A Minor Project (AI3270) **PROJECT REPORT** on

VoyageVista

Submitted to Manipal University Jaipur

Towards the partial fulfillment for the Award of the Degree of

B. Tech Computer Science and Engineering (Artificial Intelligence and Machine Learning)

2024-2025

By

Aditya Raj

229310470

Under the guidance of

Ms. Rishika Singh

____

Signature of Guide

Department of Artificial Intelligence and Machine Learning

School of Computer Science and Engineering

Manipal University Jaipur

Jaipur, Rajasthan

Table of Contents

ACKNOWLEDGMENT

We are deeply grateful to the Dean, Faculty of Engineering; the Associate Dean, Department of Computer Science; and the Head of the Department of Artificial Intelligence and Machine Learning, Manipal University Jaipur, for providing us with the opportunity to undertake this project as part of our curriculum.

We would like to express our sincere gratitude to our project supervisor, Ms. Rishika Singh, for her constant support, expert guidance, and valuable suggestions throughout the course of this project. Her encouragement kept us focused and motivated at every stage.

We also extend our thanks to the department project guide and the laboratory in-charge, whose infrastructure and technical support played a vital role in the smooth execution of our work. Their assistance made it possible to practically implement and test our ideas.

Finally, we wish to thank all the faculty members of the Department of Artificial Intelligence and Machine Learning who supported us directly or indirectly during the various phases of our project. Their expertise and feedback were highly appreciated.

ABSTRACT

In today's competitive global economy, obtaining an H-1B work visa to the United States has become increasingly difficult due to the high number of applications and strict approval criteria. As skilled professionals from across the world seek job opportunities in the U.S., the uncertainty surrounding visa approval creates anxiety and inefficiencies in the application process. This project addresses that challenge by aiming to build a predictive system that can estimate the likelihood of H-1B visa approval based on historical trends and applicant attributes.

To achieve this, we employed a data-driven methodology. Historical H-1B visa data (from 2011–2016) was sourced from Kaggle and preprocessed to extract relevant features including job title, employer name, prevailing wage, location, and full-time status. We explored various supervised learning algorithms such as Logistic Regression, Decision Trees, and Random Forests to model the approval patterns. After evaluating performance using accuracy and F1-score, the best-performing model was selected for deployment.

We developed a web application using Flask that integrates the trained machine learning model to offer real-time predictions to users. The interface allows users to input key attributes of their visa profile and receive a probability estimate of approval. The UI was designed to be simple, informative, and accessible to a wide range of users.

This system provides valuable insights to applicants, helping them understand and potentially improve their visa profiles before applying. It also serves as a foundation for future expansions, such as covering other visa types, integrating real-time data, or extending the application to immigration systems of other countries.

Introduction

What:

This project aims to predict the outcome of someone who is going to apply or has ever applied for work visa for U.S. 1.e. H-113 visa using machine learning techniques. H- I B is a special visa given to foreign workers by the government for U.S. Every year thousands of applications are submitted, the approval process is very complex hence predicting the chances of visa approval is difficult and time taking. This project uses a model that can predict the approval likelihood of an
H-1 B visa application based on historical data using different machine learning techniques like Decision Tree, Random Forest, and Logistic Regression.

Why:

To apply for H-1B visa, an U.S employer must offer an job and petition for H-1B visa with the U.S. immigration department. This is the most common and legal visa status and for international students who complete their college / higher education (Master, PhD) and work in a full-time position. The status of II- I B visa will influence the life and work, and even the career of
the international students. Visa applicants often have a little insight into the process of approval, which is influenced by factors such as wage, employer, job title, location, etc. A predictive model can help to assess the chances of their approval, which can help applicants in decision making. This way, applicants can understand the factors better that affect the approval rate and optimize
their applications accordingly.

