**Text

Description automatically generated**

IE7374 FINAL PROJECT REPORT

Adult Census Income

Group-8

|  |  |  |  |
| --- | --- | --- | --- |
| 1. | Ziqi Li | 33.33% |  |
| 2. | Zihao An | 33.33% |  |
| 3. | Mir Shezan Ahmed | 33.33% |  |

**PROBLEM SETTING & DEFINITION:**

Our group uses this data with the primary goal of wanting to analyze studies on income classification in adult census data. Analyze what conditions an adult has to be in to become a group of people who earn more than 50K. In this project we will be using various machine learning algorithms and techniques to solve the above defined problems. Apart from this we will also explore the data and find out the most contributing features. Finally, we will demonstrate several models and evaluate the best models.

**DATA SOURCES:**

Link to the UCI repository:

<https://archive.ics.uci.edu/ml/datasets/Adult>

**DATA DESCRIPTION:**

Predict whether income exceeds $50K/yr based on census data. Also known as the "Census Income" dataset. Extraction was done by Barry Becker from the 1994 Census database.

Listing of attributes : income:>50K, <=50K.

Age: continuous.

Workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt: continuous.

Education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

Education-Num: continuous.

Marital-Status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

Occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspect, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

Relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

Sex: Female, Male.

Capital-gain: continuous

Capital-loss: continuous.

Hours-per-week: continuous.

Native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

**Exploratory Data Analysis:**

Preview dataset

A screenshot of a computer

Description automatically generated with medium confidence

Printing a short summary of the data frame with the info() function shows that there are 32561 instances and 15 attributes in the dataset. The summary of this dataset shows no missing values. However, the preview shows the dataset package as '? ' values.

**Data Cleaning:**

Graphical user interface, application, Teams

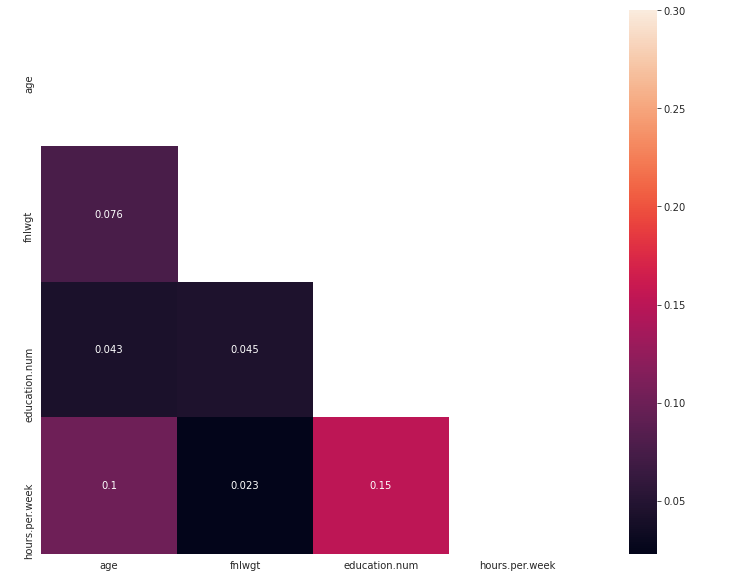
Description automatically generatedGraphical user interface, text, application

Description automatically generated

So, we need to deal with the `? ` values. Put the '? as the empty set and handle duplicate data values.

**Data Visualization**

**Correlation:**

****

Age:

Chart, histogram

Description automatically generated

Workclass:

Chart, bar chart

Description automatically generated

Education Level:

Chart

Description automatically generated

Income distribution：

Chart, bar chart

Description automatically generated

We can see many different distributions, some with Gaussian-like distributions and others with seemingly exponential or discrete distributions. We can also see that they all seem to have very different scales. Depending on the choice of modeling algorithm, we would expect that it would be useful to scale the distributions to the same range and possibly use some power transformations.

We remove the native.countries column because it has strong correlation with race column and it has too much features which could cause overfitting.

The fnlwgt column is meaningless and the values are huge so we decide to remove this column.

