JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JETIR
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

H1B VISA APPROVAL USING MACHINE LEARNING ALGORITHM

Mrs. A. Durga Bhavani[1], Guddeti Bharath[2], Dubbaka Tharun Reddy[3]

[1] *Assistant professor, Department of Computer Science & Engineering, Anurag University, Telangana, India.*

[2,3] *Student, Department of Computer Science & Engineering, Anurag University, Telangana, India.*

Abstract:

H1-B Visa is the maximum sought-after non-immigrant visa that lets in overseas employees to paintings in United States in uniqueness career. In 2019, greater than 1 million candidates implemented to get an H-1B visa together with new programs, renewals and switch of H1-B to every other company. There have been greater than 180,000 new candidates for H1-B, but, simplest 80,000 programs have been picked up withinside the lottery manner for taking it in addition to USCIS for approval. Uncertainty in obtaining an H1B visa creates uncertainty in employment and the criminal reputation of a lawsuit utility and excessive felony and visa processing charges for the business enterprise over the duration of employment. We plan to apply the anonymized dataset for 2019 that the United States Department of Labor publishes and publicly observes system control strategies to improve the predictability of this company's approval.

Keywords:

Applications, immigrant, and non-immigrant, Random Forest Classifier.

I. Introduction:

The US H1-B visa is a non-migrant visa that allows US agencies to make use of graduate degree worker's in distinctiveness occupations that require hypothetical or specialized flair especially fields like IT, finance, bookkeeping, design, designing, arithmetic, technology, medication, and so forth This is one of the highly applied visa classifications, and agencies that commonly require strange capacity rely upon it with the aid of using and The improvement of IT, Research and Development and extraordinary regions influencing US financial system has restricted US installation companies to recruit strange capacity and henceforth the tempo of H1-B visa request documenting has increment generously. Visa is the manual of authorization on a tour allow that offers a allow to the holder for transportation, departure or stay in the United States for a predetermined period of time. There are exceptional forms of foreigner visas, the desired structures, and the approach withinside the employee visa manner contingent upon the state one desires to transport. Moving to America is a critical and complicated selection. The U. South America has several settler visa instructions such as H1B, L1 and J1 and so on. To be certified to use for a employee visa, an outdoor local need to be supported with the aid of using a USA situation relative, U.S. valid perpetual inhabitant, or a deliberate enterprise, with multiple unique cases. The assist starts the motion technique with the aid of using recording an hobby to for the far off inhabitant's reason with U.S. Residency and Colonization Facilities (USCIS). Among the higher piece of this H-1B are significantly tremendous beginning past due because of manufactures no of petitions and incorrect gadget for buying consent. H1B is a visa characterization in America below motion and nationality act (INA). Empowers U.S supervisors to dismiss employees at disproportionate levels and capable of "distinguishing the electricity trades". H-1B is a enterprise primarily based totally non-brief visa amassing for short far off professionals withinside the US. For an outdoor country wide to use for H1B visa, a US enterprise need to provide a career and request to for H-1B visa with America motion office. This is the maximum broadly identified visa reputation related to and held with the aid of using well-known understudies after they end school/superior education (Masters, Ph.D.) and paintings in a full-time position. The Office of Foreign Labour Certification (OFLC), creates software records this is beneficial records approximately the motion packages together with the H1-B visa.

II. Literature Survey:

It's tough to magnify the importance — and intricacy — of the H-1B visa framework withinside the U.S. It is the largest professional traveler visa software in the United States and a wide channel for high-level movement. It allows agencies to appoint strange worker's for precise positions that may be looking to fill. It has profited the tech enterprise immensely, and extraordinary regions, together with clinical services, technology, and money, have likewise applied it to fill holes of their paintings forces. But in April, speedy after U.S. Resident and Immigration Services (USCIS) led its each yr lottery for deciding on H-1B visas (it had been given 199,000 petitions for the handy 85,000 visas), President Trump marked a number one request at the manner to placed H-1B and comparative obligations underneath new examination. Called "Buy American and Rent American", it guides government companies to check whether or not now no longer contemporary arrangements enough cognizance on American devices and make sure American personnel. The request is the most cutting-edge improvement in a long-running talk over how companies employ the H-1B software program and what it technique for American specialists. A widespread part of the question encompasses whether or not or now no longer companies make the maximum this device to employ unusual specialists for lower pay, dislodging Americans from those positions. Yet, recognize the crucial components of this talk: one diploma lays on the hefty utilization of H-1B visas with the useful resource of the use of rethinking businesses; each different lays on the struggle about whether or not or now no longer this device builds companies' admittance to scant talents, or in reality assists them with restricting expenses. H-1B visas are conceded via a business enterprise-driven framework, which means bosses request the overall public authority for visas related to unique jobs.These should be called "power professions", which typically require four to 12 months of certification (or thesame) and are found in areas such as time, design, evidence-based innovation, pharmaceuticals and business. Organizations need to authenticate that they won't pay a H-1B expert brief of what they may an American, and that H1B employees will not have "antagonistic effect on the operating conditions" of the various specialists, however it it is normal that it can hardly speak and is not strictly (if with the useful resource of useof any technique) There is likewise assessment that it opens up wonderful provisos that companies can misuse.