How:

This project uses 11-1B dataset available on Kaggle, it contains past data on H-1B visa applications from 2011 to 2016. The dataset contains data of applicants like CASE_STATUS
(approved or denied), EMPLOYER NAME, SOC_NAME, JOB_TITLE, PREVAILING WAGE, WORKSITE. These features will be used to train three types of machine learning models:

- Logistic Regression: It is used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.
- Decision Tree: It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
- Random Forest: It is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

Once the models are trained, the model with highest accuracy will be used and will be displayed via web application using python library flask, where users can give their visa application details and get prediction if their application will be approved or not, the project also predicts if the application if already submitted it is approved or rejected. The project is Application-based, it involves data analysis and machine learning for prediction pf visa approval and web development for deploying the model and make the project user friendly.

Objectives

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

Literature Survey

Comparative Analysis of Existing Software and Technologies Used

S. No.	**Title/Author**	**Year**	**Summary**	**Technology Used**	**Findings**
1.	Predictive Modeling for HI B Visa Approval Using Machine Learning by: 1) C. Kavitha Santhoshi, 2) G. Srilekha, 3) B. R. Ramadevi, 4) K. Chandrika.	2022	With the help of the random forest algorithm, they were able to improve and inform the data with accuracy. 86.88%.	Random Forest classifier.	The Random Forest performance best For visa prediction

2.	H1B visa approval using machine learning algorithms by: 1) Mrs. A. Durga Bhavani, 2) Guddeti Bharath, 3)Dubbaka Thaw Reddy	2022	With the help of Random Forest, they were able to get accuracy. 83.06%.	Random Forest Classifier.	The Random Forest performs Best for visa prediction.
3.	Predicting the Outcome of H1B Visa applications by: 1) Beliz Gunel, **2)** *Onur* Cezmi Mutlu.	2017	On an unbalanced dataset and neural network gave 98.5% precision accuracy	• Naive Bayes • Logistic regression • Support vector machine • Neural network	The Machine algorithm logistic regression works best for visa prediction
4.	Data analysis of H1B visa applications by Ratmak Roy.	2021	With the help of Logistic Regression, on a balanced dataset the algo gives an accuracy of 57%.	Logistic Regression.	On a balanced Dataset Logistic Regression did not perform that well.
5.	Predicting filed H 1 -B Visa Petitions' Status by Darshit A. Pandya.	2018	Logistic Regression and ANN classifier gives the best positive F I-Score of 92% and negative Fl -Score of 84% and 85% respectively.	• Decision tree • K-NN • SVM • Logistic Regression • Random Forest • CNN • Gaussian Naïve Bayes	Logistic Regression gave better result in comparison to Other machine learning models.

Planning of Work

4.1 Requirement Engineering

1. Interface Requirements:

   Home page: The home page allows users to enter their visa details such as:

   - Job Duration
   - Occupation Field
   - Prevailing Wage
   - Year of Application

   After submitting the details, the page will show them the predicted probability of visa approval.

2. Hardware Requirements:

   The project will run on a standard web server with minimum hardware requirements:

   - 2.0 GHz dual-core processor.
   - 4 GB RAM
   - 10 GB of free disk space

3. Other functional Requirements:

   - The project uses H-1B dataset from Kaggle, which contains details all the applicants and their outcomes over 2011-2016.
   - The dataset includes details such as CASE_STATUS, EMPLOYER_NAME, SOC_NAME, JOB_TITLE, PREVAILING_WAGE, WORKSITE, and geographic variables (longitude, latitude).

4.2 Design

I. System Design:

1. Frontend: The project uses HTML. CSS and JavaScript for designing and implementing web pages. Users will be able to enter their details and predict the probability of their visa application.
2. Backend: It uses Python library Flask for creating the web server and adding the functionality of the model, it contains the logic for processing user input, send it to the model and then return predictions.

3. Machine Learning model: Project uses Python library Scikit-learn for testing and training all the different models and selecting the best out of them.

