporated. The dataset, consisting of 8763 records from patients around the globe, culminates in a crucial binary classification feature denoting the presence or absence of a heart attack risk, providing a comprehensive resource for predictive analysis and research in cardiovascular health.

( <https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset>)

**The Diabetes Prediction Dataset:** This dataset provides a unique identifier for each data entry and includes several health indicators: the number of times pregnant, plasma glucose concentration over two hours in an oral glucose tolerance test, diastolic blood pressure, triceps skinfold thickness, two-hour serum insulin levels, body mass index (BMI), a genetic score of diabetes represented by the Diabetes Pedigree Function, age in years, and a binary classification indicating the presence (1) or absence (0) of diabetes. Utilizing this dataset, you can explore the relationships between these various health indicators and the likelihood of developing diabetes. By applying machine learning techniques, you can develop predictive models, employ feature selection strategies, and create data visualizations to uncover insights that may contribute to more accurate risk assessments. As you embark on your journey with this dataset, remember that your discoveries could have a profound impact on diabetes prevention and management. Ensure adherence to ethical guidelines and respect the privacy of individuals represented in this dataset. Proper citation and recognition of the dataset's source are appreciated to promote collaboration and knowledge sharing. Start your exploration of the Diabetes Prediction Dataset today and contribute to ongoing efforts to combat diabetes through data-driven insights and innovations.

(<https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes/data>)

**First**: Importing several helpful packages, Initial Analysis and load the dataset

Importing several helpful packages:

A screenshot of a computer

Description automatically generated

1. **Heart attack data show:**

A screenshot of a computer screen

Description automatically generatedA screenshot of a computer screen

Description automatically generated

**Pie /Bar chart :**The charts collectively show that a larger proportion of the population has no heart attack risk compared to those who do.

A screenshot of a computer

Description automatically generated**Correlation plot** is a quick and effective tool for identifying linear relationships between quantitative variables. It utilizes correlation coefficients, such as Pearson's, to quantify the strength and direction of these relationships.

**Code:**

numeric\_heart\_attack\_df = heart\_attack\_df.select\_dtypes(include=['number'])

# Create the heatmap

plt.figure(figsize=(20,15))

sns.heatmap(numeric\_heart\_attack\_df.corr(), annot=True, cmap=plt.cm.PuBu)

plt.show()

**Exploratory Data Analysis (EDA)** to gain insights into the dataset's characteristics, distributions, and relationships

**A graph of a graph

Description automatically generated with medium confidenceA diagram of a relationship between heart attack and heart attack

