CS412 - KAGGLE PROJECT

Job Satisfaction of a Kaggler

Group Name: Group 29

Can Aksoy - 23596

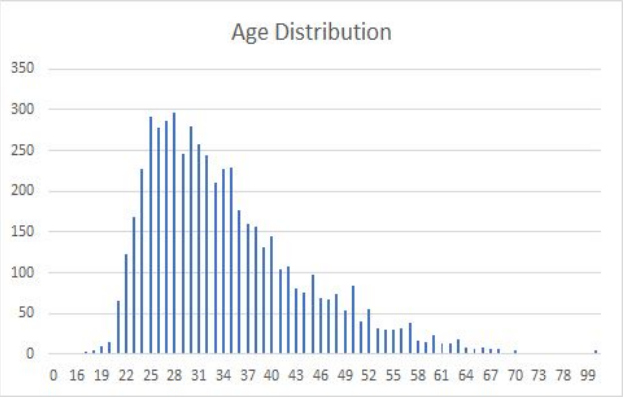
Şevval Tufan - 23679

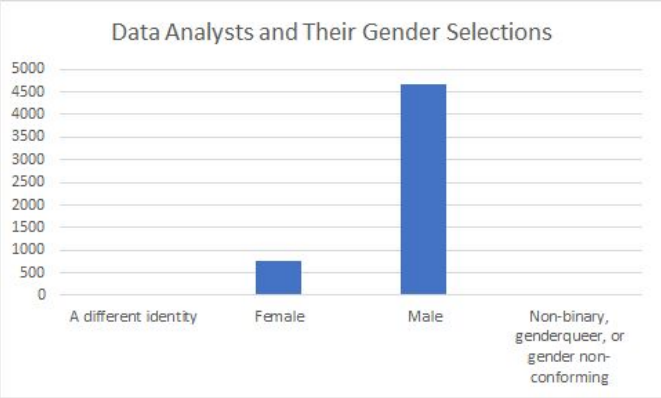
Notebook:https://colab.research.google.com/drive/1MVPR2-9NXuqI-VtDSr9tP4WK2LW5mXPV?usp=sharing

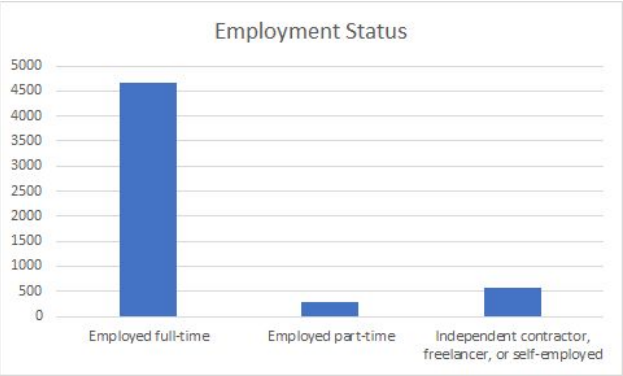
Data Visualisation and Insights

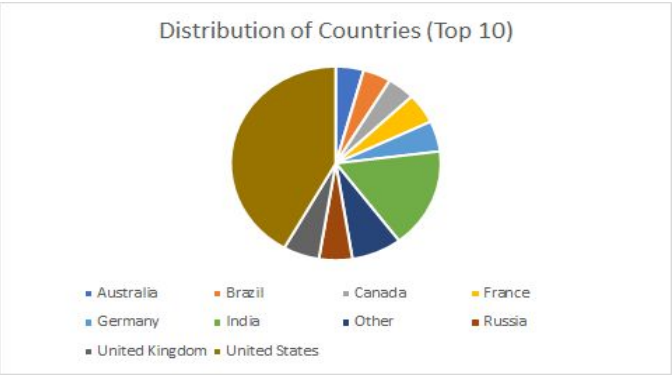
- 5529 participants (samples), 54 features.

