

## **Project Proposal**

**Project Title: To Predict Salary Income with The Help of Machine Learning Model**

**Submitted by:**

- 1. Bhagyashree Chavan**
- 2. Gulkhan Anassova**
- 3. Trineth Reddy**

**Course: DS 630 Machine learning**

**Professor: Dr. H. Wang**

**Summary:**

To perform prediction analysis, me and my other project partner has chosen the US Adult Census dataset which has 48,842 sample set with 15 features in data, extracted from US census database. We will use various techniques and methodology taught during this course to perform analysis on selected dataset.

**Source of the dataset:** UCI Machine Learning Repository

According to UCI the data extraction done by Barry Becker in 1994 from US census data.

**The dataset attributes have following information:**

1. Age: An individual age
2. Workclass: A term represent the employment of the class
3. Fnlwgt: It means Final weight or another words, the number of people the census believes the entry represents
4. Education: The high level of education obtained by individual
5. Education-num: The numerical form for high level of education by individuals
6. Marital-status: The Marital status of an individual and note down if Married-civ-spouse corresponds to a civilian spouse while Married-AF-spouse is a spouse in the Armed Forces.
7. Occupation: It is type of occupation of an individuals
8. Relationship: It is representing what this individual is relative to others (does not count in the family)
9. Race: It is descriptions of an individual's race
10. Sex: the biological sex of the individual
11. Capital-gain: Capital gains for an individual (Money gained outside the salary)
12. Capital Loss: Capital loss for an individual (Money loss outside the salary)

13. Hours-per-week: Counted the hours an individual has reported to work per week

14. Native country: country of origin

15. Income: The individual makes more than 50 K or not annually

**Objective:** Predicting salary income class with help of machine learning model

### **Our Action Plans:**

- At the beginning of the project, we will clean dataset and find out anomalies such as missing values, outliers and check data types.
- In second step we will apply Exploratory data analysis to discover data dimensions and to find out internal insights of the data.
- To find out internal analysis we are going to use plotting techniques with help of plotting python library.
- We are planning to use Decision tree and Regression machine learning techniques, and these are the supervised learning techniques.
- We are also planning to use Grid search CV to find out the best hyper parameter of the model.

### **Our Expectations:**

Apply all the knowledge of lessons learned in class, in additional try out more machine learning techniques on the data set, to find out what could be more scope of study in different area. The main aim of project to gain experience in versatile machine learning subject.

### **Using Tools:**

We have decided to use **Jupyter notebook** and **google colab**, machine learning library to ease programming work.