DATA MINING & DATA WAREHOUSING PROJECT

Prediction of Visa Application

Under the guidance of:
Prof. S. G. Bhirud

TEAM MEMBERS:
Mahek Nakhua (191071048)
Vaishnavi Shah (191071067)
Shraddha Keniya (201071901)

BRANCH: B.Tech CS Final Year
SUBMISSION DATE: 22/11/22

TITLE:

Prediction of allotment of H1B work visa in the USA using machine learning

ABSTRACT:

This project involves generation of a prediction model for predicting the allotment of the H1B Wprk Visa. The methodology used to perform this project was: obtaining data from OFLC, data cleaning, exploratory data analysis, feature scaling and selection, model training and predictions followed by comparisons of various models on the basis of performance. It was observed that Decision Tree was the model with the best performance.

INDEX:

INTRODUCTION:

The H-1B is an employment-based, non-immigrant visa category for temporary foreign workers in the United States. Every year, the US immigration department receives over 200,000 petitions and selects 85,000 applications through a random process. The application data is available for public access to perform in-depth longitudinal research and analysis.
For a foreign national to apply for H1-B visa, a US employer must offer them a job and submit a petition for a H-1B visa to the US immigration department. This is also the most common visa status applied for and held by international students once they complete college or higher education and begin working full-time.

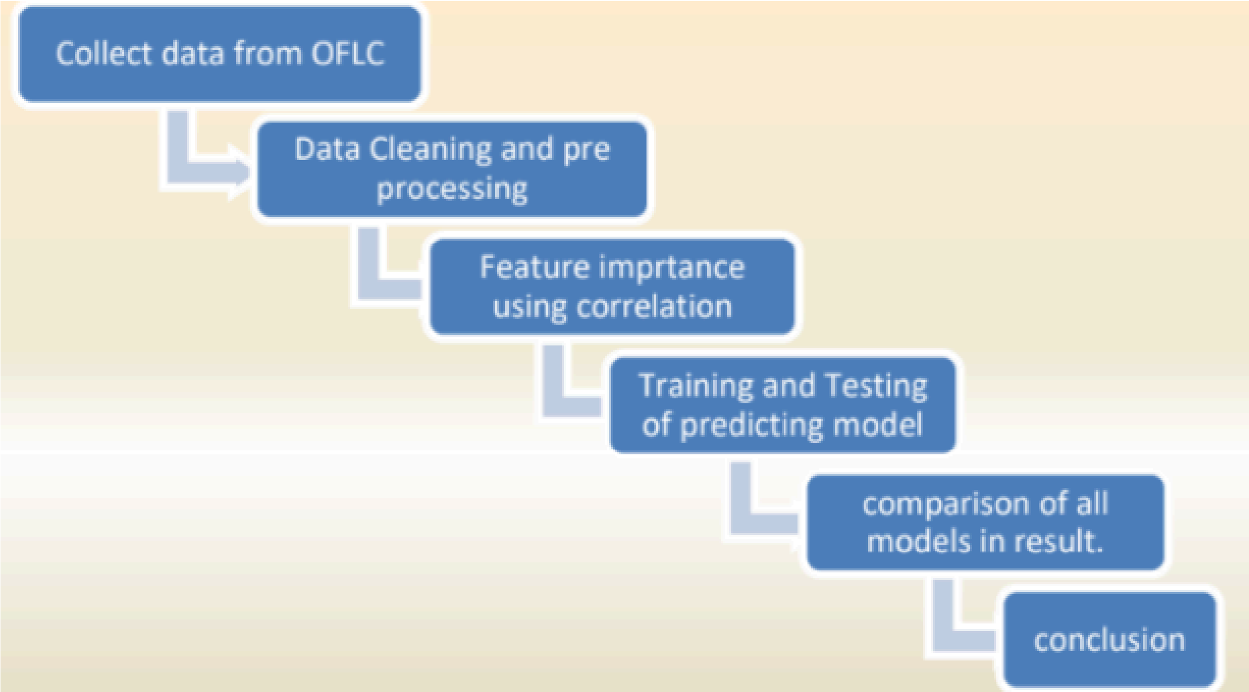
LITERATURE SURVEY (THEORY):

Visa is the guide of authorization on a travel permit that gives a permit to the holder to move in, leave or stay in the country for a predetermined timeframe. There are distinctive kinds of foreign visas, the required structures, and the means in the worker visa process contingent upon the nation one needs to move to. Moving to America is a vital and complex decision. The U.S of America has numerous classes for settler visas like H1B, L1, J1, and so on. To be qualified to apply for a worker visa, an outside native must be supported by a USA subject relative, U.S. legitimate perpetual inhabitant, or a planned business, with a couple of special cases.

H-1B is a business-based non-transient visa gathering for brief remote specialists in the US. For an outside national to apply for an H1B visa, a US business must offer an occupation and request for an H-1B visa with the US movement office. This is the most widely recognized visa status connected to and held by universal understudies once they finish school/advanced education (Masters, Ph.D.) and work in a full-time position. The Office of Foreign

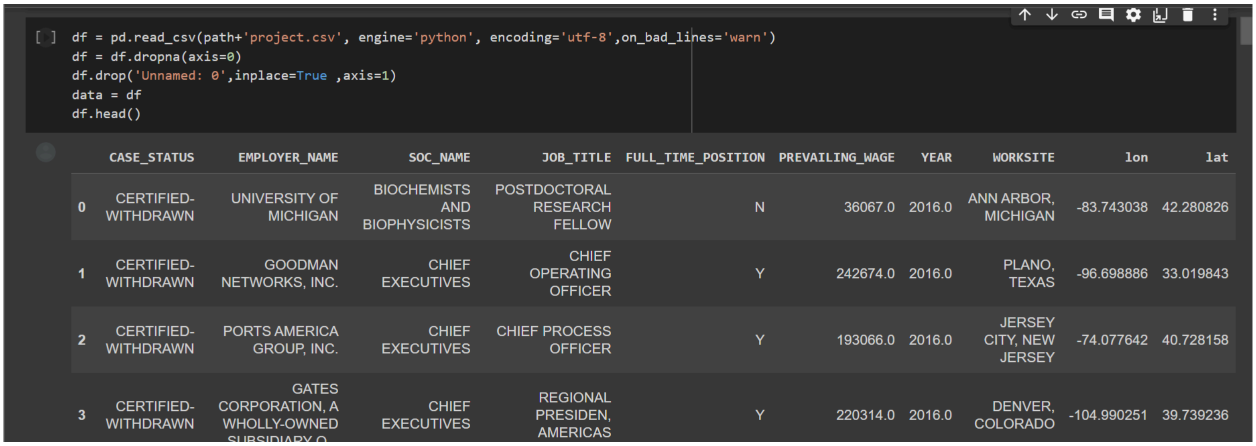
Labour Certification (OFLC) creates program information that is helpful data about the movement programs including the H1-B visa. It is intended to carry outside experts with professional educations and specific aptitudes to fill occupations when qualified Americans can't be found.

The ***methodology*** used to implement this project is as shown in the below image.


Collect data from OFLC
Data Cleaning and pre processing
Feature imprtance using correlation
Training and Testing of predicting model
comparison of all models in result.
conclusion

DATASET

The dataset used has been collected and generated by the Office of Foreign Labor Certification. The Office of Foreign Labor Certification (OFLC) generates program data that is useful information about immigration programs including the H1-B visa.



```python
df = pd.read_csv(path+'project.csv', engine='python', encoding='utf-8',on_bad_lines='warn')
df = df.dropna(axis=0)
df.drop('Unnamed: 0',inplace=True ,axis=1)
data = df
df.head()
```

	CASE_STATUS	EMPLOYER_NAME	SOC_NAME	JOB_TITLE	FULL_TIME_POSITION	PREVAILING_WAGE	YEAR	WORKSITE	lon	lat
0	CERTIFIED-WITHDRAWN	UNIVERSITY OF MICHIGAN	BIOCHEMISTS AND BIOPHYSICISTS	POSTDOCTORAL RESEARCH FELLOW	N	36067.0	2016.0	ANN ARBOR, MICHIGAN	-83.743038	42.280826
1	CERTIFIED-WITHDRAWN	GOODMAN NETWORKS, INC.	CHIEF EXECUTIVES	CHIEF OPERATING OFFICER	Y	242674.0	2016.0	PLANO, TEXAS	-96.698886	33.019843
2	CERTIFIED-WITHDRAWN	PORTS AMERICA GROUP, INC.	CHIEF EXECUTIVES	CHIEF PROCESS OFFICER	Y	193066.0	2016.0	JERSEY CITY, NEW JERSEY	-74.077642	40.728158
3	CERTIFIED-WITHDRAWN	GATES CORPORATION, A WHOLLY-OWNED SUBSIDIARY O	CHIEF EXECUTIVES	REGIONAL PRESIDEN, AMERICAS	Y	220314.0	2016.0	DENVER, COLORADO	-104.990251	39.739236

This dataset contains five years' worth of H-1B petition data, with approximately 3 million records overall. The columns in the dataset include case status, employer name, worksite coordinates, job title, prevailing wage, occupation code, and year filed.