Demographic Features

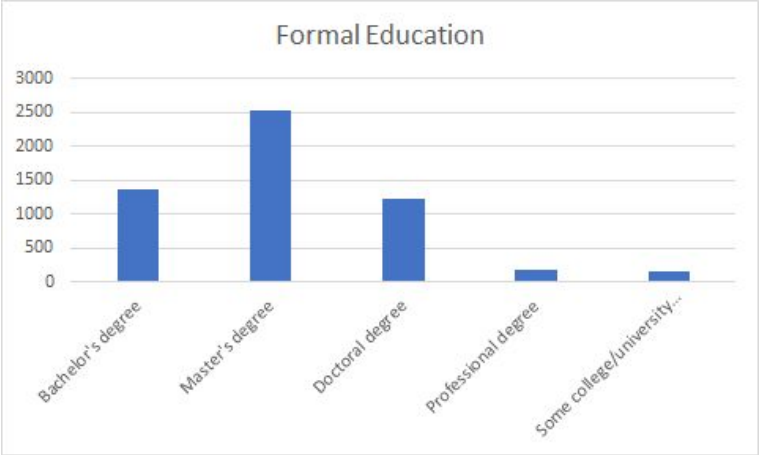


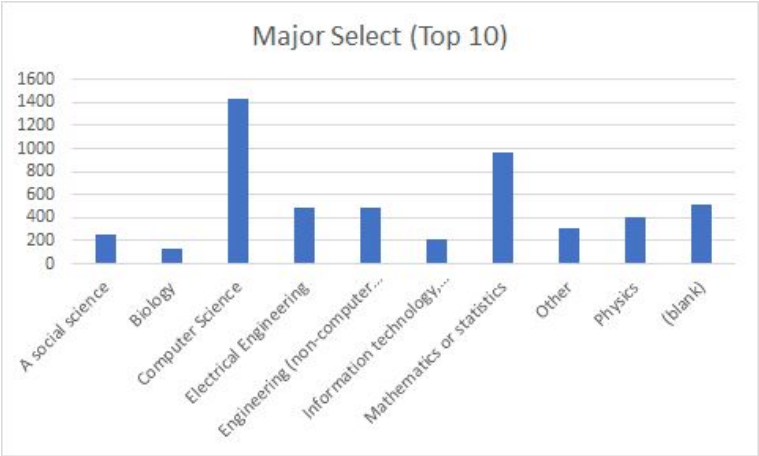




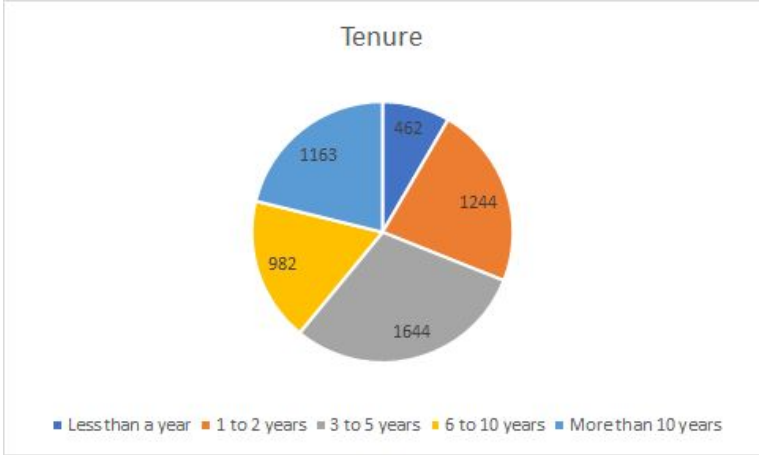


Education Features

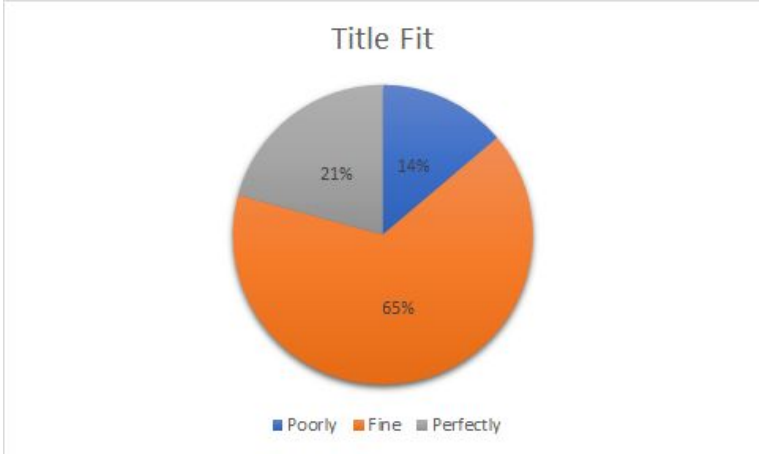