2. User Interface Design:

- A form for users to enter their details such as job duration. occupation field, prevailing wage, and year of application.
- A “Submit" button for processing the input and return the prediction(certified/denied)

3. Model Integration:

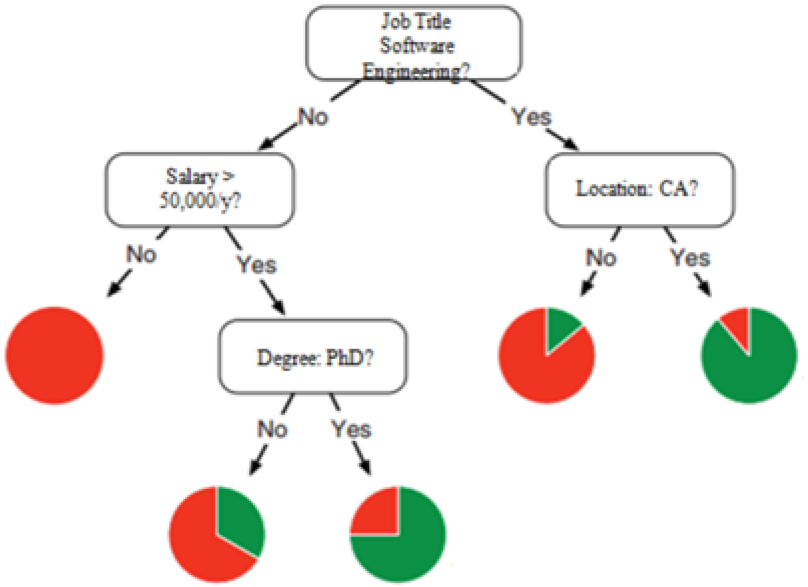
The model is integrated into the web page using python library Flask which hosts a web server and using basic web development, a web page is developed. The web page sends data to the flask and flask calls out the model, gives it all the details and then the prediction is sent forward to web page.

4.3 Theoretical Analysis:

1. Algorithms Used

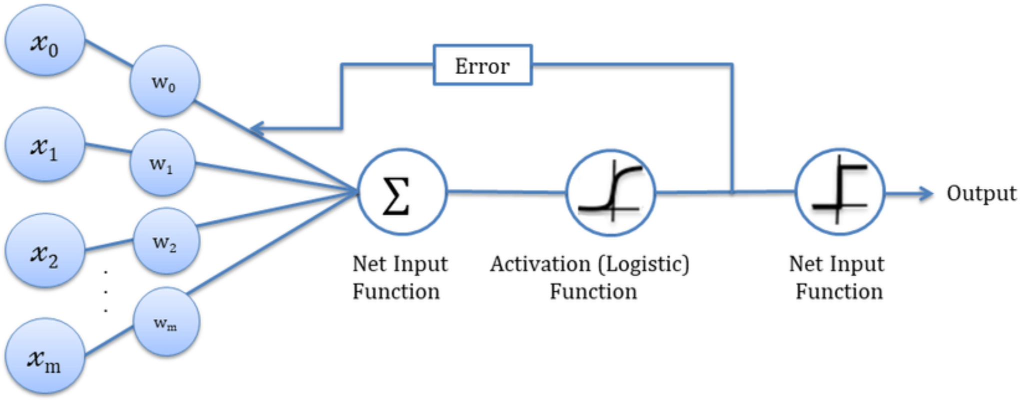
1.1 Decision Tree:

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represents the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome.



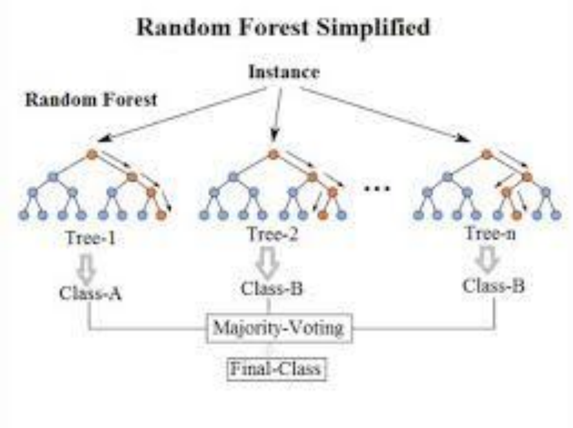
1.2 Logistic Regression:

Logistic regression is a supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not.



