**DIABETES PREDICTION USING MACHINE LEARNING**

**Submitted for**

**Statistical Machine Learning CSET211**

Submitted by:

**E23CSEU2305 SOUMIL KHANA**

**E23CSEU2294 VIVEK KUMAR**

**E23CSEU2309 PRAKHAR KESHYAP**

Submitted to

**DR. ASHIMA YADAV**

**July-Dec 2024**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

A close-up of a logo

Description automatically generated

1. **Abstract**

**Diabetes is a chronic medical condition where the body either does not produce enough insulin or cannot use it effectively. Early prediction of diabetes is crucial for preventing its progression and minimizing health risks. This project aims to predict whether a person is at risk of developing diabetes using machine learning techniques. By analyzing medical data such as age, blood pressure, glucose levels, and body mass index (BMI), we apply algorithms to classify individuals into risk categories. Several machine learning models, such as logistic regression, decision trees, and support vector machines, are employed and compared for their prediction accuracy.**

**Introduction**

**Diabetes has become a major global health problem. According to the World Health Organization, millions of people are affected, and the numbers are rising steadily. Early detection of diabetes through predictive models can help individuals take necessary precautions and seek medical intervention promptly. Machine learning (ML) offers powerful tools for predicting and diagnosing diseases based on historical data. This project focuses on diabetes prediction using ML techniques, utilizing datasets containing medical records and diagnostic information to develop a model that predicts whether an individual is diabetic or not.**

1. **Methodology**

**The methodology involves the following steps:**

1. **Data Collection: A dataset containing patient information such as age, glucose levels, BMI, and other relevant features is gathered. Commonly used datasets, such as the Pima Indians Diabetes Database, are employed for this task.**
2. **Data Preprocessing: The collected data undergoes cleaning, handling of missing values, normalization, and feature scaling to ensure consistency and improve the performance of ML models.**
3. **Model Selection: Different machine learning models are tested, including:**
   * **Logistic Regression**
   * **Decision Trees**
   * **Random Forest**
   * **Support Vector Machine (SVM)**
   * **k-Nearest Neighbors (k-NN)**
4. **Model Training: The models are trained using the preprocessed data, splitting the data into training and testing sets to evaluate performance.**
5. **Evaluation: Various metrics like accuracy, precision, recall, and F1-score are used to evaluate the effectiveness of each model.**

**Feature Extraction**

**Feature extraction is the process of selecting the most important information from the raw data that helps in predicting the target variable, i.e., whether a person is diabetic. In this project, the features include:**

* **Glucose Level: Higher glucose levels are indicative of diabetes.**
* **BMI (Body Mass Index): Obesity is a major risk factor for diabetes.**
* **Blood Pressure: Elevated blood pressure often correlates with diabetes.**
* **Age: Older individuals may be at higher risk.**
* **Insulin Level: Insulin resistance is a common feature of diabetes.**
* **Pregnancy Count: The number of pregnancies a woman has experienced.**
* **Skin Thickness: Can indicate fat accumulation and insulin resistance.**
* **Diabetes Pedigree Function: A measure of family history of diabetes.**

**These features are selected and used to train the machine learning models.**

**Similarity Computation**

**Similarity computation plays a role in algorithms like k-Nearest Neighbors (k-NN). This algorithm identifies the most similar data points (neighbors) to the target data point based on distance metrics (Euclidean, Manhattan, etc.). The classification decision is based on the majority class among the nearest neighbors. Similarity computation helps identify patterns in data that influence the prediction of diabetes.**

**Recommendation Generation**

**Once the model is trained and evaluated, recommendations can be generated based on the predictions. For example:**

* **If a person is predicted to be diabetic, recommendations can include lifestyle changes, regular monitoring of blood sugar, and consultation with healthcare providers.**
* **If a person is predicted to be non-diabetic but at risk, recommendations may involve dietary changes and preventive healthcare measures.**

**Testing**

**Testing involves evaluating the trained models on a separate test dataset that was not used during training. The following steps are part of the testing phase:**

* **Model Evaluation: We test the models on metrics like accuracy, precision, recall, and F1-score.**
* **Cross-Validation: To ensure the model’s robustness, cross-validation techniques such as k-fold cross-validation are used.**
* **Comparison: The performance of different models (Logistic Regression, Decision Trees, etc.) is compared, and the best performing model is selected for deployment.**

1. **Hardware/Software Required**

* **Hardware:** Standard computer with at least 8GB RAM (for handling large datasets).

 Python (version 3.x)

 IDE: Jupyter Notebook / PyCharm

 Libraries:

* **Pandas**: For data manipulation
* **NumPy**: For numerical computations
* **Scikit-learn**: For machine learning algorithms
* **Matplotlib** and **Seaborn**: For data visualization
* **TensorFlow or Keras** (if deep learning models are used)

1. **Conclusions**

**This project successfully demonstrates the use of machine learning techniques in predicting diabetes. By utilizing various ML algorithms and data preprocessing techniques, the model can predict whether a person is diabetic with a high level of accuracy. The models can assist healthcare professionals in early diagnosis, allowing for preventive measures. However, the results are dependent on the quality and quantity of the data used.**

**Future Scope**

1. **Integration with Wearable Devices**: Future work could involve integrating real-time data from wearable devices like glucose monitors and fitness trackers for continuous monitoring and prediction of diabetes.
2. **Deep Learning**: The use of deep learning models such as neural networks could be explored for more complex patterns in large datasets.
3. **Personalized Recommendations**: The system could be enhanced to provide personalized recommendations based on an individual's unique medical history and lifestyle.
4. **Large-Scale Data**: Utilizing a larger and more diverse dataset could improve the accuracy and generalization of the prediction model.