# VISA APPROVAL PREDICTION

A Project Report submitted to

**Jawaharlal Nehru Technological University, Hyderabad**

In partial fulfilment for the requirement for the award of Bachelor of technology Degree in

Computer Science and Engineering

Submitted By


| | |
|---|---|
| **Kotha Nithya** | **17UK1A0544** |
| **Poshala Amulya** | **17UK1A0502** |
| **Sharath Kumar Gongalla** | **17UK1A0557** |
| **Podila Madhu** | **17UK1A0542** |


Under the Guidance of

**Dr. Rakesh Nayak**

**(Professor, CSE Dept.)**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**VAAGDEVI ENGINEERING COLLEGE**

Affiliation to JNTU, Hyderabad & Approved by AICTE, New Delhi.

Bollikunta, Warangal (T.S)-506005


**2017-2021**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**VAAGDEVI ENGINEERING COLLEGE**

**Warangal**

# CERTIFICATE

This is to certify that the Project report entitled "***VISA APPROVAL PREDICTION***" is submitted by ***Kotha Nithya***(17UK1A0544), ***Poshala Amulya***(17UK1A0502), ***Sharath Kumar Gongalla***(17UK1A0557), ***Podila Madhu***(17UK1A0542), in partial fulfilment of the requirements for the award of the Degree in Bachelor of Technology in computer science and engineering during the academic year 2020-2021.

**Guide:**

**Dr. Rakesh Nayak**

Professor

**HOD:**

**Dr. R. Naveen Kumar**

Professor

**External Examiner:**

SMARTBRIDGE

# TO WHOMSOEVER IT MAY CONCERN

This is to certify that Mr./Ms. **Kotha Nithya**, has successfully completed the internship at **SmartBridge Educational Services Private Limited** from **07/10/2020 to 08/08/2020**

During this period he/she had learned the concepts of **Machine Learning & Deep Learning** and worked under the supervision of project mentor & developed the project entitled **"VISA Approval Prediction ".**

He/she was found hardworking, punctual and inquisitive, during the tenure of internship.

We wish him/her every success in career.

**Jayaprakash. Ch**
Program Manager

**February 05, 2021**
Issued on

Powered by
**Smart Internz**
www.smartInternz.com

**SB ID: SB20200060508**

Authenticity of this certificate can be validated by going to:
**https://smartinternz.com/internships/certificates/3f900db2608fb3eecb3ee77ba9ef5f60**

SMARTBRIDGE

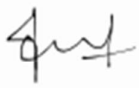
# TO WHOMSOEVER IT MAY CONCERN

This is to certify that Mr./Ms. **Poshala Amulya**, has successfully completed the internship at **SmartBridge Educational Services Private Limited** from **07/10/2020 to 08/08/2020**

During this period he/she had learned the concepts of **Machine Learning & Deep Learning** and worked under the supervision of project mentor & developed the project entitled **"VISA Approval Prediction ".**

He/she was found hardworking, punctual and inquisitive, during the tenure of internship.

We wish him/her every success in career.



**Jayaprakash. Ch**
Program Manager

**February 05, 2021**
Issued on

Powered by
**Smart Internz**
www.smartinternz.com

**SB ID: SB20200073641**

Authenticity of this certificate can be validated by going to:
**https://smartinternz.com/internships/certificates/bf5cd8b2509011b9502a72296edc14a0**

SMARTBRIDGE

# TO WHOMSOEVER IT MAY CONCERN

This is to certify that Mr./Ms. **Sharath Kumar Gongalla**, has successfully completed the internship at **SmartBridge Educational Services Private Limited** from **07/10/2020 to 08/08/2020**

During this period he/she had learned the concepts of **Machine Learning & Deep Learning** and worked under the supervision of project mentor & developed the project entitled **"VISA Approval Prediction ".**

He/she was found hardworking, punctual and inquisitive, during the tenure of internship.

We wish him/her every success in career.

**Jayaprakash. Ch**
Program Manager

**February 05, 2021**
Issued on

Powered by
**Smart Internz**
www.smartInternz.com

**SB ID: SB20200075241**

Authenticity of this certificate can be validated by going to:
**https://smartinternz.com/internships/certificates/86c51678350f656dcc7f490a43946ee5**

SMARTBRIDGE

# TO WHOMSOEVER IT MAY CONCERN

This is to certify that Mr./Ms. **Madhu Podila**, has successfully completed the internship at **SmartBridge Educational Services Private Limited** from **07/10/2020 to 08/08/2020**

During this period he/she had learned the concepts of **Machine Learning & Deep Learning** and worked under the supervision of project mentor & developed the project entitled **"VISA Approval Prediction ".**

He/she was found hardworking, punctual and inquisitive, during the tenure of internship.

We wish him/her every success in career.

**Jayaprakash. Ch**
Program Manager

**February 05, 2021**
Issued on

Powered by
**Smart Internz**
www.smartinternz.com

**SB ID: SB20200075244**

Authenticity of this certificate can be validated by going to:
**https://smartinternz.com/internships/certificates/0f3c5d0c3666eec8cd311bec6d878915**

# ACKNOWLEDGEMENT

This project has been carried out in the department of Computer Science and Engineering of Vaagdevi Engineering College, Bollikunta, Warangal. Many people have helped us in the realization of this work we would like this opportunity to express our gratitude to all of them.

We express our gratitude to our principal **Dr. P. Prasad Rao** , who permitted us to carry the project work as part of the academics.

We would like to thank **Dr. R. Naveen Kumar**, Head of Department (CSE) for his support and encouragement in completing our project. We would like to express our gratitude to our Project guide **Dr. Rakesh Nayak** Professor for his support and encouragement in completing our mini project successfully.

We express our sincere thanks and gratitude to **TheSmartBridge**, for providing internship.

Finally, we wish to take this opportunity to express our deep gratitude to our family members and all the people who have extended their cooperation in various ways during our project work.

# ABSTRACT

This work focuses on the data science challenge problem of predicting the decision for past immigration visa applications using supervised machine learning for classification. We describe an end-to-end approach that first prepares historical data for supervised inductive learning, trains various discriminative models, and evaluates these models using simple statistical validation methods.

The H1 B visa is the most demanded visa worldwide. 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.

The H-1B visa allows employers in the United States to temporarily employ foreign nationals in various specialty occupations that require a bachelor's degree or higher in the specific specialty, or its equivalents. These specialty occupations may often include, but are not limited to: medicine, health, journalism, and areas of science, technology, engineering and mathematics (STEM). Every year the United States Citizenship and Immigration Service (USCIS) grants a current maximum of 85,000 visas, even though the number of applicants surpasses this amount by a huge difference and this selection process is claimed to be a lottery system. The data set used for this experimental research project contains all the petitions made for this visa cap of the year 2016.

This project aims at using discriminative machine learning techniques to classify these petitions and predict the "case status" of each petition based on various factors. Exploratory data analysis is also done to determine the top employers, the locations which most appeal for foreign nationals under this visa cap and the job roles which have the highest number of foreign workers. I apply supervised inductive learning algorithms such as Gaussian Naïve Bayes, Logistic Regression, Decision Tree, k – Nearest Neighbors, Support Vector Machines, Random Forests to identify the most probable factors for H-1B visa certifications and compare the results of each to determine the best predictive model for this testbed.

# LIST OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF SCREENS

# CHAPTER-1
# INTRODUCTION

The H1B visa is a work visa granted by the United States department of Immigration under the immigration act of the United States constitution for highly skilled foreign workers who want to enter the US on a valid work visa issued through a selective process. The visa is validated under strict stipulations. The applications are normally made by MNC's on behalf of their employees to the US embassies in their respective countries. After considerable screening processes the applicants receive their H1B visas. The requirements set forth by the government.

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 foreigner visas, the required structures, and the means in the worker visa process contingent upon the nation one needs to move. Moving to America is a vital and complex decision. The U.S of America has numerous classes for settler visas like H1B, L1, and 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. The help begins the movement methodology by recording an interest to for the remote inhabitant's purpose with U.S. Residency and Colonization Facilities (USCIS). Among the better piece of this H-1B are greatly outstanding starting late due to manufactures no of petitions and wrong system for getting consent. H1B is a visa characterization in America under movement and nationality act (INA). Empowers U.S supervisors to yield outside workers with high degrees and capable of "distinguishing strength occupations". H-1B is a business based non-transient visa gathering for brief remote specialists in the US. 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.

The Act of 1990 set up the H-1B visa package for impermanent labourers in "claim to fame occupations". The rules defines "claim to fame occupation" as requiring hypothetical and common-sense use of a collection of exceptionally particular learning in a field of human undertaking including, yet not constrained to, design, building, arithmetic, physical sciences, sociologies, solution and wellbeing, instruction, law, bookkeeping, business fortes, religious philosophy, and expressions of the human experience. Furthermore, candidates are required to have achieved a four year certification or it's identical as a base.

In spite of the fact that H1B visa contributed a considerable measure to the economy of USA by bringing the skilled non-natives, it additionally influences American work. They lose their employment, as firms incline toward modest work when contrasted with American's. The objective of the H1B program is to connect a work hole in the U.S without influencing U.S specialists. At to

start with, the structure of H1B is to fill work hole however current structure encourages businesses to augment the work hole as there aren't any qualified U.S specialists and they are procuring modest remote labourers as H1B program.

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. Every year, the U.S. Citizenship and Immigration Services (USCIS) will officially begin accepting petitions for the H-1B Visa in April for the next fiscal year. The status of H-1B visa will definitely influence the life and work, and even the career of the international students. Since the number of applicants is very large than the number of selections and as the selection process is claimed to be as lottery there is no insight of how the attributes have influence over the outcome.

The Office of Foreign Labour Certification (OFLC) generates program data, including data about H1-B visas. The disclosure data is updated annually and is available online. The first step of the H1B application process is for the U.S. employer to file the H1B petition on behalf of the foreign worker. In second step, the prevailing and actual wages should be confirmed by the State Employment Security Agency. If the prevailing wage exceeds the offer made by the prospective employer then a wage determination will be sought. The third step of the H1B application process is to file the Labour Condition Application. The next step is to prepare the petition and file it at the proper USCIS office. Processing times for H1B application petitions are subject to vary from location to location. If you would like your petition expedited you may elect for premium processing. The final step of the H1B application process is to check the status of your H1B visa petition by entering your receipt number. Once USCIS has your application on file, they will update your status on their system.

## 1.1 Motivation

The H-1B is an employment-based, non-immigrant visa category for temporary foreign workers in the United States. For a foreign national to apply for H1-B visa, an US employer must offer a job and petition for H-1B visa with the US immigration department. This is the most common visa status applied for and held by international students once they complete college/higher education (Masters, PhD) and work in a full-time position. The Office of Foreign Labor Certification (OFLC)generates program data that is useful information about the immigration programs including the H1-B visa. Since H1-B visa petition is closely related to international students we hoped that we can get a general idea about the relation between H1-B visa application status and features like job title, prevailing wage, etc.

## 1.2 Problem definition

This project deals with the task of predicting the outcome of immigration visa applications by using machine learning methods. The prediction task is treated as one of supervised inductive learning of classification functions. Exploratory data analysis was also performed to analyze the main characteristics of the data set. A prediction model was developed which trains on historical data of the same application domain and uses machine learning classifiers to predict the outcome of visa petitions.

Our model and analysis will provide a whole picture of the different approval 2 rates by comparing different conditions based on previous data. Therefore, it will help us to predict the approval and deny rate of H-1B visa of the current year or the application in 2016.

The H-1B visa allows companies and organizations in the United States to employ foreign workers i3n specialty occupations that require major technical expertise in fields such as accounting, architecture, engineering, finance, information technology, mathematics, medicine, science, etc. A bachelor's degree or equivalent work experience are some of the basic requirements for a H-1B visa. A job offer should also be forwarded to the worker from the employer filing the petition.

Each fiscal year the United States government grants a total of 85,000 new H-1B visas; this includes 65,000 visas for overseas workers with at least a bachelor's degree and 20,000 visas available for those employees with an advanced degree in a specialty field from an accredited United States academic institution.The allotment of these visas is said to be a lottery system not depending on any specific criteria. Thus, a foreign national is not guaranteed selection for an H1B visa merely because he or she fulfills all the qualifying criteria. This work aims to use machine learning on known factors to make predictions under this uncertainty, with accuracy better than that achieved using the prior probability.

## 1.3 Objective of Project

The goal of this project was to develop a supervised inductive learning model which adopts methods of classification to predict the results of the filed applications. Various statistical evaluation methods were used to determine the accuracy of these predictive models.

The visa petitions were classified in to "Certified" or "Denied" classes formed from the "Case Status" attribute of the data set. The data present dates from the year 2016. The data from the earlier years was used to training the machine learning models and the data from years 2016 was used to testing and validating the algorithms and their performance. Discriminative models such as Logistic Regression, Decision Tree, k – Nearest Neighbors, Support Vector Machines, Naïve Bayes and

Random Forests were the main ones used for binary classification in this project. Data from the United States Bureau of Labour Statistics was used for computing the state-wise salary median, these values can be used to play an important role in classification. The overall objective was to use three main attributes of the data set and see how they influence the accuracy, precision and recall of the learning models and see which model works best for predicting the outcome of a visa petition.

## 1.4 Limitations of Project

- A small change in the data can cause a large change in the structure.
- Needs more than a single value for the prediction.
- Not robust to big-influentials.
- Requires higher time to train the model.
- Easy to overfit.

## 1.5Project Synopsis

Each petition in this data set has various factors such as 'Job Location', 'Employee Role', 'Job Category', 'Salary', etc. The supervised inductive learning algorithms were trained using examples from historical data over a chosen range of years, and predict the outcome of the attribute – "Case Status". Data from later years was used as a test data set for these learning models. The machine learning and data analysis software package Scikit-learn, written in the Python programming language, was used to implement the training and classification functions for these models. The results of the algorithms were compared using the confusion matrix and the ROC curves of these algorithms and scope of improvement was checked.

