**Predicting filed H1-B Visa Petitions’ Status**

Department of Computer Science Engineering

Geethanjali College of Engineering and Technology



Team Members

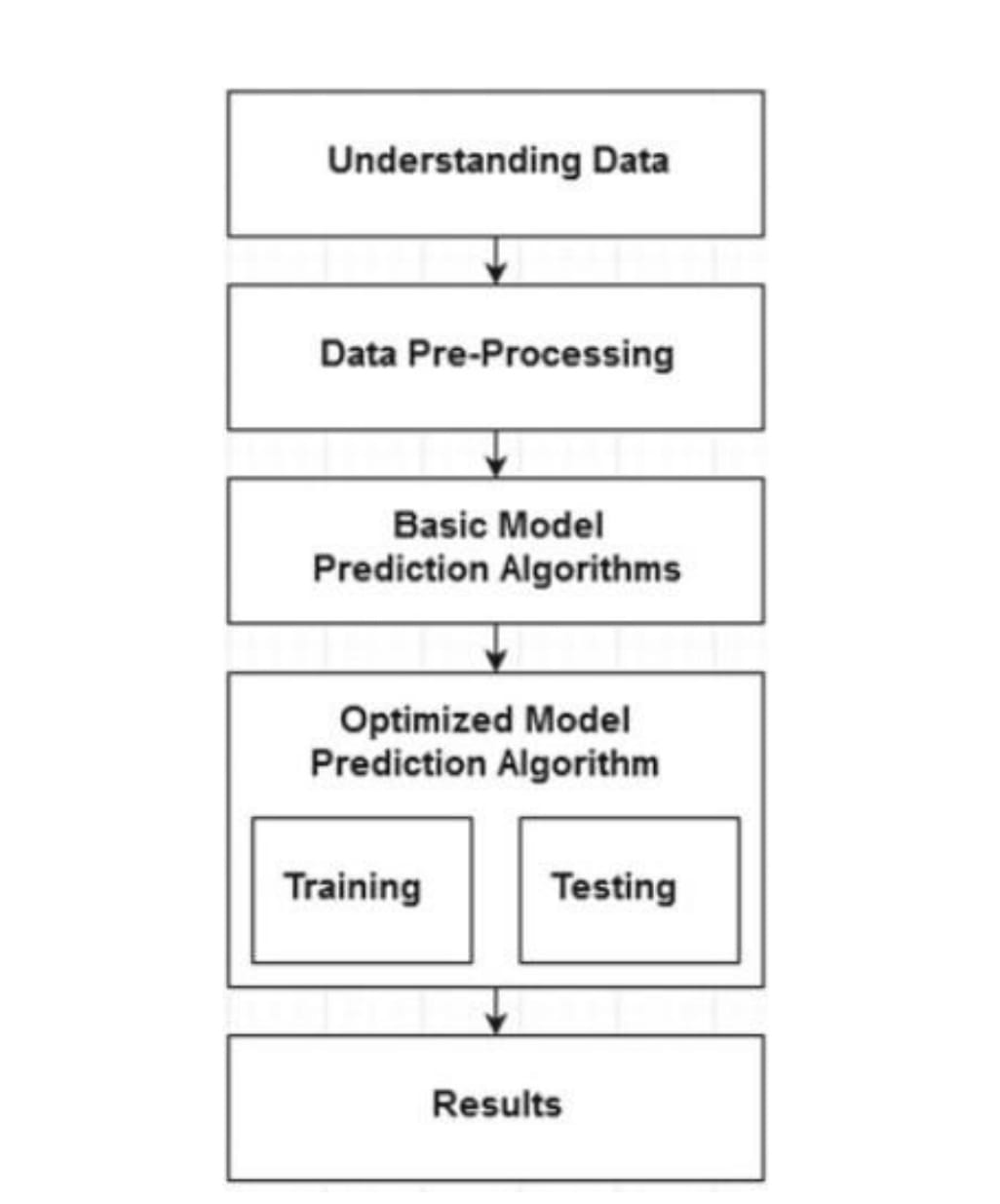
* S Pavan Sushanth (19R11A05D6)
* Sarala Vineeth (19R11A05D5)
* Samyuktha Alfred (19R11A05D4)
* Prajna S (19R11A05D2)
* Junaid AHEMED (19R11A05D1)

