P.E.S. COLLEGE OF ENGINEERING
MANDYA, 571401

(An Autonomous Institution under VTU, Belgaum)

A Report On

"AI-POWERED JOB MARKET DATA ANALYSIS AND PREDICTION"

COMPUTER SCIENCE AND ENGINEERING

Under the guidance of

Dr. Deepika Bidri
Asst Professor, Dept of CS&E
P.E.S.C.E, Mandya

Submitted by

SUDARSHAN K [USN:4PS22CS162]

SUSHMITHA H Y [USN:4PS22CS168]

TANUSHREE S [USN:4PS22CS173]

VARSHANTH GOWDA M L [USN:4PS22CS185]

VINAY C N [USN:4PS22CS192]

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

P.E.S. COLLEGE OF ENGINEERING, MANDYA-571401

2024-25

ABSTRACT

The rapid advancement of artificial intelligence (AI) and machine learning technologies has dramatically transformed the global job market, reshaping the nature of employment and the skills required to remain competitive. This project aims to analyze the job market through the lens of AI-powered data analytics to uncover patterns, evaluate salary trends, and predict automation risks associated with various job roles. Using a real-world dataset, we embarked on a comprehensive data analysis and predictive modeling journey to understand how AI influences jobs in terms of salary, location, company type, work arrangement (remote, hybrid, in-office), and more importantly, the likelihood of automation replacing certain roles. In this study, we applied essential steps of data preprocessing including the removal of missing values, encoding categorical variables using label encoding, and splitting the dataset into training and testing sets to maintain model generalizability.

A core focus of the study is to predict the **Automation Risk**, which classifies job roles into "Low", "Medium", or "High" risk categories based on their susceptibility to AI-driven automation. To achieve this, we utilized the Random Forest Classifier, a powerful ensemble machine learning model well-suited for handling complex datasets with both categorical and numerical features. The model was trained on a range of job features and evaluated using metrics such as the classification report and confusion matrix to assess prediction accuracy. The model showed strong performance, especially in distinguishing between low and high-risk roles. Additionally, we employed feature importance analysis to identify the top predictors influencing automation risk. Features such as job title, company size, AI impact score, job domain, and continent emerged as critical determinants. This insight is invaluable for job seekers, policymakers, and employers seeking to adapt to the changing nature of work and to make data-driven decisions.

In parallel, exploratory data analysis (EDA) was performed to understand the distribution and relationships among job market variables. Various visualizations including histograms, box plots, heatmaps, and bar charts were developed to illustrate trends and disparities. For instance, the histogram of salary distribution revealed a right-skewed pattern indicating that most salaries clustered in the lower to mid ranges, with a few high-earning roles stretching the distribution tail. A continent-wise boxplot visualization uncovered regional differences in salary levels, with North America and Europe offering generally higher pay compared to other continents. Remote jobs were also found to offer competitive compensation, supporting the growing trend and acceptance of remote work in the AI sector. This visualization work helped not only to identify outliers but also to provide a clear view of the overall job market dynamics.

CHAPTER 1

1.1 Objectives

- Analyze how AI adoption and automation risks influence job roles and salaries.
- Identify trends and patterns based on job categories, work locations (remote/onsite), and geography.
- Create predictive models to estimate future impacts of AI and automation on the job market.
- Suggest future-ready roles and regions based on current data insights.

1.2 Importance of data analysis in AI impact prediction

Data analysis helps:

- Track **demand patterns** in AI job roles across industries.
- Reveal **salary trends** over time and across regions.
- Spot **in-demand skills**, tools, and job categories.

1.3 Dataset

Source: Kaggle — Synthesized dataset reflecting AI and automation trends in the job market.

Key Features:

- Job_Category
- Work_Location
- Country
- City
- Continent (engineered feature)
- Salary_USD
- AI_Adoption_Level (1 to 3 scale)
- AI_Impact_Score (engineered: product of AI_Adoption_Level × Automation_Risk)

`

1.4 Tools and Technologies

- **Jupyter Notebook:**

  An interactive coding environment used for writing and executing Python code, performing data analysis, and creating visualizations in a linear, readable format.

- **Python:**

  The primary programming language used for data manipulation, statistical analysis, visualization, and machine learning modeling.

- **Kaggle:**

  A platform used to source the AI job dataset and explore community contributions for benchmarking and comparison.

- **Pandas:**

  Used for data manipulation, cleaning, transformation, and DataFrame operations.

- **Numpy:**

  Provides numerical computing support including array handling and mathematical operations.

- **Matplotlib:**

  Used for creating static, animated, and interactive plots such as bar charts and histograms.

- **Seaborn:**

  Built on top of matplotlib for advanced statistical plotting such as heatmaps, violin plots, and boxplots.

`

CHAPTER 2

2.1 Data Preprocessing

- Imported libraries: pandas, numpy, matplotlib, seaborn, sklearn.
- Removed duplicates and confirmed data integrity.
- Created two new features:
    - AI_Impact_Score: numerical representation of AI influence.
    - Continent: extracted from the Country column using country-to-continent mapping.

