2022

IOMP Project Report on

PREDICTIVE MODELING FOR H1B VISA APPROVAL USING IBM WATSON

SUBMITTED BY

PRATHYUSHA B	**19UK1A0579**
UDAY KIRAN REDDY V	**19UK1A0578**
MOHAMMED TANVEER ABDUL BARI	**19UK1A0571**
POOJITHA U	**19UK1A05D2**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VAAGDEVI ENGINEERING COLLEGE

(Affiliated to JNTUH, Hyderabad)

Bollikunta, Warangal _ 506005

2019_ 2023

CONTENTS

CHAPTERS **PAGE NO'S**

CHAPTER 1

INTRODUCTION

OVERVIEW

The H-1B is an employment-based visa in the United States, which allows U.S. employers to temporarily employ foreign workers in specialty occupations. To apply for H-1B visa, an U.S employer must offer an job and petition for H-1B visa with the U.S. immigration department. This is the most common and legal visa status and for international students who complete their college / higher education (Master, PhD) and work in a full-time position. The status of H-1B visa will definitely influence the life and work, and even the career of the international students.

So, this project tries to use algorithm learned in machine learning class, analyze historical H1B data to produce helpful information. Briefly, In this project, we apply machine learning algorithms including Decision Tree, Random forest and Logistic Regression to analyze the conditions (or attributes) of the foreign workers, such as SOC_NAME, WAGE, etc. We utilized the 2011-2016 H-1B petition disclosure data to predict the outcome of H-1B visa applications that are filed by many high-skilled foreign nationals every year. We framed the problem as a classification problem and applied Decision Tree, Random Forest and Logistic Regression in order to output a predicted case status of the application.

In addition, our analysis will also provide some statistic data to answer some questions. Such as: What is the top companies that have apply to the H-1B for employees? What is the trend of total number of H-1B application is? What is the top popular Job Title and Worksites for H-1B Visa holders? What is the salary mean values of respective Job Titles? As H-1B visa is the most common and legal status for the international student, these data might help to guard them to choose the most easier way to work in the United State and accomplish their American Dream.

1.2 PURPOSE

1. The project's goal is to extract the libraries for machine learning for Visa prediction using Python's pandas, matplotlib, and seaborn libraries.
2. Next step is to do an exploratory analysis of the dataset to answer questions like: What are the top companies that have applied to the H-1B for employees? What is the trend of the total number of H-1B applications? What is the top popular Job Title and Worksites for H-1B Visa holders?
3. Third step is to deploy a web application that predicts visa status based on the best performing machine learning algorithms. This feature will help employees to get a realtime prediction based on previous years data.

CHAPTER 2

LITERATURE SURVEY

2.1 EXISTING PROBLEM

The practise of evaluating data from many viewpoints and extracting meaningful knowledge from it is known as data mining. It is at the heart of the process of knowledge discovery. Classification, clustering, association rule mining, prediction and sequential patterns, neural networks, regression, and other data mining techniques are examples. The most widely used data mining technique is classification, which uses a group of pre-classified samples to create a model that can categorise the entire population of information. The categorization technique is particularly well suited to fraud detection and credit risk applications. This method often employs a classification algorithm based on decision trees. A training set is used to develop a model as a classifier that can categorise data objects into their respective classes in classification. The model is validated using a test set.

2.2 PURPOSED SOLUTION

Our model and analysis will provide a whole picture of the different approval rates by comparing different conditions based on previous data. In addition, our analysis will also provide some statistic data to visualize the characteristics of the application case and trends. In order to predict the status, we will be training the model with occupation category, prevailing wage, Year of application and Job duration after removing the outliers and applying label encoding to all the categorical data. For analysis part we'll be plotting different graphs to get a relevant inference and eye appealing layout.

The method or solution is Jupiter notebook and spyder we used to complete this project. and you will use this jupiter notebook for you recommended.

To build Machine learning models you must require the following packages

Sklearn: Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms.

NumPy: NumPy is a Python package that stands for 'Numerical Python'. It is the core library for scientific computing, which contains a powerful n-dimensional array object

Pandas: pandas is a fast, powerful, flexible, and easy to use open source data analysis and manipulation tool,built on top of the Python programming language.

Matplotlib: It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits

CHAPTER 3

THEORETICAL ANALYSIS

While selecting the algorithm that gives an accurate prediction, we gone through lot of algorithms like Decision tree, Random Forest etc., which gives the results abruptly accurate and from them we selected only one algorithm for the prediction problem that is Random Forest (because it gave a better accuracy). The peculiarity of this problem is collecting the customers details real time and working with the prediction at the same time, so we developed a user interface for the people who'll be accessing for the Visa status prediction.