**Scaling and Encoding:**

From the above we can see that the output variables are a mixture of numeric and categorical or ordinal data types, where non-numeric columns are represented using strings and categorical variables need to be ordinal encoded.

A picture containing text, white

Description automatically generated

Use scaling codes to make the data close to each other or simpler so that the distance between them will be lower. This helps the algorithm have the opportunity to get better and faster training, rather than data points or feature values that are more different from each other requiring more time to understand the data and accuracy will be reduced. This can help us to use the algorithm more simply.

**MODEL IMPLEMENTATION:**

There are some machine learning algorithms that can be used on this classification dataset to help us understand this data well:

1. Multi-Class Logistic Regression

2. Support Vector Machine

3. Decision Tree

**Logistic Regression:**

Logistic regression analysis, a generalized linear regression analysis model, belongs to supervised learning in machine learning. Its derivation process and calculation is similar to that of regression, but it is actually mainly used to solve dichotomous and multi classification problems. The model is trained by a given training set and classifies a given set or sets of data at the end of the training.

Algorithm:

Uses a logistic function or sigmoid function. This S-curve accepts numerical inputs and outputs results in the form of 0 and 1. The sigmoid function is used because it gives a maximum or minimum value when estimating and optimizing the cost function, and it is part of a larger class of algorithms known as generalized linear models. Logistic Regression is Simple Classification algorithm that make use of sigmoid function to return the output/response variable probabilities. These probabilities are then compared with threshold value and a label are assigned to data point which we want to classify.

Logistic Regression is simple and intuitive algorithm which is easy to implement. In Logistic Regression the response variable needs to be binary, and data should be free from missing values and outliers. The logistic function is a sigmoid function, which takes any real input t, and outputs a value between zero and one.

The logistic function is of the form:

Chart

Description automatically generated

Chart, line chart

Description automatically generated

The loss of an example is calculated using the categorical cross-entropy loss function by the following code to calculate the following sum

A picture containing chart

Description automatically generated

Use gradient descent to optimize the maximum likelihood estimate: Since we know that the logistic model is a probabilistic model, in order to obtain the maximum probability of y(i), we need to optimize the equation given above.

Text

Description automatically generated

The graph below shows the Model's accuracy, training error, and test error:

Table

Description automatically generated

**Support Vector Machine:**

The SVM is a binary classification model whose basic model is a linear classifier defined by maximizing the interval on the feature space, and the SVM training algorithm creates a model that assigns new instances to one of the two classes, making it a non-probabilistic binary linear classifier. separated, and the learning strategy of SVM is interval maximization.

Diagram

Description automatically generated

Soft Margin SVM allows the SVM to err on a small number of samples by relaxing the previous hard interval maximization condition a bit, which allows a small number of samples to not satisfy the constraint：

A picture containing chart

Description automatically generated

Since we are dealing with a multi-class problem, the support vector machine takes longer than expected to execute the model and consumes a lot of time.

A support vector machine (SVM) is a machine learning algorithm that analyzes data for

classification and regression analysis. SVM is a supervised learning method that looks at data

and sorts it into one of two categories. An SVM outputs a map of the sorted data with the

margins between the two as far apart as possible while keeping the distance between the data

point and the margin minimum. Here, we used Soft Margin SVM. A soft-margin SVM modifies the constraints from the hard-margin SVM by allowing some points to violate the margin. It introduces slack variables ξi, one for each training point, into the constraints:

A picture containing text, clock, watch

Description automatically generated

Putting the objective and constraints together, the soft-margin SVM optimization problem is:

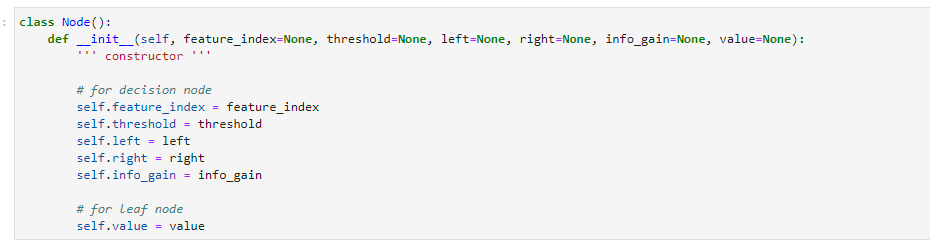
Text

Description automatically generated

**Decision Tree:**

Decision trees are a classical machine learning algorithm that can be used to deal with classification problems in a structure similar to a constant iterative dichotomous classification. Regulatory learning is required. Regulatory learning is to give a bunch of samples, each with a set of attributes and a classification result, that is, the classification result is known, then a decision tree is obtained by learning the samples. The internal nodes of the decision tree represent a judgment on an attribute, each branch represents the output of a judgment result, and finally each leaf node represents a classification result.

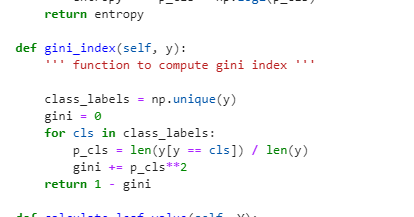
First construct Node class



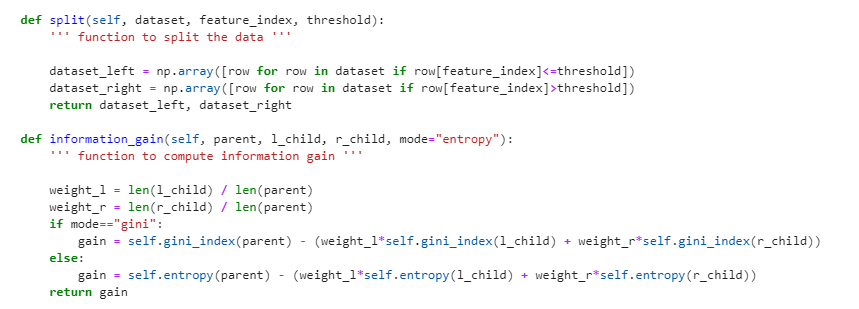
Build the tree

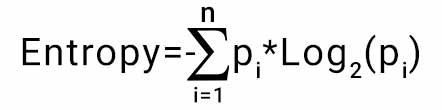


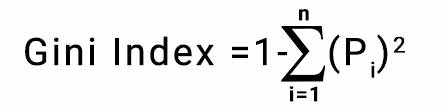
Calculate the Gini index



Split the data for the tree and calculate the weight





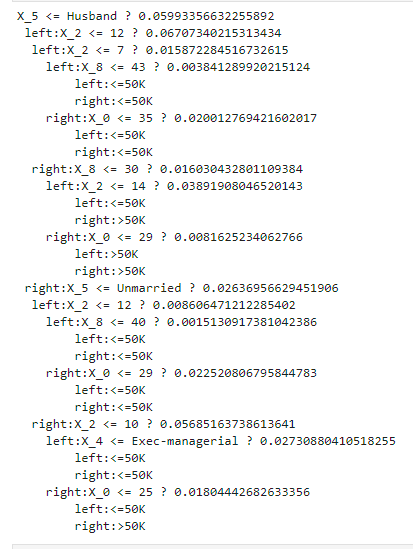


Accuracy:

{\displaystyle Accuracy=(TP+TN)/(TP+TN+FP+FN)}(T

Find the best split then build the tree, make prediction.

The tree we build:



**Naïve Bayes:**

It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Text

Description automatically generated

We have Implemented the below Naïve Bayes algorithm:

**Gaussian Naive Bayes:**

When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution.

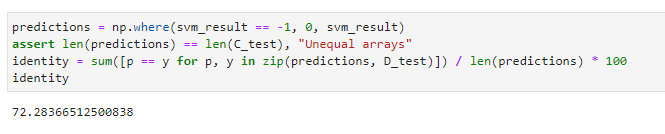
A picture containing text, clock

Description automatically generated

We will be using the Gaussian naive Bayes model for training the continuous features of data and later we will multiply the probability of classes with the probability of Bernoulli naive Bayes which will be training on the discrete features of the dataset. We can multiply the probabilities with each other since each feature is independent.