III. Methodology:

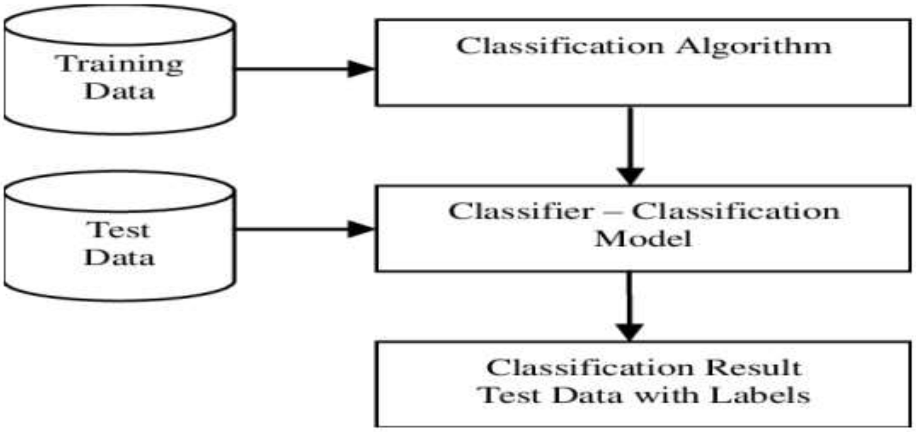


Fig 1.suggests the block diagram of a proposed H1B visa approval system.

Here we are going to assemble a class model using Random Forest Classifier. Random forests is a supervised studying set of regulations. It can be used every for class and regression. It is also the most flexible and smooth to use set of regulations. A random wooded area is a meta estimator that matches a number of chosen tree classifiers various sub-samples of the dataset and uses averaging to decorate the predictive accuracy and manage over-fitting. Random forests create choice timber on randomly determined on facts samples, gets prediction from each tree and selects the first-rate solution through voting. The requirement is to offer you with novel features based completely on the useful statistics of the dataset. It is crucial to hold in mind to avoid correlated features in some unspecified time in the future of this way. Each characteristic must handiest decorate the facts contained withinside the dataset. Figure 1: Block diagram of the proposed H1B visa approval framework

