Serge Nane

Student ID: 21415521

DSC 680 – T302\_2261\_1

**Project domain:** Healthcare / Medicine (Predictive Analytics in Chronic Disease Management)

**Milestone 1**: Proposal and Data Selection

**Topic: Predicting Diabetes Risk Using Patient Health Indicators**

**Business Problem**

Diabetes is one of the leading chronic diseases worldwide, associated with life-threatening complications such as cardiovascular disease, kidney failure, and blindness. Early identification of individuals at high risk is critical for prevention and intervention. Currently, healthcare practitioners rely on diagnostic tests administered after symptoms appear, which may be too late for effective prevention.

This project proposes the development of predictive models using patient health indicators to estimate the likelihood of diabetes onset. By leveraging data-driven insights, the project aims to demonstrate how machine learning tools can support healthcare professionals in making earlier, more informed decisions, ultimately improving patient outcomes and reducing healthcare costs.

**Datasets**

* Name: Pima Indians Diabetes Database

**Source:**

* UCI Machine Learning Repository:

<https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes>.

* Kaggle: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>.

**Description:**

* **768 records** of female patients of Pima Indian heritage
* 8 predictor variables: **pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, serum insulin, body mass index (BMI), diabetes pedigree function, and age**
* Target variable: Outcome (1 = diabetes, 0 = no diabetes)

**Research Questions**

1. Can machine learning models accurately predict diabetes onset based on patient health indicators?
2. Which variables (e.g., glucose, BMI, age) have the strongest influence on diabetes prediction?
3. How do different algorithms (Logistic Regression, Decision Tree, Random Forest, XGBoost) compare in terms of performance?
4. What actionable insights can healthcare professionals gain to support preventative care strategies?

**Methods**

* Data preprocessing: handle missing values, normalize or scale features, and address potential class imbalance.
* Exploratory Data Analysis: visualize feature distributions, correlations, and class balance.
* Modeling: implement and compare Logistic Regression, Decision Tree, Random Forest, and XGBoost.
* Evaluation metrics: Accuracy, Precision, Recall, F1-score, and ROC-AUC.
* Visualization tools: correlation heatmaps, ROC curves, confusion matrices, feature importance charts, and distribution plots.

**Ethical Considerations**

* **Data privacy:** all dataset records are anonymized, but continued emphasis on protecting sensitive health data is necessary.
* **Bias and fairness:** predictive bias could occur if the model overemphasizes certain features (e.g., age or pregnancy). The analysis must discuss fairness and limitations.
* **Interpretability:** While advanced algorithms may achieve higher accuracy, they may lack transparency. Models should be explained carefully to ensure trust among clinicians.
* **Appropriate use:** predictions should not replace clinical judgment but rather serve as decision-support tools.

**Challenges/Issues**

* **Small dataset size:** only 768 records, which may limit model generalizability.
* **Class imbalance:** fewer positive diabetes cases could reduce predictive performance.
* **Interpretability vs. accuracy:** complex models may perform better but could be harder to explain to non-technical stakeholders.

**References**

* Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care, 261–265.*
* Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
* UCI Machine Learning Repository. (n.d.). *Pima Indians Diabetes Database*. Retrieved from <https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes>.