# CHAPTER-2

# LITERATURE SURVEY

## 2.1 INTRODUCTION

Since the beginning it was decided that the domain of our project would be in data analytics and machine learning. With these fields currently in high demand we were eager to work and study on them. Every IT engineer's goal is to secure an H1B visa and travel to US and work on good projects. Upon a casual discussion on the same with respect to current visa issues, an idea struck, what if we develop a software which could tell the user his/her chances of getting an H1B visa after analyzing parameters such as salary, job title, degree, company profile etc. Further reading on the same, it was found that we would have to extensively use Machine Learning algorithm like Logistic Regression. While studying about them we concluded that all the coding will have to be done in Python. Python is a programming language with high level interpreted features intended for general development use. It is built with the philosophy that puts an emphasis on readability via white spaces. Machine learning (ML) is a category of algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available.

The dataset that we are studying is available on Kaggle under the name 'H-1B Visa Petitions 2011-2016 dataset' which is processed dataset from the original data available on Office of Foreign Labour Certification (OFLC) website after performing various data transformations on the data. From data analysis performed on this data allow us to finding top Occupations, States, Employers and Industries that contribute to highest number of H1B visa application.

### 2.1.1 Classification for Decision Support: General Methodology

A common scenario in machine learning tasks involves building a classification model that is trained on a training data set, producing a model that is then used to predict the probability of data belonging to one of the classes of the test data set. The trained model solves a classification task, producing predictions of a target variable over previously unseen inputs, which represent new cases. Since historical data is used to train the model this falls under the supervised learning techniques. This classification methodology can predict the classes in two kinds of situations; one, when predicting the decisions made by a third party is involved. For example, some common applications of classification would be loan approval prediction, medical diagnosis, etc. And the second case is when classification is based on some historical ground truths. Example spam email filtering.

The main aim of this project is to predict the certification status of the visa applications. For this purpose, the concept of classification is used. That is, when all the attributes are provided, the

classifying model is trained on historical data of this application domain and probability of rows belonging to one class or status over the other is determined. This aids in predicting decisions made by USCIS and to recognize if there is any specific pattern behind their decisions.

### 2.1.2 Application Domain: Immigration Visas

One of the crucial foundations for economic growth and a key input to the knowledge economy is highly qualified and proficient personnel. Several approaches were adopted by countries in the past to achieve this, they can either obtain the desired level of expertise through their education system or obtain workforce from overseas. To be on the competitive edge in world economy, countries need to possess high skilled employees in industries such as health care, engineering, science and information technology. The demand and supply of workers in fields such as the above-mentioned ones is unevenly distributed and poorly matched. Despite all these factors, United States has a competitive edge over the other nations for being a global centre for academic training, and international students comprise of a considerate percentage of U.S. higher education enrollment.

Various U.S. employers aim to retain these international students and immigrate skilled workers from foreign nationalities for many reasons; they also face a lot of legal hurdles in doing so. According to the Immigration Act of 1990, the H-1B visa program allows companies based in U.S. to hire foreign nationals to work for a temporary amount of time in specialized occupations. Initially, the number of visas to be issued under this cap was set to 65,000, but it rose drastically to 195,000 between the years 2001 and 2003 and from the year 2004, it has been set to 65,000 per year plus an additional 20,000 for individuals with advanced degrees from U.S. accredited institutions. The H-1B visa once issued is valid for three years with one chance for renewal for another three years. Each initial petition is counted under the cap, but renewals do not account for this number. Over the past couple of years, this visa cap has been the talk of many political scenarios and cause for unrest among many individuals. Despite all this, the presence of other work-related permits such as L-1 and the heavy competition one must go through to obtain a visa under H-1B, it remains the most popular one and the one most sought after by foreign students and non-immigrants alike. In this project, I try to device an algorithm which best predicts the outcome of an individual's visa petition given various socio-economic factors such as their salary, location, job type and the state median income.

## 2.2 EXISTING SYSTEM

The previous models have high time complexity and space complexity whereas this model is constrained with the lot of advantages and with a higher accuracy than any other model is constrained with the lot of advantages and with a higher accuracy than any other model already proposed. In this model we used Machine learning algorithm named Classification algorithms which give an accuracy more then previously predicted problem.

Although this area currently doesn't seem to be well-studied, we were able to find a report that is relevant. Th reports used the same dataset we had used. The report conducted a detailed data analysis and visualization for H-1B application distribution based on different input features such as location, salary, year and job type. Although they had a prediction algorithm based on K-means clustering, it provided prediction accuracies for only a small subset of job types instead of an average one. Overall, this report gave us good insight on the distribution of our data.

Our research says that some independent analysis performed on this data provided some insightful facts and also to predict some details. A study by Andrew Shikair explored a way to predict wages of the visa recipients by performing text analysis of application attributes and their study concludes that Occupational Classification and Job Title were the two most important fields to predict the applicants wage as accurately as possible. A similar project has been done at UC, San Diego for predicting the decision of the file H1- B visa petition. A project done by the students of UC Berkley aims to predict the waiting time to get a work visa for a given job title and for a given employer. They used K-Nearest Neighbours as the primary model to predict 'Quickest Certification Rate' across both occupations and companies. Although some studies provide some insightful facts about our data, my study is to predict the final outcome of the applicants by determining the influence of the attributes over the outcome.

## 2.3 PROPOSED SYSTEM

### 2.3.1 Classification Algorithms

Classification is a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories. The classification predictive modelling is the task of approximating the mapping function from input variables to discrete output variables. The main goal is to identify which class/category the new data will fall into.

<u>Classification Terminologies In Machine Learning</u>

- **Classifier** – It is an algorithm that is used to map the input data to a specific category.
- **Classification Model** – The model predicts or draws a conclusion to the input data given for training, it will predict the class or category for the data.
- **Feature** – A feature is an individual measurable property of the phenomenon being observed.
- **Binary Classification** – It is a type of classification with two outcomes, for eg – either true or false.
- **Multi-Class Classification** – The classification with more than two classes, in multi-class classification each sample is assigned to one and only one label or target.

- **Multi-label Classification** – This is a type of classification where each sample is assigned to a set of labels or targets.
- **Initialize** – It is to assign the classifier to be used for the
- **Train the Classifier** – Each classifier in sci-kit learn uses the fit (X, y) method to fit the model for training the train X and train label y.
- **Predict the Target** – For an unlabelled observation X, the predict(X) method returns predicted label y.
- **Evaluate** – This basically means the evaluation of the model i.e. classification report, accuracy score, etc.

## Types Of Learners In Classification

- **Lazy Learners** – Lazy learners simply store the training data and wait until a testing data appears. The classification is done using the most related data in the stored training data. They have more predicting time compared to eager learners. Eg – k-nearest neighbor, case-based reasoning.
- **Eager Learners** – Eager learners construct a classification model based on the given training data before getting data for predictions. It must be able to commit to a single hypothesis that will work for the entire space. Due to this, they take a lot of time in training and less time for a prediction. Eg – Decision Tree, Naive Bayes, Artificial Neural Networks.

In machine learning, classification is a supervised learning concept which basically categorizes a set of data into classes. The most common classification problems are – speech recognition, face detection, handwriting recognition, document classification, etc. It can be either a binary classification problem or a multi-class problem too. There are a bunch of machine learning algorithms for classification in machine learning.

There are three types of classifications:

- Binary classifier
- Zero-R classifier
- One-R classifier

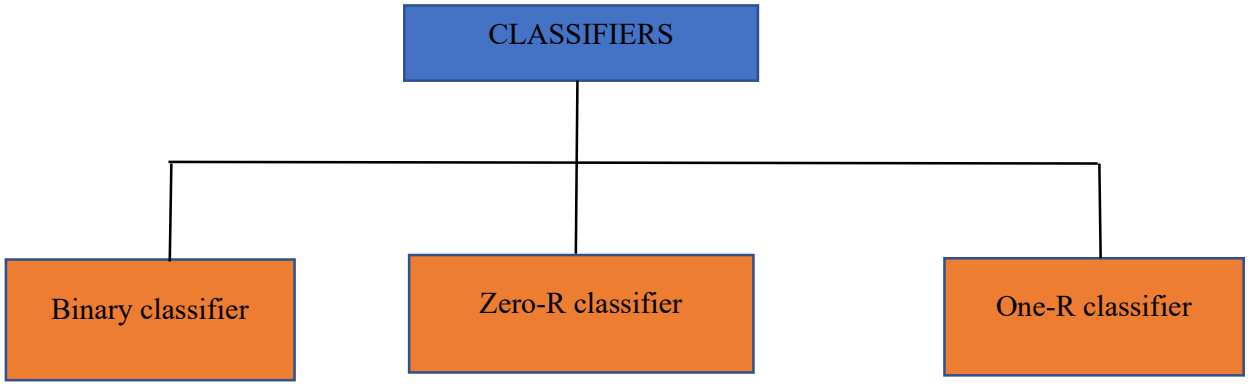



**Fig 2.1** Flowchart of classifiers

**Binary classifier:**

Binary Classification comes under Supervised Learning where the training dataset is labelled and it consists of two classes. Where in several possible binary analysis difficulties, if two combinations are not symmetric rather than overall correctness, then the corresponding symmetry are of varies according to the types of failures concern. For example, in medicinal tests, a false positive (identifying a virus if it is not present) is viewed uniquely from a false negative (not identifying a virus when it is present).

**Zero-R Classifier:**

Zero-R is the easiest method which relies on the target also overlooks all its features. It identifies the popular group but Zero -R does not have the predictability power, it is just useful to define a baseline execution for another group systems.

**Process Build:**

A frequency statistics for the targeted class and choose the common repeatedly value.

**Predictors Role:**

The contribution of features in the zero-R algorithm is not much because of it doesn't use any one of them.

**Model Evaluation:**

The Zero-R only identify the largest class correctly. As mentioned earlier, it is just useful to define a baseline execution for another group systems

**One-R Classifier:**

One -R classifier is also defined as “One Rule” with an accurate algorithm which generates one rule for each identified data set further it choose a rule with least complete error as the “one rule”.

When the goal of a problem is to learn how to map inputs from X to outputs from Y, where $Y \in \{1, \ldots C\}$, where C is the number of classes. When the number of output classes is two, $C = 2$, then it called Binary Classification. If the value of $C > 2$, then the classification is called Multiclass Classification. We use binary classification in this project. There are many classification methodologies present for binary classification such as Support Vector Machines, Naïve Bayes, Decision Tree, k – Nearest Neighbors, Random Forest, Logistic Regression. Not all the algorithms are suited for all problems and some result in over-fitting or under-fitting. In this project, Logistic Regression, Random Forests, Gaussian Naïve Bayes and k – Nearest Neighbors, Support Vector Machines, Decision Tree are implemented and their results are compared.

# CHAPTER-3
# ANALYSIS

## 3.1 THEORETICAL ANALYSIS

While selecting the algorithm that gives an accurate prediction we gone through lot of algorithms which gives the results abruptly accurate and from them we selected only one algorithm for the prediction problem that is Decision Tree. Decision Tree and Random Forest accuracies are very close but based on performance and accuracy we have considered Decision Tree as the best algorithm for this dataset.

The pecularity of this problem is collecting the real time details and working with the prediction at the same time, so we developed an user interface for the people who'll be accessing for the visa status prediction. Accuracy is defined as the ratio of the number of samples correctly classified by the classifier to the total number of samples for a given test data set. The formula is as follows

$$Accuracy=(TP+TN)/(TP+TN+FP+FN)$$

- **Accuracy**
    - Accuracy is a ratio of correctly predicted observation to the total observations
    - True Positive: The number of correct predictions that the occurrence is positive.
    - True Negative: Number of correct predictions that the occurrence is negative.

At first we got like lot of accuracies because we tried lot of algorithms for the best accurate algorithm, finally after all of that we tried the best suitable algorithm which gives the prediction accurately is Decision Tree. And developed it to use as a real time prediction problem for the visa case status i.e; visa approval prediction. In statistics, a receiver operating characteristic(ROC), is a two dimensional graphical plot that illustrates the performance of a binary classifier system. The curve is created by plotting the true positive rate(TPR) against the false positive rate(FPR) at various threshold settings. ROC curve can intuitively represent the performance of classifier.

## ROC CURVE:

Receiver operating characteristics or ROC curve is used for visual comparison of classification models, which shows the relationship between the true positive rate and the false positive rate. The area under the ROC curve is the measure of the accuracy of the model.

The curve is plotted between two parameters

- TRUE POSITIVE RATE
- FALSE POSTIVIE RATE

Before understanding, TPR and FPR let us quickly look at the confusion matrix.

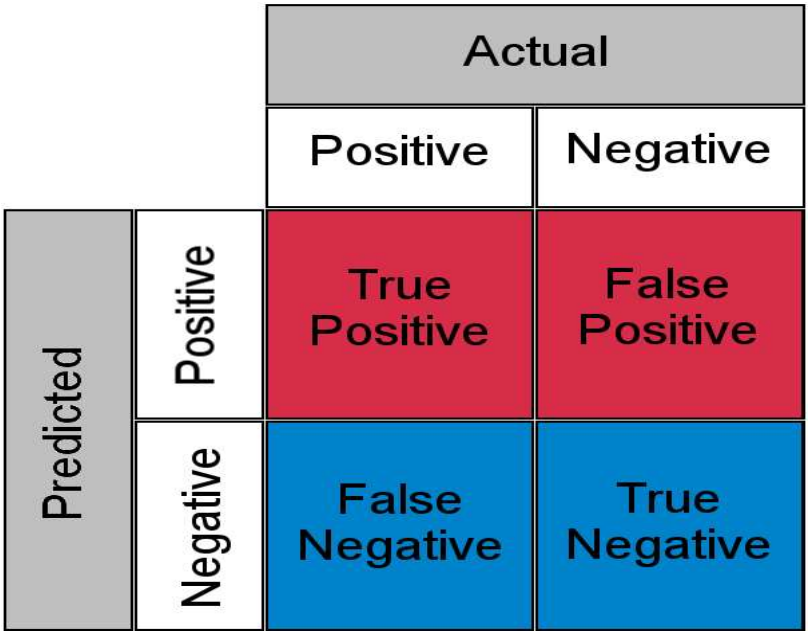



**Fig 3.1** Confusion Matrix

- **True Positive**: Actual Positive and Predicted as Positive
- **True Negative**: Actual Negative and Predicted as Negative
- **False Positive(Type I Error)**: Actual Negative but predicted as Positive
- **False Negative(Type II Error)**: Actual Positive but predicted as Negative

In simple terms, you can call False Positive as **false alarm** and False Negative as a **miss**. Now let us look at what TPR and FPR.