**Result:**

**SVM: Accuracy:72%**

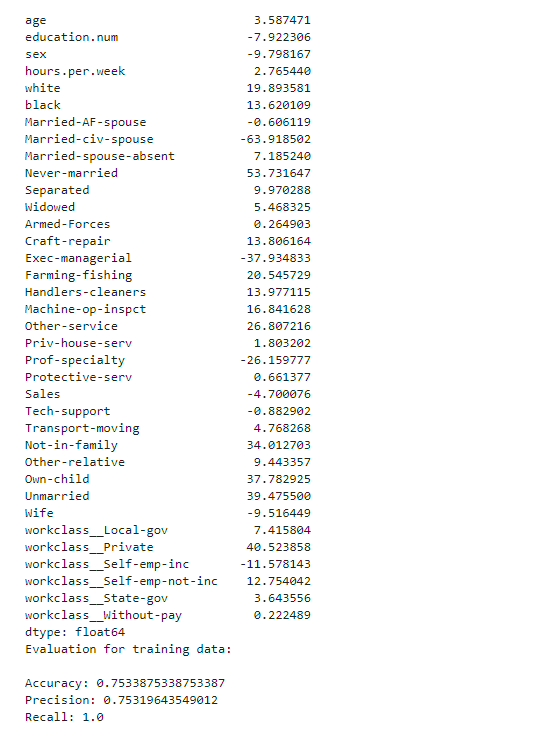


The dataset has too much categories features, this made SVM run very slowly and we have to make training set as small as possible (0.01) to 300 records. The accuracy is absolutely not good. The lack of numeric variables also makes the vector easy to overfitting.

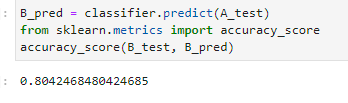
**Logistic Regression: Accuracy:75%**

This dataset has lots of category column and some discrete features, in this case the logistic regression performs good for our data set and have good result.

It is also easy to train and get result fast. We have lots of records so that it will help reduce the overfitting. We only keep independent variables which also improve our model performance.



**Decision Tree: Accuracy:80%**

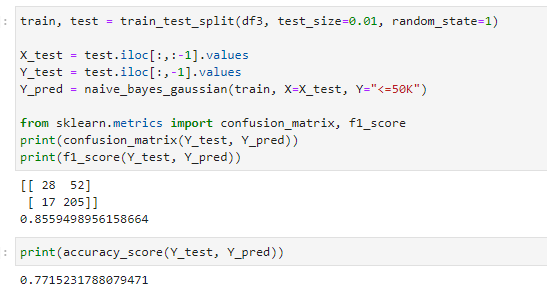


The decision tree is a perfect match for our dataset. We have lots of non-linear relationship, the decision tree can deal these separate data well.

Based on so much categories column, it is much more effective than other classification algorithms, it gets the highest accuracy and speed among our test. To apply each algorithm, we need to make a lot of preparation to modify the data. The decision tree does not need to worry about outsider or missing data. This algorithm also do not need to make assumptions so the result is convinced.

**Naive Bayes:**

**Calculate P(X=x|Y=y) using Gaussian dist**

****

## Calculate P(X=x|Y=y) categorically

## 

## Because we can only use three numeric columns to apply Gaussian dist, the second categorical method will be more helpful with more than 10 columns. The naive bayes also run fast with so many independent variables. The most part of our input are categorical, only few numeric variables make sure naive bayes perform better. The f1 score is around 83 percent and the accuracy is around 74 percent. However, in the case of naïve bayes assume all features are independent, although most variables are independent, there are also variables looks dependent like sex and relationship. These will affect naive bayes performance. We assume this is the result of naive bayes perform not as good as decision tree.

**Summary:**

The decision tree performs best among all four algorithms which has more than 80 percent accuracy. The naive bayes perform close to logistic regression as 74 percent. The SVM perform worst which has only 72 percent accuracy.