The meaning of the columns are as follows:

- **CASE_STATUS** - status of the application
- **EMPLOYER_NAME** - the name of the employer (submitting labor condition application) as registered in the H-1B Visa application
- **SOC_NAME** - the occupation code for the employment (Occupational name associated with the `SOC_CODE`. `SOC_CODE` is the occupational code associated with the job being requested for temporary labor condition, as classified by the Standard Occupational Classification (SOC) System)
- **JOB_TITLE** - the job title for the employment
- **FULL_TIME_POSITION** - whether the application is for a full-time position or for a part-time position (`Y` = Full-Time Position; `N` = Part-Time Position.)
- **PREVAILING_WAGE** - the most frequent wage for the corresponding role as filled in the Visa application. The wage is listed on an annual scale in USD.

- ➔ **YEAR** - the year in which the H-1B visa petition was filed
- ➔ **WORKSITE** - the address of the employer's worksite
- ➔ **lon** - longitude of the employer’s worksite
- ➔ **lat** - latitude of the employer's worksite

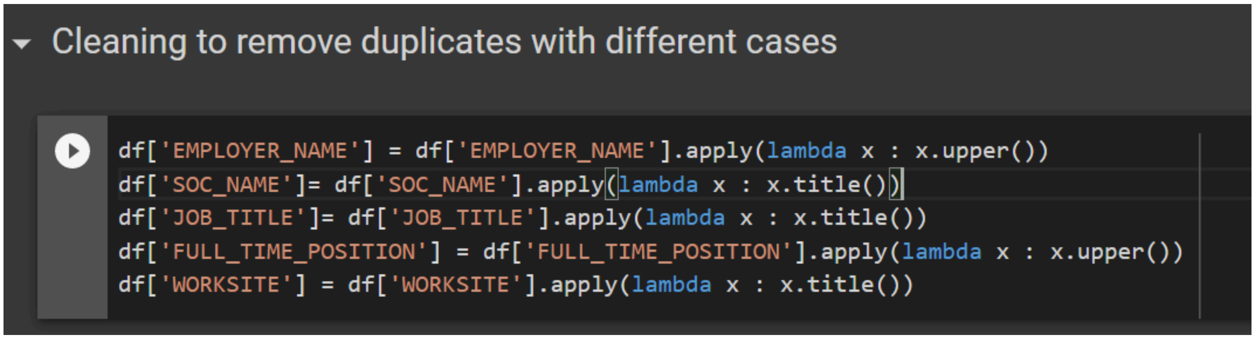
SCOPE OR NEED:

The motivation behind this project and studying this dataset is to help US visa applicants gauge their chances of getting their visa approved based on various factors that have been taken into account in the dataset. We have drawn inferences about which factors play a major role in deciding the status of the applicant’s decision.

Speaking from a business perspective, profound knowledge is required with the goal that the businesses comprehend the procedure of the visa appeal, to stop the outsourcing firms from backtracking applications, amending the framework, and utilizing better techniques like compensation-based, justify-based, or encounter-based for conceding of visas. By means of this project, we try to foresee the negative and positive consequences of the applications and find for which sort of occupation, the number of petitions is high or low with the goal that contracting of economical work is extremely dense by utilizing machine learning techniques.

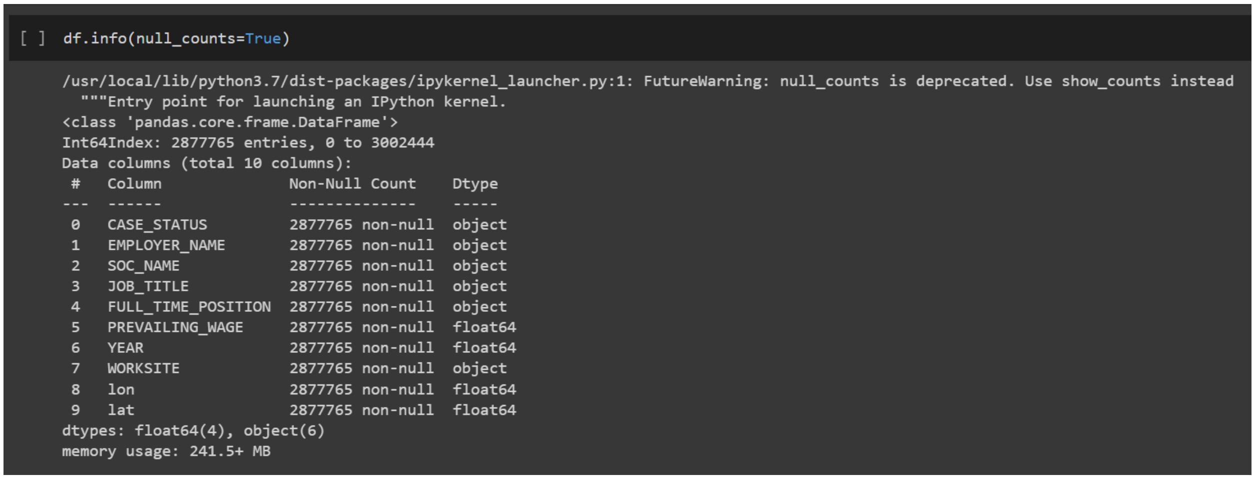
IMPLEMENTATION:

DATA CLEANING

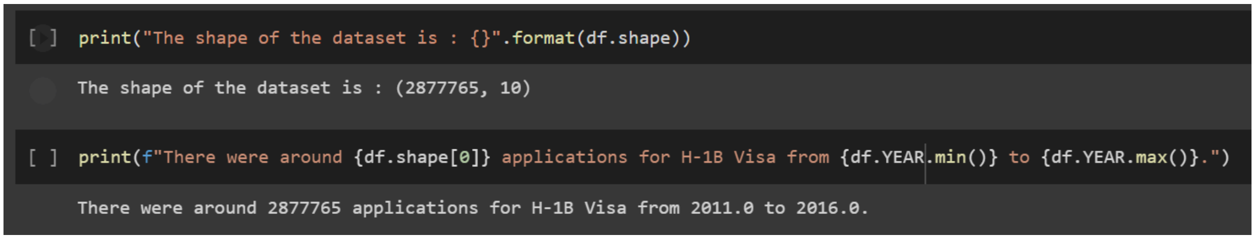


Cleaning to remove duplicates with different cases

```
df['EMPLOYER_NAME'] = df['EMPLOYER_NAME'].apply(lambda x : x.upper())
df['SOC_NAME']= df['SOC_NAME'].apply(lambda x : x.title())
df['JOB_TITLE']= df['JOB_TITLE'].apply(lambda x : x.title())
df['FULL_TIME_POSITION'] = df['FULL_TIME_POSITION'].apply(lambda x : x.upper())
df['WORKSITE'] = df['WORKSITE'].apply(lambda x : x.title())
```



```
[ ]  df.info(null_counts=True)

     /usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: FutureWarning: null_counts is deprecated. Use show_counts instead
       """Entry point for launching an IPython kernel.
     <class 'pandas.core.frame.DataFrame'>
     Int64Index: 2877765 entries, 0 to 3002444
     Data columns (total 10 columns):
      #   Column              Non-Null Count    Dtype
     ---  ------              --------------    -----
      0   CASE_STATUS         2877765 non-null  object
      1   EMPLOYER_NAME       2877765 non-null  object
      2   SOC_NAME            2877765 non-null  object
      3   JOB_TITLE           2877765 non-null  object
      4   FULL_TIME_POSITION  2877765 non-null  object
      5   PREVAILING_WAGE     2877765 non-null  float64
      6   YEAR                2877765 non-null  float64
      7   WORKSITE            2877765 non-null  object
      8   lon                 2877765 non-null  float64
      9   lat                 2877765 non-null  float64
     dtypes: float64(4), object(6)
     memory usage: 241.5+ MB
```