- FPR=FP/FP+TN
- TPR=TP/TP+FP

Basically TPR/Recall/Sensitivity is **ratio of positive examples that are correctly identified** and FPR is the **ratio of negative examples that are incorrectly classified** and as said earlier ROC is nothing but the plot between TPR and FPR across all possible thresholds and AUC is the entire area beneath this ROC curve.

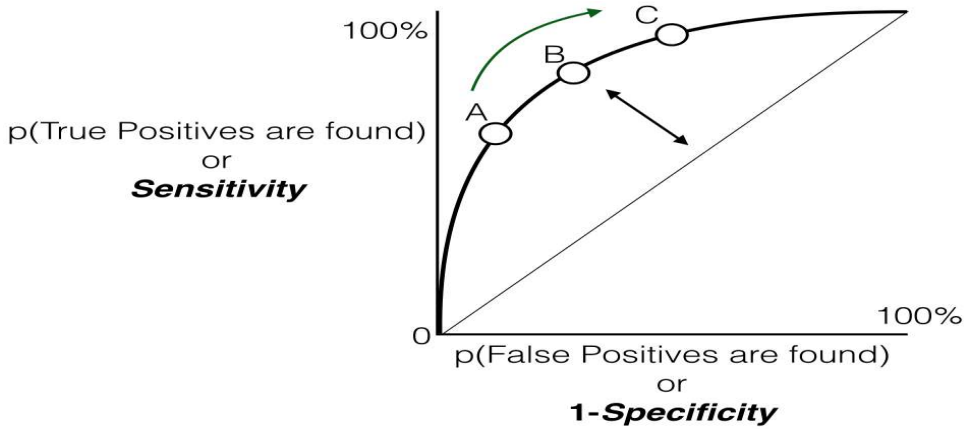



**Fig 3.2** auc_roc curve

## 3.2 SOFTWARE REQUIREMENTS

- Jupyter Notebook Environment
- Spyder Ide
- Machine Learning Algorithms
- Python(pandas, numpy, matplotlib, seaborn, sklearn)
- HTML
- Flask

We developed this visa approval prediction by using the Python language which is a interpreted, dynamically typed and highlevel programming language and using the Machine Learning algorithms.

For coding we used the Jupyter Notebook environment of the Anaconda distributions and the Spyder, it is an integrated scientific programming in the python language.

For creating an user interface for the prediction we used the Flask. It is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions, and a scripting language to create a webpage in HTML by creating the templates to use in the functions of the Flask and HTML.

## 3.3 HARDWARE REQUIREMENTS

- Windows 7,8 or 10(32 or 64 bit)
- RAM-8GB
- Processor:1.5HZ or above
- HDD:100GB or above

## 3.3 BLOCK DIAGRAM

Machine learning is a field of computer science that gives computers the ability to learn without being programmed explicitly. The power of machine learning is that you can determine how to differentiate using models, rather than using human judgment. The basic steps that lead to machine learning and will teach you how it works are described below in a big picture.

The link is to create a system that answers a particular question. This question answering system called a model is created via a process termed as training. The main goal of training is to create an accurate model that answers our questions correctly, at least for most of the times. But in order to train a model, you also need to collect data on what you'd want to train on. This is where you start and then the rest follows.

After the data collection Data cleaning and pre-processing steps were followed on the data set. Depending on the task at hand, the features were divided into subsets. Python was used as the major programming language for this program and using in-built libraries, the data set was split into training and testing data in 70 and 30 percentages respectively. Scikit-learn was used to train the inductive learners and apply them to the data sets. The training data was fed to the learning models and the prediction scores of the classification task were obtained. These models were then / tested on the testing data. The results are compared to find the best suited classifying model.

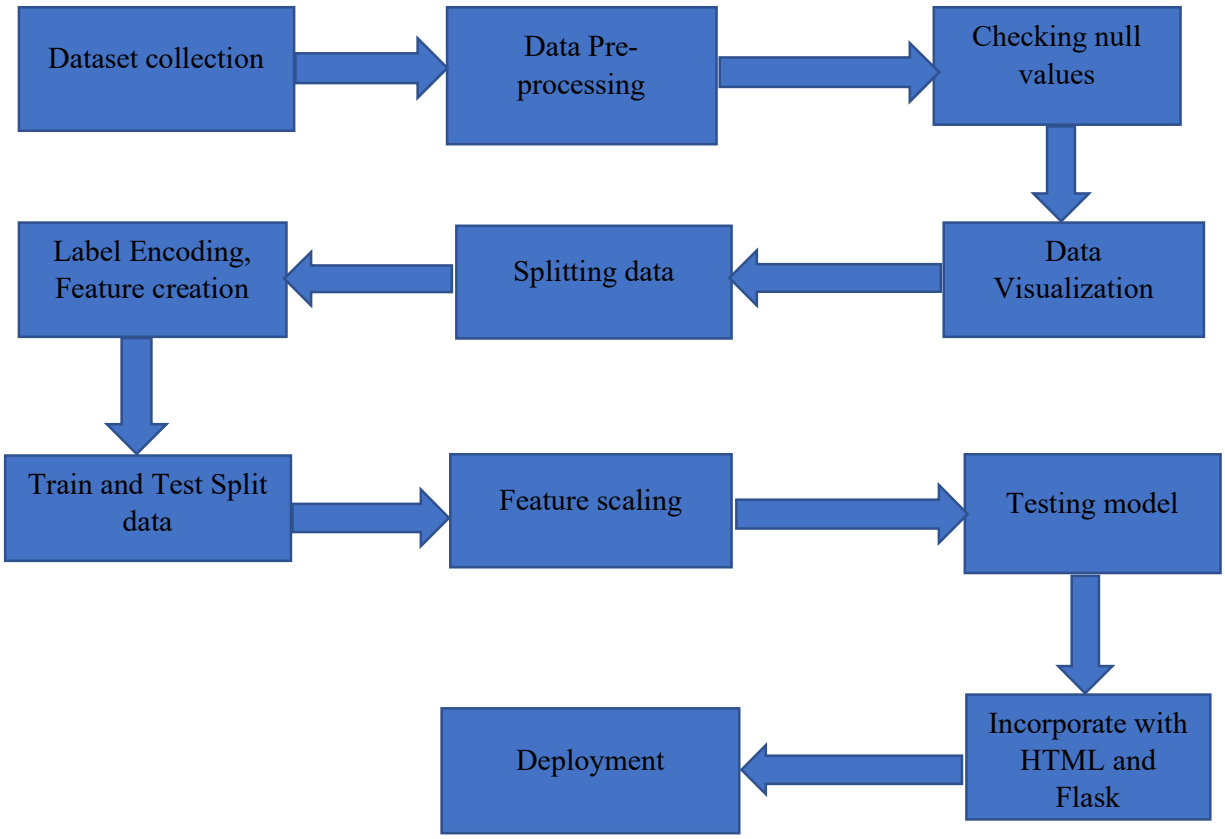



**Fig 3.3** Block diagram

## 3.4 ALGORITHMS & FLOWCHART

The following are the algorithms which we have used in this project.

### 3.4.1 Logistic regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. This algorithm assigns observations to discrete sets of classes and uses a sigmoid function to return a value to map two or more classes. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. Logistic regression uses the concept of predictive modeling as regression; therefore, it is called logistic regression, but is used to classify samples; Therefore, it falls under the classification algorithm. There are various types of Logistic Regression such as:

a. Binary

b. Multi

c. Ordinal Binary

Logistic Regression is used in this project. Sigmoid Function is used in Logistic Regression and this maps a value to another value and these values range from 0 to 1.

Logistic Regression measures the relationship between the dependent variable (our label, what we want to predict) and the one or more independent variables (our features), by estimating probabilities using it's underlying logistic function.

These probabilities must then be transformed into binary values in order to actually make a prediction. This is the task of the logistic function, also called the sigmoid function. The Sigmoid-Function is an S-shaped curve that can take any real-valued number and map it into a value between the range of 0 and 1, but never exactly at those limits. This values between 0 and 1 will then be transformed into either 0 or 1 using a threshold classifier.

The sigmoid function is given by:

$$S(x)=1/(1+e^{-x})$$

Here S(x) is the output between 0 and 1, x is the function's input. And e is the natural log's base. A threshold

value called the decision bound is selected in order to map the probability score which the classification function returns to a discrete class.

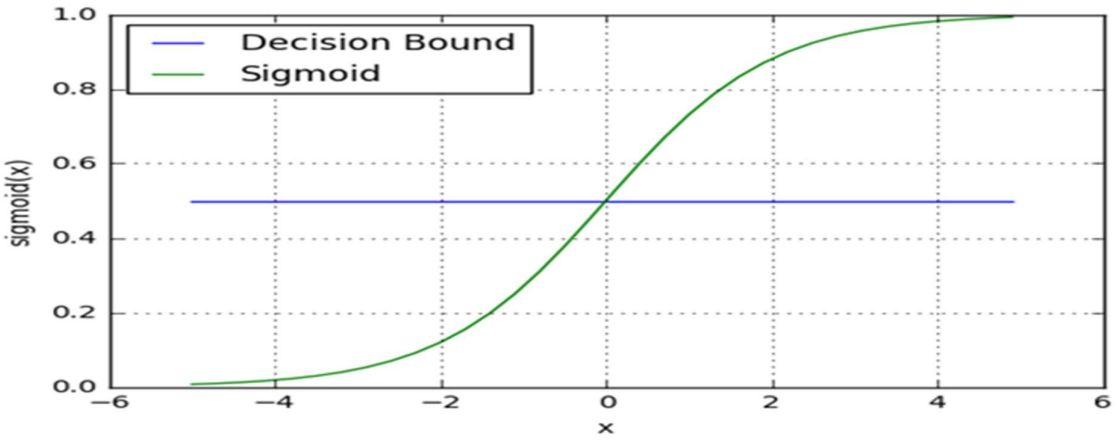



**Fig 3.4** Logistic Regression

From these sigmoid functions and decision boundaries, we can compute the prediction outcome of the classification by the Logistic Regression model.

**Advantages and Disadvantages**

Logistic regression is specifically meant for classification, it is useful in understanding how a set of independent variables affect the outcome of the dependent variable.

The main disadvantage of the logistic regression algorithm is that it only works when the predicted variable is binary, it assumes that the data is free of missing values and assumes that the predictors are independent of each other.

### 3.4.2 k-Nearest Neighbor

The k-Nearest Neighbors (k-NN) classifier can be used to classify data of various natures and one of its most important traits is its versatility and robustness. k-NN is a supervised machine learning algorithm and is of discriminative nature. k-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. Let us assume x and y are feature and label respectively. The classification learning task is to find a function:

$$h:X \rightarrow Y$$

This function is consistent with training data (or is empirically good according to the evaluation criteria,

e.g., a minimal loss function over labelled training data). k-NN Classifier adopts the technique of forming a majority vote among the K most similar instances of a data set. Similarity between these data points is usually measured using the Euclidean distance formula. There are however other distance formulas too such as Manhattan, Hamming etc.

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2 + \ldots + (x_n - x'_n)^2}$$

**Fig 3.5** Euclidean Distance

The value of K can be decided by the user. The higher the value of K, the more the system is resilient to outliers and the smoother the function is. The conditional probability of the classifier can be given by:

$$P(y = j|X = x) = \frac{1}{K} \sum_{i \in \mathcal{A}} I(y^{(i)} = j)$$

**Fig 3.6** Conditional Probability of k-NN Classifier

An illustration of how the outcome of the classifier is shown below.

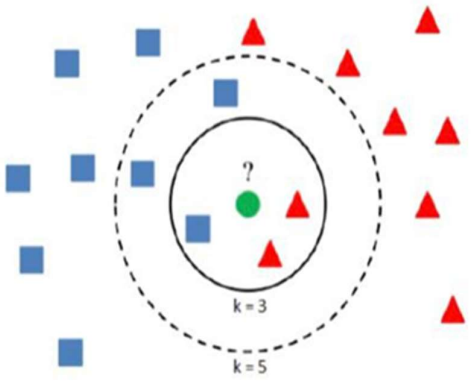



**Fig 3.7** k-NN Classification

**Advantages And Disadvantages**

This algorithm is quite simple in its implementation and is robust to noisy training data. Even if the training data is large, it is quite efficient. The only disadvantage with the KNN algorithm is that there is no need to determine the value of K and computation cost is pretty high compared to other algorithms.

### 3.4.3 Decision Tree

Decision trees model sequential decision problems under uncertainity. A decision tree describes graphically the decisions to be made, the events that may occur, and the outcomes associated with combinations of decisions and events. Probabilities are assigned to the events, and the values are determined for each outcome. A major goal of the analysis is to determine the best decisions.

The decision tree algorithm builds the classification model in the form of a tree structure**.** It utilizes the if-then rules which are equally exhaustive and mutually exclusive in classification. The process goes on with breaking down the data into smaller structures and eventually associating it with an incremental decision tree. The final structure looks like a tree with nodes and leaves. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covering the rules are removed. The process continues on the training set until the termination point is met.

Decision tree calculation is a standout between the most vital classification measures in information mining. Decision tree classifier as one sort of classifier is a stream diagram like a tree structure, where each inside hub indicates a test on a characteristic, each branch speaks to a result of the test, and each leaf hub speaks to a class. The technique that a decision tree demonstrate is utilized to group a record is to discover a way that from root to a leaf by estimating the characteristics test, and the trait on the leaf is classification result. The decision tree is the fundamental innovation utilized for classification and expectation.

In Decision Tree Classification a new example is classified by submitting it to a series of tests that determine the class label of the example. These tests are organized in a hierarchical structure called a decision tree. The tree is constructed in a top-down recursive divide and conquer approach. A decision node will have two or more branches and a leaf represents a classification or decision. The topmost node in the decision tree that corresponds to the best predictor is called the root node, and the best thing about a decision tree is that it can handle both categorical and numerical data.