3.1 BLOCK DIAGRAM

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model's prediction

The fundamental concept behind random forest is a simple but powerful one: A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don't constantly all err in the same direction).

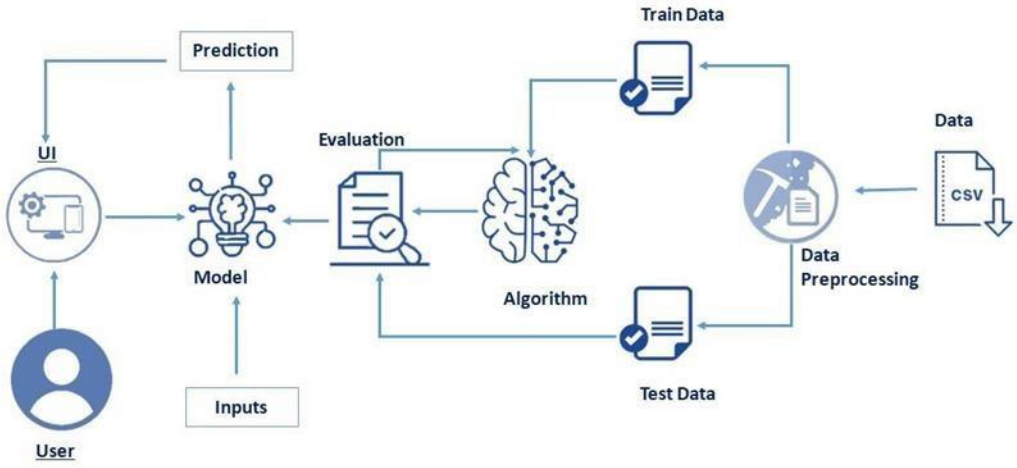


Fig 3.1. Architecture

3.2 HARDWARE / SOFTWARE DESIGNING

The hardware required for the development of this project is:

Processor	: Intel CoreTM i5-9300H
Processor speed	: 2.4GHz
RAM Size	: 8 GB DDR
System Type	: X64-based processor

SOFTWARE DESIGNING:

The software required for the development of this project is:

Desktop GUI	: Anaconda Navigator
Operating system	: Windows 11
Front end	: HTML, CSS
Programming	: PYTHON

CHAPTER 4

EXPERIMENTAL INVESTIGATION

IMPORTING AND READING THE DATASET

Importing the Libraries

First step is usually importing the libraries that will be needed in the program.

Pandas: It is a python library mainly used for data manipulation.

NumPy: This python library is used for numerical analysis.

Matplotlib and Seaborn: Both are the data visualization library used for plotting graph which will help us for understanding the data.

csr_matrix() :A dense matrix stored in a NumPy array can be converted into a sparse matrix using the CSR representation by calling the csr_matrix() function.

Train_test_split: used for splitting data arrays into training data and for testing data.

Pickle: to serialize your machine learning algorithms and save the serialized format to a file.

Reading the Dataset

For this project, we make use of data set ’H-1B Visa Petitions 2011-2016 dataset’We will be selecting the important features from the dataset that will help us in predicting the h1b visa approvalThe next step is to read the dataset into a data structure that’s compatible with pandas. Let’s load a .csv data file into pandas. There is a function for it, called **read_csv().**

We will need to locate the directory of the CSV file at first (it’s more efficient to keep the dataset in the same directory as your program).If the dataset in same directory of your program, you can directly read it, without any path. After the next Steps we made following bellow:

1.Data visualization
2.Collabrative and filtering
3.Creating the Model
4.Test and save the model
5.Buil Python Code
6.Build HTML Code
7.Run the Application
We are the following above sections we did and investigate it.