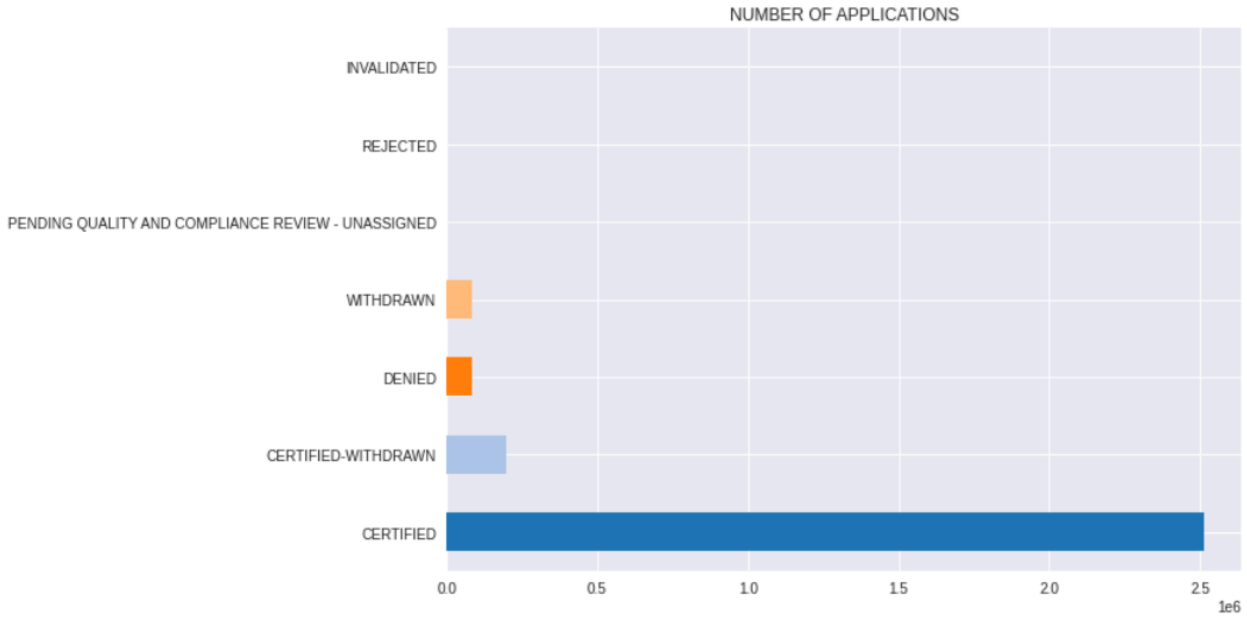


```
[ ]  print("The shape of the dataset is : {}".format(df.shape))

     The shape of the dataset is : (2877765, 10)

[ ]  print(f"There were around {df.shape[0]} applications for H-1B Visa from {df.YEAR.min()} to {df.YEAR.max()}.")

     There were around 2877765 applications for H-1B Visa from 2011.0 to 2016.0.
```

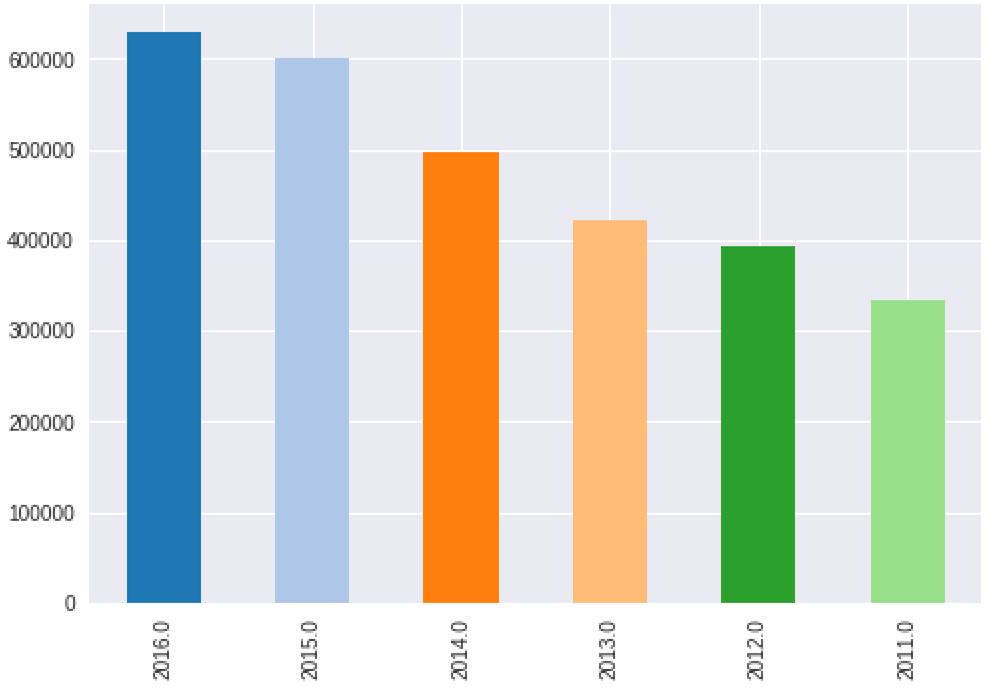
EXPLORATORY DATA ANALYSIS (EDA)

We performed exploratory data analysis to understand the various trends in the data set and to observe in patterns in the data.

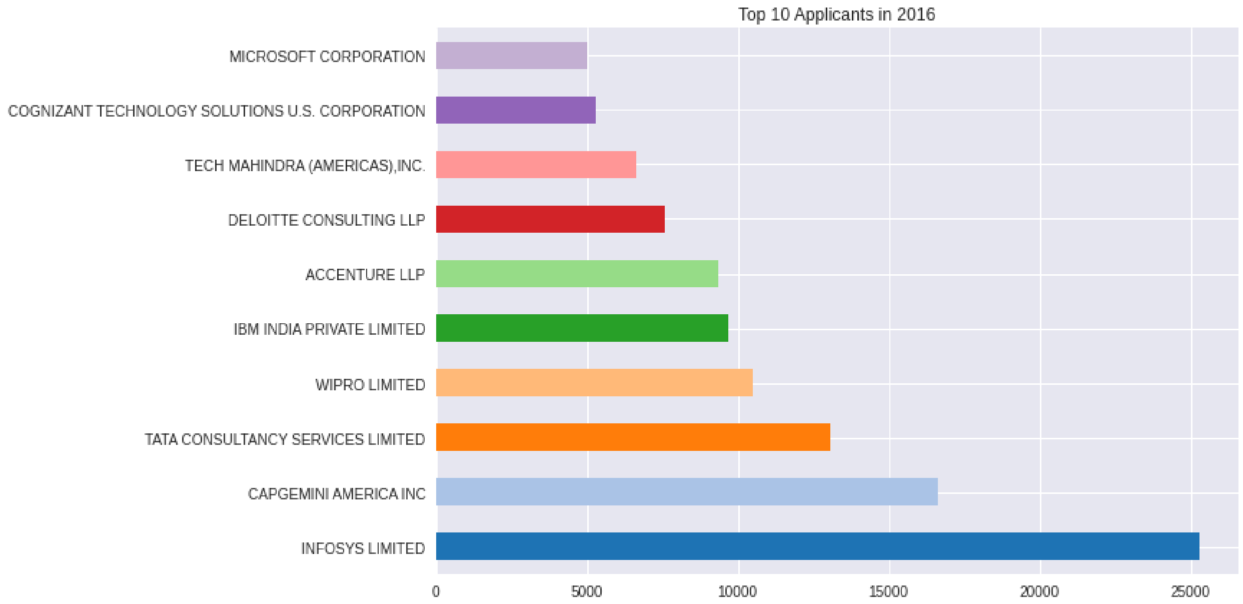
We plotted a graph of the counts of various case statuses:

