Project Proposal

Group 7

Graham, David

Mudarth, Ramila

Rakaj, Sivi

Saffarzadeh, Panthea

Smetanin, Maxim

Tang, Jing

**Topic: Classification Models to Predict Salary Over or Under 50,000 USD Dollars.**

Data Set: <https://archive.ics.uci.edu/ml/datasets/Census+Income>

(Extraction was done by Barry Becker from the 1994 Census database)

Problem Statement:

We are going to use given dataset to develop a suitable model to predict personal annual income – over or under 50,000 dollars per year.

Project Process:

1. Data preparation:

* Detect missing values in the dataset
* Replace and clear missing values
* Create new classifications for ‘Hours\_Per\_Wk’ column
* Clean column headers

1. Data visualization

Overview of target column:



Overview of other feature columns:

Chart

Description automatically generated

Figure 1 Occupation distribution

Chart, bar chart

Description automatically generated

Figure 2 Education count

A picture containing chart

Description automatically generated

Figure 3 Martial Status

Chart, bar chart, waterfall chart

Description automatically generated

Figure 4 Sex and Race Distribution

Pair plots of entire dataset:

Diagram

Description automatically generated

Figure 5 Dataset Overview

Correlation heatmap of dataset:

Chart, treemap chart

Description automatically generated

Figure 6 Correlation Matrix

1. Prepare for training models

* Definition of Target Column and Predictors
* Using pipeline to clear data

**In planning processes:**

1. Select proper classification model
2. Training model
3. Hyperparameter Tuning
4. Evaluate the model accuracy