ABSTRACT

* The H1-B visa in United States allows employers to employ foreign workers in specialty occupations.
* The report addresses the approach to predict the case status of the filed H1-B Visa petitions using various input data such as employer name, job category, job title, location of job, filing year, and prevailing wage.
* The trained models help to relate the decision with the attributes of the application.
* ***Keywords***— **H1B, status, classification, evaluation, analysis**

***INTRODUCTION***

* The US H1-B visa is a non-immigrant visa that allows US companies to employ graduate level workers in specialty occupations that require theoretical or technical expertise in specialized fields such as IT, finance, accounting, architecture, engineering, mathematics, science, medicine, etc.
* This is one of the highly used visa categories, and companies that usually require foreign talent rely on it to a great extent.
* This report shows an efficient approach to solve the problem of foreseeing the decision before filing or after filing, and before receiving the decision on the filed petition.

FLOW CHART

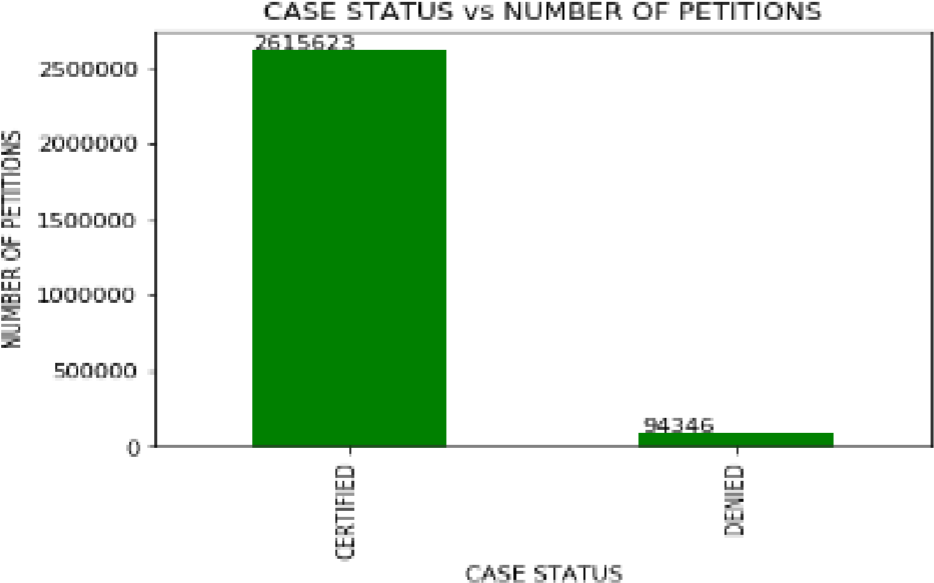
**PROBLEM FORMULATION**

The report tries to show the dependency of the decision on the attributes of the application.

* The attributes of the application here serve as an input and the output is the predicted decision.
  + **Name of the employer**: Name of employer submitting labor condition application.
  + **Category of the job or SOC Name**: Standard Occupational Classification system defines the codes and the names associated with them.
  + **Job title**: The requested job title in the petition.
  + **Employment Type**: Full time employment (Y) or a part-time employment (N).
  + **Year of Filing**: Year when petition is filed (in between 2011-2016).
  + **Prevailing wage**: Prevailing Wage for the job being requested for temporary labor condition.

SYSTEM DESIGN

***1.Data Cleaning and Filtering:*** The raw data considered for the problem statement needs some pre-processing and cleaning depending on its attributes' value. Firstly, all the records with NULL value in either of the attributes were pruned as they can't be handled with any random values. The problem discusses about the decision prediction as either "CERTIFIED" or "DENIED", so all the records containing value as "CERTIFIED" or "DENIED" in their case status were considered for further steps.