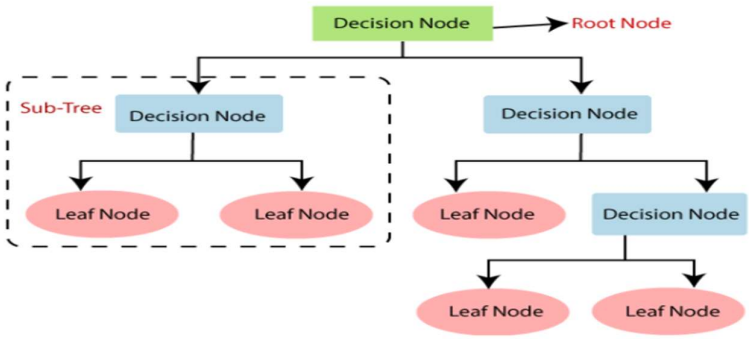



**Fig 3.8** Structure of Decision Tree

**Advantages and Disadvantages**

A decision tree gives an advantage of simplicity to understand and visualize, it requires very little data preparation as well. The disadvantage that follows with the decision tree is that it can create complex trees that may bot categorize efficiently. They can be quite unstable because even a simplistic change in the data can hinder the whole structure of the decision tree.

### 3.4.4 Naïve Bayes

Naïve Bayes methods are supervised learning methods based on the Bayes' Theorem and is perhaps one of the most widely used classification model to deal with real world problems such as spam mail filtering, classifying text articles into classes. It is mainly used in *text classification* that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. To put the description in layman's terms, Naïve Bayes assumes the features are all independent of each other and calculates the probabilities of each attribute and it selects the outcome with the highest probability.

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

- **Naïve**: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.

- **Bayes**: It is called Bayes because it depends on the principle of Bayes' Theorem. Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

  Following is the Bayes theorem to implement the Naive Bayes Theorem.

$$P(C_i|\, x_1, x_2 \ldots, x_n) = \frac{P(x_1, x_2 \ldots, x_n|C_i).P(C_i)}{P(x_1, x_2 \ldots, x_n)} \; for\; 1 < i < k$$

**Fig 3.9** Bayes theorem

**Advantages and Disadvantages**

The Naive Bayes classifier requires a small amount of training data to estimate the necessary parameters to get the results. They are extremely fast in nature compared to other classifiers.The only disadvantage is that they are known to be a bad estimator.

### 3.4.5 Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning**,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting**.**

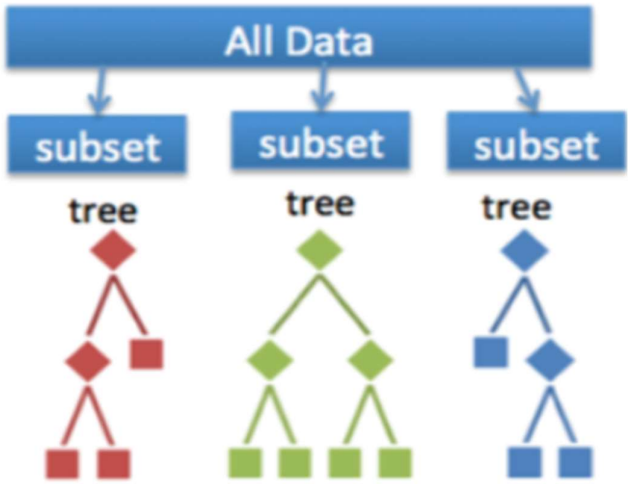



**Fig 3.10** Structure of Random Forest

**Advantages and Disadvantages**

The advantage of the random forest is that it is more accurate than the decision trees due to the reduction in the over-fitting. The only disadvantage with the random forest classifiers is that it is quite complex in implementation and gets pretty slow in real-time prediction.

### 3.4.6 Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

Support Vector Machine tries to find a hyperplane that separates two different classes such that the distance from the closest data point to that hyperplane is maximized.

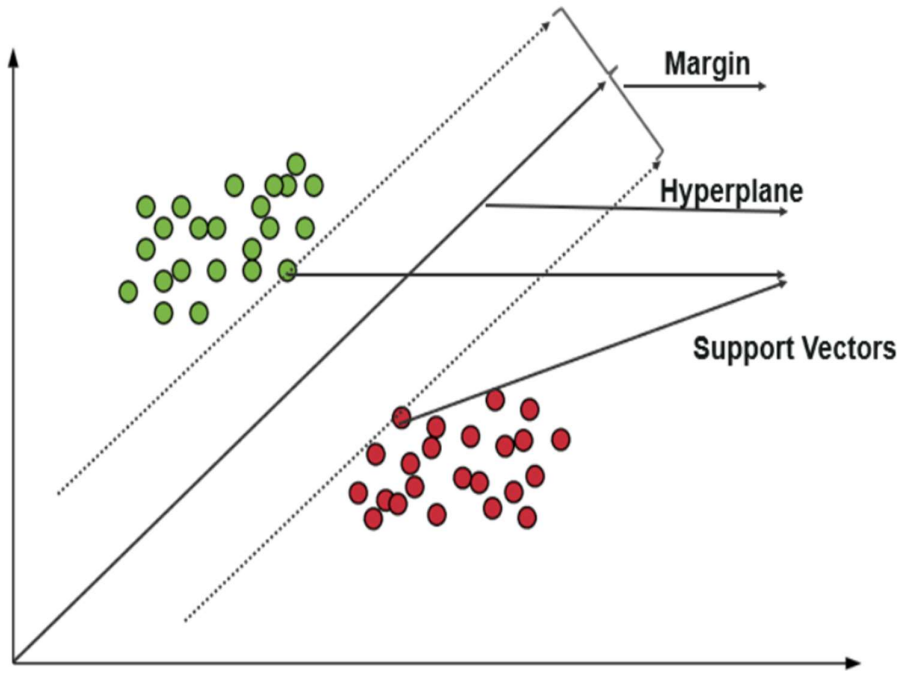



**Fig 3.11** Support Vector Machine

**Advantages and Disadvantages**

It uses a subset of training points in the decision function which makes it memory efficient and is highly effective in high dimensional spaces. The only disadvantage with the support vector machine is that the algorithm does not directly provide probability estimates.

### 3.4.7 FLOWCHART

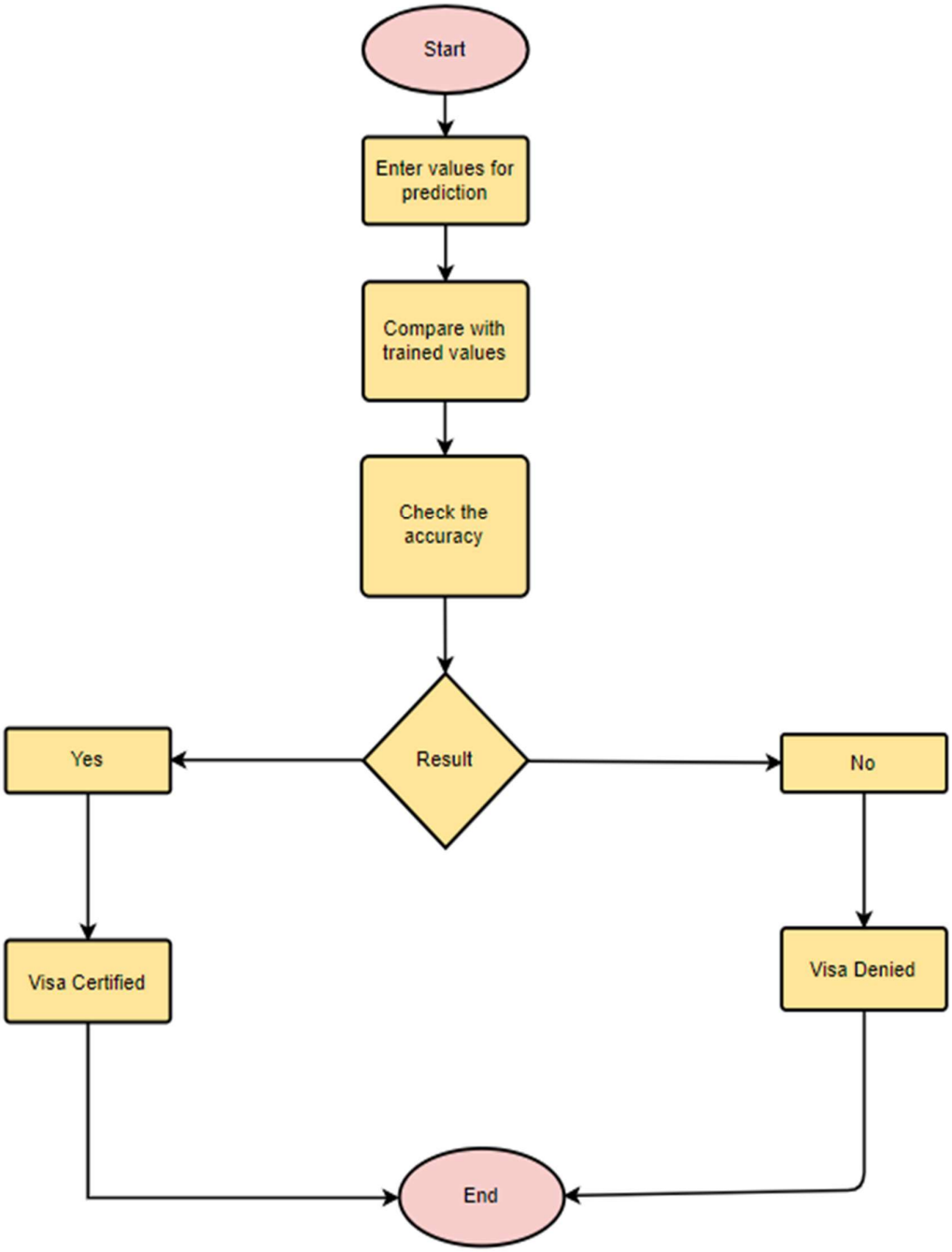



**Fig 3.12** Flow chart

# CHAPTER-4

# DESIGN

An in-depth explanation of the experiments conducted on the data set is given in this section.

## 4.1 DATA PREPARATION

For this project, supplemental data was downloaded from the U.S Bureau of Labor Statistics. It contains the state-wise median salary for each household. The percentile value of the "Prevailing Wage" column and the state-wise median salary was computed and this was used for further experiments. The attribute "Case Status" formed the target feature of this project. This attribute contains four valid entries – Certified, Denied, Certified-Withdrawn and Withdrawn. Further research of the application indicated that the values Certified-Withdrawn and Withdrawn were not beneficial for the analysis or the classification part of the project, hence the tuples with these values were removed from further processing. This converts the problem into a binary classification problem; the valid values of Case Status then became Certified and Denied. Apart from this, for data preparation data cleaning was performed to ensure the quality of further predictions.

Then the cleaned and compiled data set was split into training and testing data in a 70%- 30% basis. The classifier was built upon the training features and its performance was measured by the test data.The main goal of this project was to predict the "CASE STATUS" of a visa petition. All the other attributes of the data set form the array of features for training. The accuracy of the classification model was tested upon the "CASE STATUS" value.

