**Comparison between predictive models to determine the income level for people in US.**

**Project report**

**COSC 6342: Machine Learning**

**Submitted by**

**S M Salah Uddin Kadir (1800503)**

**Rubayat Jinnah (1891217)**

The dataset for our study was accessed from the University of California Irvine (UCI) Machine Learning Repository. It was extracted by Barry Becker using the 1994 census database.

“Adult ” dataset has been used in different analysis multiple times to predict the income level in US. Multiple disciplinary approach like Statistical Analysis and Machine Learning were taken to predict income level as well as the influencing factors of income level. This dataset has been used in papers where different algorithms were implemented for prediction. In our study we tried to enhance the prediction using adult dataset and give a comparison of predictions based on their accuracy.

**Relevant Work**

The dataset has a “Data Set Description” link that contains a list of 17 algorithms ran on the dataset and the results. One of those algorithms is naive bayes with 16.12% error rate. Another study used naïve bayes algorithm but omitted the entries with unknown values and had a higher error rate of 20.43%. [2]. In Ron Kohavi’s paper “Scaling Up the Accuracy of Naïve Bayes Classifier: a Decision Tree Hybrid”, the NBTree is described. Kohavi described NBTree as a hybrid of naïve bayes and decision tree. In this approach the error rate was 14.01% and this is the best among 17 algorithms.[3]

Another approach was taken in the paper “Ensemble Selection from Libraries of Models” by Rich Caruana. Here he concluded that ensemble selection consistently finds ensembles that outperform all other models, including models trained with bagging, boosting, and Bayesian model averaging.[4] In a study it was concluded that **a Log Transformed Linear Regression is best model for predicting Income Per Capita.[5]**

**Data set description**:

Data set Link: The dataset is available at e UC Irvine Machine Learning Repository <http://archive.ics.uci.edu/ml/datasets/Adult>

The dataset has 48,842 instances with 14 attributes. We can define this attributes as categorical and continuous attributes.

**Categorical attributes are**:  
**1) marital-status**: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.  
**2) occupation**: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.  
**3) relationship**: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.  
**4) race**: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

**5) sex**: Female, Male.  
**6) 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.

**7) workclass**: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

**8) 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.

**Continuous attributes has an integer value. Continuous attributes are:**

1. **capital-gain**

**2)capital-loss**

**3) hours-per-week**

**4) age**

**5) fnlwgt**

**6) education-num**

**Data Cleaning Process:**

To do our project we needed to clean the data set. Our data set was not clean. There were some missing data also. So, we took two approaches to clean and structure our dataset.

1. Replace missing data with 0.
2. Replace missing data with highest occurrence of feature for an attribute

**Algorithms applied:**

**Final result:**

Referrences:

1. <https://susanli2016.github.io/Census-Income/>
2. <http://cseweb.ucsd.edu/classes/sp15/cse190-c/reports/sp15/048.pdf>
3. <http://robotics.stanford.edu/~ronnyk/nbtree.pdf>
4. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.2859&rep=rep1&type=pdf>
5. <https://medium.com/alien-status/using-regression-models-to-predict-per-capita-and-median-household-income-in-nyc-91d5ca899509>