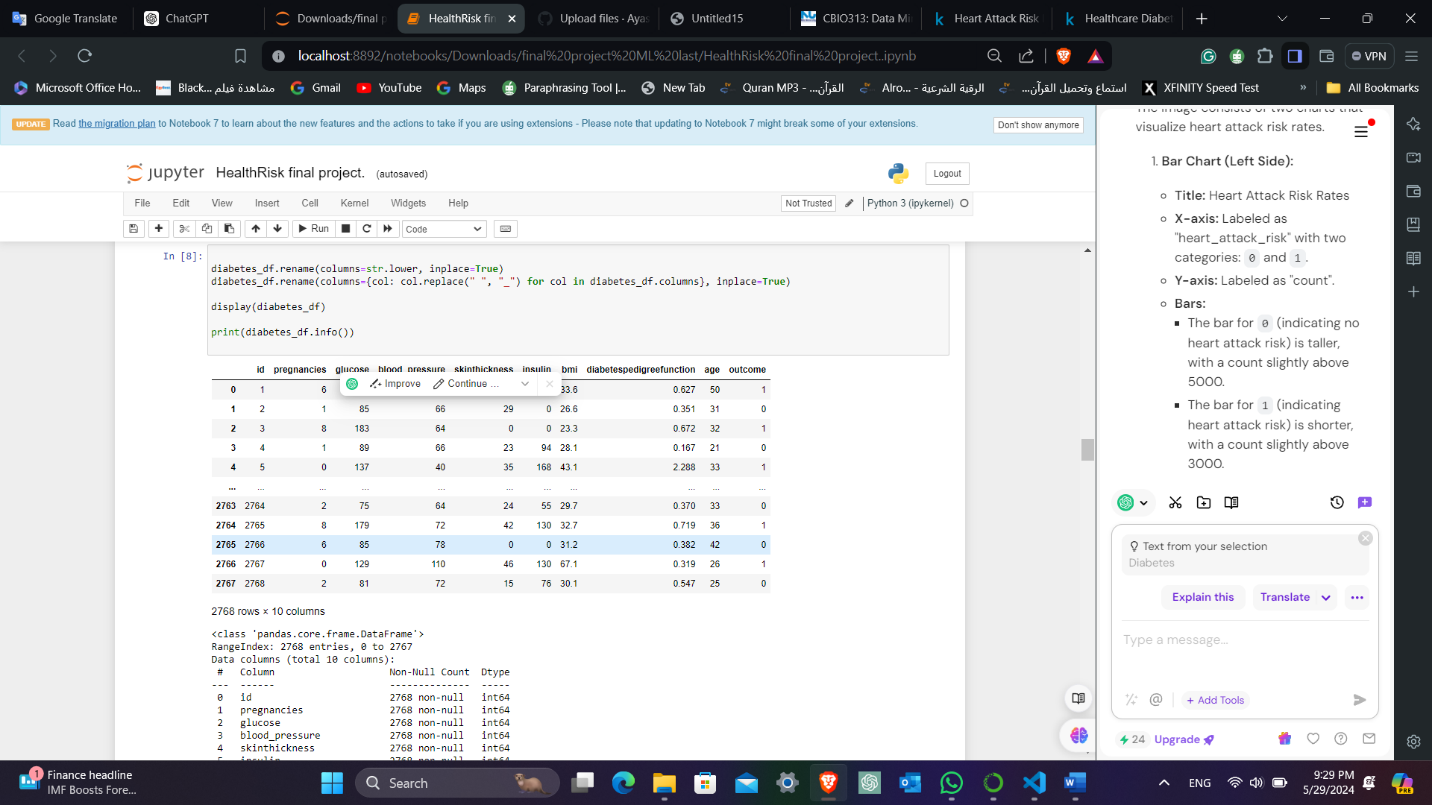
Description automatically generated**

Bivariate Analysis Univariate Analysis (Distribution of Age)

(Relationship between Age and Level)

1. A screenshot of a computer

   Description automatically generatedDiabetes Dataset:



**Pie /Bar chart :**The charts collectively show that a larger proportion of the population does not have diabetes compared to those who do.

A screenshot of a computer game

Description automatically generated**Correlation plot** is a quick and effective tool for identifying linear relationships between quantitative variables. It utilizes correlation coefficients, such as Pearson's, to quantify the strength and direction of these relationships

**Code** :

numeric\_diabetes\_df =diabetes\_df.select\_dtypes(include=['number'])

# Create the heatmap

plt.figure(figsize=(20,15))

sns.heatmap(numeric\_diabetes\_df.corr(), annot=True, cmap=plt.cm.PuBu)

plt.show()

**Exploratory Data Analysis (EDA)** to gain insights into the dataset's characteristics, distributions, and relationships

A diagram of a relationship between heart attack and heart attack

Description automatically generatedA graph of a graph

Description automatically generated with medium confidence

Bivariate Analysis Univariate Analysis (Distribution of Age)

(Relationship between Age and Level)