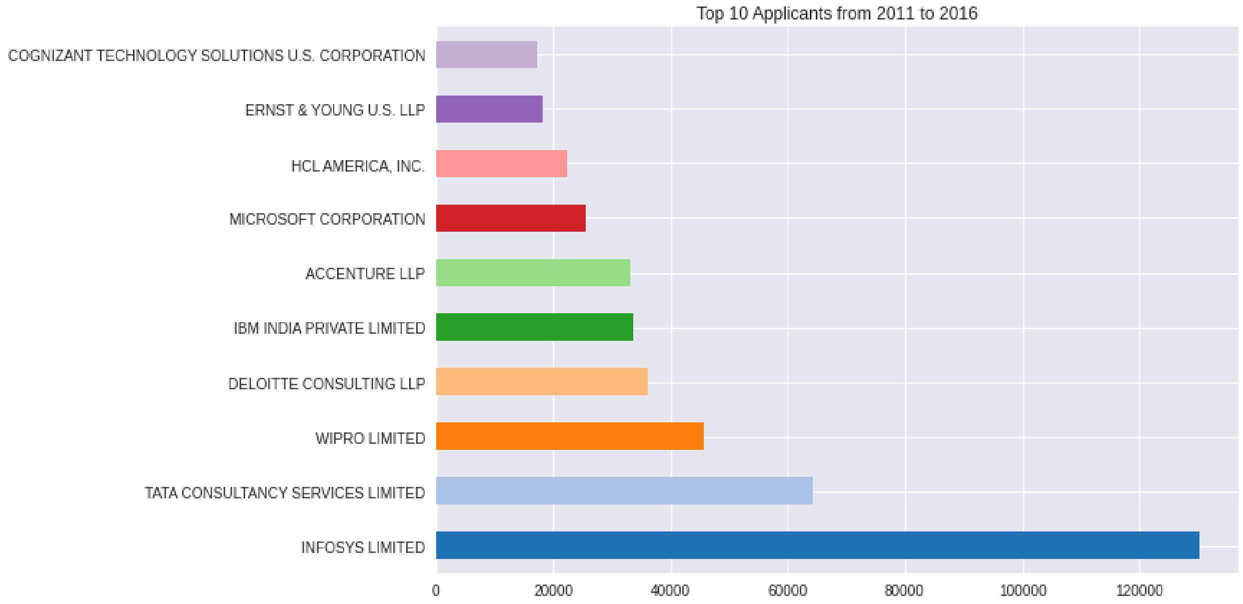


From the above graph we can say that the employees who have applied for the H-1B Visa were more than 2500000 whose application got certified and there were more than 200000 application's were certified and withdrawn and there were around 90000 whose application's were denied and there we around 80000 were withdrawn.

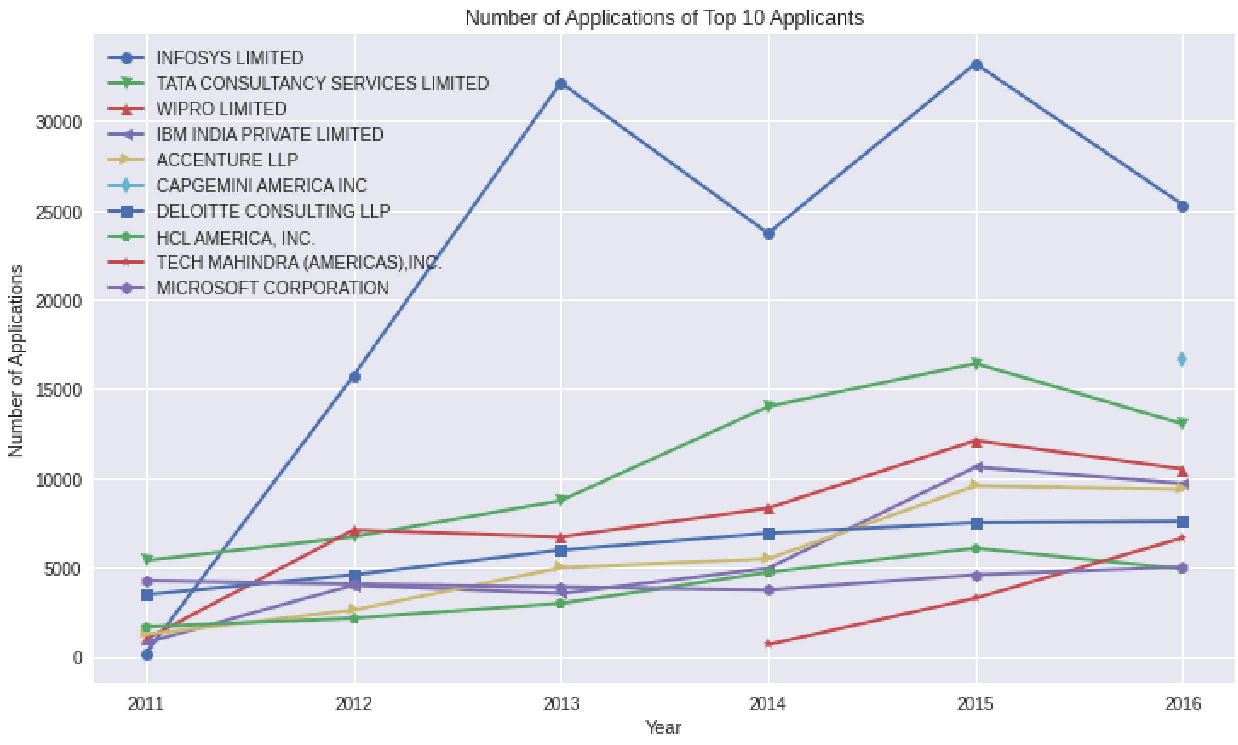


We can that there is a massive increase in the number of applications as the year passes



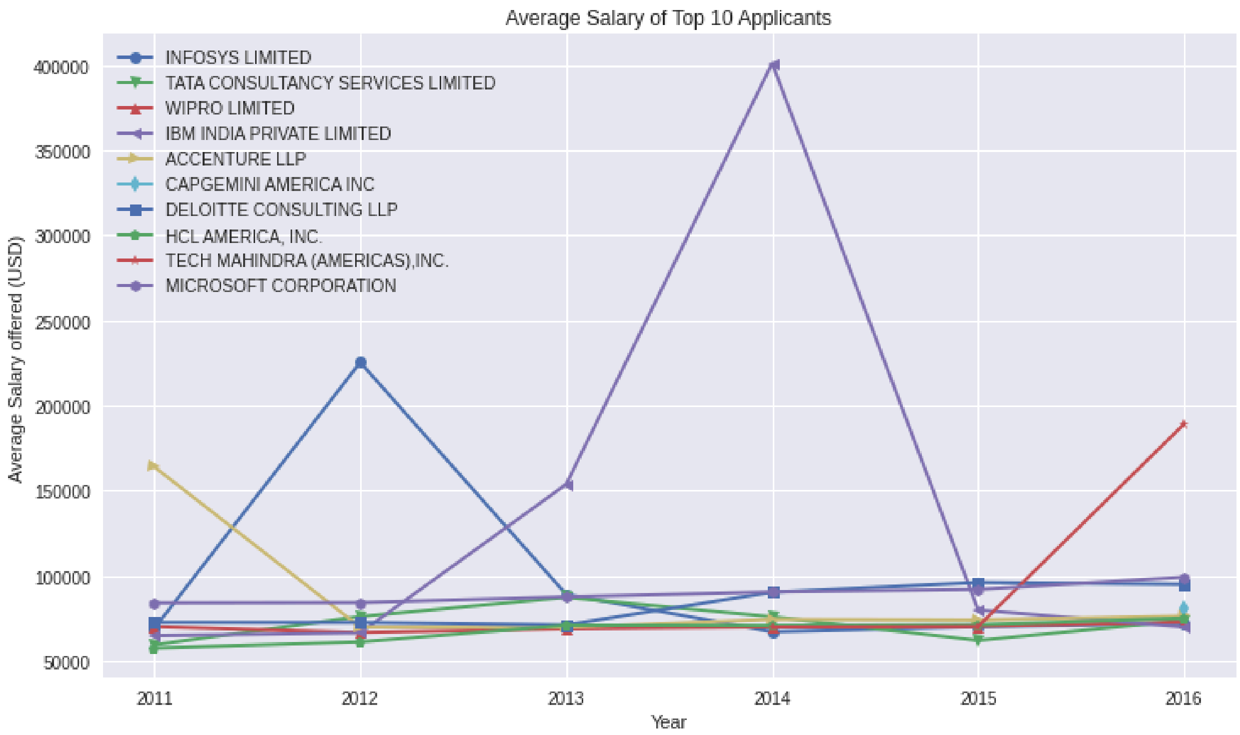


As seen from these graphs we can see the top companies which have been applying for H1-B Visas.

