Income Prediction

Using Support Vector Machines

Kumar Katyayan Jaiswal

Data Science Department

Roll No.-20155

Abstract

The objective of this project is to predict whether the income of a person is more or less than a specific amount or not.
This is achieved using two discriminative classifiers of Machine Learning called **Support Vector Machine** and **K-Nearest Neighbours.**

The Dataset

The Dataset used for the prediction is extracted attached in the Zip file. All the details about the dataset are given in the python notebook.

There are three datasets -

1. Train.csv
2. Train_class_labels.csv
3. Test.csv

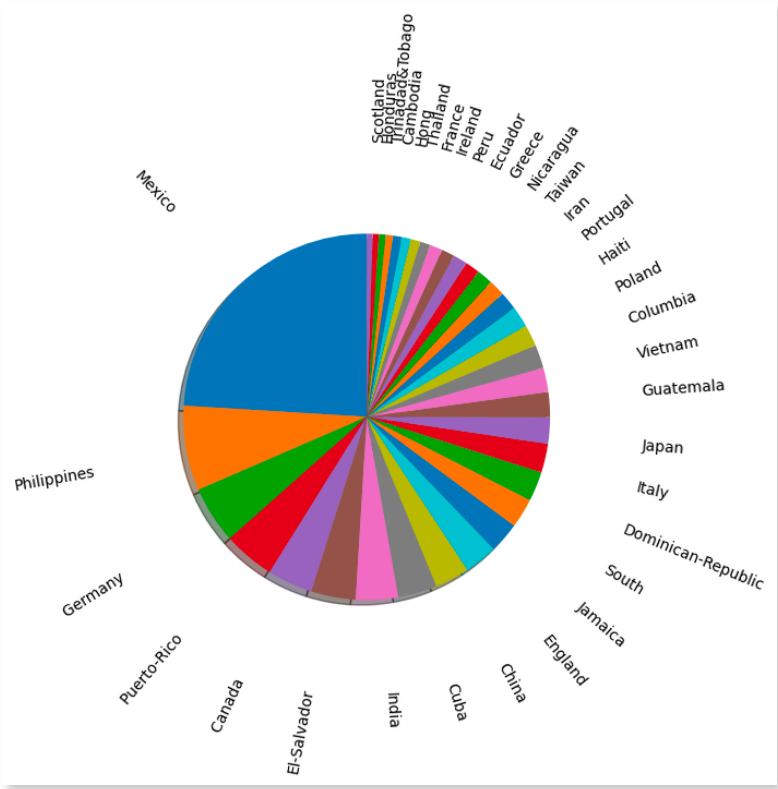
The Train.csv file contains various feature vectors which have been used to train the classifier. The Train_class_labels.csv file contains all the corresponding labels . Please note that this is a binary Classification problem.

The third file contains the Feature Vectors of the data points whose class labels have been predicted by the two classifiers.

The results have been displayed in the form of an array in the part of the python notebook. It took a significant amount of time to **clean** the dataset as well.

Exploratory Data Analysis

The features whose values were categorical .i.e. non- numerical have been plotted in the python notebook for eg. -



The share of each nation towards the nationality feature of the training instances.

All the categorical columns had to be converted to numerical representations to feed them to the classifiers.

Feature Selection

The best 4 features were selected for prediction order to de-noise the data. This was done by finding the correlation between different features and the income label.

Income labels are **0** and **1.**

- **0 -** Income below $50K
- **1-** Income above $ 50K

Conclusion

All the procedure regarding the steps done in prediction are given in the python notebook.

Thank You