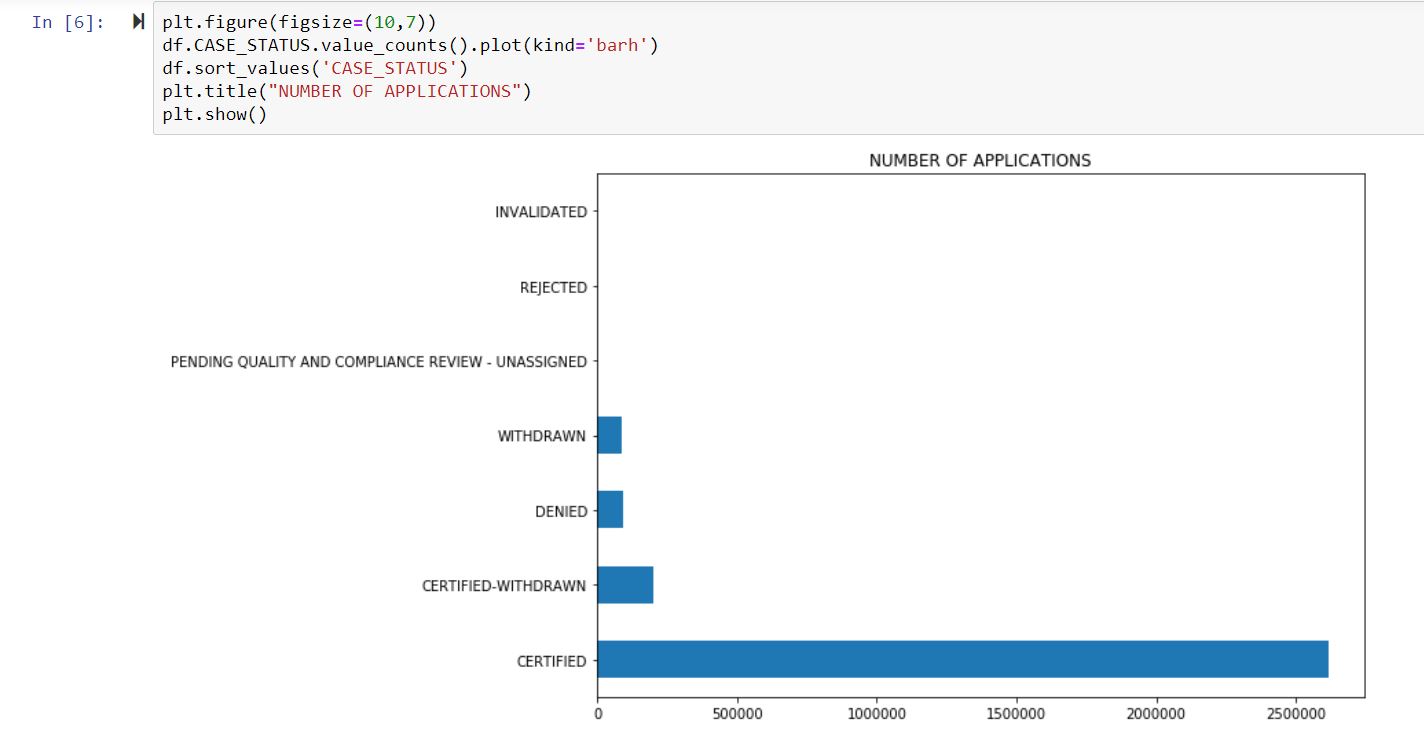


DISTRIBUTION OF STATUS LABELS

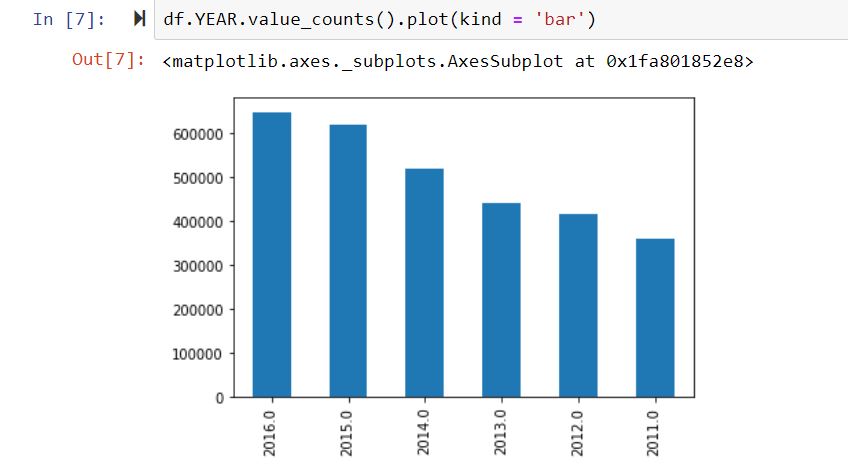
|  |  |
| --- | --- |
| **CERTIFIED** | 2615623 |
| **CERTIFIED-WITHDRAWN** | 202659 |
| **DENIED** | 94346 |
| **WITHDRAWN** | 89799 |
| **PENDING QUALITY AND COMPLIANCE REVIEW - UNASSIGNED** | 15 |
| **REJECTED** | 2 |
| DATA DOWN SAMPLING   * The data is highly imbalanced as the samples with certified status are far more than with the ones with denied status. Hence, data balancing becomes the most important step. * We down sampled the certified samples to the count of denied samples, to match with the sample numbers.   The distribution turns out to be as below: |
|  |
| *DATA CONSISTENCY THROUGH REFERENCING:*   * + *Considering the down sampled data, in SOC names, the data has values like “Computer Systems Analysts” and “Computer Systems Analysts” and many more, which are actually one and the same.*   + *Hence, we took the reference of the job categories as an external file input and the job categories in data were mapped and refined using the ratio threshold of string match as 0.94.*   + *The threshold was derived after calculating the mean of the string match ratio of various random almost similar strings.* |

DATA CONVERSION AND FEATURES EXTRACTION:

* We decided to use the one-hot encoding approach to convert the categorical values to numerical.
* Since there are chances that after applying hash function on the values of the attributes, they might be converted into the same values.
  + *Data Analysis*: Looking at the unique values of the attributes in the below table, it is difficult to convert them to numerical values using one-hot encoding. Hence, we decided to partition the values into categories according to their values.
  + Case Status: As it has only 2 unique values, it will be easy to apply one-hot encoding



* One-Hot Encoding: Case status, employer acceptance category, SOC acceptance category, job title acceptance category, filing year, wage category, employment type, worksite are passed as the column names in the one-hot encoding function and after application, in total 73 columns are created.