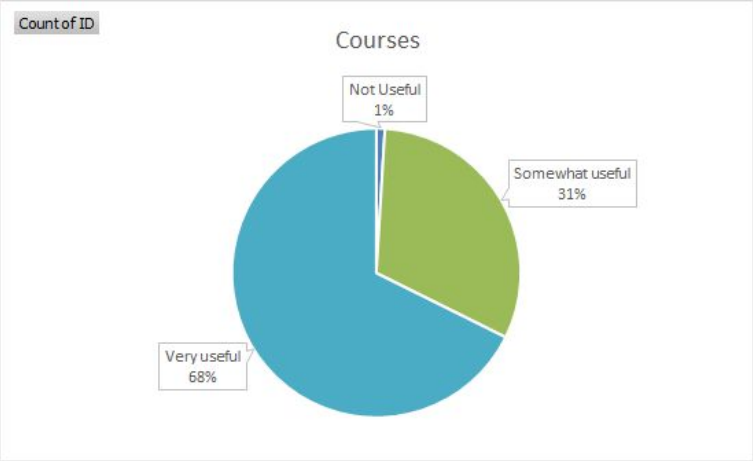
Data Science Experience

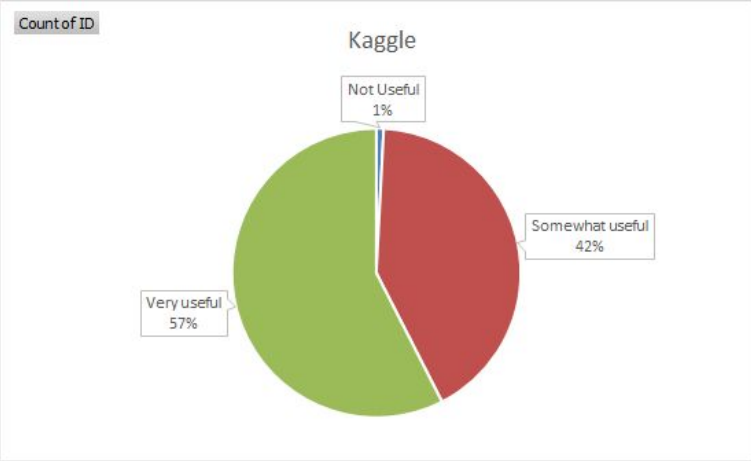


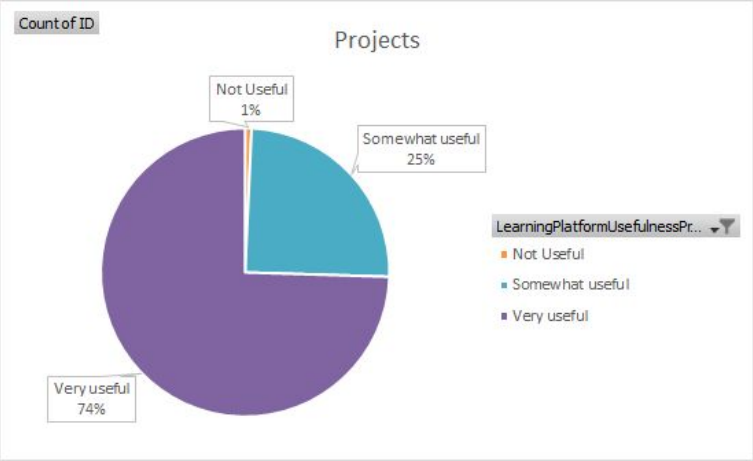
Workspace

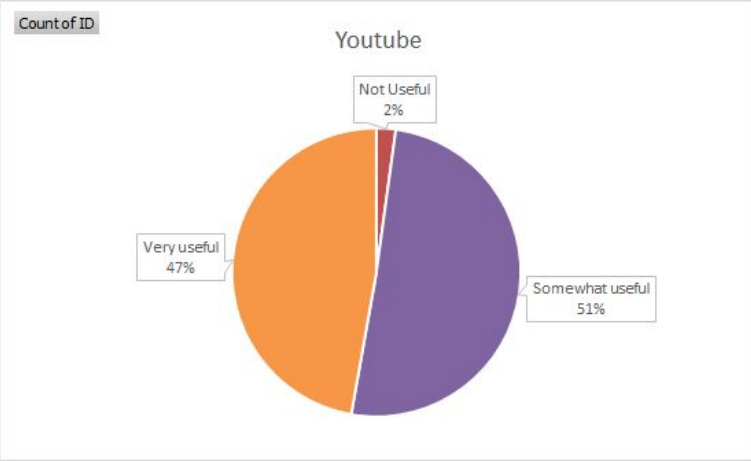


Learning Platform Usefulness

