Predicting Income

By Dian Todorov, 61670

Motivation

Problem
Мотивация, Задача на курсовата работа
Кратък обзор
Извършвана ли и някаква работа в тази област преди и ако да, каква?
Какви методи могат да се използват за решаването на задачата?
Вашето решение
Програмна реализация
Резултати от експерименти (ако има)
Заключение
Какво сте направили
и какво още може да се направи
Литература – цитирате всички използвани източници

Problem

Classify if the income is above or below 50K.

Solution

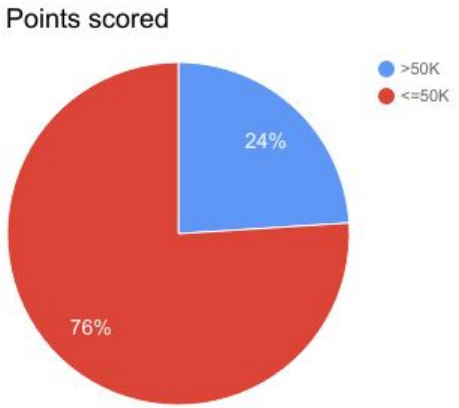
Classify if the income is above or below 50K.

- **Naive Bayes**
- **KNN**
- **Decision trees. Random forests.**
- **Neural networks.**

Details about the Data

- 45000 instances
- Multivariable
- Missing Values
- Many Outliers

Proportion

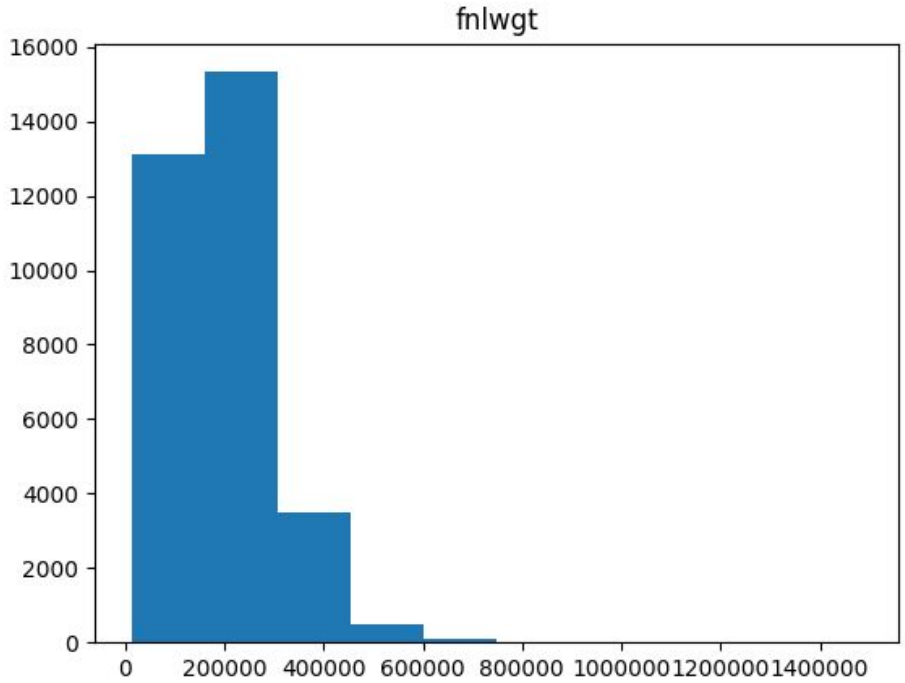


Attributes

- Age
- Worklass
- Fnlwgt
- Education
- Relationship
- Sex
- Capital-loss
- Capital-gain
- Other……

Data Cleaning

- Remove missing variable
- Discretize Data
- Normalize
- Trim outliers



Model testing

- 50 % train
- 25 % test
- 25% validation

Used Methods

- KNN
- Naive Bayes
- Decision Trees

Naive Bayes

- Intro
- Tuning
  - increase buckets
  - Play around
- Results - 84%

KNN

- Intro
- Problems
  - Lazy algorithm
  - Really slow
  - Unreliable, better off saying that every time the data will <50K
- Tune - play with K, use 6
- Results 79 %

Decision Trees

- Could not implement this algorithm
- Really useful for the case

Future Development

- Implement Decision Tree Algorithms
- Do ensemble method
- Neural Networks

Question?