## 4.2 UML DIAGRAMS

### 4.2.1 CLASS DIAGRAM

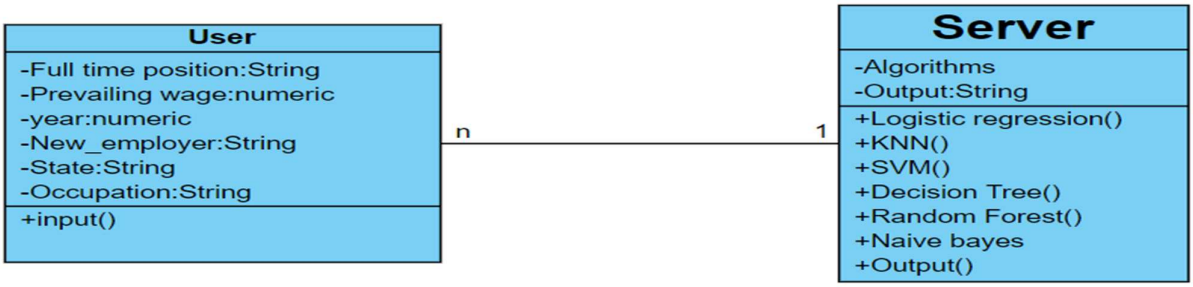



**FIG 4.1:**Class diagram

### 4.2.2 USE CASE DIAGRAM

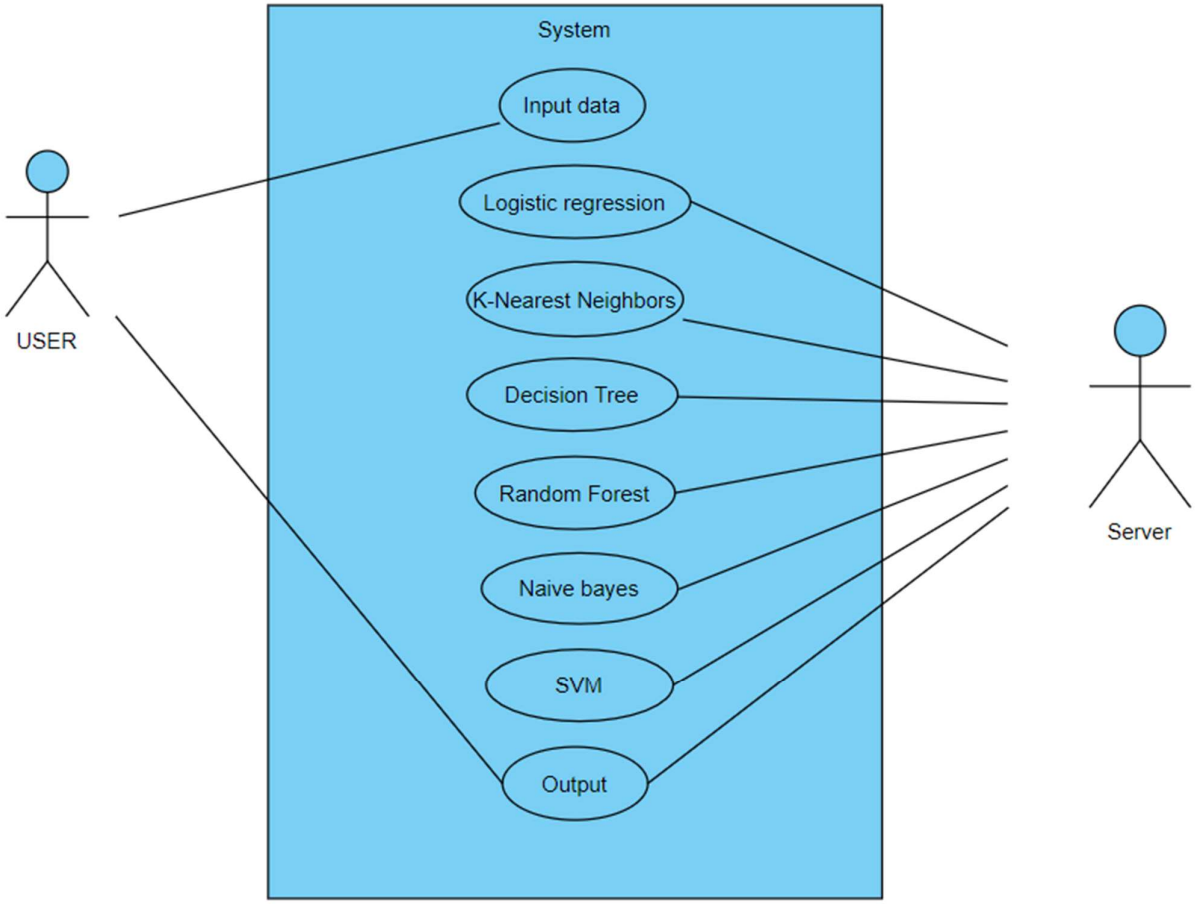



**FIG 4.2:**Use case diagram

### 4.2.3 ACTIVITY DIAGRAM WITH SWIMLANES

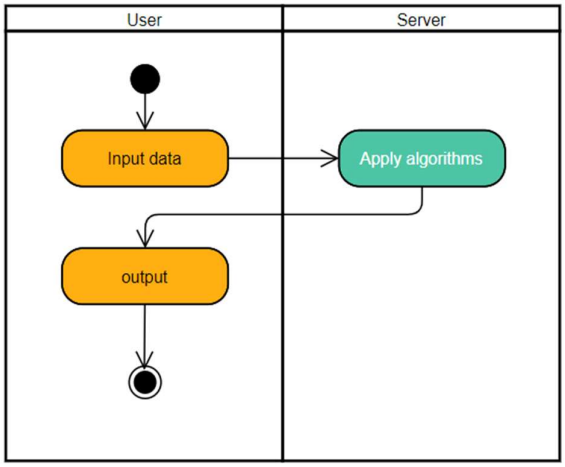



**FIG 4.3:**Activity diagram with swimlanes

### 4.2.4 SEQUENCE DIAGRAM

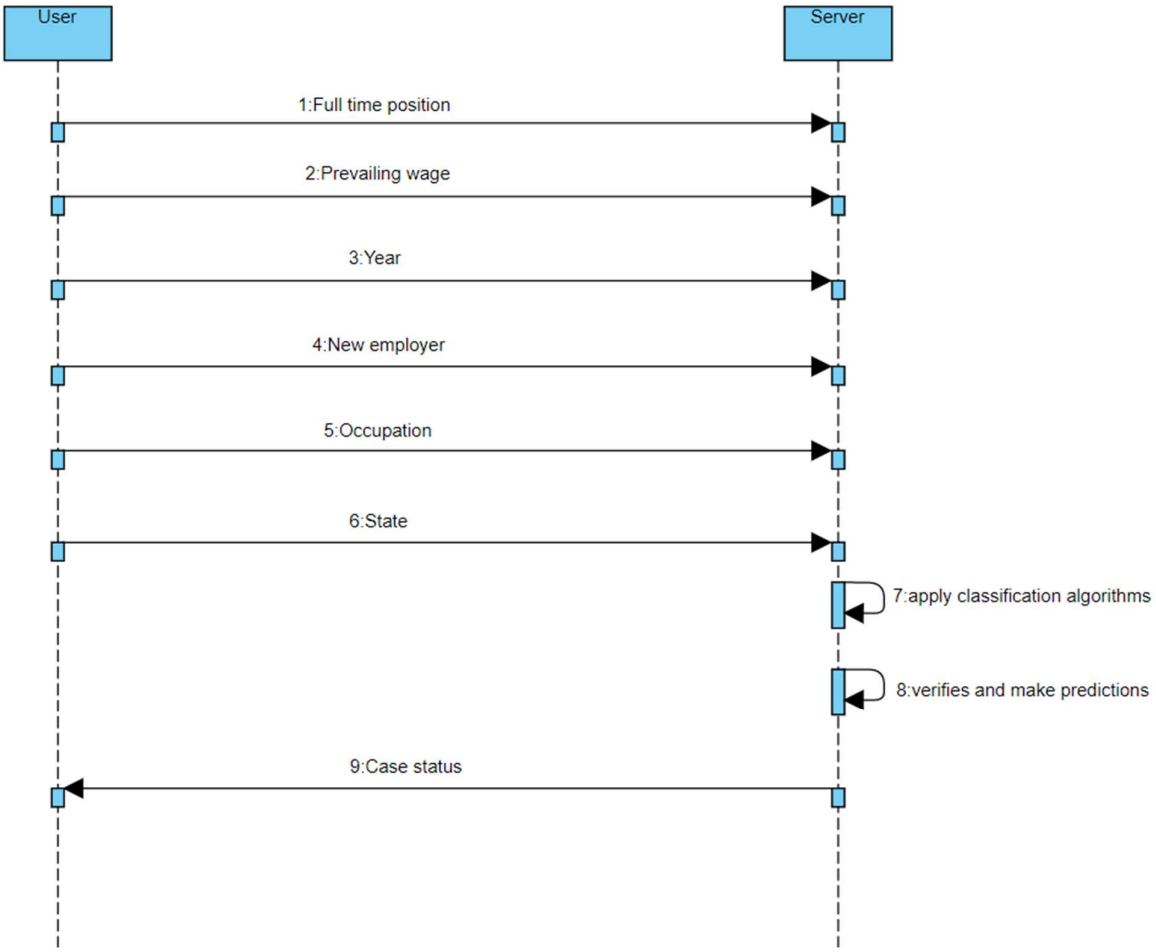



**FIG 4.4:**Sequence diagram

### 4.2.5 COMPONENT DIAGRAM

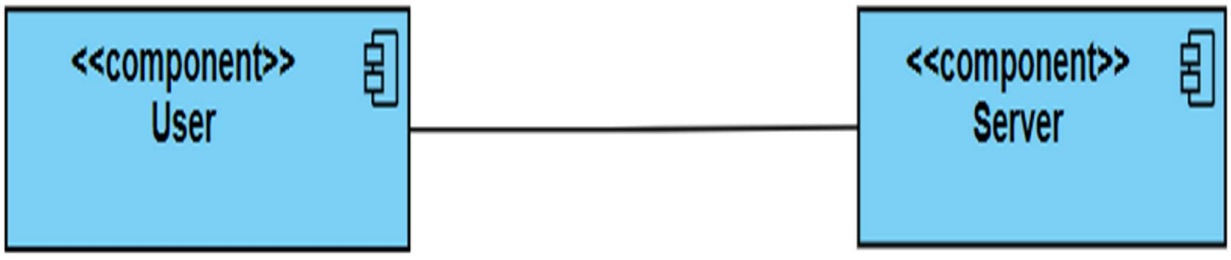



**FIG 4.5:**Component diagram

### 4.2.6 DEPLOYMENT DIAGRAM

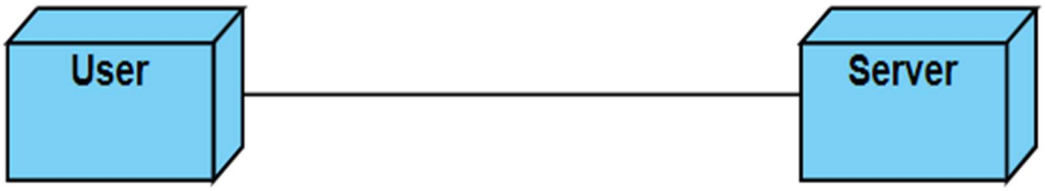



**FIG 4.6:**Deployment diagram

## 4.3 DESIGN OF PROJECT

Various machine learning algorithms discussed in Chapter 3 are used as the basis for the experiments conducted in this project. They are described in detail below. Python offers a package called Scikit-learn which has been used to implement all the inductive learning models. The evaluation criteria for the trained classifiers were: test set. These can help determine the accuracy of the classification task. These results were later manually compared to estimate which was the best classifier for this test bed.

#### 4.3.1 Logistic Regression

This algorithm was implemented by importing the library Logistic Regression from Scikitlearn in this way: from sklearn.linear_model import LogisticRegression. The classifier was then fit on the training features and labels. The
function predict was used to make the actual predictions for class labels.

#### 4.3.2 Random Forest

The algorithm Random Forests was implemented by importing the library in Scikit-learn as follows: from sklearn.ensemble import RandomForestClassifier. The depth, estimators and random state fields were specified before fitting the model and predicting the scores.

#### 4.3.3 Naïve Bayes

Similar methods were used to import the Naïve Bayes library from Scikit-learn, using: from sklearn.naive_bayes import GaussianNB. The function fit was used to fit the learning model on the data and the function score was used to find out the F-score of this algorithm and to assess its performance.

### 4.3.4 K-Nearest Neighbors

For implementing this algorithm, from sklearn.neighbors import KNeighborsClassifier is used.Methods stated earlier were used to assess the performance of this model as well.

### 4.3.5 Decision tree

For implementing this algorithm, from sklearn.tree import DecisionTreeClassifier is used. In the acode, we use classifier object, in which we will pass two main parameters;

- **"criterion='entropy':** Criterion is used to measure the quality of split, which is calculated by information gain given by entropy.
- **random_state=0":** For generating the random states.

### 4.3.6 Support Vector Machine

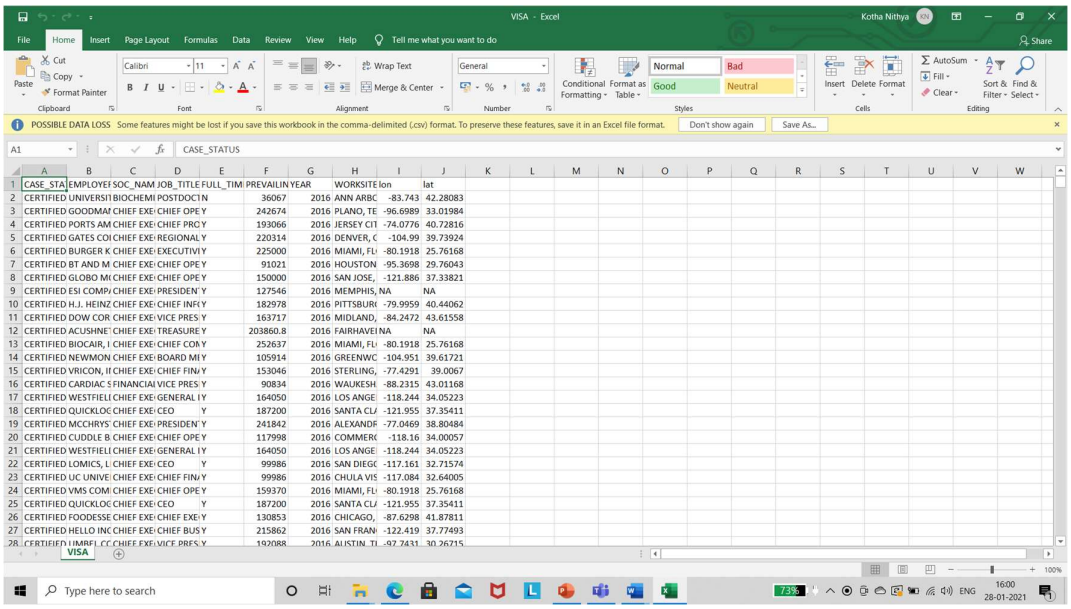
To create the SVM classifier, we will import SVC class from Sklearn.svm library. In the code, we use kernel='linear', as we are creating SVM for linearly separable data.

## 4.4 MODULE DESIGN AND ORGANIZATION

Our dataset is from Kaggle listed under the name “H-1B Visa Petitions 2011-2016 dataset”, and it initially included about 3 lakh data points. It contained 9 features and 1 class attribute which can be examined in following figure, and it was submitted by Sharan Naribole. All the data was originally extracted from the U.S. Office of Foreign Labor Certification’s iCERT Visa Portal System. This is an electronic system for filling and processing of applications of H-1B non-immigrant workers requested by their employers. The information in this data set consisted of data collected from all the fifty states in USA.

It is highly probable that a dataset in its raw form can include missing values, inconsistent entries and noise. These data entries affect the quality of the results. Thus, to improve the performance efficiency of the machine learning models, it is important to train and test the system on clean and consistent data. Redundant and unnecessary fields and values were also removed. Some of the steps followed in this / project for this purpose are: Data Cleaning, Feature Selection. We processed some of the existing features, created new features that we thought could be useful for prediction and discarded some features using the library Pandas. In particular, we noticed that the JOB_TITLE feature represents highly redundant information with the SOC_NAME feature, therefore we discarded JOB_TITLE from the beginning. Also, we transformed both SOC_NAME and COMPANY_NAME features into the corresponding forms of success rate and total number of

applications. Finally, we normalized all the features such that they had zero mean and unit variance. Each visa petition in the data set contains the following fields:



**Screen 4.1** Dataset

CASE_STATUS: We excluded the cases 'CERTIFIED-WITHDRAWN' and 'WITHDRAWN', since 'WITHDRAWN' decisions are either made by the petitioning employer or the applicant, therefore not predictive of USCIS's future behavior. We labeled 'CERTIFIED' cases as 0 and 'DENIED' cases as 1.