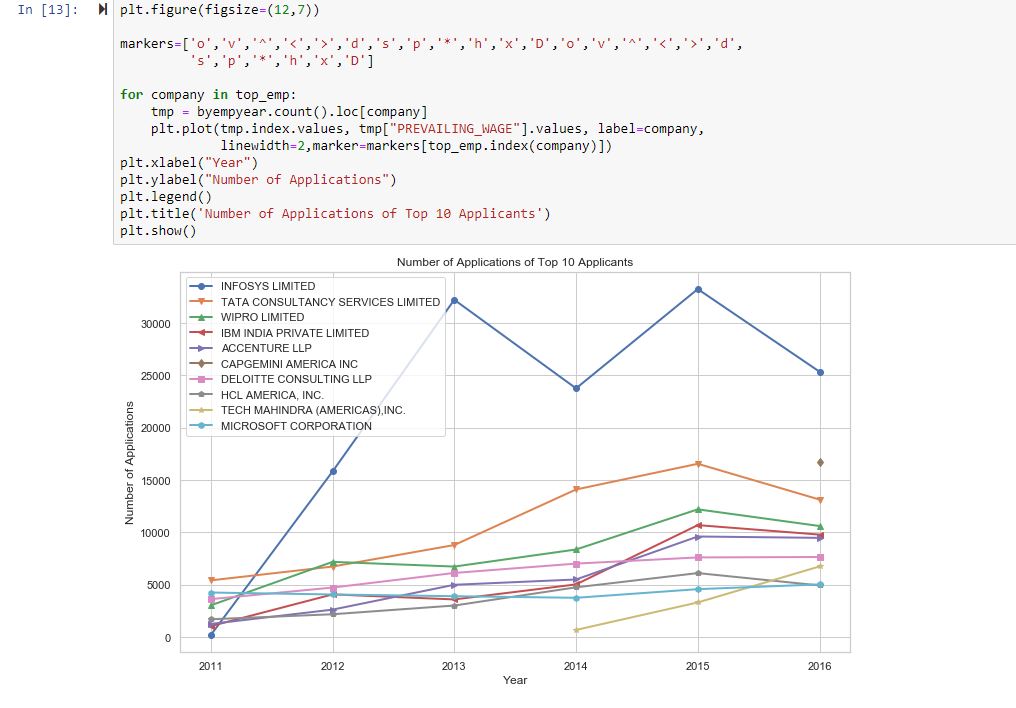


FEATURE SELECTION

* Relative effect of the attributes' values on the final decision making, a feature elimination model called Recursive Feature Elimination(RFE) was employed. In RFE, weights are assigned to the features and then least important features are eliminated recursively from the current set of features until the desired number of features are eventually reached.
* Considering the output of RFE, attributes employer acceptance level, job acceptance level, wage category, worksite state and filing year. Top attributes selected are: {'EMPLOYER', 'FILING', 'FULL', 'JOB', 'SOC', 'WAGE', 'WORKSITE'}
* Hence, employer acceptance category, filing year, job title acceptance category, wage category, work state, job type, soc acceptance category is considered further.

EXPERIMENTAL EVALUATION

* Data Splitting :In classification, we divide the data into two sets viz. 1. Training Data 2. Test Data. We considered 20 percent of the total records for test data and the rest of 80 percent for training the classifiers.
* Models Applied : We trained total 3 different classifiers for predication task and evaluated each based on the evaluation metrics.
* **k-Nearest Neighbour(k-NN)**
* **Logistic Regression**
* **Random Forest**



MODEL EVALUATION METRICS

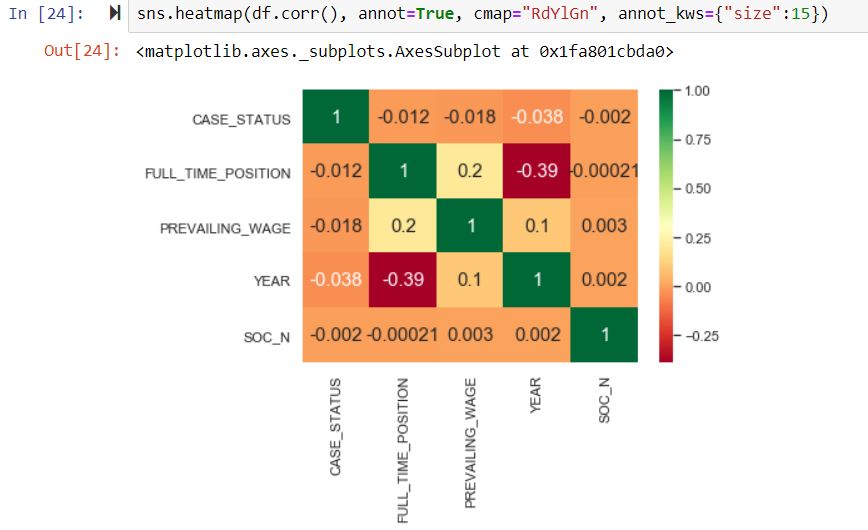
* All the classifier models were evaluated using three types of metrices.
  + **Precision Score**: Precision Score is calculated as the ratio of correctly predicted positive observations to the total predicted positive observations.

Precision = TP/ TP + FP

* + **Recall Score**: Recall is calculated as the ratio of correctly predicted positive observations to the all observations in actual class – yes.

