**📑 EasyVisa Visa Approval Prediction – Final Report**

**1. Introduction**

Every year, thousands of foreign workers apply for employment-based visas to work in the United States. These applications go through a labor certification process, where each case is either **Certified** (approved) or **Denied** by the U.S. Department of Labor. The decision depends on many factors, including the applicant’s education, job experience, wage offered, and the company’s profile.

This project uses machine learning to predict whether a visa application will be approved or denied. By analyzing historical data, we aim to uncover patterns that influence approval decisions and build a model that can make accurate predictions. This can help employers prepare stronger applications and guide applicants toward better opportunities.

**2. Methodology**

**2.1 Data Collection and Preparation**

* **Dataset Used:** The EasyVisa dataset, which contains real labor certification cases. Each row represents one visa application.
* **Key Columns:**
  + education\_of\_employee: Highest education level of the applicant.
  + has\_job\_experience: Whether the applicant has prior work experience.
  + prevailing\_wage: Wage offered to the applicant.
  + region\_of\_employment: U.S. region where the job is located.
  + yr\_of\_estab: Year the sponsoring company was established.
  + case\_status: Final decision (Certified or Denied).
* **Cleaning Steps:**
  + Filled missing numeric values (e.g. wage, employee count) with median values.
  + Standardized Yes/No fields to binary format (Y = 1, N = 0).
  + Created new features:
    - company\_age: How old the company is.
    - wage\_yearly: Wage normalized to yearly scale.
  + Encoded categorical variables (e.g. education, continent, region) using label encoding.
  + Converted case\_status to binary: Certified = 1, Denied = 0.

**2.2 Exploratory Data Analysis (EDA)**

We explored the data to understand how different features relate to visa approval:

* **Class Distribution:** About 65% of applications were Certified, while 35% were Denied.
* **Education Level:** Applicants with Master’s or Bachelor’s degrees had higher approval rates.
* **Wage Offered:** Higher wages were strongly associated with approval.
* **Company Size:** Larger companies (more employees) tended to have more Certified cases.
* **Region of Employment:** Some regions (like Northeast) had slightly higher approval rates.

Visualizations like histograms, box plots, and scatter plots helped us spot these patterns.

**2.3 Data Preprocessing**

To prepare the data for machine learning:

* **Feature Engineering:**
  + company\_age was calculated from the year of establishment.
  + wage\_yearly was created to standardize wage comparisons.
* **Encoding:**
  + Categorical variables (education, continent, region) were converted to numbers using label encoding.
  + Yes/No fields were converted to 1 and 0.
* **Scaling:**
  + All numeric features were scaled using StandardScaler to ensure consistent ranges.
* **Train-Test Split:**
  + The dataset was split into 80% training and 20% testing to evaluate model performance fairly.

**2.4 Model Training and Evaluation**

We trained four different machine learning models:

| **Model** | **Description** |
| --- | --- |
| Logistic Regression | A simple model that predicts based on linear relationships. |
| Decision Tree | A model that splits data into branches based on feature values. |
| Random Forest | A powerful model that combines many decision trees for better accuracy. |
| Gradient Boosting | A model that builds trees sequentially to correct previous mistakes. |

Each model was evaluated using five key metrics:

* **Accuracy:** How often the model was correct.
* **Precision:** How many predicted approvals were actually approved.
* **Recall:** How many actual approvals were correctly identified.
* **F1 Score:** A balance between precision and recall.
* **AUC-ROC:** Measures how well the model separates approved from denied cases.

**3. Results**

**3.1 Best-Performing Model**

The **Random Forest Classifier** gave the best overall performance:

| **Metric** | **Score** |
| --- | --- |
| Accuracy | 83% |
| Precision | 80% |
| Recall | 76% |
| F1 Score | 78% |
| AUC-ROC | 0.86 |

This means the model correctly predicted visa outcomes 83% of the time and was especially good at identifying approved cases.

The model was saved as best\_model.pkl using joblib, making it ready for deployment in a Streamlit app.

**3.2 Important Features**

The model identified the following as the most important factors in predicting visa approval:

| **Rank** | **Feature** | **Why It Matters** |
| --- | --- | --- |
| 1 | Wage (wage\_yearly) | Higher wages often meet labor standards and signal job quality. |
| 2 | Company Age | Older companies may be more stable and trusted by immigration authorities. |
| 3 | Education Level | Applicants with advanced degrees are more likely to qualify. |
| 4 | Full-Time Position | Full-time roles are preferred over part-time for visa approval. |
| 5 | Region of Employment | Some regions may have more favorable labor market conditions. |

**4. Discussion**

The results align with real-world expectations:

* **Wage and company profile** are strong indicators of approval. Employers offering competitive wages and having established businesses are more likely to succeed.
* **Education level** plays a key role. Applicants with Master’s or Bachelor’s degrees are more likely to be approved.
* **Full-time positions** are favored, possibly due to their stability and long-term potential.
* **Regional differences** may reflect local labor demand or processing efficiency.

These insights can help employers design stronger applications and guide applicants toward better opportunities.

**5. Conclusion and Recommendations**

**Conclusion:**

We successfully built a machine learning model that predicts visa approval outcomes using features like wage, education, job experience, and company profile. The Random Forest model achieved strong performance and highlighted key drivers of approval.

**Recommendations for Employers and Applicants:**

1. **Offer competitive wages** that meet or exceed prevailing standards.
2. **Highlight company stability** by showcasing years of establishment and employee count.
3. **Encourage full-time roles** for foreign workers to improve approval chances.
4. **Support applicants with advanced education** through targeted recruitment.
5. **Use the prediction model** to assess risk before submitting applications and improve documentation for borderline cases.