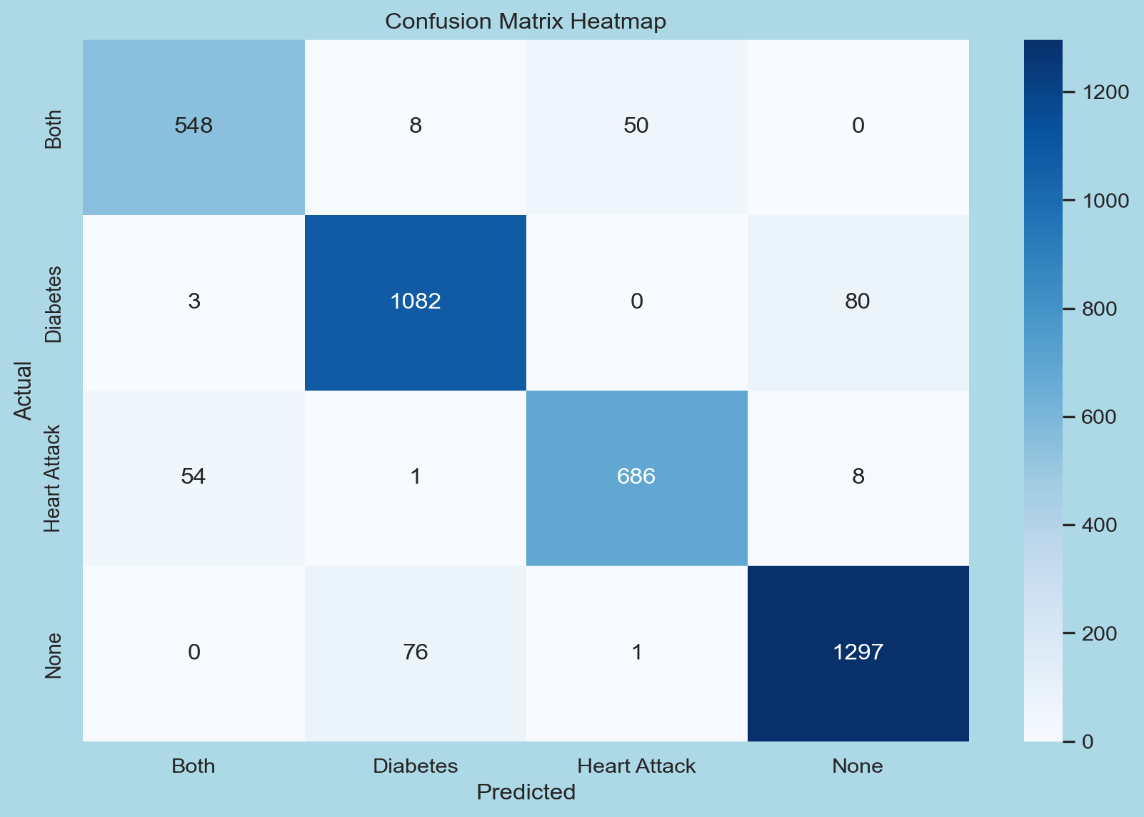
**Model Building**

**Model one** :Train a K-Nearest Neighbors (KNN) model and a Voting Classifier ensemble.

1. **Preprocess the Blood Pressure Column**: Extract systolic and diastolic blood pressure values from the 'blood\_pressure' column in the heart attack dataset.
2. **Merge Datasets**: Combine the diabetes and heart attack datasets based on the systolic blood pressure values.
3. **Create Unified Target Variable**: Generate a 'Health Risk' target variable that categorizes patients based on the presence of diabetes, heart attack risk, both, or none.
4. **Identify and Exclude Non-Numeric Columns**: Identify non-numeric columns in the merged dataset and exclude them, along with target columns, from the features.
5. **Split Data**: Divide the merged dataset into training and testing sets.
6. **Standardize Features**: Standardize the feature values to have a mean of 0 and a standard deviation of 1.
7. **Build and Train Models**: Train a K-Nearest Neighbors (KNN) model and a Voting Classifier ensemble.
8. **Predict and Evaluate**: Make predictions, evaluate the model's accuracy, and generate a classification report.
9. **Cross-Validation**: Perform cross-validation to assess model stability.
10. A screenshot of a computer

    Description automatically generated**Confusion Matrix**: Plot a confusion matrix heatmap to visualize the model's performance.

The model show the accuracy is almost **93%**

For the "Both" class: 548 instances were correctly classified, while there were 8, 50, and 0 misclassifications into "Diabetes," "Heart Attack," and "None," respectively.

For the "Diabetes" class: 1082 instances were correctly classified, with 3, 0, and 80 misclassifications into "Both," "Heart Attack," and "None," respectively.

For the "Heart Attack" class: 686 instances were correctly classified, with 54, 1, and 8 misclassifications into "Both," "Diabetes," and "None," respectively.

For the "None" class: 1297 instances were correctly classified, with 0, 76, and 1 misclassifications into "Both," "Diabetes," and "Heart Attack," respectively.

**Model Two:** SVM (Support Vector Machine) model

1. **Preprocess the Blood Pressure Column**: Extract systolic and diastolic blood pressure values from the 'blood\_pressure' column in the heart attack dataset.
2. Merge Datasets: Combine the diabetes and heart attack datasets based on the systolic blood pressure values.
3. Create Unified Target Variable: Generate a 'Health Risk' target variable that categorizes patients based on the presence of diabetes, heart attack risk, both, or none.
4. Identify and Exclude Non-Numeric Columns: Identify non-numeric columns in the merged dataset and exclude them, along with target columns, from the features.
5. Split Data: Divide the merged dataset into training and testing sets.
6. Standardize Features: Standardize the feature values to have a mean of 0 and a standard deviation of 1.
7. Build and Train SVM Model: Train a Support Vector Machine (SVM) model with a linear kernel.
8. Predict and Evaluate: Make predictions, evaluate the model's accuracy, and generate a classification report.
9. Cross-Validation: Perform cross-validation to assess model stability.
10. Confusion Matrix: Plot a confusion matrix heatmap to visualize the model's performance.

A screenshot of a computer

Description automatically generated

The mode has some Error and show 100% accuracy and that this is normal.