1.3 Random Forest classification:

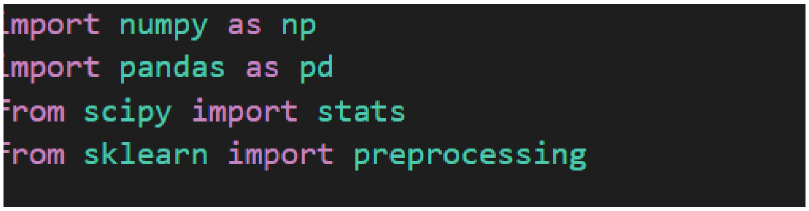
Random Forest algorithm is a powerful tree learning techniquein Machine Learning. It works by creating several Decision Trees during the training phase. Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition. This randomness introduces variability among individual trees, reducing the risk of overlitting and improving overall prediction performance.



4.4 Coding

1. Data Loading and Preprocessing:

- Import all the necessary libraries.



```
mport numpy as np
mport pandas as pd
rom scipy import stats
rom sklearn import preprocessing
```

- Load the dataset using numpy library.

- Drop unnecessary columns including employer name, job title, worksite, longitude and latitude,

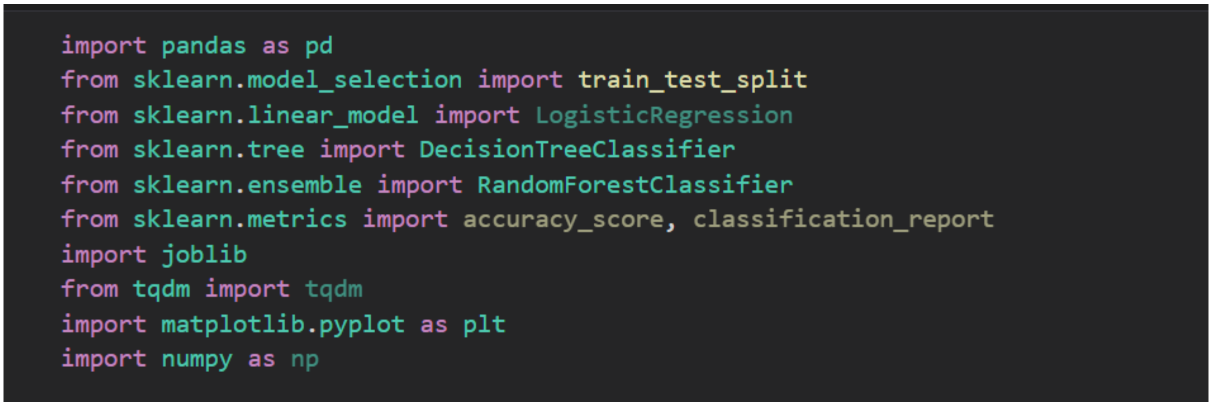
- Handle all the missing values and remove outliers.

Encode categorical variables using LabelEncoder.

- Save the pre-processed data for training.

2. Model Training:

- Import all the necessary libraries.