CHAPTER 5

FLOWCHART

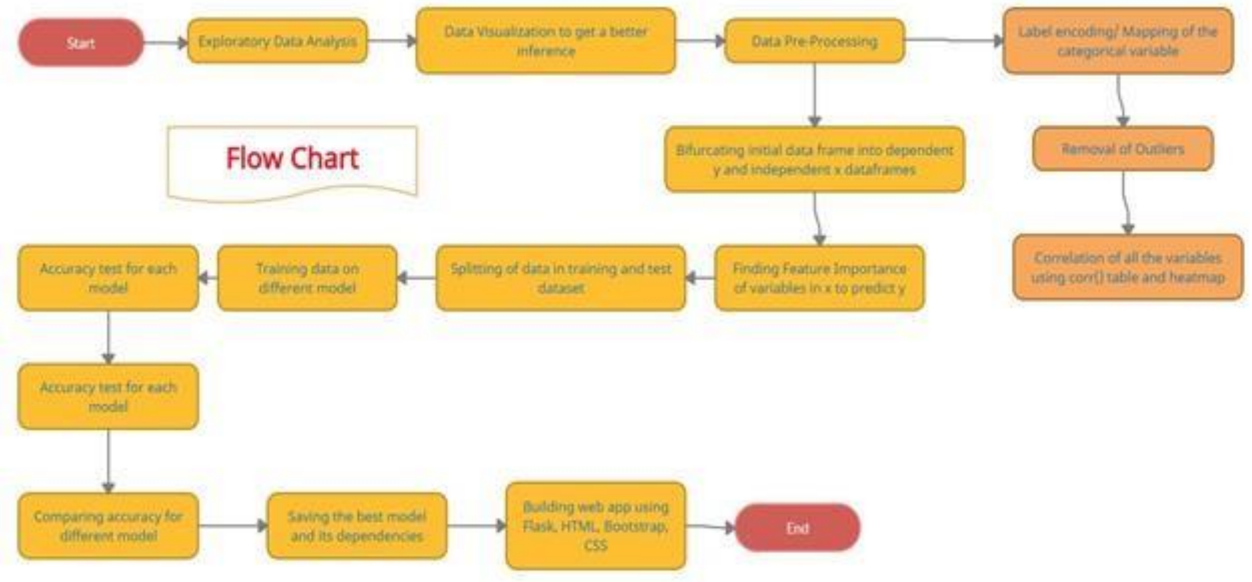

Start
Exploratory Data Analysis
Data Visualization to get a better inference
Data Pre-Processing
Label encoding/ Mapping of the categorical variable
Flow Chart
Bifurcating initial data frame into dependent y and independent x dataframes
Removal of Outliers
Correlation of all the variables using corr() table and heatmap
Accuracy test for each model
Training data on different model
Splitting of data in training and test dataset
Finding Feature Importance of variables in x to predict y
Accuracy test for each model
Comparing accuracy for different model
Saving the best model and its dependencies
Building web app using Flask, HTML, Bootstrap, CSS
End

Fig 5.1 Flowchart of the project

Project Flow:

- User interacts with the UI (User Interface) to upload the input features.
- Uploaded features/input is analyzed by the model which is integrated
- Once model analyses the uploaded inputs, the prediction is showcased on the UI.

1. Data Collection.

- ML depends heavily on data, without data, a machine can't learn. It is the most crucial aspect that makes algorithm training possible. In Machine Learning projects, we need a training data set. It is the actual data set used to train the model for performing various actions.
- You can collect datasets from different open sources like kaggle.com, data.gov; UCI machine learning repository etc. The dataset used for this project was obtained from Kaggle

2. Data Pre- processing.

Data Pre-processing includes the following main tasks

- Import the Libraries.
- Importing the dataset.
- Exploratory Data Analysis
- Data Visualization

3. Collaborating Filtering

- Merging datasets
- Creating the Model
- Predicting the results
- Saving our model and dataset

4. Application Building

After the model is built ,we will be integrating it to a web application so that normal users can also use it to know if any website is [hishing or safe in a no-code manner.

In the application ,the user provides any website URL to check and the corresponding parameter values are generated by analysing the URL using which legitimate websites are detected.

- Create an HTML file.
- Build a Python Code.
- Execute and test your model.
- The final output.

