Advanced Machine Learning (MDS) - Project 1 Proposal

A Comparative Analysis of Part I Models for Breast Cancer Diagnosis

Team Members:
Adrià Espinoza Gómez
Biel Manté Peñalba
Vendrix Alexis

October 31, 2025

1. Problem

The problem is the binary classification of breast tumors as either **malignant (M)** or **benign (B)**. We will use the "Breast Cancer Wisconsin (Diagnostic)" dataset, which provides 30 real-valued features (e.g., cell radius, texture, compactness) computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

The dataset will be sourced directly from the UCI Machine Learning Repository:

- **URL:** `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`

2. Motivation

We chose this problem for three primary reasons:

1. **Suitability for Part I:** It is a classic benchmark problem. Its binary nature and real-valued features make it an ideal case study for applying and rigorously analyzing the classification algorithms from Part I.
2. **Model Comparison:** It allows for a direct investigation of model assumptions, performance, and theoretical properties in a tangible, real-world context.
3. **Pre-processing Practice:** The data requires analysis and normalization, fulfilling the pre-processing requirements of the practical work.

3. Planned Techniques

Our work will focus on implementing, analyzing, and comparing the following models from the course syllabus:

- **Generalized Linear Model (GLM):** We will implement **Logistic Regression**. We will analyze it from the GLM perspective, justifying its choice of link function (logit) for a Bernoulli-distributed target.
- **Bayesian Generative Method:** We will implement **Gaussian Naive Bayes (GNB)**. This will require us to study its strong conditional independence assumption.
- **Perceptron:** We will implement the classic **Perceptron algorithm**. We will investigate its convergence properties and analyze the linear separability of the dataset.

A key part of our analysis will be to rigorously compare these models based on classification **accuracy** (using appropriate metrics and resampling) and their **training time**.

4. Resources

The models are not computationally intensive. All development will be conducted on our computers using the **R programming language**.

5. Learning Goals

We expect to gain a deep understanding of the differences between discriminative (Logistic Regression, Perceptron) and generative (Naive Bayes) models. We aim to rigorously analyze the impact of their underlying assumptions on model performance for a real-world problem.