# CS 5228 Final Project Report

|  |  |
| --- | --- |
| Saurav Sircar | A0198873E |
| Siddharth Satish | A0198875B |

We have opted for Option 1 involving the prediction of whether a person’s income exceeds $50,000 per year.

## Pre-processing and Exploratory Data Analysis

The training dataset consists of 24421 data entries consisting of 13 input columns and the output column “exceeds50K” where 0 indicates an annual income less than $50,000 and 1 indicates an annual income higher than $50,000.

We performed pre-processing on some of the 13 input columns as follows:

### Education

This field indicates the highest education qualifications of a person. We have converted specific educational qualifications to be more general in the following way:

|  |  |
| --- | --- |
| Old Values | New Values |
| ‘Preschool’, ‘1st-4th’, 5th-6th’, 7th-8th’, ‘9th’, ‘10th’, ‘11th’, ‘12th’ | ‘dropout’ |
| ‘HS-Grad’, ‘HS-grad | ‘HighGrad’ |
| ‘Some-College’, ‘Assoc-adm’, Assoc-voc’ | ‘CommunityCollege’ |
| ‘Bachelors’ | ‘Bachelors’ |
| ‘Masters’, ‘Prof-school’ | ‘Masters’ |
| ‘Doctorate’ | ‘Doctorate’ |

### Marital Status

This field indicates the marital status, where we have grouped certain categories together in the following way:

|  |  |
| --- | --- |
| Old Values | New Values |
| ‘NeverMarried’, ‘Married-spouse-absent’ | ‘NotMarried’ |
| ‘Married-AF-spouse’, ‘Married-civ-spouse’ | ‘Married’ |
| ‘Separated’, ‘Divorced’ | ‘Separated’ |
| ‘Widowed’ | ‘Widowed’ |

### Age

This field represents the age. We have used the cut feature to create 20 equal bins for the age.

### Hours-per-week

This field represents the number of working hours per week. Again, the cut feature has been used here to create 10 equal bins for the work hours per week.