CHAPTER 6

RESULT

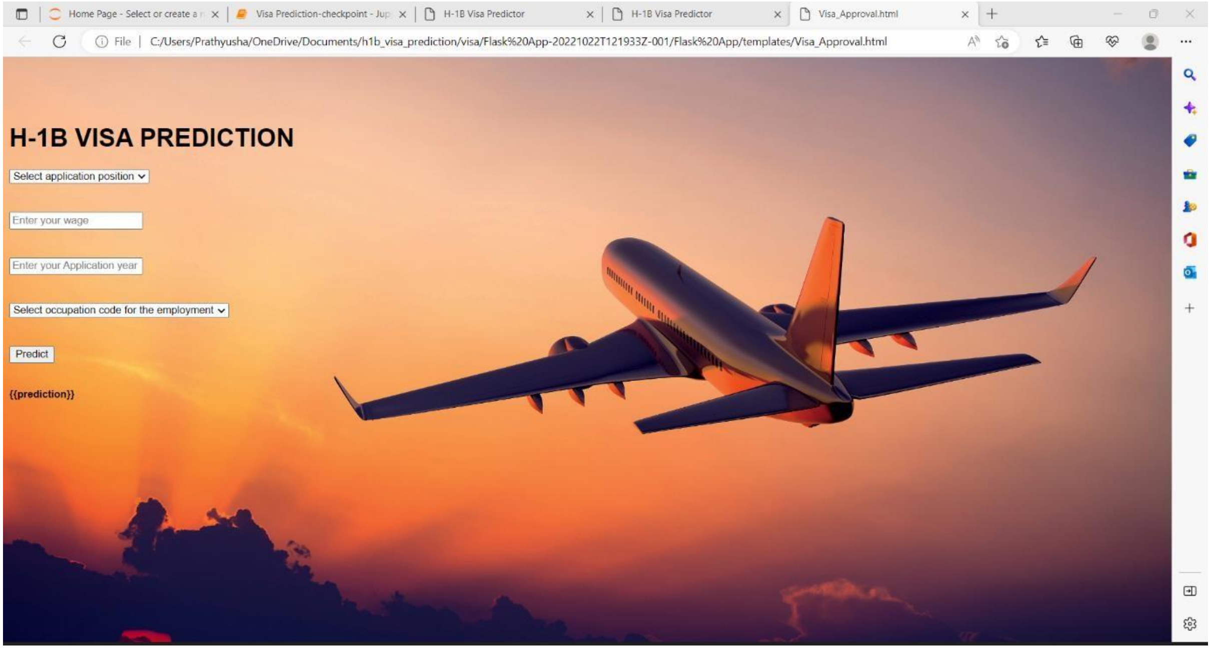
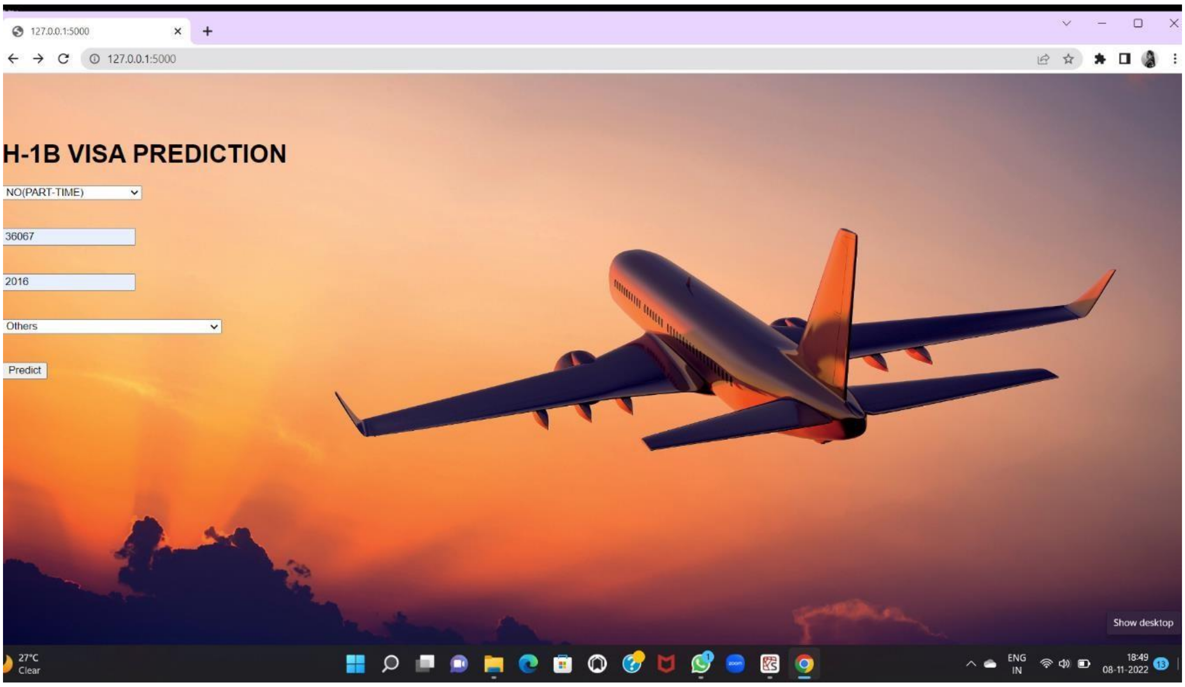