A. System Architecture:

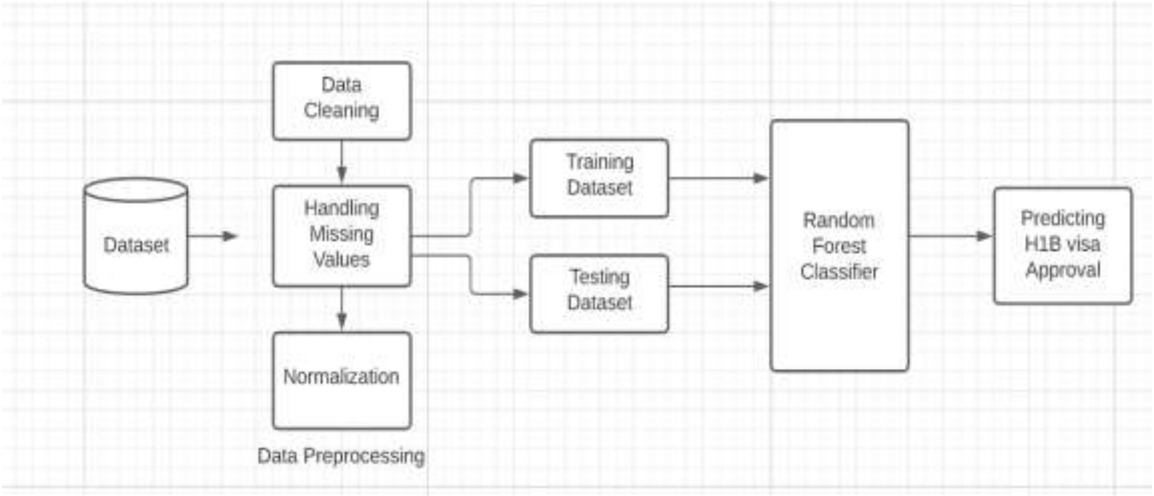


Figure 2: Architecture

Collecting facts is step one of the way drift which includes defining the project, installing location the device environment suitable for the development requirements and later statistics the facts using wonderful Python libraries and device studying techniques. Data Cleaning wants to be achieved on the facts accumulated just so the assessment be very accurate for satisfactory results.

B. Algorithm:

Random Forest : It's an ensemble set of regulations which means internally it will use multiple classifier algorithms to assemble accurate classifier model. Internally this set of regulations will use choice tree set of regulations to generate it educate model for class. Random Forest is a classifier that consists of a number of choice timber on various subsets of the given dataset and takes the uncommon place to decorate the predictive accuracy of that dataset. The larger wood type in the wooded area leads to greater accuracy and avoids the inconvenience of over-adjustment. How does Random Forest set of regulations artwork? Random Forest works in -segment first is to create the random wooded area with the useful resource of the using the combination of the N choice tree and the 2nd is to make predictions for each tree created in the main segment.

The Working way can be described withinside the beneath steps and diagram:

Step-1: Select random K facts elements from the training set.

Step-2: Build the choice timber associated with the selected facts elements (Subsets).

Step-3: Choose the range N for choice timber that you want to assemble.

Step-4: Repeat Step 1 & 2.

Step-5: For new facts elements, find out the predictions of each choice tree, and assign the today's facts elements to the elegance that wins the majority votes.

IV. Implementation:

Modules:

To placed into impact this project, I surely have used following modules:

A. **Data Collection –** Dataset is downloaded from kaggle.com.
B. **Data Pre-processing –** As we don't have any null values, we've got were given processed for Standard Scaler.
C. **Data Splitting -** We divided the records into education (70%) and testing (30%).

Module Description:

i. We have made the next reading modules of the Random Forest Classifier tool As the selection shows, "Random forest is a classifier that includes some of woods selected from diverse subsets of the given dataset and takes the now no longer unusual place to beautify the predictive accuracy of that dataset".

ii. Decision Tree Classifier-The desire tree classifier creates the elegance version with the beneficial useful resource of using constructing a desire tree. Each node withinside the tree specifies a check on an characteristic, every department descending from that node corresponds to one of the feasible values for that characteristic.

D. Introduction to Technologies Used :

Python

Python is presently the maximum extensively used multi-motive, immoderate-degree programming language. Python allows programming in any element-oriented and procedural paradigm. Python packages are typically smaller than precise programming languages consisting of Java. Programmers need to kind plenty less, and the language's indentation requirement makes them readable all of the time. The Python language is used by almost all companies of the time such as Google, Amazon, Facebook, Instagram, Dropbox, Uber...etc. The largest power of Python is its big series of modern libraries which may be used for the subsequent –

- Machine Learning
- GUI Application Tkinter
- Sklearn
- Seaborn
- Keras

NumPy

It is a general-motive array-processing bundle deal deal. It offers a immoderate-everyday usual overall performance multidimensional array item and system for strolling with those arrays. It is the vital bundle deal for medical computing with Python. It includes several skills as well as essential skills: A effective N-dimensional array item. Sophisticated (broadcasting) competencies .C/C++ and Fortran code integration gear Useful abilities in linear algebra, Fourier remodel and random c language Moreits apparent medical makes use of, NumPy moreover can be used as an green 18 multidimensional challenge of common data. Arbitrary data sorts may be described the usage of NumPy which lets in NumPy to seamlessly and short combine with a big form of databases.