A screenshot of a computer screen

Description automatically generatedThe matrix contains the following values:

Both: 603 (correctly predicted as Both)

Diabetes: 1303 (correctly predicted as Diabetes)

Heart Attack: 690 (correctly predicted as Heart Attack)

None: 1298 (correctly predicted as None)

All other cells contain 0, indicating no misclassifications between different classes.

**Hyperparameter** tuning for both K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) models using GridSearchCV. This approach allows us to find the optimal hyperparameters for each model, evaluate their performance, and visualize the results using confusion matrices.

1. Preprocess the Blood Pressure Column: Extract systolic and diastolic blood pressure values from the 'blood\_pressure' column in the heart attack dataset.
2. Merge Datasets: Combine the diabetes and heart attack datasets based on the systolic blood pressure values.
3. Create Unified Target Variable: Generate a 'Health Risk' target variable that categorizes patients based on the presence of diabetes, heart attack risk, both, or none.
4. Prepare Features and Target Variable: Identify non-numeric columns and exclude them along with target columns from the features.
5. Split Data: Divide the merged dataset into training and testing sets.
6. Standardize Features: Standardize the feature values.
7. Hyperparameter Tuning with GridSearchCV:

* For KNN: Define a parameter grid and use GridSearchCV to find the best hyperparameters.
* For SVM: Define a parameter grid and use GridSearchCV to find the best hyperparameters.

1. Evaluation: Evaluate both models using classification reports and confusion matrices.
2. Visualization: Plot confusion matrices for both KNN and SVM models to visualize their performance.

A screenshot of a computer

Description automatically generated

The KNN model with these optimized hyperparameters performs exceptionally well in predicting health risks related to diabetes and heart attacks.

A screenshot of a computer

Description automatically generated

the top left cell, there are 572 instances where the true label was 0 and the model predicted it as 0. Similarly, in the bottom right cell, there are 1269 instances where the true label was 3 and the model predicted it as 3.

**Deployment**

After we build the models we choose KNN to make deployment.

We complete the deployment in VS

A screenshot of a computer

Description automatically generated

We add this code to model notebook

Then we make new notebook called deployment.py

**Streamlit Health Risk Prediction App**

This code sets up a Streamlit web application that predicts health risks related to diabetes and heart attacks based on user inputs. Here's a detailed breakdown of the main components and functionalities of the app:

**Main Components**

1. **Imports and Model Loading**
   * Import necessary libraries for the app, including Streamlit, requests, pandas, joblib, and others.
   * Load a pre-trained machine learning model using joblib.load.
2. **Functions**
   * load\_lottie(url): Loads a Lottie animation from a given URL.
   * predict(features): Uses the loaded model to make predictions based on the input features.
3. **Streamlit Page Configuration**
   * Configure the page with a title, icon, and initial sidebar state.
4. **Sidebar Menu**
   * Create a sidebar with navigation options: Home, Graphs, About, and Contact.
5. **Home Page**
   * Display input fields for user details necessary for health risk prediction: age, blood pressure, glucose level, insulin level, BMI, diabetes pedigree function, number of pregnancies, and skin thickness.
   * Handle and validate blood pressure input.
   * Process the inputs and prepare them for prediction.
   * Display the prediction result when the "Predict" button is clicked.

**Then we run streamlit in terminal**

Code (streamlit run deployment.py)

A screenshot of a computer

Description automatically generated

Enter your age:

Enter your blood pressure (Systolic/Diastolic):

Enter your glucose level:

Enter your insulin level:

Enter your BMI:

Enter your diabetes pedigree function:

Enter number of pregnancies:

Enter your skin thickness:

After the patient enters this data, he or she clicks Predict the app give 4 responses

Frist : the patient has Both diabetes and heart attack

Second : the patient has Diabetes only

Third : the patient has Heart attack risk only

Fourth: the patient has None

**Conclusion**

In conclusion, our study aimed to develop and optimize machine learning models for predicting diabetes and heart attack risk using a merging dataset containing various patient attributes. We explored the performance of two models,KNN and SVM classifier, leveraging techniques like hyperparameter tuning with GridSearchCV to enhance predictive accuracy. Our findings revealed that both models achieved high accuracy, precision, recall, and F1-scores, indicating their effectiveness in predicting diabetes and heart attack risk.

The KNN, after hyperparameter tuning, demonstrated exceptional performance, achieving an accuracy of 96% on the test dataset. These results highlight the potential of machine learning algorithms in HealthRisk settings, offering valuable insights into factors influencing predicting diabetes and heart attack riskand paving the way for future research to refine and improve predictive models further.

Overall, our study underscores the importance of leveraging advanced analytics to augment medical decision-making and improve patient outcomes in preventive healthcare.