# BITS F464 : Machine Learning Assignment – 2 Report

## Team Members

1. Kushagra Verma (2020A7PS0225H)
2. Sai Panda (2020A7PS0080H)
3. Khooshrin Aspi Pithawalla - 2020A7PS2067H

# Part A: Naive Bayes Classifier to predict income

## Task 1

### The Dataset

The dataset provided had 32561 examples, with each example containing 14 features and a label indicating whether the annual income was 50K. Out of the 14 features, 6 features had continuous values (namely, "age", "fnlwgt", "education-num", "capital-gain", "capital-loss" and "hours-per-week") and the remaining 8 features had discrete values (namely, "workclass", "education", "marital-status", "occupation", "relationship", "race", "sex" and "native-country"). There were 24720 examples with income 50K and 7841 examples with income 50K.

#### Fill Missing Values

There were 2399 examples which had one or more missing values. We filled the missing values by going over each column and replacing the missing values with the most commonly occurring non-missing value in case of discrete features and with the mean of all the non-missing values in case of continuous features.

#### Create Training and Testing Sets

We split the dataset and used 67% of the dataset for training and 33% for testing. Our testing split contained 21816 examples and training split contained 10745 examples.

## Task 2

### Calculating Prior Probability

We calculated prior probability of each of two classes using maximum likelihood estimator.

### Calculating Conditional Probability

We calculated conditional probabilities of every feature assuming that features are independent of each other given the income class. We binned the continuous features into 10 bins of equal size for calculating likelihood.

### Predicting Class

We made predictions for the testing split using prior and likelihood that we had calculated earlier. We ignored the denominator since it is common to both income classes and directly compared the numerators to make our prediction.

## Task 3

### Evaluation

We evaluated our predictions by calculating the accuracy, precision, recall and F1-score. We calculated them using the following formulae:

### Smoothing

We applied Laplacian smoothing with values as 1, 10, 50 and 100. We used the following formula for smoothing:

where is the number of discrete values that education can take.

### Results

Here are the results averaged over 10 different test-train splits:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **α** | TP | FP | TN | FN | Accuracy | Precision | Recall | F1-Score |
| N/A | 6821.0 | 587.8 | 1992.0 | 1344.2 | 0.689511 | 0.920660 | 0.835379 | 0.875944 |
| 1.0 | 6821.0 | 591.2 | 1988.6 | 1344.2 | 0.689828 | 0.920238 | 0.835378 | 0.875753 |
| 10.0 | 6853.1 | 621.8 | 1958.0 | 1312.1 | 0.695663 | 0.916818 | 0.839310 | 0.876347 |
| 50.0 | 7036.6 | 772.9 | 1806.9 | 1128.6 | 0.726803 | 0.901034 | 0.861786 | 0.880967 |
| 100.0 | 7243.1 | 948.4 | 1631.4 | 922.1 | 0.762355 | 0.884225 | 0.887078 | 0.885641 |