Home Page - Select or create a
Visa Prediction-checkpoint - Jup
H-1B Visa Predictor
H-1B Visa Predictor
Visa_Approval.html
File
C:/Users/Prathyusha/OneDrive/Documents/h1b_visa_prediction/visa/Flask%20App-20221022T121933Z-001/Flask%20App/templates/Visa_Approval.html
H-1B VISA PREDICTION
Select application position
Enter your wage
Enter your Application year
Select occupation code for the employment
Predict
{{prediction}}

Fig 6.1.Home Page of Predictive Modeling for H1B visa Approval



H-1B VISA PREDICTION

NO(PART-TIME)

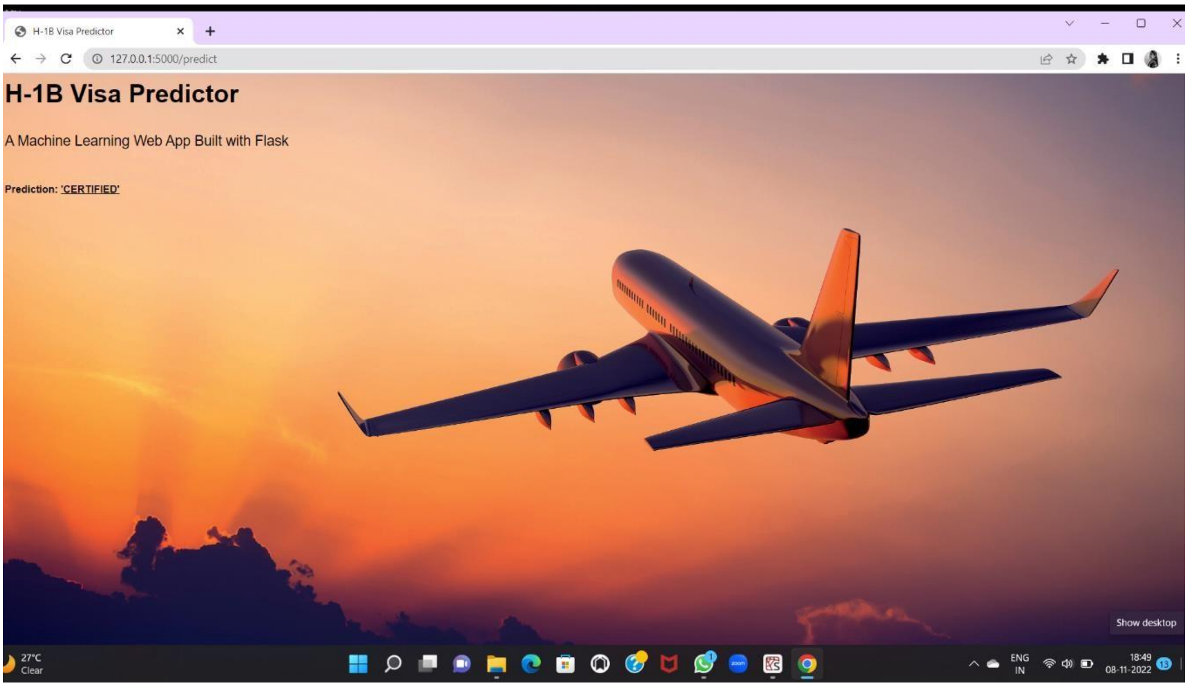
36067

2016

Others

Predict

Fig 6.2.Input Page of Predictive Modeling for H1B visa Approval



H-1B Visa Predictor

A Machine Learning Web App Built with Flask

Prediction: 'CERTIFIED'

Fig 6.3.output Page of Predictive Modeling for H1B visa Approval

CHAPTER 7 ADVANTAGES AND DISADVANTAGES

ADVANTAGES

- The H1 B visa is the most demanded visa world-wide.
- The H1 B visa applications are very heavily varied across many fields i.e. job, job title, year of petition, accountable wages, city of work etc. The purpose of this research is to estimate the likelihood of visa approval on the basis of metadata provided. We shall consider all aspects by which the petition may be approved or otherwise, strictly working on the data provided in the application.
- The designed classifier in the after mentioned report serves a dual purpose of H1B applicants and hopeful employers to measure the probability of getting certified prior to and after applying the petition.

DISADVANTAGES

Lack of Data. Perhaps the biggest issue facing predicting systems is that they need a lot of data to effectively make predictions. ...

