DEPI IBM Data Science

Graduation Project Proposal

Group code:ALX1_AIS3_S1e

Name :Omar mohamed mashaly

Project Title:
Predictive Analytics for Diabetes Diagnosis and Risk Management Using Machine Learning

Project Overview:
Diabetes is a chronic condition that affects millions of people worldwide, leading to severe health complications if not managed properly. Early diagnosis and proper risk assessment can significantly improve patient outcomes. This project aims to leverage machine learning and data analytics techniques to predict the likelihood of diabetes and identify key risk factors using a large clinical dataset. By building predictive models, we can assist healthcare providers in decision-making and designing effective intervention strategies for at-risk patients.

Data Source:
The project will use the *Diabetes Clinical Dataset* from Kaggle, which consists of 100,000 patient records, including features such as patient demographics, medical history, blood test results, and diagnosis labels for diabetes.

Dataset Link: Kaggle - 100,000 Diabetes Clinical Dataset

Objectives:

1. **Data Exploration and Preprocessing**:
    - Conduct exploratory data analysis (EDA) to understand the dataset's structure, identify missing values, outliers, and data imbalances.
    - Clean the dataset by handling missing values, outliers, and normalising the features where necessary.
2. **Feature Selection**:
    - Perform feature selection to identify the most significant factors contributing to diabetes, that would be done by using techniques like correlation matrix.
3. **Model Building and Evaluation**:
    - Build classification models to classify patients based on their likelihood of having diabetes.
    - Compare model performance using metrics such as accuracy, precision, recall and F1-score.
    - Perform hyperparameter tuning to optimise model performance.
4. **Risk Factor Analysis**:

    - Use feature importance techniques and correlation to identify the most critical factors contributing to diabetes risk.
5. **Model Interpretation and Visualization**:
    - Visualise the relationship between risk factors and diabetes through data visualisations.
    - Create visualisations for healthcare providers to input patient data and receive diabetes risk assessments.
6. **Deployment and Real-World Application**:
    - Develop a simple web application or interface where healthcare providers can input patient data and obtain diabetes predictions.
    - Ensure the model is interpretable and explainable for practical use by medical professionals.

Methodology:

1. **Data Exploration**:
    - Analyse the distribution of various clinical features (age, BMI, glucose levels, etc.).
    - Identify any correlations between independent variables and the target variable (diabetes diagnosis).
2. **Modelling Approach**:
    - Train multiple classification models(e.g., Logistic regression , naive bayes and Decision tree) and use cross-validation to evaluate generalisation.
    - Implement feature scaling for algorithms sensitive to feature magnitude.
3. **Risk Factor Identification**:
    - Apply feature importance methods and correlation to identify which clinical factors contribute the most to diabetes.
4. **Model Validation**:
    - Use a test set to validate model performance and assess potential overfitting.
    - Tune models with grid search for optimal performance.
5. **Deployment**:
    - Use a lightweight framework (e.g., Flask or Streamlit) to develop an accessible tool for healthcare practitioners.

Expected Outcomes:

- Development of an accurate diabetes prediction model with real-world applicability.
- Identification of key clinical risk factors contributing to diabetes diagnosis.
- A user-friendly dashboard or tool for healthcare providers to input patient data and get real-time predictions of diabetes risk.

Tools and Technologies:

- Programming Languages: Python (pandas, NumPy, scikit-learn)
- Data Visualization: Matplotlib, Seaborn
- Deployment Framework: Flask/Streamlit/ML Flow
- Other: Jupyter Notebooks for analysis and experimentation

Conclusion:
This project will provide valuable insights into diabetes diagnosis and patient risk assessment. By leveraging data science and machine learning techniques, it aims to contribute to the medical field's ability to make informed decisions regarding diabetes management and prevention.