PREDICTING ADULT ANNUAL SALARY

RAFEED SULTAAN (S3763175)
VISHAL BENIWAL(S3759790)

GOAL OF THE PROJECT

The goal of the project is to use data modelling techniques to model the Adult data set and to predict the annual salary of an adult.

DESCRIPTION OF THE ADULT DATA SET

- Created by Ronny Kohavi and Barry Becker
- Collected from UCI Repository
- Has 2 Target Labels: >=$50K/year and <$50K/year
- Contains 48842 Observations and 14 Features.
- Features contain both Categorical and Continuous Values.
- Features include: Age, Race, Gender, Education, Marital Status and other features
- Suitable for **Classification** Technique to model the data

DATA CLEANING

- **White Spaces** were removed
- **Unknown values (?)** in the data which were first **replaced with the Null** type value and then finally dropped.
- Fixed **Typos** to overcome data redundancy using replace function
- Similar types of **values were grouped together** to have better understanding of the data distribution.

HYPOTHESIS AND RESEARCH QUESTIONS

- The **main hypothesis** is that **can we predict the annual income** of a random adult based on the adult's features, such as Age, Marital Status, Working Class, Work Hours, etc.
- Some of the Research Questions are:

  - Does Age have an effect on Capital Loss or Capital Loss?

  - Does Marital Status have an impact on Annual Income?

  - Is Working Hours contributing to Annual Income?

  - Is Education a contributing factor for Annual Income?

RESULTS

- Example of **Data Exploration**

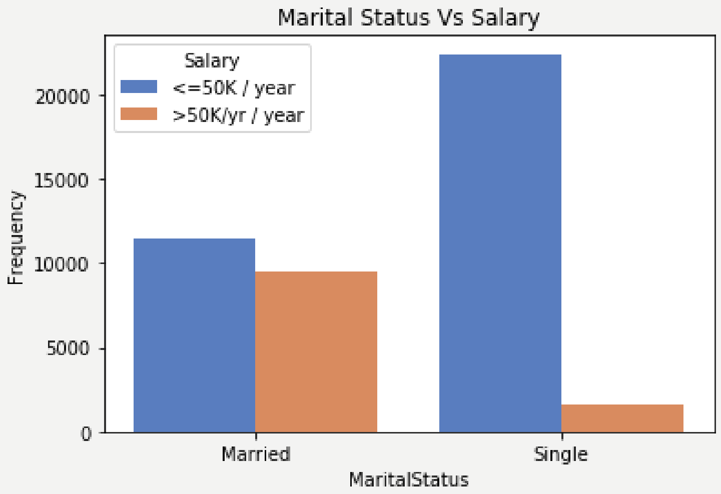
  As marital status changes

  from single to married,

  we can see that amount of adults

  earning salary >=50K per year

  is increasing.



- After **feature selection** we have found that , the

following **8 features** are believed to be the he most important for prediction. They are:

Sex, Relationship, Education, Capital loss, Occupation, Capital Gain, Work Class, Marital Status

- For **Data Modelling**,

we use **KNN** and **Decision Tree** algorithm,

combined with Feature Selection.

Algorithm	Accuracy
KNN	83.01%
Decision Tree	83.25%

CONCLUSION & RECOMMENDATIONS

- The dataset was **biased** towards Adults from USA. Since the only 2 races of people : **white and black races are over-represented** and **other races are under-represented**.
- The **accuracy of around 83%** is a relatively high percentage, so we can safely say the 4 features mentioned before can be used to predict **ANNUAL INCOME**.

- We need to source data so that **minorities** from other countries can be represented.
- We can use **K-Cross Validation** in splitting of our data, to check if we can get better results.
- We can also compare **other classification models** to see increase the accuracy of the data.

REFERENCES

- Becker, B. and Kohavi,R. 2017, *Adult Data Set*, electronic dataset, UCI Machine Learning Repository, viewed 27 May 2019, <https://archive.ics.uci.edu/ml/datasets/adult>.
- Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.