- Changing Data. ...
- Changing User Preferences. ...
- May be get a wrong prediction..

CHAPTER 8

APPLICATIONS

- Can be applied in each and every individual’s Daily Life.
- The web app made by using ML models could be used by students/employee to check whether their application will be certified or not based on previous years data..
- To apply for H-1B visa, an U.S employer must offer an job and petition for H-1B visa with the U.S. immigration department. This is the most common and legal visa status and for international students who complete their college / higher education (Master, PhD) and work in a full-time position. The status of H-1B visa will definitely influence the life and work, and even the career of the international students. So, this project tries to use algorithm learned in machine learning class, analyze historical H1B data to produce helpful information.

CHAPTER 9

CONCLUSION AND FUTURESCOPE

CONCLUSION

Over the beyond decade, the selection for H greater each year, so the Scope of this challenge is to gather a tool on the way to deliver a desire to every character who're suffering for H1B visa magnificence method and predicting the recognition of software with top accuracy Supplemental information regarding the Standard Occupational Classification (SOC) may be accumulated and implemented in coordination with of the way the H-1B Visa preference manner works. By the use of the income opinions and levels beneath SOC, the income feature on this information set may be correctly mounted to a range of salaries which could then be used to categorize the visa petitions primarily based totally totally on career roles in desire to region In addition, wonderful magnificence algorithms other than the discriminative fashions may be experimented with this examined and their performances moreover can be The Random Forest classifier works right proper right here with the greater accuracy in evaluation to all of the wonderful algorithms which might be gift to carry out the assessment operations Here we had been given an accuracy of 87% even as the information is boosted and informed with the beneficial aid of the use of the Random Forest Algorithm and as our effects this set of suggestions is the awesome in form for the Prediction of H1B visa approval.

With the growing fashion of candidates of H1B visa, it has emerge as obligatory to increase a tool to are looking for the approval of H1B visa accurately. Therefore, with the beneficial aid of the use of the use of several gadget learning magnificence algorithms we're capable of are looking for the H1B visa approval reputation. This can be very beneficial for overseas employees traveling to the United States.

FUTURESCOPE

Further Random forest can be applied on other data sets available for visa approvals to further investigate its accuracy. Other machine learning algorithms can also be implemented in the project like Naïve bayes model or the SVM model. In further study, we will try to conduct experiments on larger data sets or try to tune the model so as to achieve the state -of-art performance of the model and a great UI support system making it complete web application model. The project can also probe deeper in the process of predicting visa for an individual by including Job title, location and also through categorisation of the individuals..

CHAPTER 10

BIBILOGRAPHY

1. "H-1B Fiscal Year (FY) 2018 Cap Season," USCIS. [Online]. Available: https://www.uscis.gov/going for walksunited-states/temporary-employees/h-1bspecialty-occupationsandfashion-fashions/h-1b-fiscal-365days-fy-2018cap-season.
2. The highly professional visa programs had an immeasurable impact on the relationship," CNNMoney said. [In line]. Available: http://money.cnn.com/2016/04/12/technology/h1b-capvisa-fy-2017/index.html
3. "Use Text Analytics to Predict H1B Salaries", BigML.com Official Blog, October 1, 2013. [Online]. Available: https://blog.bigml.com/2013/10/01/the use of-textual content-analysisistopredicth1-b-wages/.
4. Predicting the Case Status of H1B Visa Applications". [In line]. Available: https://cseweb. ucsd.edu/commands/wi17/cse258a/reports/a054.pdf.
5. Predicting the Case Status of H1B Visa Applications". [In line]. Available: https://cseweb. ucsd.edu/commands/wi17/cse258-a/reports/a054.pdf.

APPENDIX

A Source Code of Flask:

```
 import numpy as np
import pandas as pd
from flask import Flask, request, render_template
import pickle
import os


app = Flask(__name__)
model = pickle.load(open('Visarf.pkl', 'rb'))

@app.route('/')
def home():
    return render_template('Visa_Approval.html')

@app.route('/predict',methods=['POST'])
def predict():
    input_features = [float(x) for x in request.form.values()]
    features_value=[np.array(input_features)]
    features_name=['FULL_TIME_POSITION','PREVAILING_WAGE','YEAR','SOC_N']

    df=pd.DataFrame(features_value,columns=features_name)
    output=model.predict(df)
    output=np.argmax(output)
    print(output)


    return render_template('resultVA.html', prediction_text=output)

if __name__ == '__main__':
  app.run(debug=False)
```