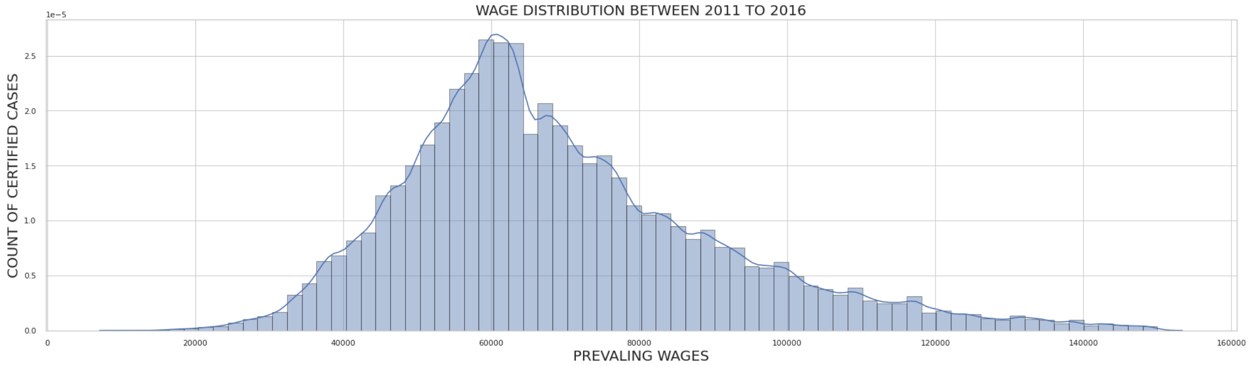


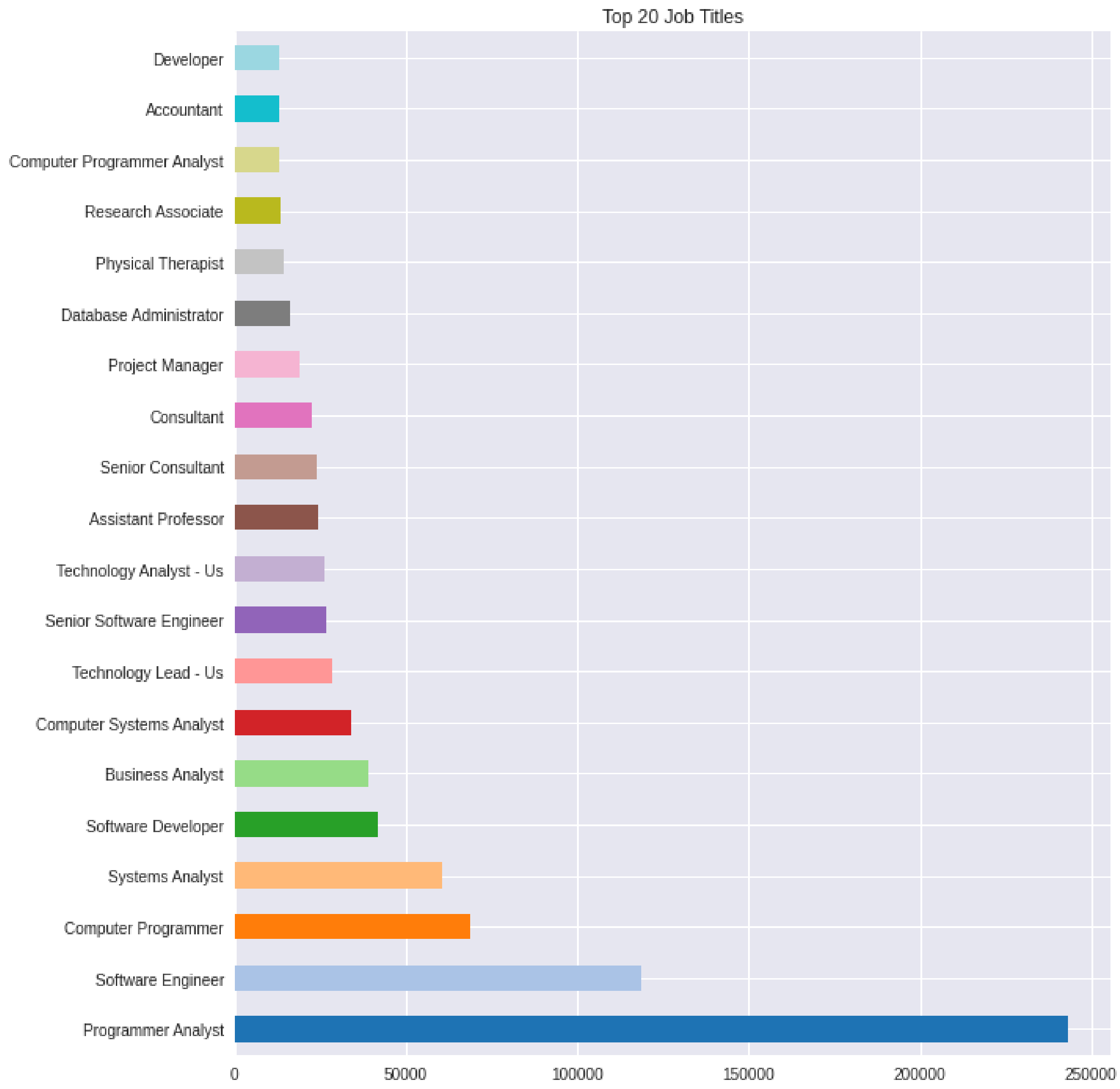
Furthermore, we analyzed the applications and concluded,

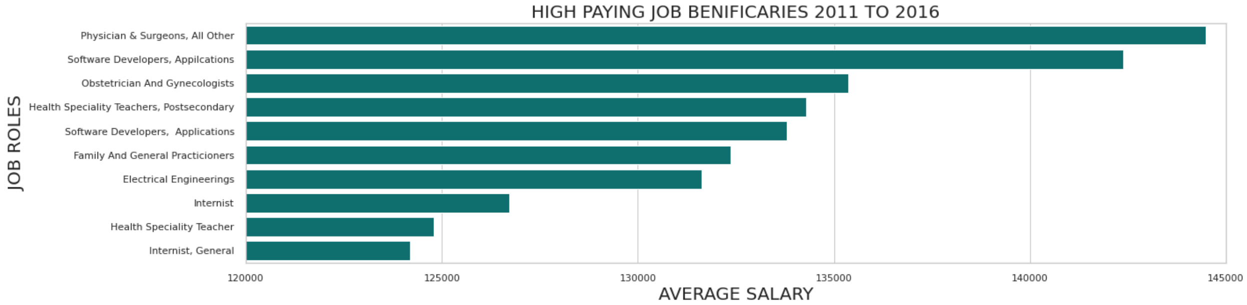
- We can clearly see that there are 2 new companies which are TECH MAHINDRA(AMERICAS), INC. & CAPGEMINI AMERICA.
- INFOSYS showed rapid growth between the years 2011 and 2013 where it came from 0 applications to more than 30k applications.
- TATA also showed significant growth.
- From the above plot except for the 2 newcomers we can say that the number of applications received by the top 10 employers started decreasing from the year 2015.
- These are the companies who filed the most number of applications.
- It is predominately dominated by Indian companies.