Pandas

Pandas is an open-supply Python bundle deal this is maximum extensively used for data era/data evaluation and system reading responsibilities. It is constructed on pinnacle of every extraordinary bundle deal named NumPy, which offers assist for multi-dimensional arrays. As one of the maximum famous data-wrangling packages, Pandas works well with many precise era modules withinside the Python environment and is usually covered in each Python distribution, from those who include your walking tool to company supplier distributions like Active State's lively python. Pandas make it easy to do a number of the time consuming, repetitive responsibilities related to strolling with data, together with:

- Data cleansing
- Data fill
- Data normalization
- Merges and joins
- Data visualization
- Statistical evaluation
- Data inspection
- Loading and saving data
- And a good deal extra

In truth, with Pandas, you can do everything that makes world-foremost data scientists vote Pandas because of the truth the first rate data evaluation and manipulation

Sklearn

Scikit-take a look at (Sklearn) is the maximum beneficial and sturdy library for system reading It offers a preference of green system for system reading and statistical modeling together with elegance, regression, clustering, and dimensionality good deal via a regular interface in Python. This library, which is primarily written in Python, is built on **NumPy, SciPy, and Matplotlib.**

Scikit-take a look at library is centered on modeling the data. Some of the maximum famous businesses of fashions supplied with the beneficial useful resource of using Sklearn are as follows –

Supervised Learning algorithms − Almost all of the famous supervised reading algorithms, like Linear Regression, Support Vector Machine (SVM), Decision Tree, etc. , are a part of sci-kit-take a look at. **Unsupervised Learning algorithms** − On the opposite hand, it furthermore has all of the famous unsupervised reading algorithms from clustering, factor evaluation, PCA (Principal Component Analysis) to unsupervised neural networks.

Clustering − This version is used for grouping unlabeled data.

Cross-Validation − It is used to test the accuracy of supervised fashions on unseen facts.

Dimensionality Reduction − It is used for lowering the fashion of attributes in facts which may be further used for summarization, visualization, and function preference.

Ensemble strategies − As the selection shows, it's far used for combining the predictions of more than one supervised fashions.

Feature extraction − It is used to extract the competencies from facts to outline the attributes in image and textual content facts. Feature preference − It is used to emerge as aware of beneficial attributes to create supervised fashions.

Jupyter

The Jupyter Notebook is an open-supply internet software that you could use to create and percentage files that incorporate stay code, equations, visualizations, and textual content. Jupyter Notebook is maintained with the beneficial useful resource of the use of the human beings at Project Jupyter.

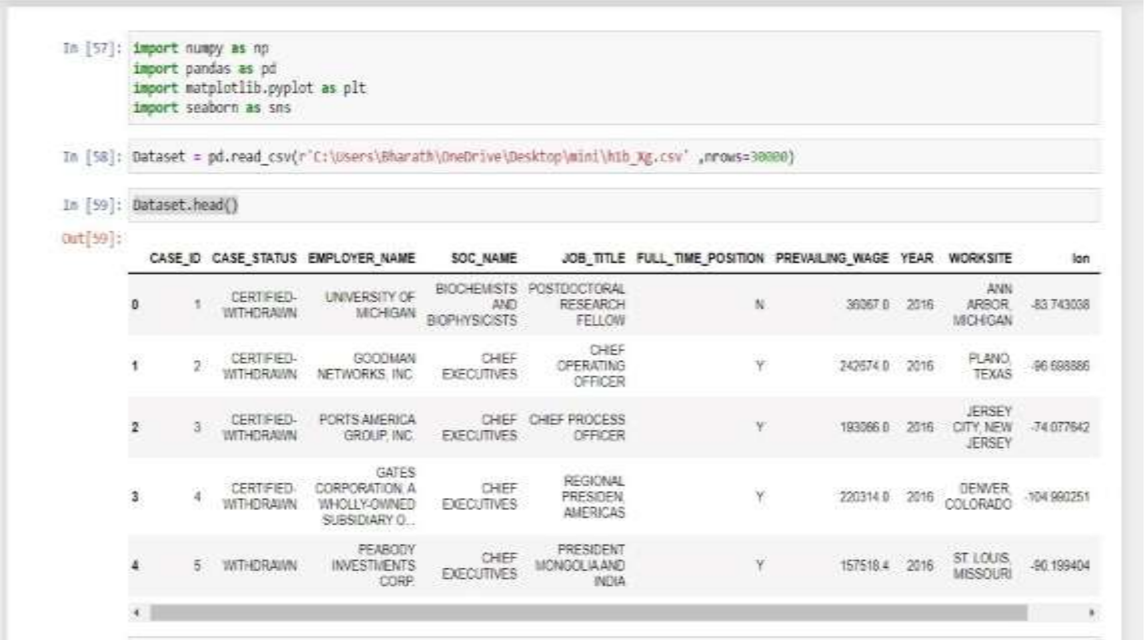
Jupyter Notebooks are a spin-off assignment from the Python assignment, which used to have an Python Notebook assignment itself. The call, Jupyter, comes from the middle-supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the Python kernel, which allows you to put in writing your packages in Python, however there are presently over one hundred awesome kernels that you could furthermore use. Jupyter Notebook isn't covered in Python, so in case you need to strive it, you may want to put in Jupyter. There are many distributions of the Python language. This article will interest on truly of them to put in Jupyter Notebook. The maximum famous is Python, 20 it's miles the reference model of Python that you could get from their website. It is likewise assumed which you are the usage of Python3.