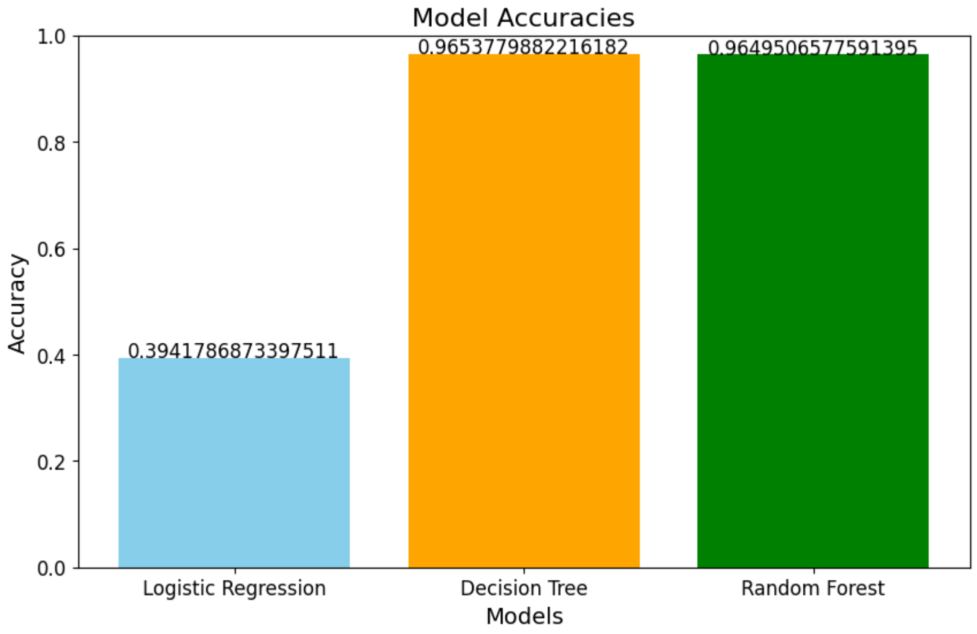
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report
import joblib
from tqdm import tqdm
import matplotlib.pyplot as plt
import numpy as np
```

- Load the pre-processed data.

- Split the dataset into training and test sets by the ratio of 75% and 25%.

- Train and save all the models including Logistic Regression , Decision Tre and Random Forest using Scikit-learn.

3. Model Evaluation:



- Evaluate the performance of all the models using accuracy, precision. recall. sl- score and support.
- Amongst all the trained and saved models, select the model having the highest accuracy.

4. Web Deployment:

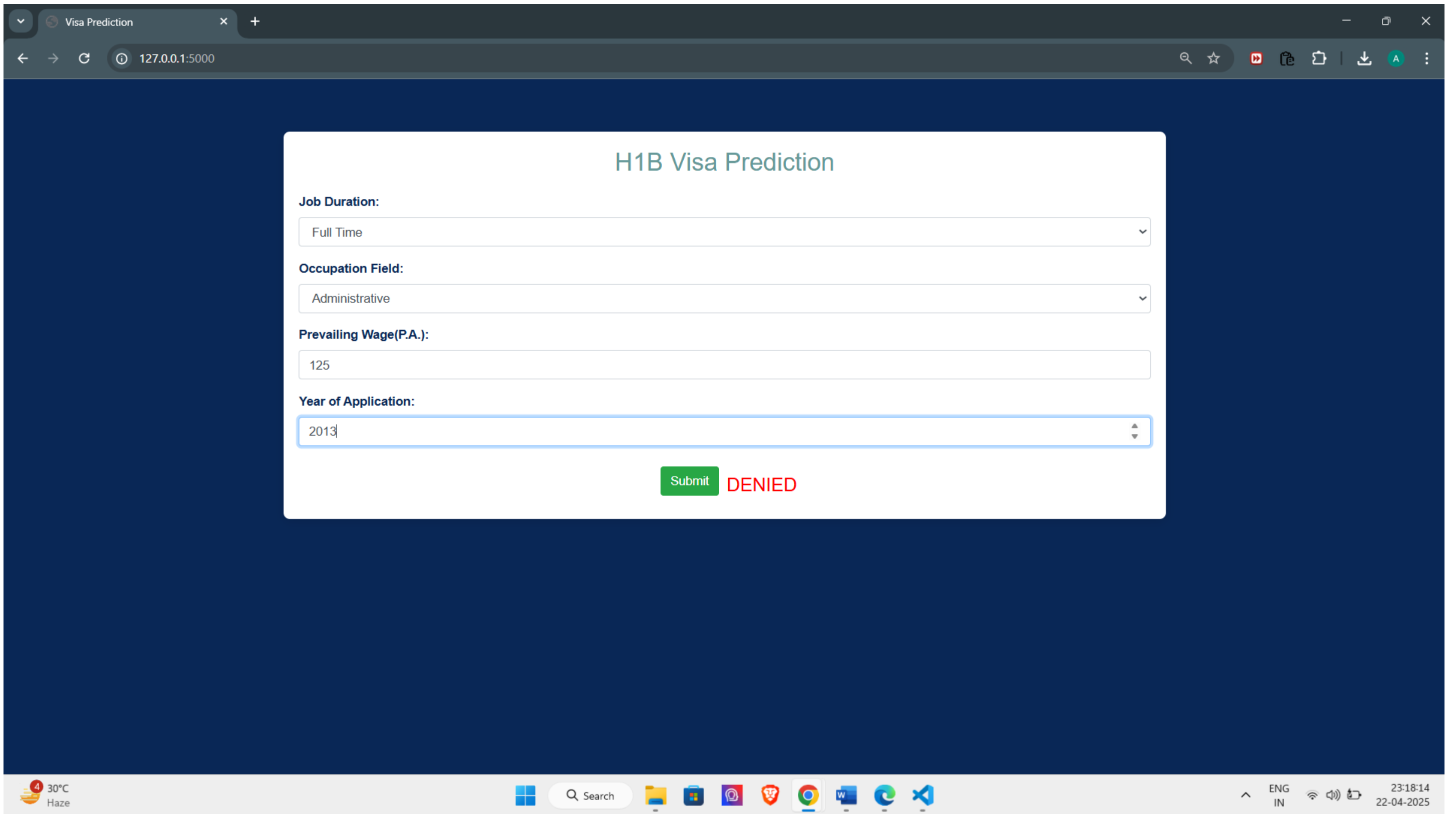
- Create a Flask app to handle the processing of user input, sending the user input fetched From the web page to the model, getting the prediction and then displaying it onto the web page.

- Now run app.py tile, and open http://127.0.0.1:5000/ on the browser to see your project running.

5. Testing and Debugging:

- To test the project open http://127.0.0.1:5000/ on your web browser and test the web application by applying various inputs to ensure that the predictions are correct, and the product is Functioning as expected.

6. UI:



H1B Visa Prediction

Job Duration: Full Time

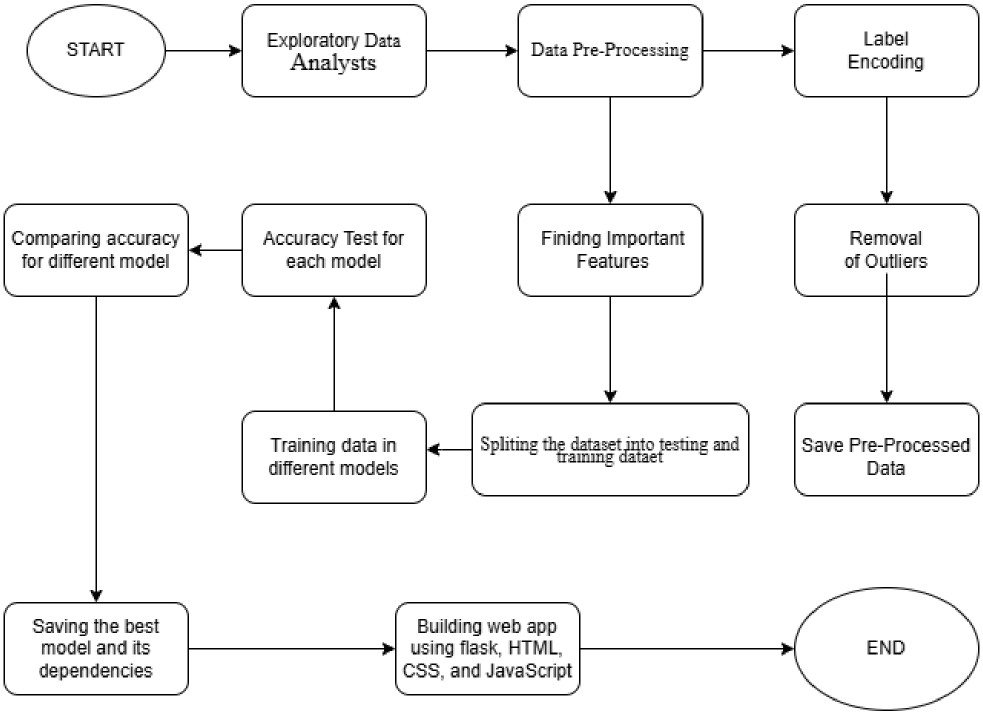
Occupation Field: Administrative

Prevailing Wage(P.A.): 125

Year of Application: 2013

Submit DENIED

4.5 Flow Chart:


START
Exploratory Data Analysts
Data Pre-Processing
Label Encoding
Comparing accuracy for different model
Accuracy Test for each model