- ➔ We can see that the Average Salary offered by Infosys was very high compared to the rest of the companies in the year 2012.
- ➔ It's very interesting to see a huge peak in 2014 by IBM INDIA PRIVATE LIMITED looking like something went wrong.
- ➔ More sudden peaks were observed by ACCENTURE LLP in the year 2011 and by TECH MAHINDRA in the year 2016.



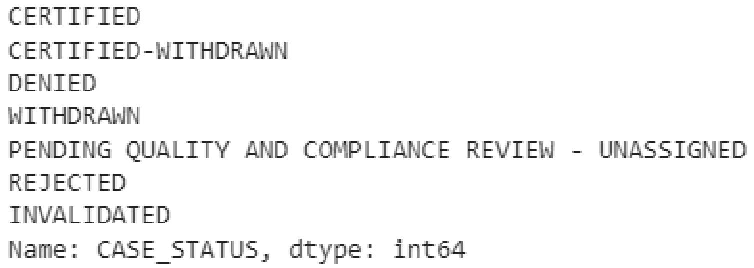




Using these two graphs we can see which job positions have the most number of applications and the distribution of wages amongst them.

FEATURE ENGINEERING

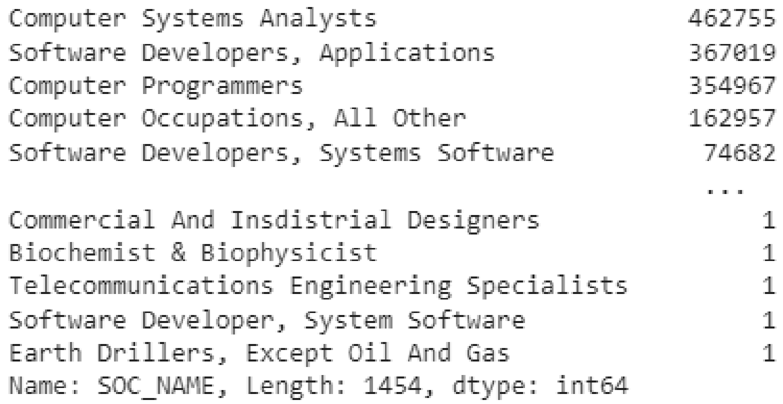
The data has several categories of case status:



```
CERTIFIED
CERTIFIED-WITHDRAWN
DENIED
WITHDRAWN
PENDING QUALITY AND COMPLIANCE REVIEW - UNASSIGNED
REJECTED
INVALIDATED
Name: CASE_STATUS, dtype: int64
```

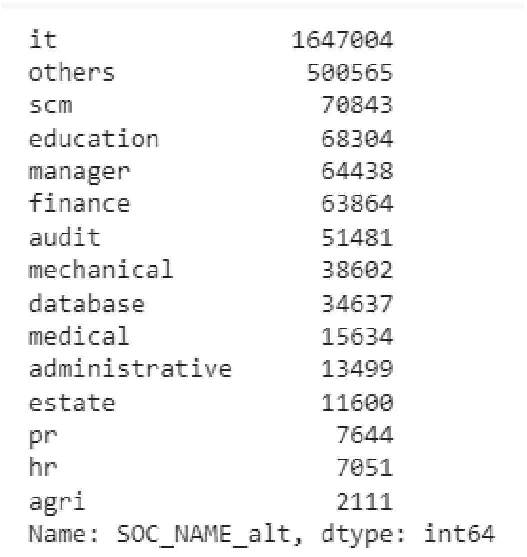
We combined this into two categories: Certified (Y) or Denied (N).

There were several types of soc_name



```
Computer Systems Analysts                        462755
Software Developers, Applications                367019
Computer Programmers                             354967
Computer Occupations, All Other                  162957
Software Developers, Systems Software             74682
                                                   ...
Commercial And Insdistrial Designers                  1
Biochemist & Biophysicist                             1
Telecommunications Engineering Specialists            1
Software Developer, System Software                   1
Earth Drillers, Except Oil And Gas                    1
Name: SOC_NAME, Length: 1454, dtype: int64
```

We generalized these categories as:



```
it                 1647004
others              500565
scm                  70843
education            68304
manager              64438
finance              63864
audit                51481
mechanical           38602
database             34637
medical              15634
administrative       13499
estate               11600
pr                    7644
hr                    7051
agri                  2111
Name: SOC_NAME_alt, dtype: int64
```

We also categorized the range of wage into these categories:

```
wage <=50000: "VERY LOW"
wage >50000 and wage <= 70000: "LOW"
wage >70000 and wage <= 90000 return "MEDIUM"
wage >90000 and wage<=150000: "HIGH"
wage >=150000: "VERY HIGH"
```

Further, we generalised the worksite column by extracting the state.

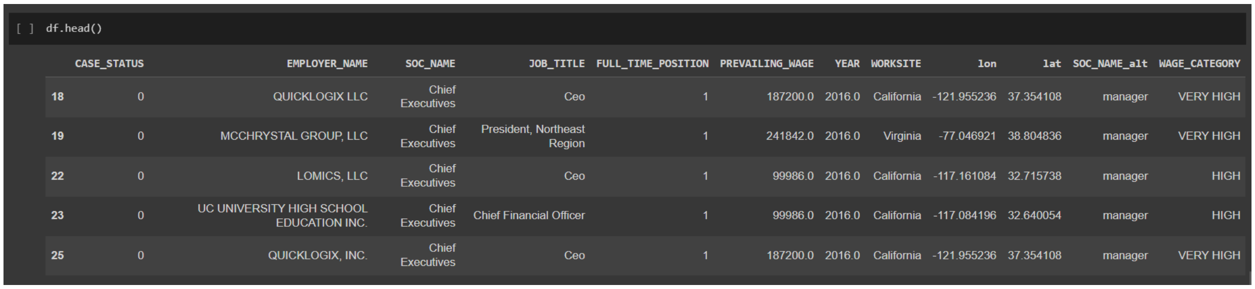
We decided to perform Label encoding on the following features:

- ➔ CASE_STATUS
- ➔ FULL_TIME_POSITION
- ➔ SOC_NAME

What is Label Encoding ?

Label Encoding refers to converting the labels into numeric form so as to convert it into machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

The updated dataframe is as shown below



```
[ ] df.head()
```

	CASE_STATUS	EMPLOYER_NAME	SOC_NAME	JOB_TITLE	FULL_TIME_POSITION	PREVAILING_WAGE	YEAR	WORKSITE	lon	lat	SOC_NAME_alt	WAGE_CATEGORY
18	0	QUICKLOGIX LLC	Chief Executives	Ceo	1	187200.0	2016.0	California	-121.955236	37.354108	manager	VERY HIGH
19	0	MCCHRYSTAL GROUP, LLC	Chief Executives	President, Northeast Region	1	241842.0	2016.0	Virginia	-77.046921	38.804836	manager	VERY HIGH
22	0	LOMICS, LLC	Chief Executives	Ceo	1	99986.0	2016.0	California	-117.161084	32.715738	manager	HIGH
23	0	UC UNIVERSITY HIGH SCHOOL EDUCATION INC.	Chief Executives	Chief Financial Officer	1	99986.0	2016.0	California	-117.084196	32.640054	manager	HIGH
25	0	QUICKLOGIX, INC.	Chief Executives	Ceo	1	187200.0	2016.0	California	-121.955236	37.354108	manager	VERY HIGH

We decide to drop certain columns deemed unnecessary for prediction. These features are:

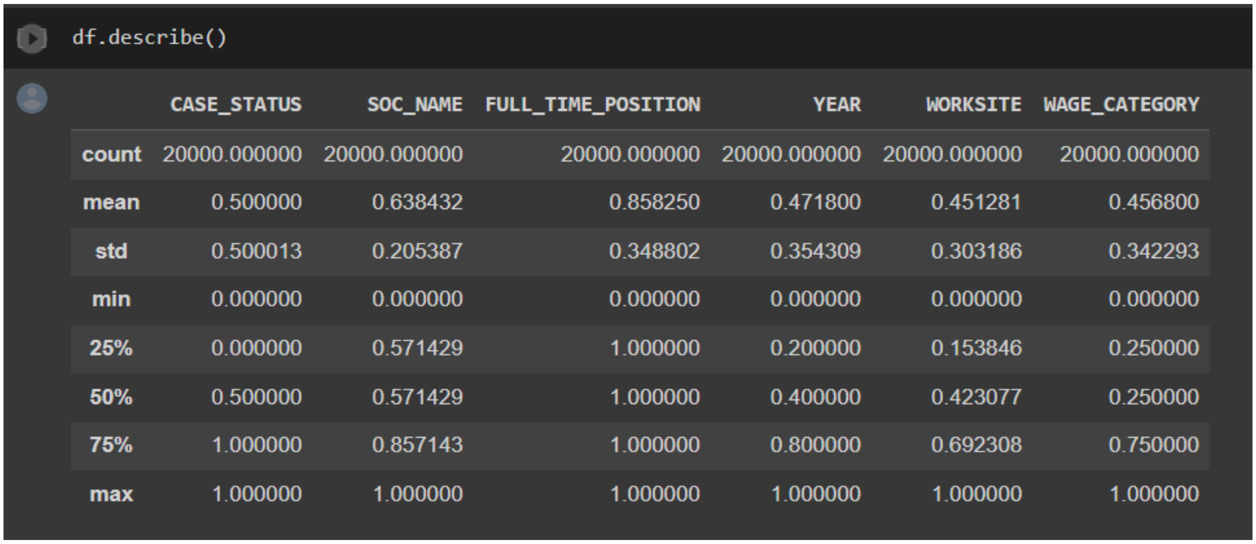
- ➔ SOC_NAME_alt
- ➔ JOB_TITLE

- ➔ EMPLOYER_NAME
- ➔ Lon
- ➔ Lat
- ➔ PREVAILING_WAGE
- ➔ JOB_TITLE

To ensure accurate predictions and to avoid skewing of the predictions due to differences in data ranges we performed **data scaling** using the MinMaxScaler on the following features:

- ➔ SOC_NAME
- ➔ FULL_TIME_POSITION
- ➔ YEAR
- ➔ WORKSITE
- ➔ WAGE_CATEGORY

Scaling means that you're transforming your data so that it fits within a specific scale, like 0-100 or 0-1. You want to scale data when you're using methods based on measures of how far apart data points are, like support vector machines (SVM) or k-nearest neighbors (KNN). With these algorithms, a change of "1" in any numeric feature is given the same importance. By scaling your variables, you can help compare different variables on an equal footing.

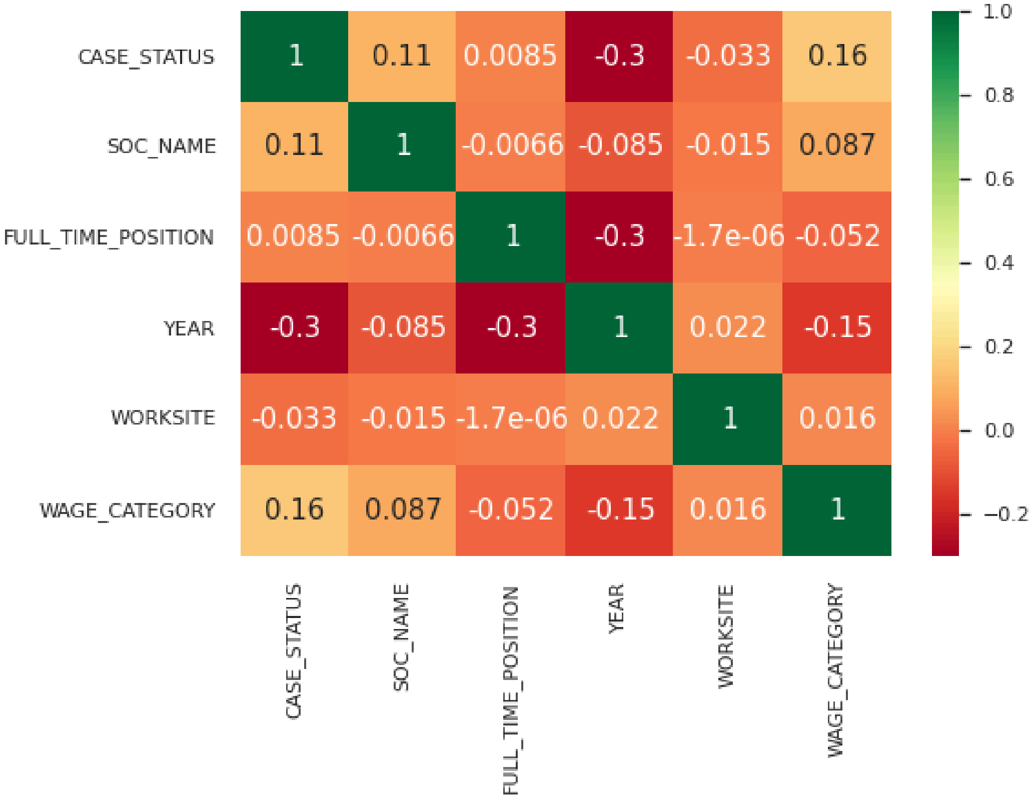


```
df.describe()
```

```
       CASE_STATUS      SOC_NAME  FULL_TIME_POSITION          YEAR      WORKSITE  WAGE_CATEGORY
count  20000.000000  20000.000000        20000.000000  20000.000000  20000.000000   20000.000000
mean       0.500000      0.638432            0.858250      0.471800      0.451281       0.456800
std        0.500013      0.205387            0.348802      0.354309      0.303186       0.342293
min        0.000000      0.000000            0.000000      0.000000      0.000000       0.000000
25%        0.000000      0.571429            1.000000      0.200000      0.153846       0.250000
50%        0.500000      0.571429            1.000000      0.400000      0.423077       0.250000
75%        1.000000      0.857143            1.000000      0.800000      0.692308       0.750000
max        1.000000      1.000000            1.000000      1.000000      1.000000       1.000000
```

Next, we performed feature selection.

Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.

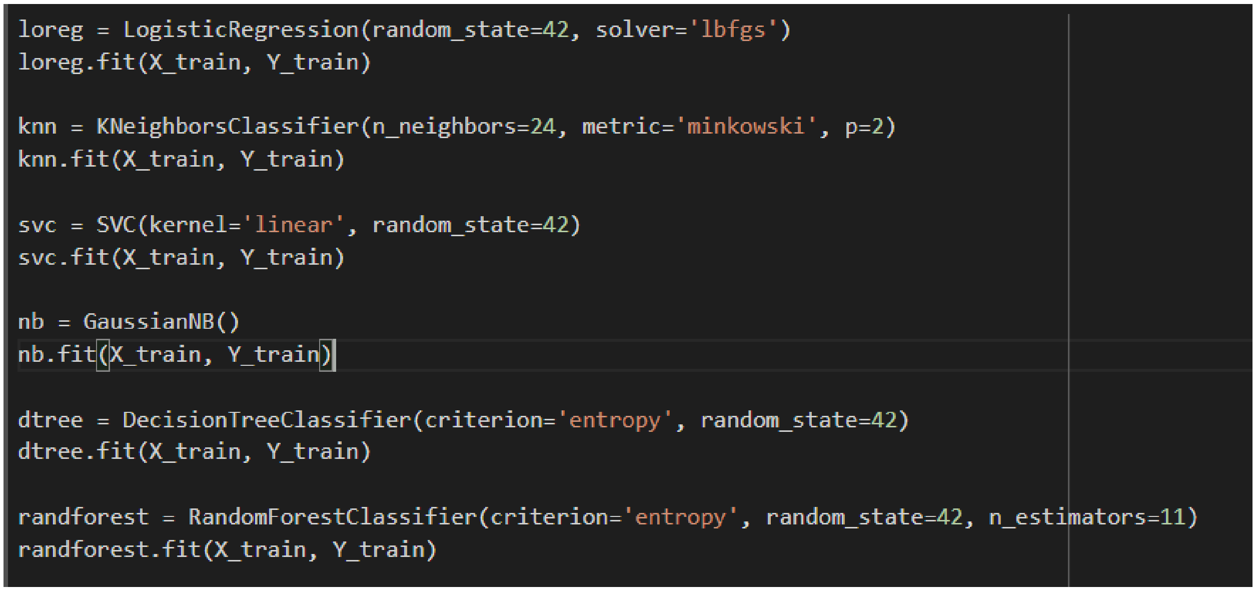


MODEL SELECTION AND TRAINING

Our overall data size included 25 lakh records, we downsampled this data to 30k records and split the dataset into a training set(70%) and a testing set (30%).

