 

**Cairo University Credit Hours System**

**Faculty of Engineering CCE-C**

**Big Data report**

**“US elections voting”**

**Submitted by:**

1. Assem Amr Sadek
2. Lana Tarek Shafik
3. Mohamed Mohsen El-Samanoudy

**Description of the problem:**

In this problem, we used data from thousands of users and one hundred different questions to see for which party, Democrat or Republican, a person is going to vote.

**Analysis:**

As it is a supervised learning problem and we are trying to predict binary classes (Democrat or Republican) so, we decided to try Logistic Regression, Decision Tree, Boosted Decision Tree and Random Forest models.

**Pipeline Solution:**

1. Data Cleaning (Preprocessing):

First, we removed ‘USER\_ID’ feature as it wasn’t an indication for anything just a numbering.

Second, we removed some training examples with inconvenient Year Of Birth ‘YOB’ such as people below 18 years old and above 100. Also, there were some training examples with Year Of Birth greater than 2016 and they were also excluded.

Finally, we have the problem of missing data. By observing the number of examples with missing data, we discovered that there are about 3000 from 5568 examples with missing data. So, removing these examples were not a good choice as it will decrease our data and we decided to use imputation to fill all null values. Also, we found that most probably each generation have similar features so, we decided to group people to generations of 10 years each.

1. Model Building:

Our last model is binary logistic regression. We used the built-in function in R to know which independent variables are significant/useful to keep. The final model contains these variables:

Income, Q118232, Q116881, Q116953, Q115611, Q115899, Q109244, Q106993, Q106997, Q101162 and Q98869

1. Results and evaluation:

The resulting accuracy is about 80% on test data.

1. Visualization:

We used pie chart technique, because it’s the common understandable way for public people when we want to represent a voting result.

**Trials:**

After using many techniques, like decision tree, boosted decision tree, random forest, logistic regression, we decided to use logistic regression since it gives the highest accuracy on test data.

While trying to enhance the performance of other models, we noticed that the models tend to overfit that’s why chose logistic regression at the end.

**Enhancement in the future:**

The data contains a lot of missing values and it’s partially virtual, so we need to make a real questionnaire on real people without the interference of any simulation tool.