The Jupyter Notebook has numerous menus that you could use to have interaction together alongside aspect your Notebook. The menu runs alongside the pinnacle of the Notebook much like menus do in awesome applications. Here is a list of the current menus:

- File
- Edit
- View
- Insert
- Cell
- Core
- Widget.

V. Results and Discussion:

Sample code

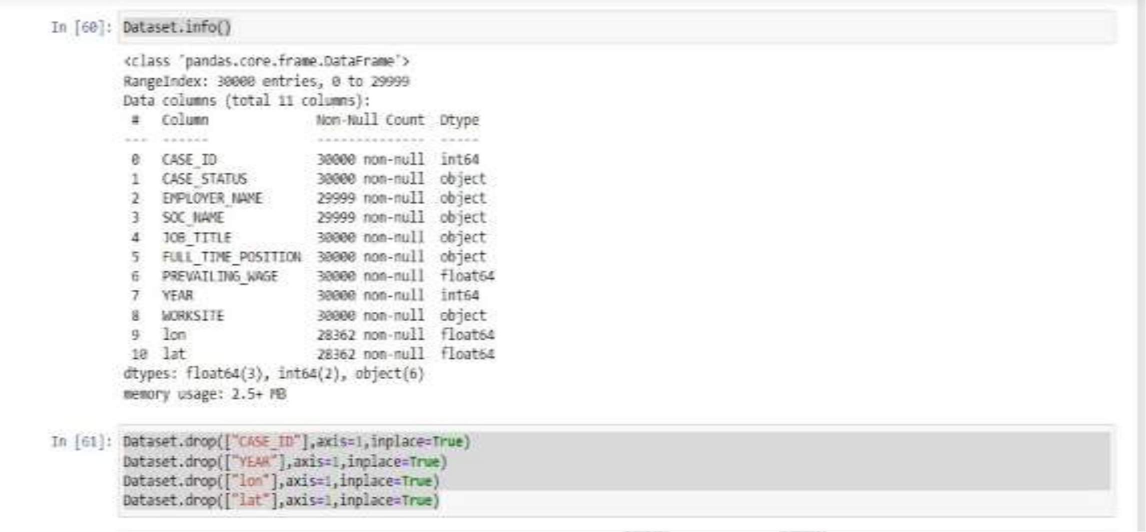


```
In [57]: import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         import seaborn as sns

In [58]: Dataset = pd.read_csv(r'C:\Users\Bharath\OneDrive\Desktop\mini\h1b_Xg.csv' ,nrows=30000)

In [59]: Dataset.head()
Out[59]:
```