FULL_TIME_POSITION: Positions are given in "Full Time Position = Y; Part Time Position = N" format. We converted them to "Full Time Position = 1; Part Time Position = 0" format.

PREVAILING_WAGE: Prevailing wage is the average wage paid to employees with similar qualifications in the intended area of employment. We discarded the outlier terms and used the rest of the data as it was.

YEAR: Year in which application was filed.

NEW_EMPLOYER: EMPLOYER_NAME contains the names of the employers and there are lot of unique employers. It is the company which submits the application for its employee. We cannot use EMPLOYER_NAME directly in the model because it has got many unique string values or categories; more than 500 employers. These employers act as factors or levels. It is not advisable to use this many factors in a single column. We have created a new feature called NEW_EMPLOYER: If the employer name contains the string 'University' (for instance if a US university is filing a visa

petition, then it has more chances of approval for the employee). So, if the EMPLOYER_NAME contains 'University', then NEW_EMPLOYER contains the university value.All the strings in EMPLOYER_NAME containing the keyword university will have 'university' as value in the NEW_EMPLOYER column. All the remaining empty rows will be filled with 'non university'.

OCCUPATION: SOC_NAME, it consists of an occupation name. There are lot of values associated with SOC_NAME, so we have created a new feature that will contain the / important occupation of the applicant, mapping it with the SOC_NAME value. We created a new variable called OCCUPATION. For example computer, programmer and software are all computer ocupations. This will cover the top 80% of the occupations, and minor and remaining occupations will be categorized as others.

state: Since visa applications majorly depend on State location, you should split the state information from the WORKSITE variable.

Lat : This field denotes the latitude coordinates of the worksite. since it is not useful for prediction we have dropped that column.

Lon : This field denotes the longitude coordinates of the worksite. since it is not useful for prediction we have dropped that column.

# CHAPTER 5

# IMPLEMENTATION AND RESULTS

## 5.1 INTRODUCTION

Project implementation (or project execution) is the phase where visions and plans become reality. This is the logical conclusion, after evaluating, deciding, visioning, planning, and finding the resources of a project. Technical implementation is one part of executing a project. Results basically refer to any particular output or end point that comes as a result of the completion of the activities and or processes that have been performed as part of the project or as part of a particular project component.

This chapter encompasses data pre-processing methods and implementation steps of this work.

## 5.2 METHOD OF IMPLEMENTATION

The task that I used as a testbed for supervised inductive learning of classification was from a Kaggle data science challenge, uploaded by Sharan Naribole, 2017. Thus, the primary data set was downloaded from Kaggle and the supplementary data from the U.S. Bureau of Labor Statistics. In the below diagram, the process flow of the project can be found.

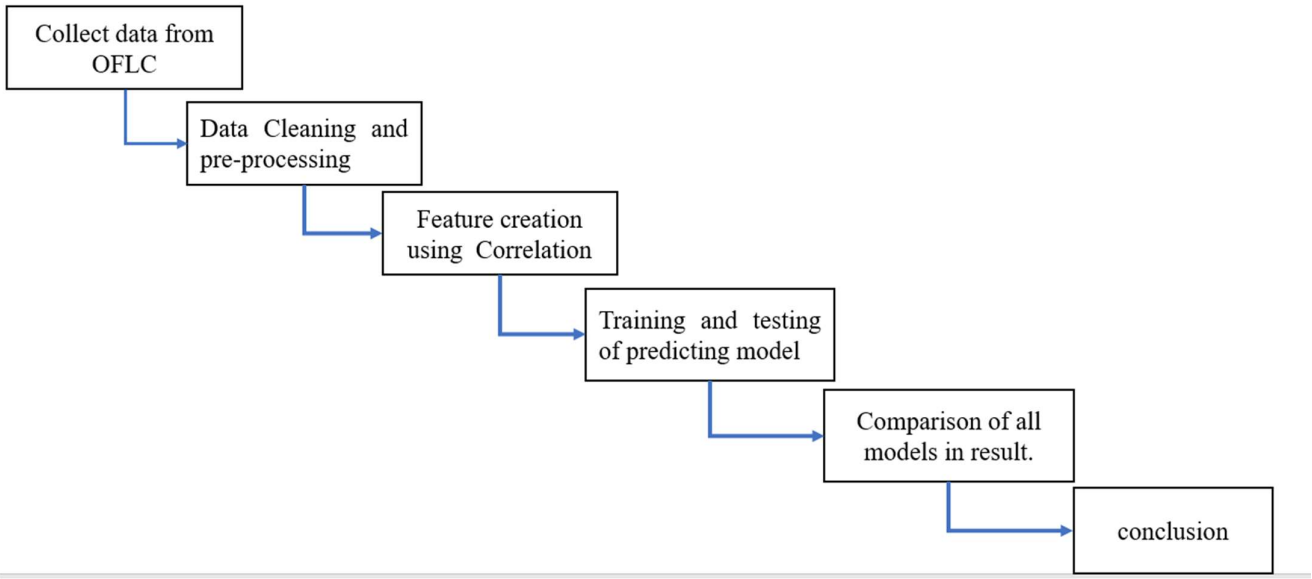



**Fig 5.1** Flowchart of implementation

In the initial step, the H1B visa dataset is taken from The Office of Foreign Labor Certification (OFLC) official site with attributes. It is likewise accessible on Kaggle site . The second step is data cleaning and pre-handling

under which initially copy esteems and missing esteems are evacuated. We used Microsoft Excel to clean adjust and standardize our dataset. The initial phase in cleaning the dataset was to evacuate accentuations to sustain a tactical Space from mistakes. We expelled the accompanying accentuations from our dataset: Subsequent to evacuating accentuations, we also divided the dataset and continue to our next ventures to cleaning it. Irrelevant qualities can delude the classifier into building an off-base model for foreseeing exactness. Qualities with exceptional esteem, It would not help us in examining the informational index. Such a characteristic is additionally called a False Predictor. We kept Employer_State and Worksite_State since there are 50 states and 5 US regions, we can examine class variable in light of that. Job_Title, SOC_Code, and SOC_Name were all giving the same word related data. In this way, we expelled Job_Title and SOC_Code and kept OCCUPATION. Dataset had 20 irrelevant properties that were expelled.

In the next step, forecasting models are prepared and tried on the date set with insights measure that are utilized for models correlations and help in anticipating best model out of the considerable number of models. In this paper, we utilize seven machine learning models. Built-in models are available in tool and are utilized straight forwardly on the dataset just by introducing required model from the library.

### 5.2.1 DATA PRE-PROCESSING

It is highly probable that a data set in its raw form can include missing values, inconsistent entries and noise. These data entries affect the quality of the results. Thus, to improve the performance efficiency of the machine learning models, it is important to train and test the system on clean and consistent data. Redundant and unnecessary fields and values were also removed.

Some of the steps followed in this project for this purpose are:

**Data Cleaning**

Some petition rows in the data set contain some missing values, to handle such situations, the missing fields have either been filled with dummy values or the entire row has been discarded, depending on the best suited task. Some features such as ID are not necessary for any computations, so they have been completely ignored. Some computing tasks required features to be pre-processed by binary thresholding.

**Feature Selection**

The machine learning models were trained on the most important features of the data set, such as SOC_NAME, PREVAILING_WAGE and WORKSITE to see if selecting this subset of features improves the performance metrics of the learning models. The models were trained on various subsets of features before the above-mentioned subset was deemed to be the better one from the lot. Some

functionalities of the project required only a subset of the features, for such cases, the rest of the features were ignored.

**Training:**

After the before steps are completed, you then move onto what is often considered the bulk of machine learning called training where the data is used to incrementally improve the model's ability to predict.

The training process involves initializing some random values for say A and B of our model, predict the output with those values, then compare it with the model's prediction and then adjust the values so that they match the predictions that were made previously.This process then repeats and each cycle of updating is called one training step.

**Evaluation:**

Once training is complete, you now check if it is good enough using this step. This is where that dataset you set aside earlier comes into play. Evaluation allows the testing of the model against data that has never been seen and used for training and is meant to be representative of how the model might perform when in the real world.

**Prediction:**

Machine learning is basically using data to answer questions. So this is the final step where you get to answer few questions. This is the point where the value of machine learning is realized. Here you can Finally use your model to predict the outcome of what you want.

The above-mentioned steps take you from where you create a model to where you Predict its output and thus acts as a learning path.

## 5.3 INPUT & OUTPUT SCREENS

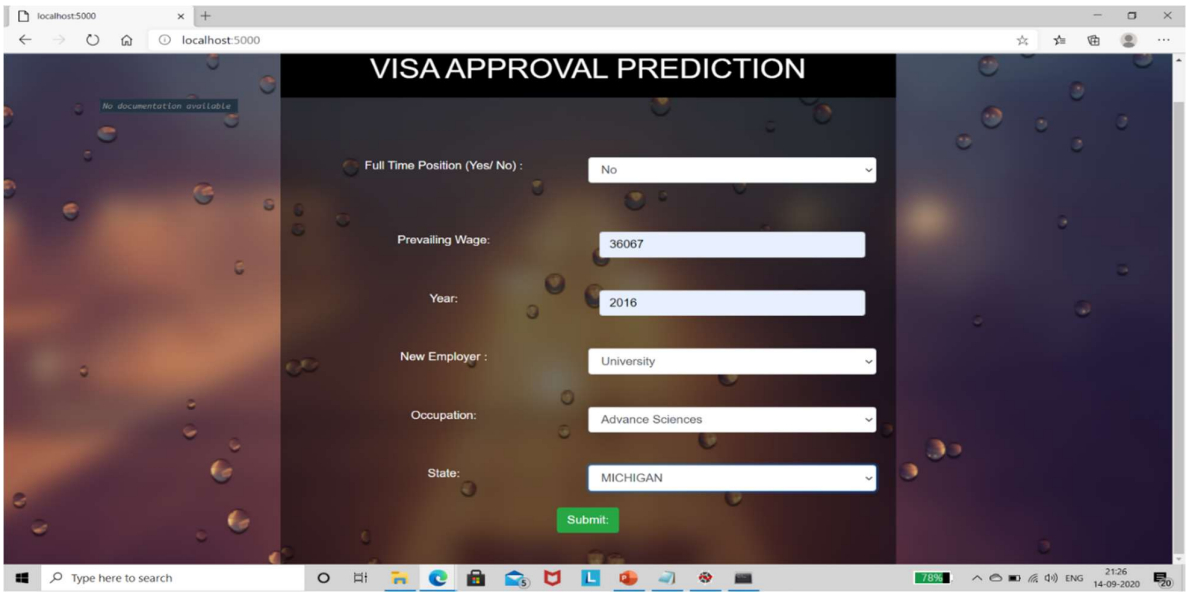
Flask is a web application framework written in python, it makes the process of designing a web application simpler. Flask lets us focus on what the users are requesting and what sort of response to give back. Flask depends on the Jinja template engine and the Werkzeug WSGI toolkit.

Flask is a lightweight WSGI web application framework. It is designed to make getting started quick and easy, with the ability to scale up to complex applications. It began as a simple wrapper around Werkzeug and Jinja and has become one of the most popular Python web application frameworks.

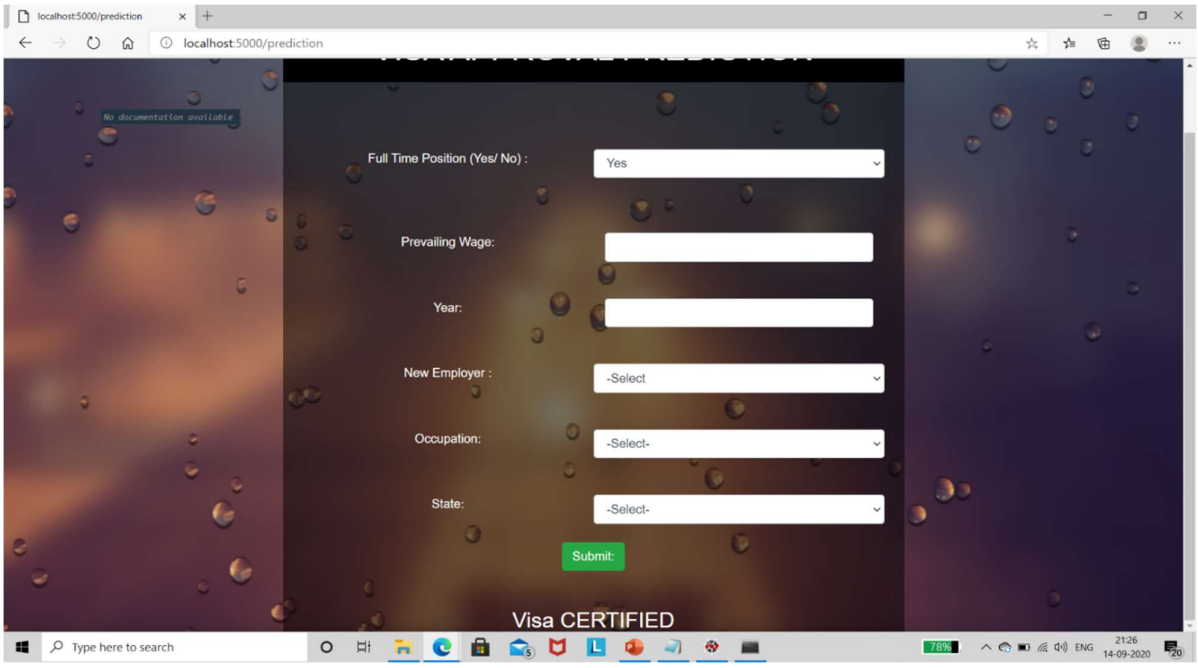
Flask offers suggestions, but doesn't enforce any dependencies or project layout. It is up to the developer to choose the tools and libraries they want to use. There are many extensions provided

by the community that make adding new functionality easy.

