Group
Mezun Tayfa

Canberk Keleş 25393
Kaan Koç 20680
Ömer Köse 25224

https://colab.research.google.com/drive/1hzEMI4XRoUKL_AEWMUTJp_Er8gQ_T4QW?authuser=1#scrollTo=5CDtKh4ilNdj

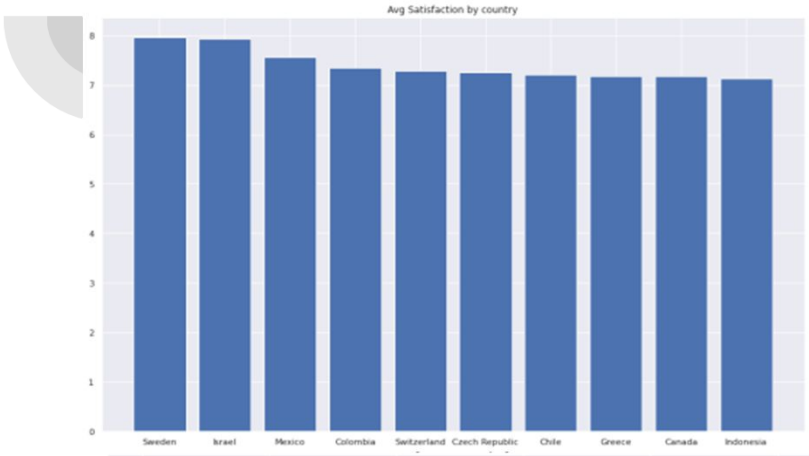
Challenge Description

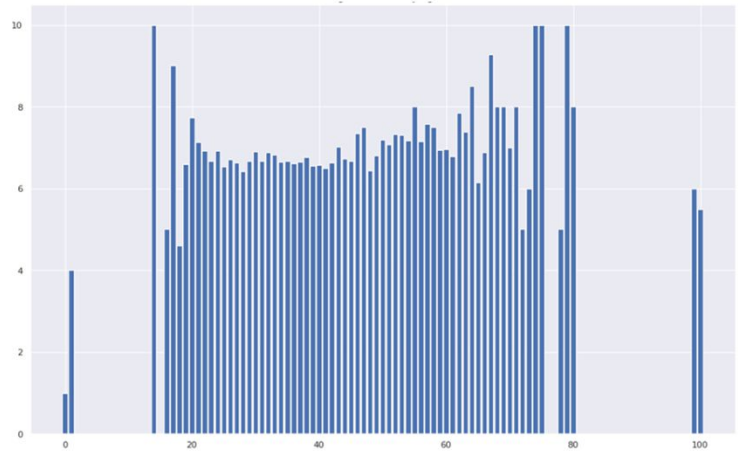
In this problem, we were to predict job satisfaction of a Kaggler given the dataset. The dataset consists of different categorical and ordinal values that needed to be processed. And the challenge is to come up with such and ML code or algorithm that can predict the job satisfaction with the given training data and obtain the highest accuracy score with the test data.

Data Analysis

We had to visualize the data for further insight and to decide what do with the given data. So before processing any further we have used code pieces like **.shape()** so that we could understand what the data includes and what it doesn't. Afterwards we have obtained some main graphs that could help us during the process, and how we could handle the data that was given. We have obtained **54** different columns or attributes for the data and the result being the job satisfaction level were rated between **0** and **10**, **10** being the highest satisfaction level.

Data Visualization

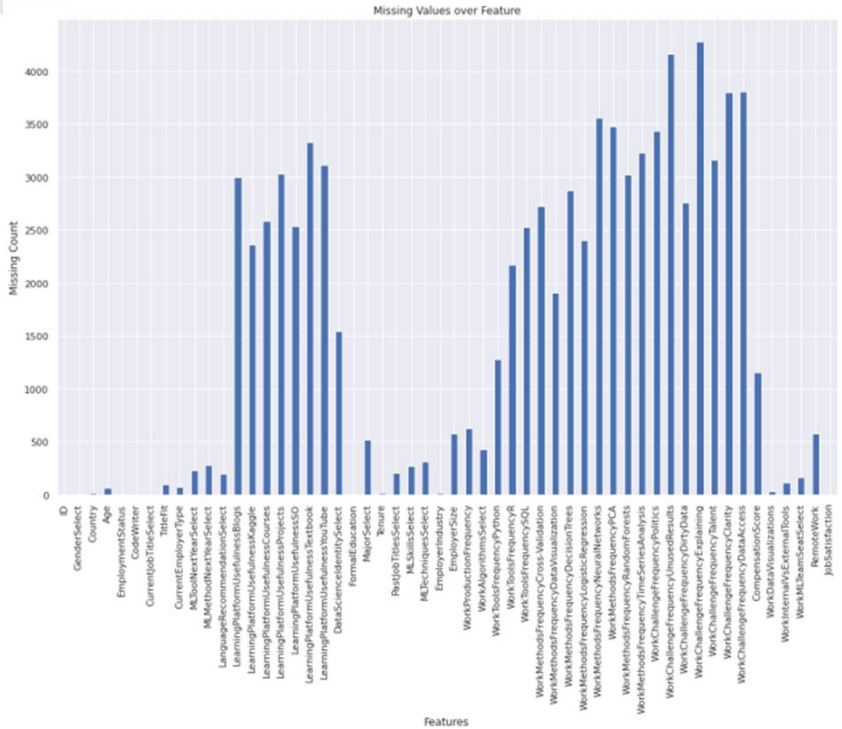




- From the graph left we have printed out the top 10 countries with the highest job satisfactions sot that we can have further insight about the data. And as can be seen in the graph, the average satisfaction is higher in western countries such as Sweden, Switzerland and Czech Republic.

