Team 13

Team Members and Roles

1. Carl Patel (Team Leader) - Training the Model
2. Marcello Battista - Preprocessing Data
3. Makhtar Thioune - Model Testing
4. Benjamin Tang - Front-End Creation
5. Jordan Tecson - Plotting Models

Problem Statement:

Our project aims to develop a predictive model using the Adult dataset from the UCI Machine Learning Repository to determine whether an individual earns more than $50,000 annually. This classification will be based on various demographic and employment-related attributes such as age, education, occupation, and hours worked per week.

Tentative Timeline/Roadmap:	
Write 1-Pager	04/25
Prepare Data	4/30
Train Model	4/30
Midterm Report Due	05/06
Develop api and front end	5/19
Work on final report	5/28
Final Report Due	06/06

Dataset Description

The Adult dataset from the UC Irvine Machine Learning Repository consists of 48,842 instances and 14 attributes, with the target variable being income. These attributes represent socioeconomic indicators, including work class, education, marital status, occupation, race, sex, and country of origin.

Model Selection:

We plan to use logistic regression for each model to classify whether a given person makes more than $50k in a year. More concretely, we will define a probability threshold to determine class labels and will classify them as greater than $50k or less than $50k a year. Since logistic regression is best for modelling binary output, we plan to employ this model on our training data.

Goal of Project and Deliverables:

Our main goal for the project is to develop an accurate classification model for the estimated income of a given person. This means we will be training multiple different models to see which one is the most accurate fit for our data. Our deliverables will include 3 trained models, visualizations and performance metrics on both training and testing.