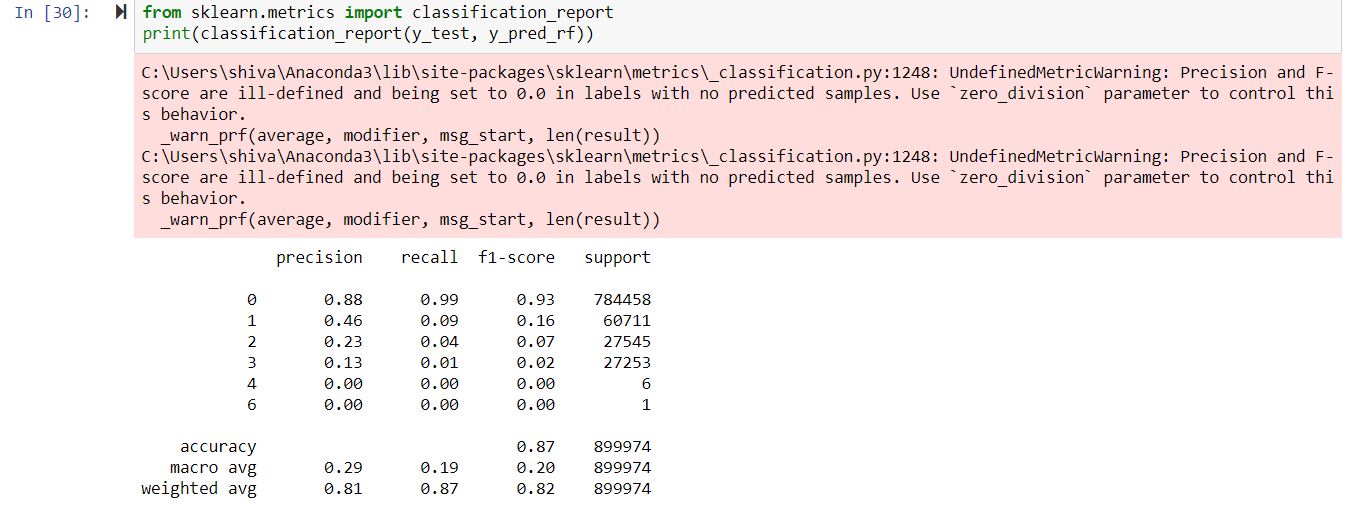
Recall = TP/TP+FN

* + **F1 Score**: F1-Score is calculated as a weighted average of both precision and recall score.
* F1 Score=2\*(Recall\*Precision)/ (Recall +
* Precision)
* Here,
* FP = False Positive: Predicted Yes when Actual is No
* FN = False Negative: Predicted No when Actual is Yes
* TP = True Positive: Predicted Yes when Actual Yes
* TN = True Negative: Predicted No when Actual No

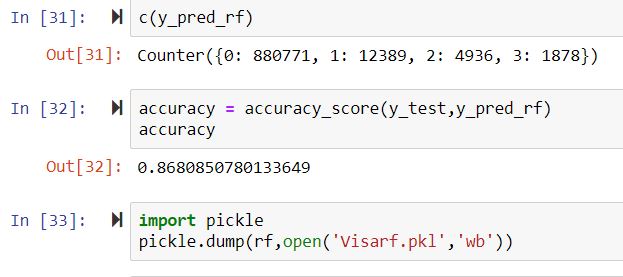


MODEL EVALUATION RESULTS

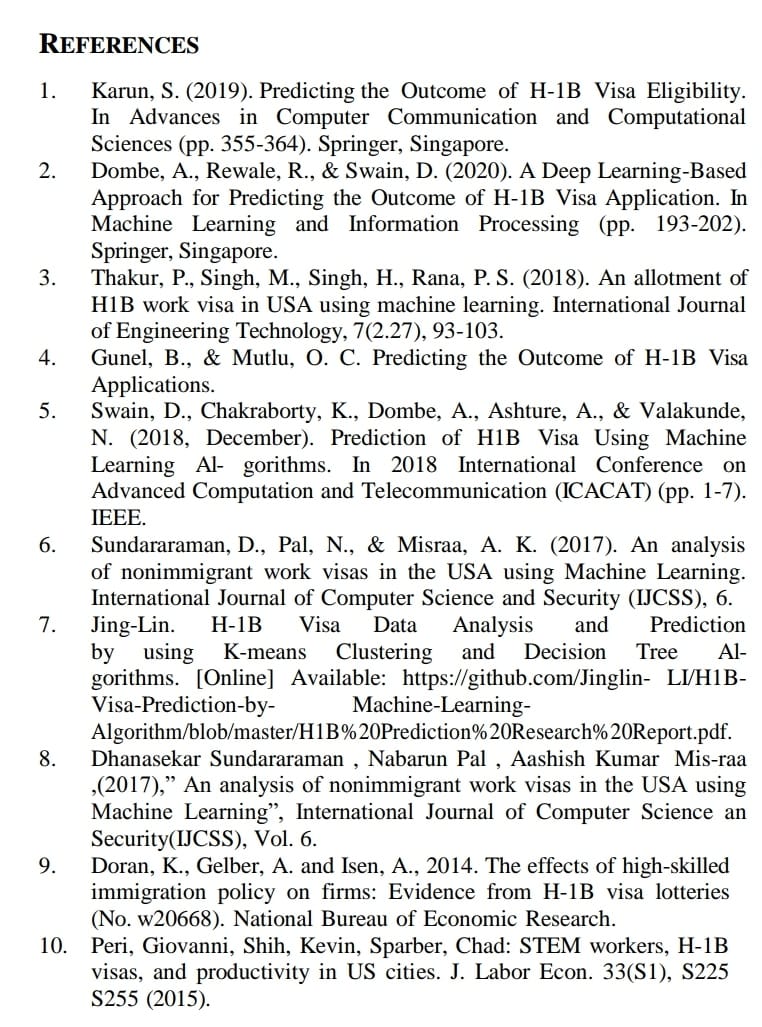
* Compared to all approaches presented till date, we have achieved the highest accuracy of prediction task through application of Artificial Neural Network and Logistic Regression along with Random Forest.
* Hence we chose to predict with Random forest.
* Data balancing played a major role in achieving the high accuracy for both the classes. We also infer that job title, employer name/employer acceptance ratio, wages, worksite and filing year play an important role in inferring the value of the case status.