Finidng Important Features
Removal of Outliers
Training data in different models
Spliting the dataset into testing and training dataet
Save Pre-Processed Data
Saving the best model and its dependencies
Building web app using flask, HTML, CSS, and JavaScript
END

Future scope:

In future, other ML and DL algorithms can be tested and used that give us more accuracy and precision, dataset used in this project is not recent due to unavailability of recent data, but by using new dataset with latest information, the output can be maximised. Also, this project currently supports H-lB visa prediction only, but with enough data available models can be trained on other type of visa also such as travel visa, student visa, etc. also we can train our model on the data of different countries also not just US. The future scope of this project is very broad.

Conclusion:

The project VoyageVista successfully predicts if a user's visa application will be approved or rejected with the accuracy of 89.8%, there is room for improvement but for now with available knowledge and data the project has been made successfully, The project takes the data from the user using web page, and at backend feeds it to the model trained and then the model predicts the possibility of visa being accepted or rejected.

REFERENCES:

Journal / Conference Papers

[1] A. Mishra and S. Sharma, "Prediction of H1B Visa Using Machine Learning Algorithms," International Journal of Recent Technology and Engineering, vol. 9, no. 1, 2020, pp. 59-120.

[2] V. Sharma and R. Kumar, "An allotment of H1B work visa in USA using machine learning," International Journal of Engineering and Advanced Technology, vol. 8, no. 1, 2018, pp. 488-339.

[3] P. Singh, K. Sharma, and A. Gupta, "H-1B Visa Approval Prediction Using Random Forest Algorithm," Proceedings of the International Conference on Computing and Communication Systems, NIT Silchar, India, March 18-20, pp. 49-54, 2019.

[4] J. Wang, S. Li, and T. Chen, "Predicting H-1B Visa Application Outcomes with Ensemble Methods," IEEE Transactions on Computational Social Systems, vol. 6, no. 5, 2019, pp. 56-67.

[5] M. Patel and S. Desai, "A Predictive Analysis of H1B Visa Petitions using Machine Learning," International Journal of Innovative Research in Science, Engineering and Technology, vol. 7, no. 10, 2018, pp. 37-45.

[6] D. Liu and K. Singh, "Machine Learning Approach for Visa Application Status Prediction," Journal of Emerging Technologies and Innovative Research, vol. 9, no. 4, 2022, pp. 278-285.

Reference / Handbooks

[1] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction," MIT Press, 2nd Edition, ISBN 978-0262039246.

[2] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning," Springer, 2nd Edition, ISBN 978-0387848570.

Web

[1] H-1B Visa Dataset, Kaggle, Last Accessed April 2025.

[2] H-1B Employer Data Hub, USCIS Official Website, Last Accessed April 2025.

[3] Machine Learning Models for Immigration Cases, Stanford CS229 Course Projects, Last Accessed March 2025.

[4] US Department of Labor Foreign Labor Certification Data Center, Office of Foreign Labor Certification, Last Accessed April 2025.