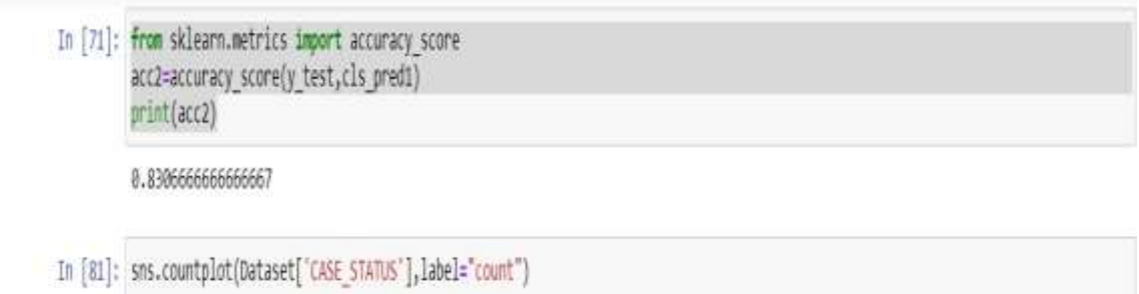
	CASE_ID	CASE_STATUS	EMPLOYER_NAME	SOC_NAME	JOB_TITLE	FULL_TIME_POSITION	PREVAILING_WAGE	YEAR	WORKSITE	lon
0	1	CERTIFIED-WITHDRAWN	UNIVERSITY OF MICHIGAN	BIOCHEMISTS AND BIOPHYSICISTS	POSTDOCTORAL RESEARCH FELLOW	N	36067.0	2016	ANN ARBOR, MICHIGAN	-83.743038
1	2	CERTIFIED-WITHDRAWN	GOODMAN NETWORKS, INC.	CHIEF EXECUTIVES	CHIEF OPERATING OFFICER	Y	242674.0	2016	PLANO, TEXAS	-96.698886
2	3	CERTIFIED-WITHDRAWN	PORTS AMERICA GROUP, INC.	CHIEF EXECUTIVES	CHIEF PROCESS OFFICER	Y	193066.0	2016	JERSEY CITY, NEW JERSEY	-74.077642
3	4	CERTIFIED-WITHDRAWN	GATES CORPORATION, A WHOLLY-OWNED SUBSIDIARY O...	CHIEF EXECUTIVES	REGIONAL PRESIDEN, AMERICAS	Y	220314.0	2016	DENVER, COLORADO	-104.990251
4	5	WITHDRAWN	PEABODY INVESTMENTS CORP.	CHIEF EXECUTIVES	PRESIDENT MONGOLIA AND INDIA	Y	157518.4	2016	ST. LOUIS, MISSOURI	-90.199404



```
In [60]: Dataset.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   CASE_ID             30000 non-null  int64
 1   CASE_STATUS         30000 non-null  object
 2   EMPLOYER_NAME       29999 non-null  object
 3   SOC_NAME            29999 non-null  object
 4   JOB_TITLE           30000 non-null  object
 5   FULL_TIME_POSITION  30000 non-null  object
 6   PREVAILING_WAGE     30000 non-null  float64
 7   YEAR                30000 non-null  int64
 8   WORKSITE            30000 non-null  object
 9   lon                 28362 non-null  float64
 10  lat                 28362 non-null  float64
dtypes: float64(3), int64(2), object(6)
memory usage: 2.5+ MB

In [61]: Dataset.drop(["CASE_ID"],axis=1,inplace=True)
         Dataset.drop(["YEAR"],axis=1,inplace=True)
         Dataset.drop(["lon"],axis=1,inplace=True)
         Dataset.drop(["lat"],axis=1,inplace=True)
```