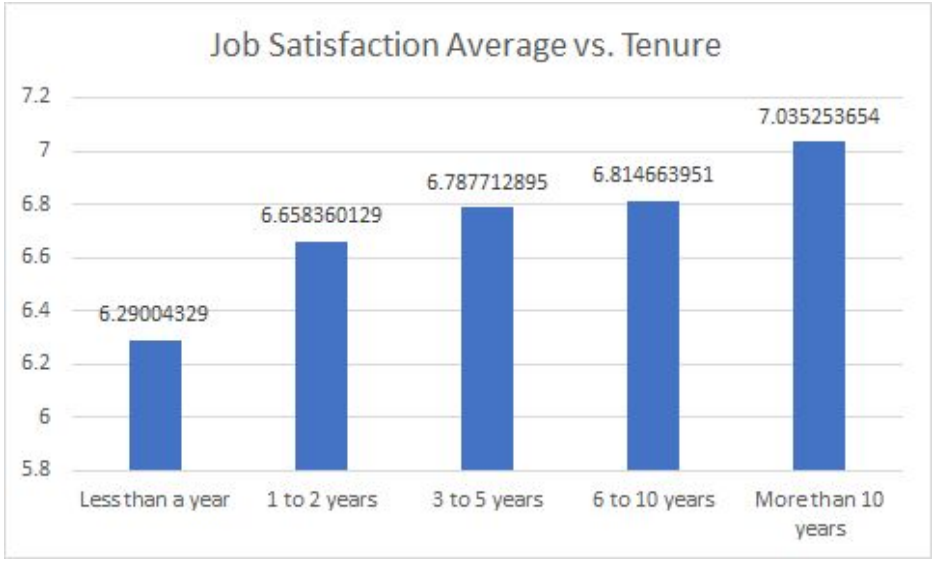


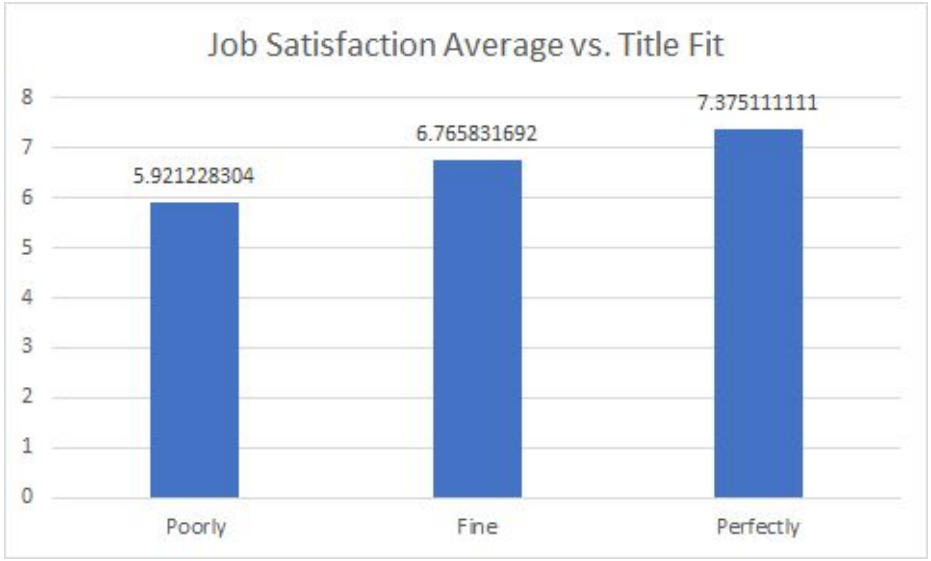






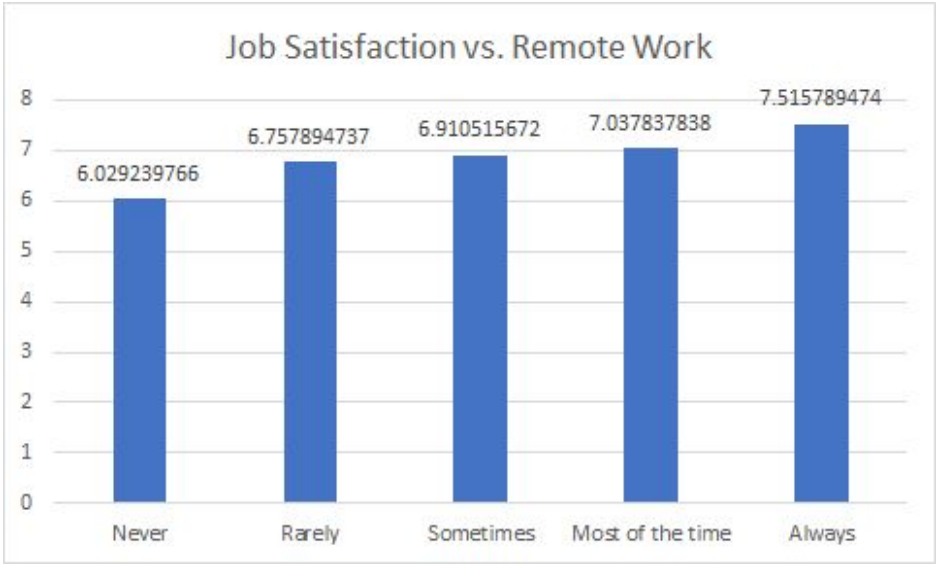
Time to Draw Insights...

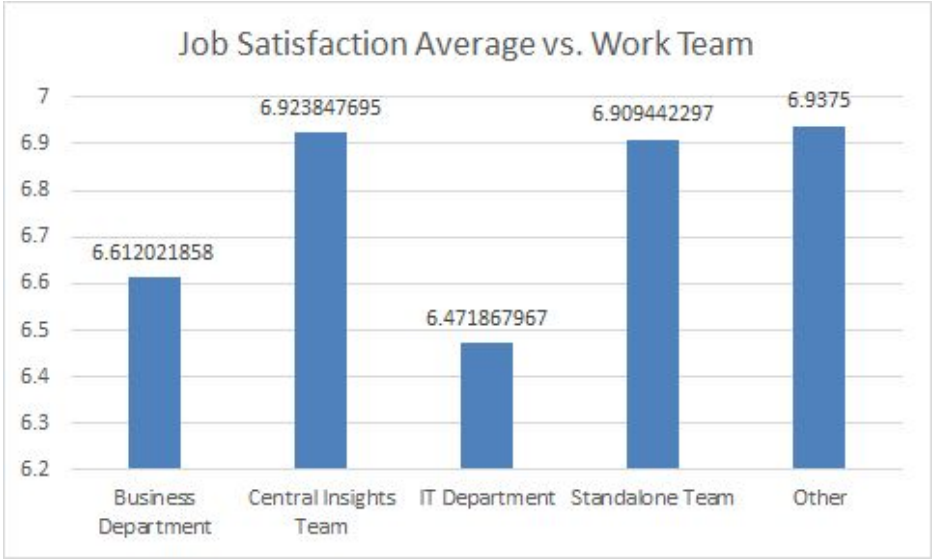




- The longer people write code to analyze data, the higher job satisfaction they have.
- Especially in the first year, job satisfaction of people is considerably less.
- If people feel that their title fit what they do, job satisfaction level increases.