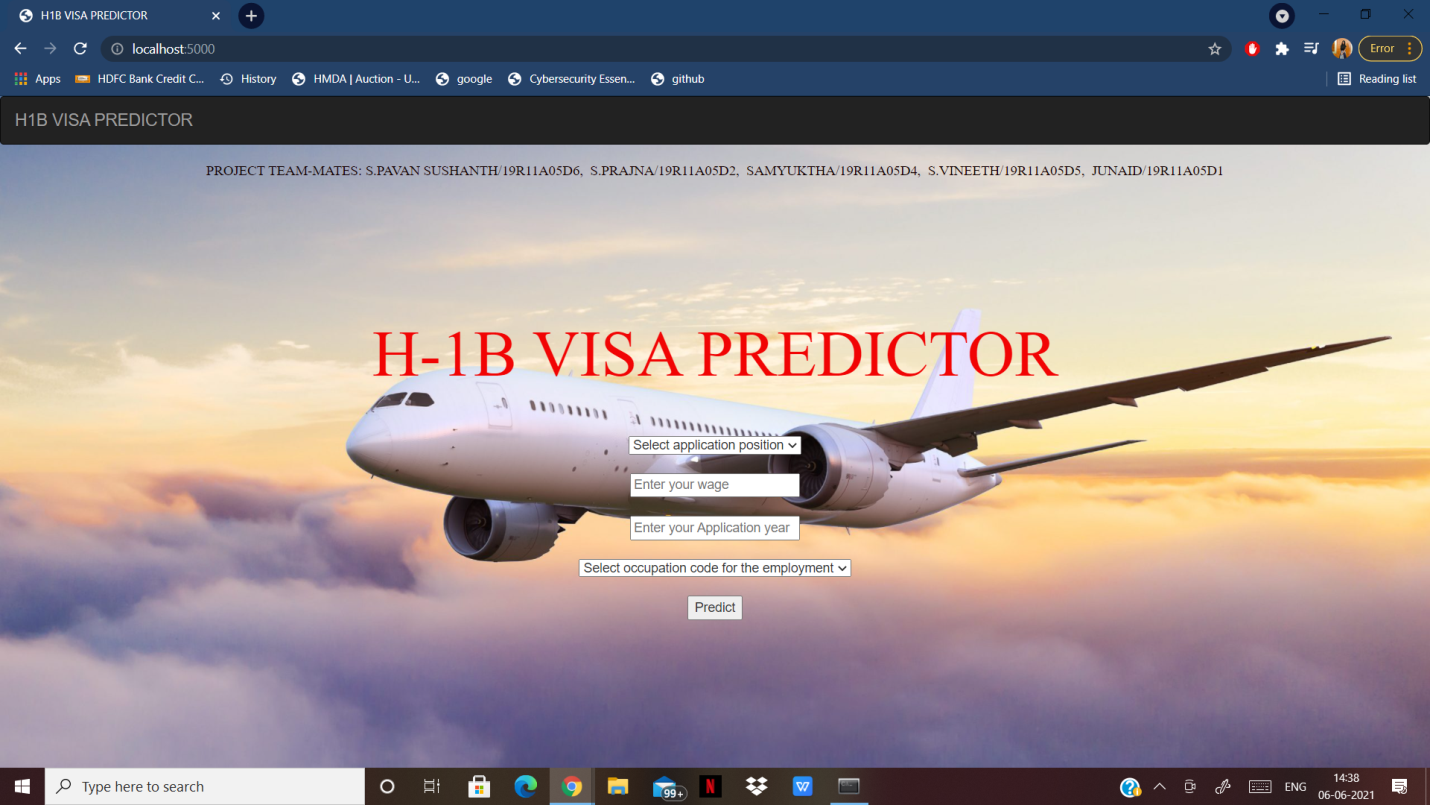
Final Prediction with accuracy



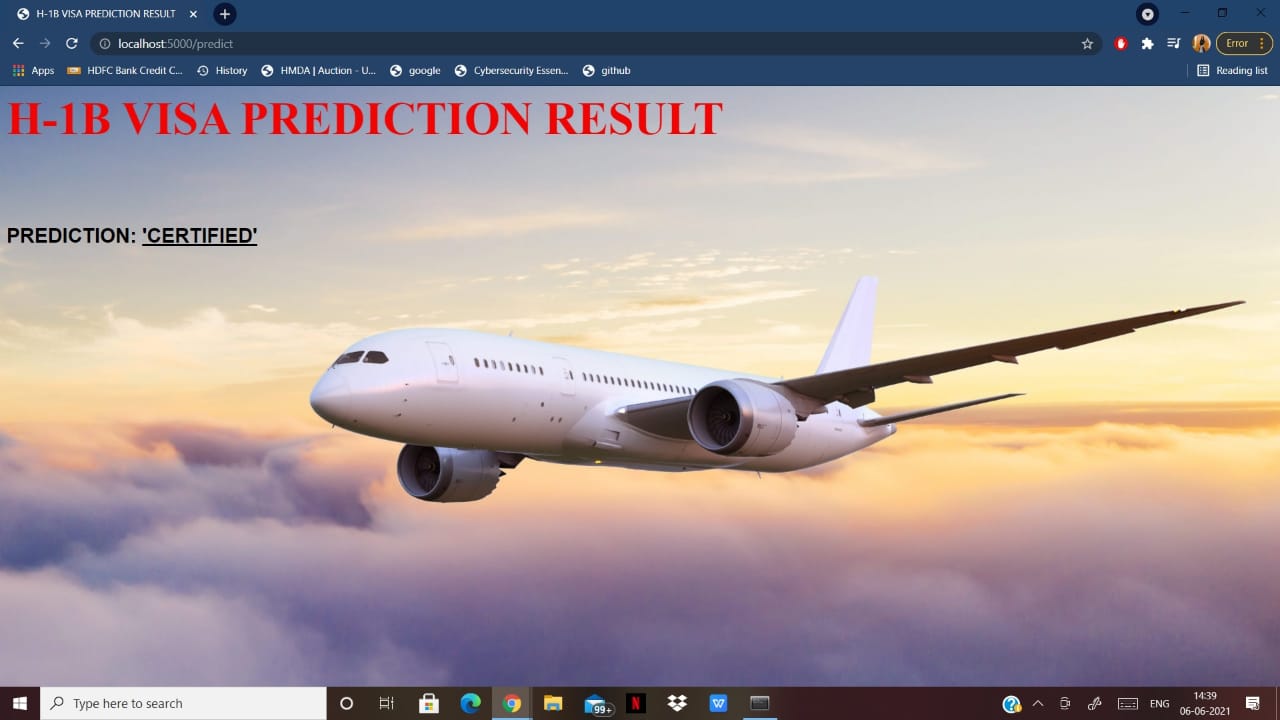
BIBILOGRAPHY



WEBSITE PAGE



PREDICTION PAGE



**LEARNING EXPERIENCES**

* As a team, we learnt that
  + Data balancing is an important step when the data is imbalanced else the results will be inclined towards the class with larger number of records
  + Categorizing feature values helps to draw meaning out of the categorical variables
  + Artificial Neural Network’s performance reduces if the number of hidden layers are comparatively high and the data is not complex. Because, it converges easily and then the rest of layers adds unnecessary overhead.
  + Thus concluding….as this has taught us that data managing makes it easier for us to predict and analyze any future decisions.

THANK YOU

CSE-C

TEAM-04