The algorithms used to train the model are:

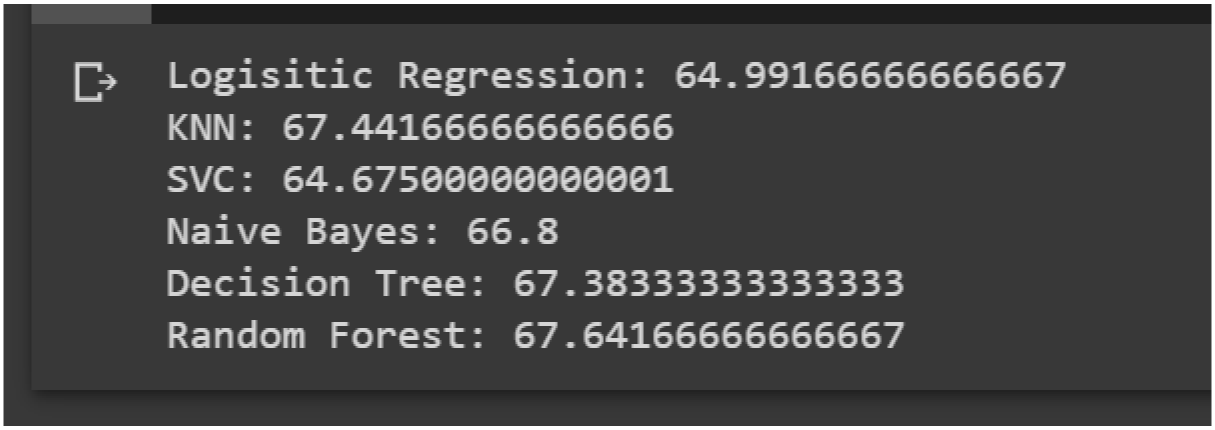
- Logistic Regression
- KNN
- SVC
- Naive Bayes
- Decision Tree
- Random Forest



```
loreg = LogisticRegression(random_state=42, solver='lbfgs')
loreg.fit(X_train, Y_train)

knn = KNeighborsClassifier(n_neighbors=24, metric='minkowski', p=2)
knn.fit(X_train, Y_train)

svc = SVC(kernel='linear', random_state=42)
svc.fit(X_train, Y_train)

nb = GaussianNB()
nb.fit(X_train, Y_train)

dtree = DecisionTreeClassifier(criterion='entropy', random_state=42)
dtree.fit(X_train, Y_train)

randforest = RandomForestClassifier(criterion='entropy', random_state=42, n_estimators=11)
randforest.fit(X_train, Y_train)
```

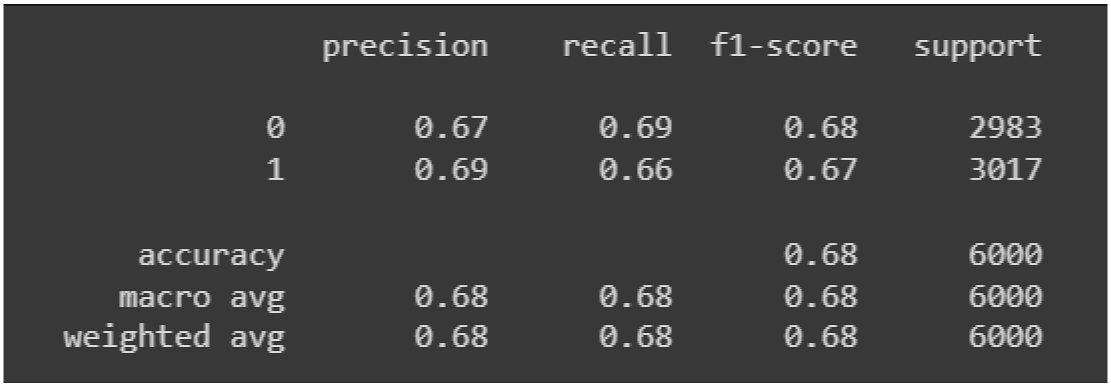
RESULTS AND OUTPUT:

We then ran the models on the test dataset and the observed accuracies were:



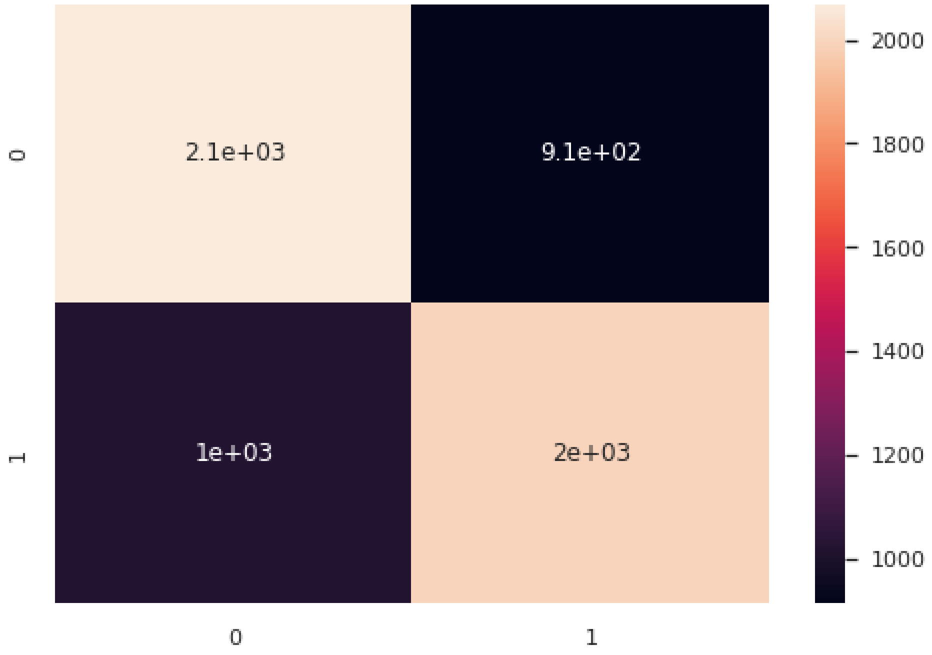
```
Logisitic Regression: 64.99166666666667
KNN: 67.44166666666666
SVC: 64.67500000000001
Naive Bayes: 66.8
Decision Tree: 67.38333333333333
Random Forest: 67.64166666666667
```

We also looked at other metrics to measure performance,



```
              precision    recall  f1-score   support

           0       0.67      0.69      0.68      2983
           1       0.69      0.66      0.67      3017

    accuracy                           0.68      6000
   macro avg       0.68      0.68      0.68      6000
weighted avg       0.68      0.68      0.68      6000
```

The confusion matrix generated:



As seen, the decision tree model had the best prediction out of all the models used.

CONCLUSION:

By means of this project, we made the following conclusions:

- It was observed that most of the applications for visas are of the H-1B type, and most of the jobs are in the computer industry.
- We performed data preprocessing, data cleaning, exploratory data analysis, feature selection, feature scaling on the dataset before training the prediction models.
- We used multiple models to predict the approval of the H1B Visa Application namely, Logistic Regression, KNN, SVC, Bayes, Decision Tree, and Random Forest model.
- The decision tree model performed the best out of all the models used with the accuracy of 80.26%.

IMPROVEMENTS:

- Logistic Regression
  - Grid Search: 65.0%
- SVC
  - Grid Search: 65.9%
- Decision Tree
  - Grid Search: 68.5%
  - Bagging: 67.9%
  - Boosting: 65.5%
- Random Forest
  - Hyperparameter Tuning: 69.3%
  - Stratified K-fold cross validation: 70.4%
- Ensemble
  - Voting Classifier: 66.3%
  - Stacking: 66.9%

REFERENCES:

Sklearn documentation: https://scikit-learn.org/stable/
Pandas documentation: https://pandas.pydata.org/docs/
OFLC Data Center: https://www.flcdatacenter.com/