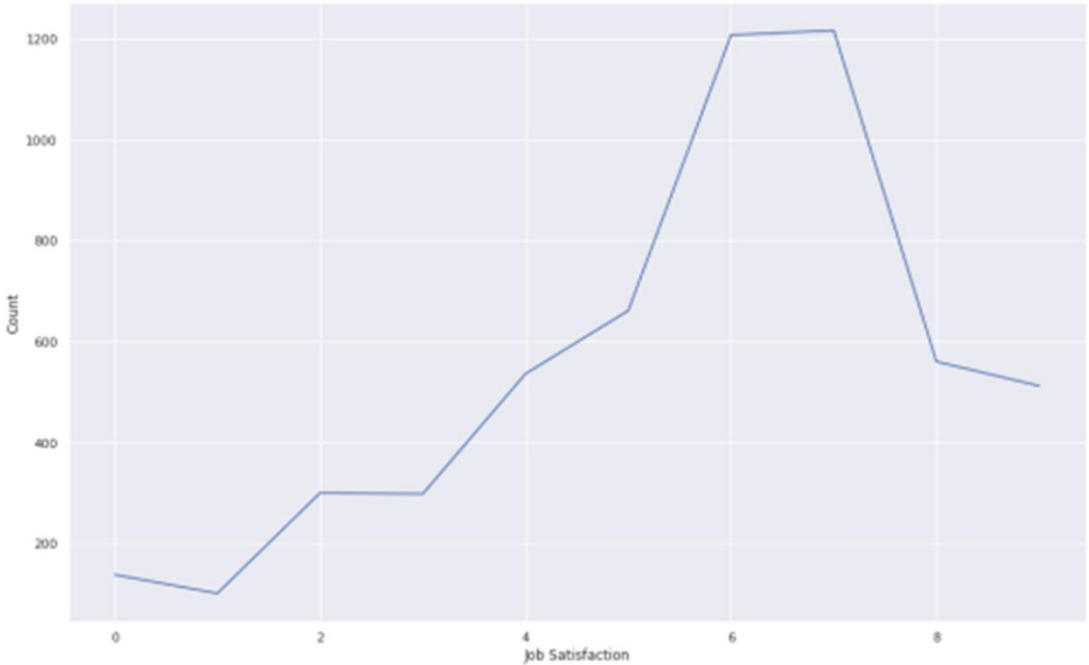
-Similarly from the graph on the left, we have obtained average satisfaction by age graph. As can be observed highest satisfaction level belongs to those with 18-24 and 45-65. Similarly, we can see that there is missing features for the data since there are unrelated values in the edges.

Data Visualization Cont.



We have also obtained how much irrelevant or missing data we have, and printed it out. For most categories, the data had missing values that needed to be reshaped, either filled or to be dropped. We have resolved the issue by imputing NaN into mean of the column. Likewise, dropping some columns has helped to get rid of missing values.

Data Visualization Cont.



From the graph left, we can see that there is a shape similar to bell curve, only it's mean is not centered in the middle but at the higher part of the values.

Preprocessing

From the previous slides we have gained insight that there were a lot of empty values for specific data sets. Also some of the categorical values were nearly same and were causing trouble while we move any further with our modelling and data. So we have continued with the following steps;

- We have dropped some irrelevant features or features we thought irrelevant such as ID, Job Title and Gender
- Similarly we have changed the categorical attributes with the numerical ones for our model to process better.
- At the end of the processing we have obtained a data removed from all the non values, either dropped or filled with other values, and nearly possessing numerical values for the attributes.
- After we have continued with the reshaped data.

Model Description

The problem can be either classified as a classification problem or a regression problem, since model is supposed to predict integer values between 0 and 10 inclusive. We trained different models including Linear Regression, Decision Tree,Naive Bayes, Ridge Regression and Neural Network. We had different MSE and accuracy rates on different ML techniques. The highest accuracy for the reshaped data was with the neural network reaching up-to 90% accuracy rate. But we have obtained high MSE results. So we have picked another ML technique which is Decision Tree Classifier.

Summary

Real world problems require very complex datasets the integrity of which should be questioned, and therefore, preprocessed before training a model. During this project Preprocessing the data with 54 column values was a challenge.

Best Accuracy was achieved when Decision Tree Classifier was used with 5.66 Mean Square Error.