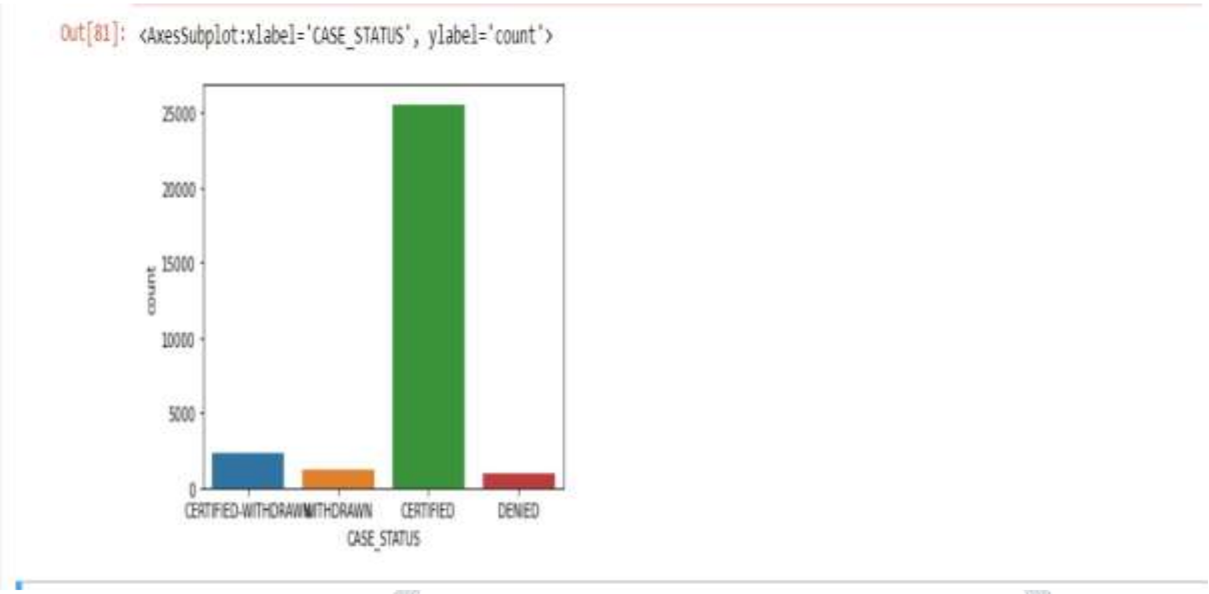


```
In [71]: from sklearn.metrics import accuracy_score
         acc2=accuracy_score(y_test,cls_pred1)
         print(acc2)

         0.8306666666666667

In [81]: sns.countplot(Dataset['CASE_STATUS'],label="count")
```

Test cases :



```
Out[81]: <AxesSubplot:xlabel='CASE_STATUS', ylabel='count'>
```

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 83. 06% 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.

Future Enhancement:

In in addition Random Forest set of suggestions may be implemented on wonderful information gadgets to be had for visa approvals to in addition take a look at its accuracy. A rigorous evaluation of diverse gadget learning algorithms other than those six moreover can be completed in destiny to research the energy of gadget learning algorithms for visa reputation prediction. In in addition take a look at, we can attempt to behavior experiments on big information gadgets or attempt to tune the version which will benefit the kingdom -of-paintings regular basic overall performance of the version and a great UI assist tool making it entire internet software In the destiny we can attempt to use a few greater strategies and strategies in-order to are looking for the recognition of H1B visa we can look at several possible answers to predict the candidate's Finally, we can attempt to achieve the awesome feasible answer.

VI. Bibliography:

[1].“H-1B Fiscal Year (FY) 2018 Cap Season,” USCIS. [Online]. Available: https://www. uscis.gov/going for walks-united-states/temporary-employees/h-1b-specialty-occupationsand-fashion-fashions/h-1b-fiscal-365 days-fy-2018-cap-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-cap-visa-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-analysisto-predicth1-b-wages/. [4].Predicting the Case Status of H1B Visa Applications". [In line]. Available: https://cseweb. ucsd.edu/commands/wi17/cse258-a/reports/a054.pdf. [5]. Mallikarjuna Reddy, V. Venkata Krishna, L.Sumalatha, "Facial recognition based primarily on Full Diagonal Cross Matrix", International Journal of Image, Graphics and Signal Processing (IJIGSP),Vol. 10,No.3, pp. 59-66, 2018.DOI: 10.5815/ijigsp.2018.03.07. [6]. Ayaluri MR, K. SR, Konda SR, Chidirala SR. 2021. Efficient analysis of the use of a community of vehicular convolutional cellular encoders to ensure adequate image quality. Peer J Informatica 7: e356 https://doi.org/10.7717/peerjcs.356 [7] A Mallikarjuna Reddy, Vakulabharanam Venkata Krishna, Lingamgunta Sumalatha and AvukuObulesh, “Age Classification Using Motif and Statistical Features Derived on Gradient Facial Images”, Recent Advances in Computer Science and Communications (2020) 13: 965. https://doi.org/10. 2174/2213275912666190417151247. [8].Predicting the Case Status of H1B Visa Applications". [In line]. Available: https://cseweb. ucsd.edu/commands/wi17/cse258-a/reports/a054.pdf. [9].gov/going for walks-united-states/temporary-employees/h-1b-specialty-occupationsand-fashionmodels/h-1b-fiscal-365 days-fy-2018-cap-season. [Accessed: 20-Oct-2017]. [10] Swarajya Lakshmi V Papineni, SnigdhaYarlagadda.