EMPLOYEE SALARY CLASSIFIER

Presented by Chilukuri Laxmi Prashasthi

Chaitanya Bharathi Institute of Technology
Artificial Intelligence and Machine Learning Dept.

edunet foundation

CONTENTS

edunet
foundation

PROBLEM STATEMENT

- To predict whether an employee earns more than $50K per year based on various demographic and occupational attributes.
- This classification helps in workforce analysis, HR planning, and policy-making.
- Supports data-driven decision-making in recruitment, compensation benchmarking, and organizational structuring.
- Assists policymakers and researchers in identifying patterns of income inequality across gender, race, and education.

edunet foundation

DATASET DESCRIPTION

Source: UCI Adult Census Income Dataset
Size: ~48,000 records
Features: Age, Workclass, Education Level, Marital Status, Occupation, Relationship, Race, Gender, Capital Gain, Capital Loss, Hours per Week
Target: Salary class (<=50K or >50K)

edunet
foundation

STEP-BY-STEP PROCEDURE

1. Dataset Import: Loaded UCI Adult dataset from CSV
2. Exploration & Cleaning:
    - Inspected data shape, data types, and value distributions
    - Replaced '?' with 'Others'; filtered out irrelevant entries
3. Outlier Handling:
    - Applied boxplots for age, capital-gain, education, hours-per-week
    - Filtered ranges (e.g., age 17–75, education-num 5–16)
4. Feature Selection:
    - Selected 11 columns including target
5. Label Encoding:
    - Encoded categorical variables with LabelEncoder
    - Saved encoders as label_encoders.pkl

edunet
foundation

STEP-BY-STEP PROCEDURE

6.Model Training:

- Tried 5 models: Logistic Regression, Random Forest, KNN, SVM, Gradient Boosting
- Used pipelines with scaling and fitting
- Chose Gradient Boosting as best (Accuracy: 86.47%)

7.Evaluation:

- Assessed model using accuracy, precision, recall, f1-score
- Compared across all models with a bar chart

8.Model Export:

- Saved model using joblib to best_model.pkl

9.Streamlit App:

- Developed a clean interface for live predictions
- Supports batch CSV input and output download

DATA PREPROCESSING

- Handled missing values ("?" converted to 'Unknown')
- Applied Label Encoding to categorical features using LabelEncoder
- Numerical features normalized
- Train-test split (e.g., 80-20%) for model evaluation
- Final feature set: 11 input variables

edunet
foundation

FINAL MODEL USED

- Model: Gradient Boosting Classifier
- Why this model?
  - Excellent for classification problems
  - Handles both numerical and categorical data well
  - Offers high accuracy and model explainability
- Accuracy Achieved: 86.47%
- Confidence: Calculated using predict_proba, typically 80% - 95%

edunet
foundation

RESULTS

- Best Model: Gradient Boosting
- Accuracy: 86.47% (from notebook training results)
- Predicted Classes: '>50K' or '<=50K'
- Batch Summary: Counts and percentages of salary categories
- Confidence Scores: Displayed with each prediction in the app

IMPORTANT LINKS

GitHub:

https://github.com/lp-0406/Employee-Salary-Prediction

Live Demo:

https://employee-salary-prediction-6yc6gnecxkdrr5vmdi46kz.streamlit.app/

edunet foundation

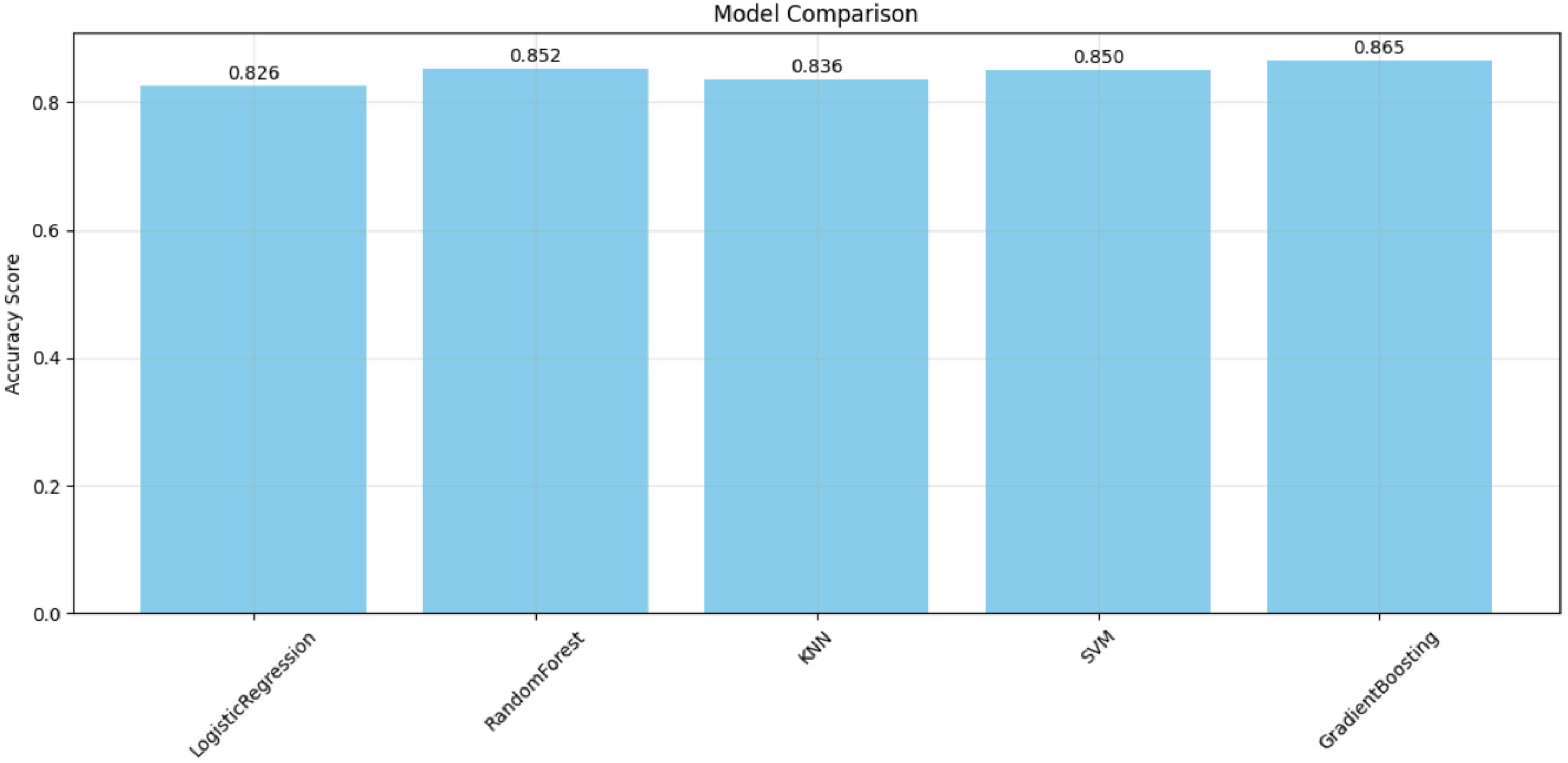
```python
best_model_name = max(results, key=results.get)
best_model = trained_models[best_model_name]
best_accuracy = results[best_model_name]

print(f"\nBest Model: {best_model_name}")
print(f"Best Accuracy: {best_accuracy:.4f}")
```

```
Best Model: GradientBoosting
Best Accuracy: 0.8647
```

```python
make_sample_prediction()
```

```
Sample Prediction: >50K
Prediction Probabilities: [0.43132704 0.56867296]
True
```

COLAB



edunet foundation

💼 Employee Salary Classifier

Advanced ML-powered salary prediction system

👤 Enter Employee Information

📊 Demographics

📅 Age

35

17 75

🎓 Education Level

6

🌍 Race

Black

⚧ Gender

Male

💼 Work Details

🏢 Work Class

Others

💼 Occupation

Exec-managerial

👥 Personal Details

💑 Marital Status

Married-civ-spouse

👫 Relationship

Other-relative

🕒 Hours per Week

40

1 80

💰 Capital Gain

5

Mana

	Age	Work Class	Education Level	Marital Status	Occupation	Relationship	Race	Gender	Capital Gain	Capital Loss	Hours/Week
0	35	Others	6	Married-civ-spou	Exec-manageri	Other-relative	Black	Male	5	3	40

🚀 Predict Salary Class

🎯 Prediction Result

Salary Class: <=50K

Confidence: 96.5%

📁 Batch Prediction from CSV

Upload a CSV file with employee data to get predictions for multiple employees at once.

Choose CSV File

Drag and drop file here

Limit 200MB per file • CSV

Browse files

CONCLUSION & FUTURE WORK

In the project I successfully created an interactive salary classification system that enables users to predict whether an employee earns more than $50K per year based on demographic and work-related attributes. This tool helps visualize salary distributions across different groups, providing valuable insights for HR and policy-making. Future enhancements may include integrating SHAP or LIME for better model explainability, incorporating resume parsers for real-world data input, and implementing bias or fairness checks across sensitive attributes such as race and gender.

REFERENCES

- UCI Adult Dataset
- Scikit-learn Documentation
- Streamlit Documentation

edunet
foundation

THANK YOU

edunet
foundation