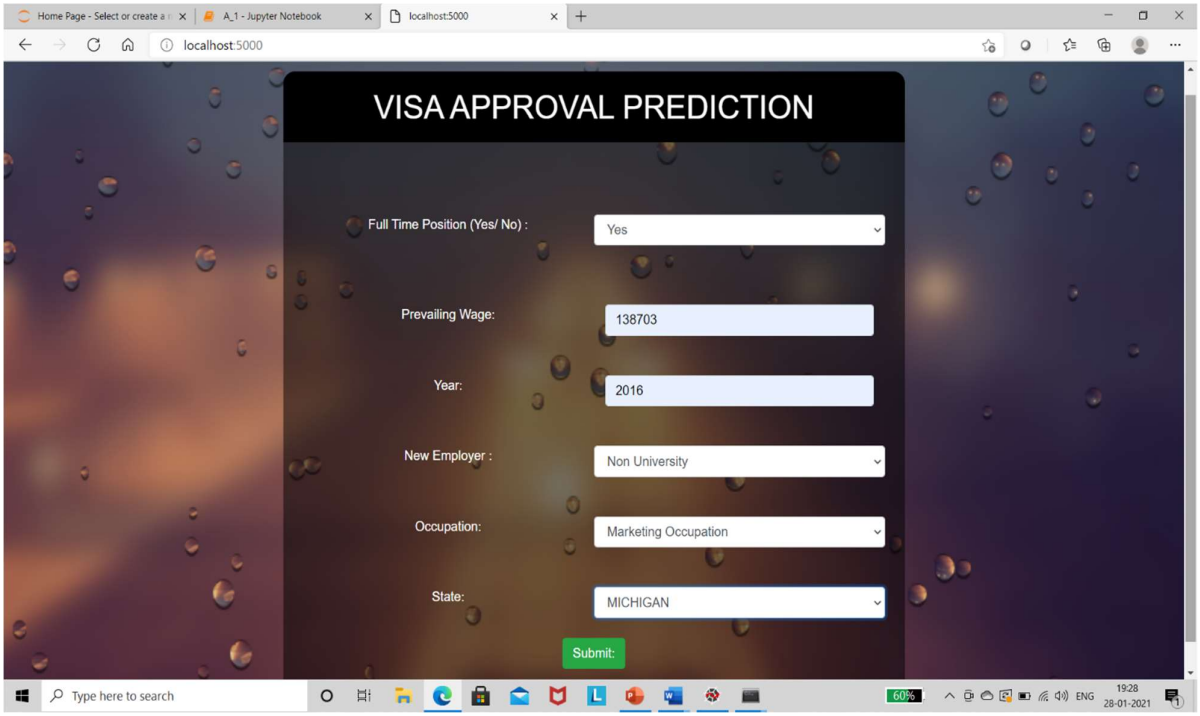
we have created an UI using the flask for visa approval prediction. The UI which we have created is shown below. When we save and run, the webpage will be opened in the local host. When the end user enter details and submit, the prediction is done. It predicts whether the visa got approved or denied. Our User interface displays all the attributes which are required for the prediction of case status. when the end user gives the necessary attribute values then based on the training data the UI will display whether the person's visa will get approved or rejected.



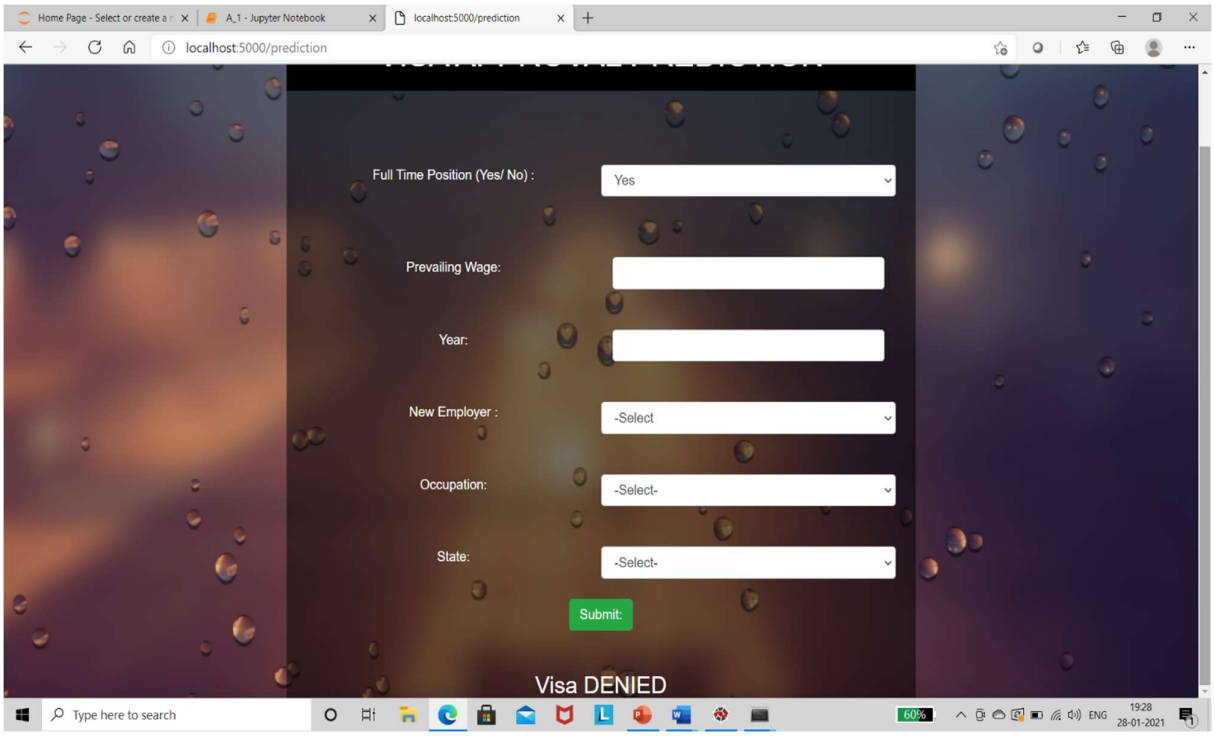

**Screen 5.1** Input-1




**Screen 5.2** Output-1




**Screen 5.3** Input-2




**Screen 5.4** Output-2

## 5.4 RESULT ANALYSIS

The results obtained for the algorithms discussed in Chapter 4 are explained here using their performance metrics.

### 5.4.1 Logistic Regression

The Accuracy scores obtained when this model was run is given in the table below:

| PERFORMANCE METRIC | VALUE |
|---|---|
| Accuracy | 60% |
| Auc_roc | 0.52 |

**Table 5.1** Performance metrics of Logistic Regression

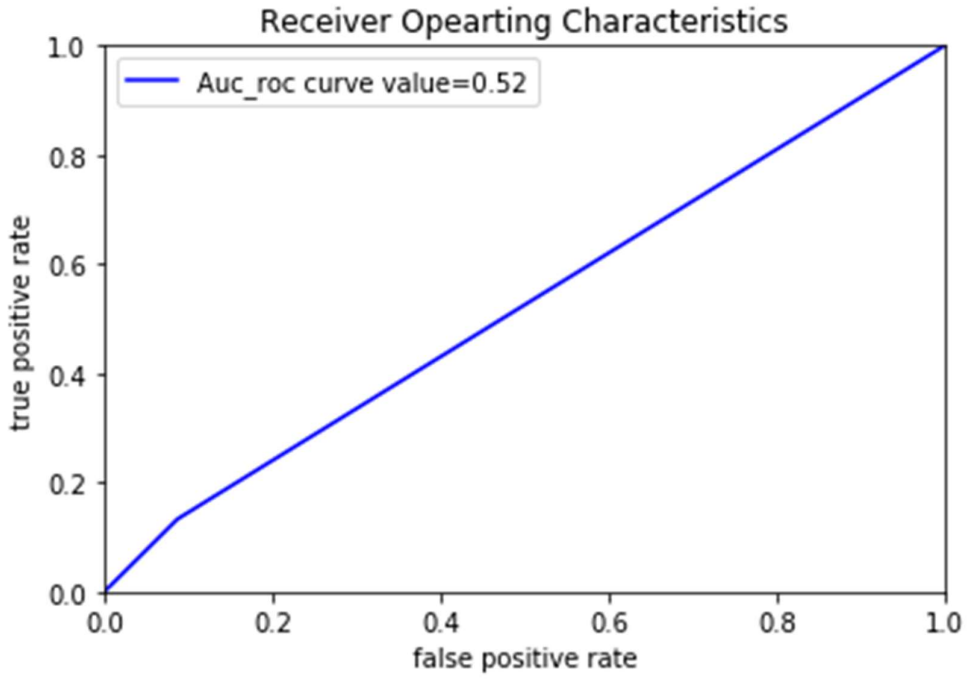



**Fig 5.2** Auc_roc curve of Logistic Regression

### 5.4.2 K – Nearest Neighbors

The Accuracy scores obtained when this model was run are given in the table below:

| PERFORMANCE METRIC | VALUE |
|---|---|
| Accuracy | 63% |
| Auc_roc | 0.59 |

**Table 5.2** Performance metrics of K-Nearest Neighbors

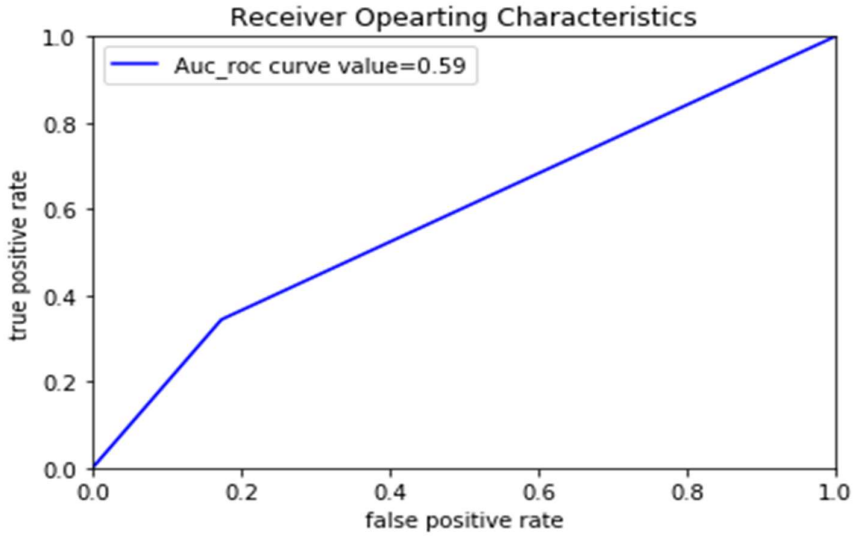



**Fig 5.3** Auc_roc curve of K-Nearest Neighbors

### 5.4.3 Decision Tree

The Accuracy scores obtained when this model was run are given in the table below:

| PERFORMANCE METRIC | VALUE |
|---|---|
| Accuracy | 74% |
| Auc_roc | 0.71 |

**Table 5.3** Performance metrics of Decision Tree

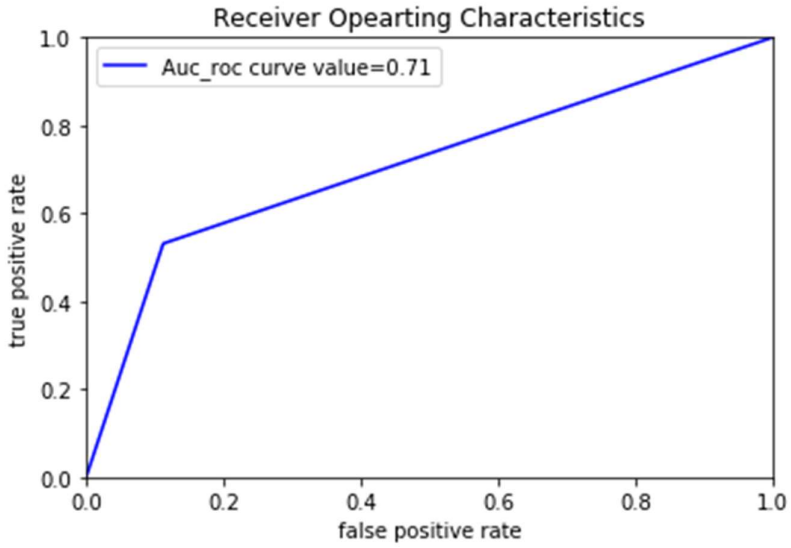



**Fig 5.4** Auc_roc curve of Decision Tree

### 5.4.4 Naïve Bayes

I have used the Gaussian Naïve Bayes classifier for this project. The Accuracy scores obtained when this model was run are given in the table below:

| PERFORMANCE METRIC | VALUE |
|---|---|
| Accuracy | 62% |
| Auc_roc | 0.57 |

**Table 5.4** Performance metrics of Naïve Bayes

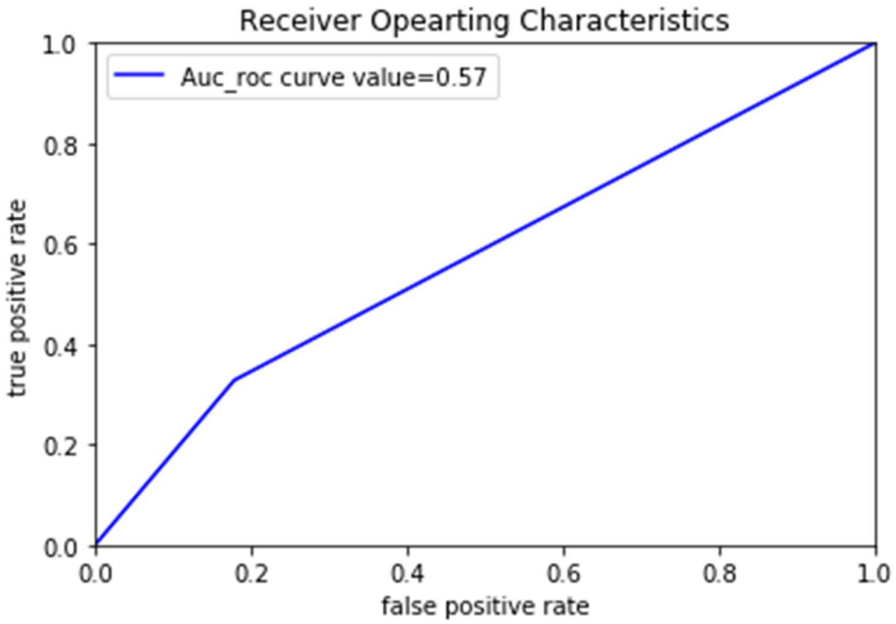



**Fig 5.5** Auc_roc curve of Naïve Bayes

### 5.4.5 Random Forest

The parameter for this algorithm is: estimators = 10. After a series of experiments, this value is chosen as this gave the best prediction accuracy. The Accuracy scores obtained when this model was run are given in the table below:

| PERFORMANCE METRIC | VALUE |
|---|---|
| Accuracy | 74% |
| Auc_roc | 0.70 |

**Table 5.5** Performance metrics of Random Forest

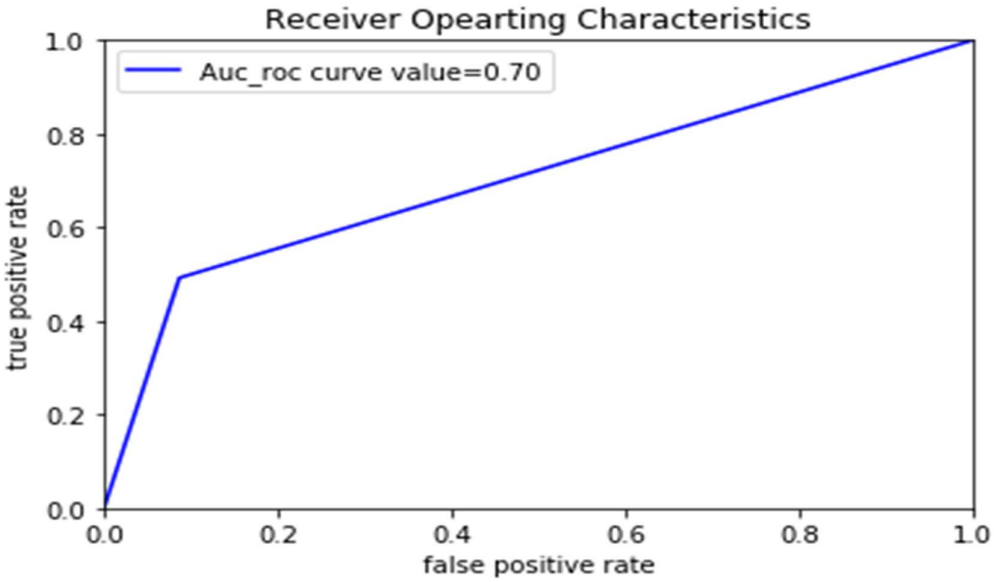



**Fig 5.6** Auc_roc curve of Random Forest

### 5.4.6 Support Vector Machine

The Accuracy scores obtained when this model was run are given in the table below:

| PERFORMANCE METRIC | VALUE |
|---|---|
| Accuracy | 62% |
| Auc_roc | 0.55 |

**Table 5.6** Performance metrics of Support Vector Machine

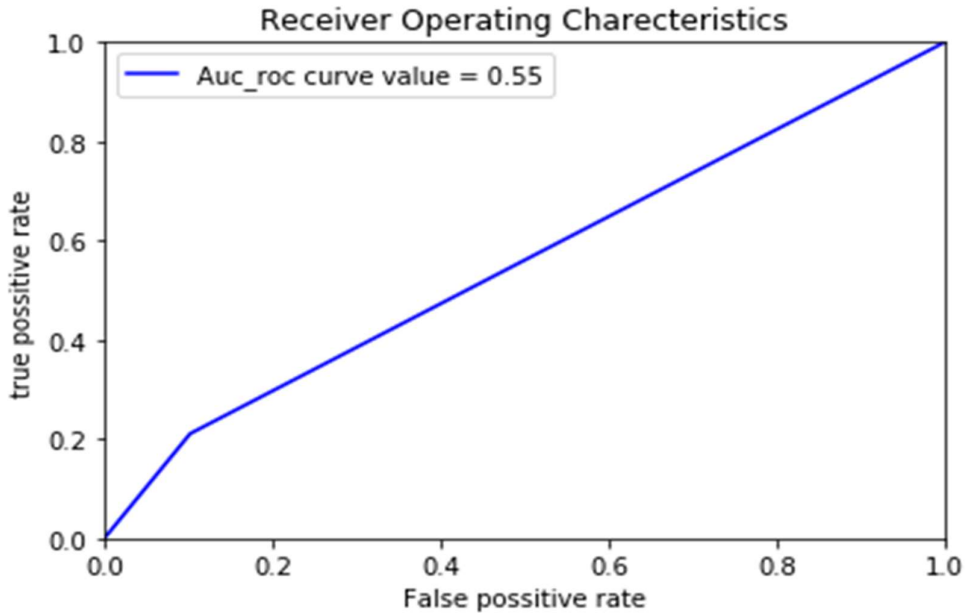



**Fig 5.7** Auc_roc curve of Support Vector Machine

## 5.5 Comparison of all the classifiers

The performance metrics for all the models were compared and it was observed that Decision Tree is best to train and predict compared to the others.

From the graphs in the above chapter it can be deduced that Random Forests and Decision Tree are the better performing algorithms for this test bed, with Decision Tree exceeding by a minute percentage. K-Nearest Neighbors, Logistic Regression, SVM and Naïve Bayes classifiers show a lower rate of performance. Listed below are the classifiers and their accuracies and Area Under Curve values obtained are as follows to predict over this data set.

| CLASSIFIERS | ACCURACY | AUC_ROC |
|---|---|---|
| Logistic Regression | 60% | 0.52 |
| K-Nearest Neighbors | 63% | 0.59 |
| Decision Tree | 74% | 0.71 |
| Naïve Bayes | 62% | 0.57 |
| Random Forest | 74% | 0.70 |
| Support Vector Machine | 62% | 0.55 |

**Table 5.7** Comparison of performance metrics

From the above analysis we can conclude that, the Decision Tree algorithm is best suitable to predict as its performance rate is higher when compared with another five machine learning methods namely the Logistic Regression, KNN, Naive Bayes, SVM and the Random Forest. The obtained results are displayed in Table below. The results show that, the performance of DecisionTree and Random Forest have comparable performance than that of Naive Bayes, K-nearest Neighbor, Logistic Regression and SVM, but the Decision Tree still performs the best, than the RandomForest. The ROC curve of the prediction model based on DecisionTree are all above 0.71, it means there is 71% chance that model will be able to distinguish between positive class and negative class, indicating that the model has ability of generalization.

# CHAPTER 6

# TESTING & VALIDATION

Steps of Training Testing and Validation in Machine Learning is very essential to make a robust supervised learning model. Training alone cannot ensure a model to work with unseen data. We need to complement training with testing and validation to come up with a powerful model that works with new unseen data.

It is essential to test our model on unseen data to check if it will generalize to new cases.
There can be two ways to check the performance of the data:

1. To generate the model and put it directly into the production, in this way we can see how it performs on new(unseen) data but if the model is not good the users will not be happy.

2. The other smart way to do it is to split the data into two parts and then use one to train the model whilst keeping the other for testing. The error rate produced by the test set is also called a generalization error.

We used the second method as it is more safe and reliable.we keep a chunk of data as training and the other for testing. Usually, we take the test data as 30% of the original data but it can be changed as per the requirement. Model Building is an iterative process and therefore once we build our model we keep improving it.

It is very common for people to train model and the test the results on test set. However, if accuracy is not coming on test set, we tend to tune hyper parameters and forcefully try to match training accuracy with test set. So, if you can realize we actually end up over fitting the model with test set indirectly. So again when it encounters new unseen data in production the model fails to deliver the accuracy it had promises on test set. In this approach, we initially do train test split like before, however from the training set we again set aside some portion – this portion is known as Validation Set. Based on the volume of available data this portion can be 10%-20% of your training data. After you feel the model is properly trained, as one final testing you can run the model on test set and obtain the performance accuracy. This final accuracy is going to be indicative of the accuracy you can expect in production.

In machine learning, model validation is referred to as the process where a trained model is evaluated with a testing data set. The testing data set is a separate portion of the same data set from which the training set is derived. The main purpose of using the testing data set is to test the generalization ability of trained model. Model validation is carried out after model training.

## 6.1 TEST CASES:

| CASE | RESULT |
|---|---|
| If all the required inputs are not given. | Then it does not display the output whether the visa is certified or not. |
| If negative values are given in year. | If negative values are given then the machine does not predict the output. |
| If prevailing wage is given in floating points. | The machine will predict the output. |
| If prevailing wage is given in negative number. | The machine will not predict the output as wage will not be in negative range. |
| If drop down is not given for full time position whether it takes numeric values. | No it will raise an error as full time position is declared as string |
| If drop down is not given for New _Employerwhether it takes numeric values. | No it will raise an error as New_Employer is declared as string |
| If the user enters the invalid data does it produce the output. | No it does not produce the output as every attribute is declared with its appropriate datatype. |

**TABLE 6.1** Test cases

# CHAPTER 7

# CONCLUSION

## 7.1 PROJECT CONCLUSION

In this work Logistic Regression, KNN, Decision Tree, Naïve Bayes, Random Forest and SVM were considered for determining the status of visa. Decision Tree Classifier performed the best in terms of accuracy and prediction. In the end, it is indeed possible to predict the outcome of H-1B visa applications based on the attributes of the applicant using machine learning.

The utilization of machine learning related algorithm enabled us to analyze the data based on training and learning steps. Therefore, it helped us to predict the approval and deny rate of H-1B visa of the testing year by Decision Tree. These data provides the situation of the H-1B visa during the recent years. The analysis of H-1B visa application might help to guard students, individuals to accomplish their American Dreams. There is no definitive guide of which algorithms to use given any situation. What may work on some data sets may not necessarily work on others. Therefore, always evaluate methods using cross validation to get a reliable estimates.

## 7.2 FUTURE ENHANCEMENT

This dataset presented interesting challenges for dealing with complex dataset. The data still needed to cleaned, and more Machine Learning algorithm, could exploited and implemented. Supplemental data concerning the Standard Occupational Classification (SOC) can be gathered and used in coordination with this data set to obtain a more comprehensive analysis of how the H-1B Visa selection process works. By using the wage evaluations and ranges under SOC, the wage attribute in this data set can be correctly put in to a range of salaries which can then be used to classify the visa petitions based on occupation roles rather than location wise. In addition, other classification algorithms other than the discriminative models can be experimented with this tested and their performances can also be analyzed.A rigorous analysis of other machine learning algorithms other than these six can also be done in future to investigate the power of / machine learning algorithms for visa approval prediction.

# REFERENCES:


1. H-1B Visa Petitions 2011-2016 — Kaggle. [Online]. Available: https://www.kaggle.com/nsharan/h-1b-visa/data. [Accessed: 20-Oct2017].
2. A. Y. Ng, M. I. Jordan. (2001). On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes. NIPS'01 Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, (pp. 841- 848).
3. F. Harrell. (2017, January 15). Classification vs. Prediction. Retrieved April 01, 2018, from Statistical Thinking: http://www.fharrell.com/post/classification/
4. Bound, J., Khanna, G. and Morales, N., (2017). Understanding the Economic Impact of the H-1B Program on the US. In High-Skilled Migration to the United States and its Economic Con-sequences. University of Chicago Press.
5. C Tom M. Mitchel, Mc GrawHil, Decision Tree Learning, Lecture slides for textbook Machine Learning,197
6. "Predicting Case Status of H-1B Visa Petitions." [Online]. Available: https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a054.pdf.
7. A. Ng, "CS229 Lecture Notes" [Online]. Available: http://cs229.stanford.edu/notes/.
8. "H-1B Visa Data Analysis and Prediction by using K-means Clustering and Decision Tree Algorithms." [Online]. Available: https://github.com/Jinglin-LI/H1B-VisaPrediction-by-Machine-LearningAlgorithm/blob/master/H1B\%20Prediction\%20Research\%20Report.pdf
9. Qing-yun Dai, Chun-ping Zhang, Hao Wu, "Research of Decision Tree Classification Algorithm in Data Mining", Vol.9, No.5 (2016), pp.1-8
10. A. Liaw, M. Wiener, "Classification and regression by random forest", R news 2 (3), (2002) 18–22.
11. S.Naribole. (2017). H-1B Visa Petitions 2011-2016. Retrieved from Kaggle: https://www.kaggle.com/nsharan/h-1b-visa
12. Dan, Lucas, Samuel Kabue, and Sarah Neff. Project Alien Worker. PDF. Berkeley: UC Berkeley School of Information, April 2016. https://www.ischool.berkeley.edu/sites/default/files/projects/project_alien_worker_-_final.pdf.

# HELP FILE

The Theme of our Project is Machine Learning Techniques for Predicting Visa Approval Using python. In this we focus on machine learning technique for predicting Visa Approval Status through which we can build the Classification model i.e., Decision tree Algorithm.

- So, to predict this we need to use 3 codes i.e., Python Decision tree code, Flask and HTML code.
- Decision tree code, Prediction is to be executed in Jupyter notebook IDE, python3.
- App.py, HTML, need to be executed in SpyderIDE.
- In Decision Tree code, first we have to import the libraries, reading dataset and next pre-processing of data can be done.
- In this, we need to check the null values and replace them with mean, mode, median values, Droping unnecessary columns present in dataset.
- Label Encoding to convert text to binary values, splitting them as X and Y, Feature creation, applying Decision Tree Algorithm with entropy criterion for Classification.
- Predict the values and finding Accuracy to Decision Tree Algorithm.
- Now we imported and dumped the pickle library to output column Case – Status which gets saved in backend of Tensorflow.

Then by using this saved model we written a code for prediction.

- Our model is predicting the Case-Status for Visa Approval with the given input values.
- But we should copy the saved model path and paste in the code.
- It will be a problem for copying the path again and again.
- To overcome this problem, we use flask- an awesome tool for model deployment.
- Flask is a web application framework written in python.
- It is very easy to make APIs and develop webpage in UI.
- By using flask web application, we have created a UI and also HTML code is written to structure a webpage and its content.