Time to Draw Insights...





- It can be interpreted as that the frequency of working remotely is proportional to the level of job satisfaction people have.
- Moreover, people who never work remotely has relatively low job satisfaction.

- People who work in IT Department and Business Department have relatively low job satisfaction compared to other departments.
- Job satisfaction of the other departments are very close to each other.

Our Approach

Preprocessing

1. **Import Dataset and Related Libraries**
   - The necessary libraries used in the program are imported such as pandas and numpy.
   - The dataset is converted into csv file and then, it is read using pandas function “pd.read”.

2. **Identify and Handle Missing Values**
   - Missing values in the dataset are found and their data types are checked.
   - If the data types are float or integer, the missing values of the corresponding columns are filled with the mean of the column.
   - If the data types are object, then the missing values of the corresponding columns are filled based on our decision.

     For example: If the answers are related to the frequency such as rarely, sometimes and always, then the missing values of these features are filled with “rarely”. Also, ,if the feature column includes “other”, then the missing values of this feature is filled with “other”.

Our Approach

Preprocessing

1. **Feature Selection/Extraction**
   - The necessary libraries used in the program are imported such as pandas and numpy.
   - The dataset is converted into csv file and then, it is read using pandas function “pd.read”.

2. **Identify and Handle Missing Values**
   - Missing values in the dataset are found and their data types are checked.
   - If the data types are float or integer, the missing values of the corresponding columns are filled with the mean of the column.
   - If the data types are object, then the missing values of the corresponding columns are filled based on our decision.

     For example: If the answers are related to the frequency such as rarely, sometimes and always, then the missing values of these features are imputed with “rarely”. Also, ,if the feature column includes “other”, then the missing values of this feature is imputed with “other”.

Our Approach

Preprocessing

3. Feature Selection/Extraction

- The corresponding answers of each feature/question are considered in detail and the ones thought as irrelevant to job satisfaction are removed from the dataset that are “MLToolNextYearSelect”, “MLMethodNextYearSelect”, and “WorkInternalvsExternalTools”.
- The features having too many different answers are dropped out of the dataset that are “PastJobTitleSelect”, “MLSkillsSelect”, and “MLTechniquesSelect”.
- Finally, ID feature that identifies each participant is removed since it has no contribution to determine job satisfaction of the participants.

4. Encoding Categorical Data

- Since the dataset are mostly based on texts, in order for the machine to understand these texts, they are converted into numerical values.
- The features that are related to frequency are converted to ordinal integer starting from 1 to 4. (rarely => 1, sometimes => 1, often => 1, most of the time => 1)
- The rest of the features that do not have integer or float values are encoded with value between 0 and n_classes-1.

Our Approach

Preprocessing

5. Splitting the Dataset

- The dataset is split into two sets that are training set and test set. The model is trained using the training set and then, the performance of the model is tested using the test set.
- 80% of the dataset is allocated to training set and the remaining 20% to test set.

Methods/Algorithms

- Since it is asked to predict the job satisfaction of a Kaggler based on a number of independent features/variables, the analysis of the dataset is done based on regression.
- 7 different regression algorithms are tried to analyse the dataset:
  - Random Forest Regression
  - Logistic Regression
  - Gaussian Naive Bayes
  - MLP Regression
  - Support Vector Regression
  - Linear Regression
  - Keras Regression

Our Approach

Methods/Algorithms

- The algorithm that resulted in best cross-validation after hyperparameter tuning, performance is Random Forest and the corresponding result is 1.98158 as root mean squared error.

The Best Method

- Random Forest is a bagging technique and it operates by constructing multiple decision trees.
- After the training set is trained, the mean of the predictions are calculated and the output is defined.
- In this method, the hyperparameter tuning is done carefully by setting minimum number of sample splits, the maximum depth of the tree, and the minimum number of samples leaf.
- Hyperparameter tuning is crucial since it prevents that the ensemble model does not heavily count on any individual feature and allows every feature to contribute fairly.
- Finally, the number of trees are chosen as 100 based on the best result obtained.