CP70066E Machine Learning – Group Project

Academic session: September 2025-26
Assignment title: Predicting Diabetes Outcomes Using Machine Learning

Assignment type: Coursework
Module weighting: 70%
Tutor: Professor Julie Wall
Issue date: 17/10/2025
Submission due date: 15/12/2025 4pm

1. Project Synopsis

Diabetes is a major global health concern, affecting hundreds of millions of people worldwide. Early detection of diabetes risk is essential for improving patient outcomes and reducing healthcare costs.

In this group project, students will build a Machine Learning (ML) model to predict whether a person is likely to develop diabetes based on diagnostic measurements such as glucose levels, blood pressure, BMI, and age.

The dataset used in this project is a modified version of the Pima Indians Diabetes Database, originally published on Kaggle. The modified dataset has been adapted for educational purposes and includes additional variables and data-quality challenges designed to enhance learning outcomes.

Students must use the version of the dataset provided on Blackboard — submissions based on the original Kaggle dataset will not be accepted.

The dataset contains medical diagnostic measurements from women of Pima Indian heritage aged 21 and older, along with an indicator of diabetes diagnosis. The goal is to apply the full ML lifecycle — from exploratory data analysis and preparation to model evaluation and ethical reflection — to understand both the technical and societal dimensions of predictive modelling in healthcare.

Dataset Description:

The dataset contains approximately 800 observations and 11 variables representing diagnostic and lifestyle measurements of women aged 21 years or older of Pima Indian heritage.

Predictor variables include Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age. Two additional variables have been included to support feature-engineering and ethical analysis:

- PhysicalActivityLevel (1 = Low, 2 = Moderate, 3 = High) — a synthetic behavioural feature inversely related to BMI and age.
- FamilyHistory (None, Moderate, Strong) — a categorical indicator derived from the Diabetes Pedigree Function to simulate genetic predisposition.

To simulate real-world data-quality issues, the dataset also contains a small number of duplicate records, missing values in several medical features (e.g. Insulin, SkinThickness, BMI), and a few mild outliers in Glucose.

Students are expected to identify and appropriately handle these issues during Exploratory Data Analysis and Data Preparation.

The target variable, Outcome, indicates whether the patient was diagnosed with diabetes (1 = positive, 0 = negative).

2. Project Tasks & Deliverables

Your team will carry out the following tasks:

1. **Perform Exploratory Data Analysis (EDA):**
    - Explore the dataset, visualise feature distributions, and investigate relationships between predictors and diabetes outcomes.
    - Identify missing data, potential outliers, and correlations.
    - Discuss potential biases (e.g. gender or ethnicity) and their implications for model generalisation.
2. **Perform Data Preparation for ML:**
    - Clean and preprocess the dataset based on EDA findings.
    - Handle missing or invalid values (e.g., zeros or NaN entries in medical features).
    - Detect and address duplicate or outlier records where appropriate.
    - Engineer new features (e.g., BMI categories, age groups).
    - Normalise or scale features as needed for ML model training.
3. **Develop an ML Model:**
    - Build and evaluate at least three ML algorithms taught in this module (e.g., Logistic Regression, Decision Trees, Support Vector Machines).
    - Compare model performance using appropriate metrics (e.g., accuracy, precision, recall, F1-score).
    - Discuss trade-offs in hyperparameter tuning, overfitting/underfitting, and model interpretability.
    - **Model Evaluation Requirements**: All models must be trained and evaluated using separate training, validation, and testing sets, or an equivalent cross-validation approach.
        - The dataset split and rationale (e.g. 70/15/15 or 80/20) must be clearly stated.
        - Results must include performance metrics for each stage (training, validation, and testing).

- Reports that only provide training accuracy will be considered incomplete and may lose marks under the ML Model Development and Jupyter Notebook criteria.

4. **Ethical Analysis of the Dataset and Model:**
   - Write an 800-word report discussing the ethical implications of predictive healthcare modelling using this dataset. Include:
     - Identification of dataset biases and their origins.
     - Potential fairness concerns (e.g., generalisation to other populations).
     - Trade-offs between accuracy and fairness.
     - Real-world implications of using predictive models in medical contexts.
     - Cite relevant ethical frameworks or academic literature where appropriate.
5. **Documentation:**
   - Use Google Colaboratory or Jupyter Notebook as your primary platform for all analysis and reporting.
   - Include:
     - Well-documented code and data visualisations.
     - Results and clear justifications for your modelling decisions.
     - A 500-700 word Executive Summary providing a formal overview of your workflow and findings.

3. Assessment Elements

The project will be marked out of **100%,** scaled to the module's **weighting**. The assessment criteria are as follows:

Assessment Elements	**Marks**
Exploratory Data Analysis (EDA)	20%
Data Preparation for ML	20%
ML Model Development	20%
Ethical Analysis	20%
Jupyter Notebook & Executive Summary	20%
Total	**100%**

4. Online Submission

Submit the final project documentation to Blackboard by the specified due date.

Your submission should be a zipped folder containing the following:

1. **Jupyter Notebook (.ipynb)** — including EDA, data preparation, model development, and evaluation.
2. **800-word Ethical Analysis** — critical reflection on fairness and bias in healthcare prediction.
3. **500-700 words Executive Summary** — a concise formal report summarising your full ML lifecycle and key findings, including:
    a. Aim and dataset used (~100 words)
    b. Main EDA findings (~150 words)
    c. Overview of data preparation (~100 words)
    d. Models used and key results (~150 words)

Submission guidance: Only one group member should upload the final zipped submission folder to Blackboard. Ensure all files open correctly before submission.

5. Assessment Grading Scheme

- Marking rubric available on Blackboard.
- All students within a group will receive the same mark based on the submitted work.