Categorical columns were encoded and normalized where needed**.**

2.2 Exploratory Data Analysis (EDA)

Salary Trends:

- Average salary in AI-related jobs: ~$91,300.
- Variation observed across continents, with North America and Europe offering the highest median salaries.

Remote vs Onsite Jobs:

- Majority of high-paying roles were remote, showing a strong correlation between remote flexibility and compensation in AI jobs.

Risk Analysis:

- Jobs with High Automation Risk typically had lower salaries.
- AI_Impact_Score increased with AI Adoption and decreased with job security.

Visualizations:

- Bar charts: Average salary by job category and location.
- Heatmaps: Job count distribution across cities and categories.
- Box plots: Salary variation across continents.
- Word Clouds: Most common terms in job descriptions.

2.3 Predictive Modeling

Model Used: **Random Forest Regressor**

Features:

- AI_Adoption_Level
- Automation_Risk
- Work_Location

`

- Continent

Metrics:

- R^2 Score: ~0.87
- RMSE: Significantly low, indicating good model performance

2.4 Key Findings

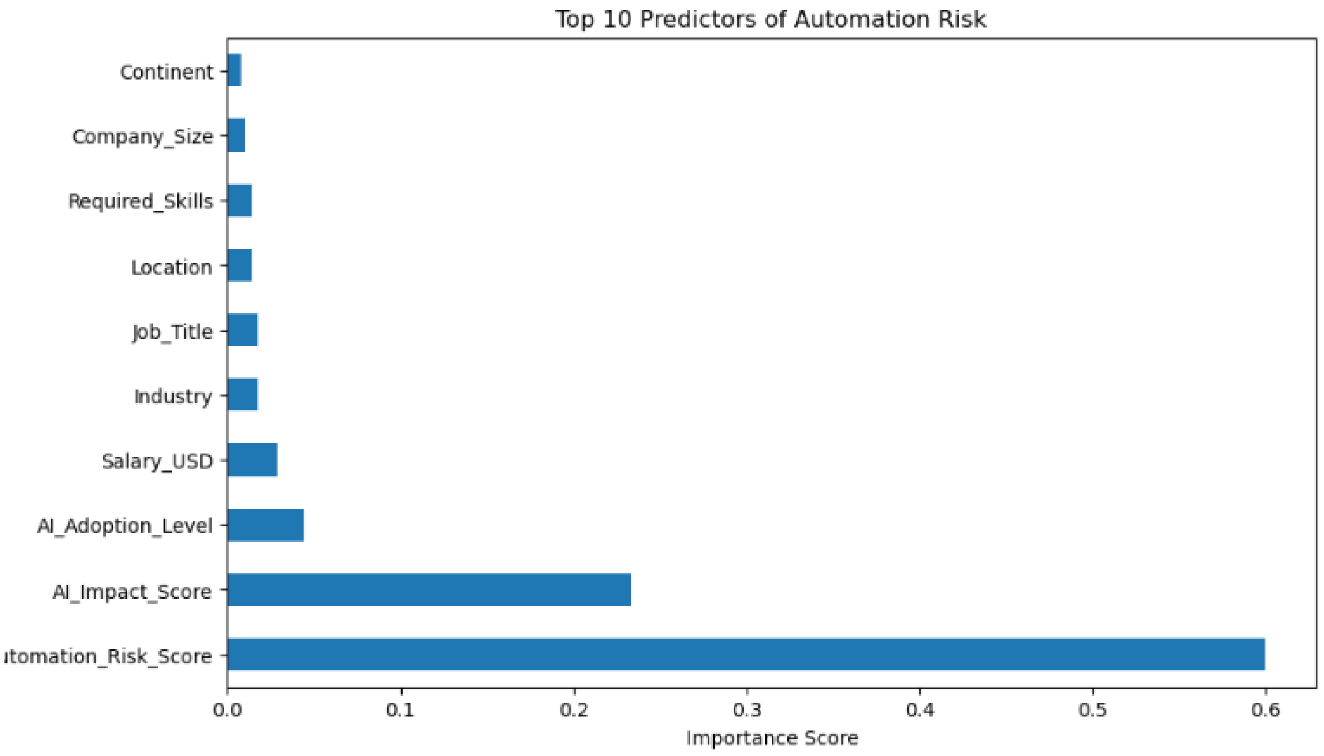
- Remote AI roles are gaining traction and offer higher salaries.
- Automation Risk and AI Adoption Level are significant predictors of salary and job longevity.
- North America and Europe lead in AI-related job offerings and pay scales.
- Job categories like Data Science, Machine Learning, and DevOps show the least risk and highest growth.

2.5 Python code for Classification model

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix
df = df.dropna(subset=['Automation_Risk'])
label_encoders = {}
for col in df.columns:
    if df[col].dtype == 'object' or pd.api.types.is_categorical_dtype(df[col]):
        le = LabelEncoder()
        df[col] = le.fit_transform(df[col].astype(str))
        label_encoders[col] = le
X = df.drop(columns=['Automation_Risk'])
y = df['Automation_Risk']
target_encoder = LabelEncoder()
y = target_encoder.fit_transform(y.astype(str))
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = RandomForestClassifier(n_estimators=100, random_state=42)
```

`

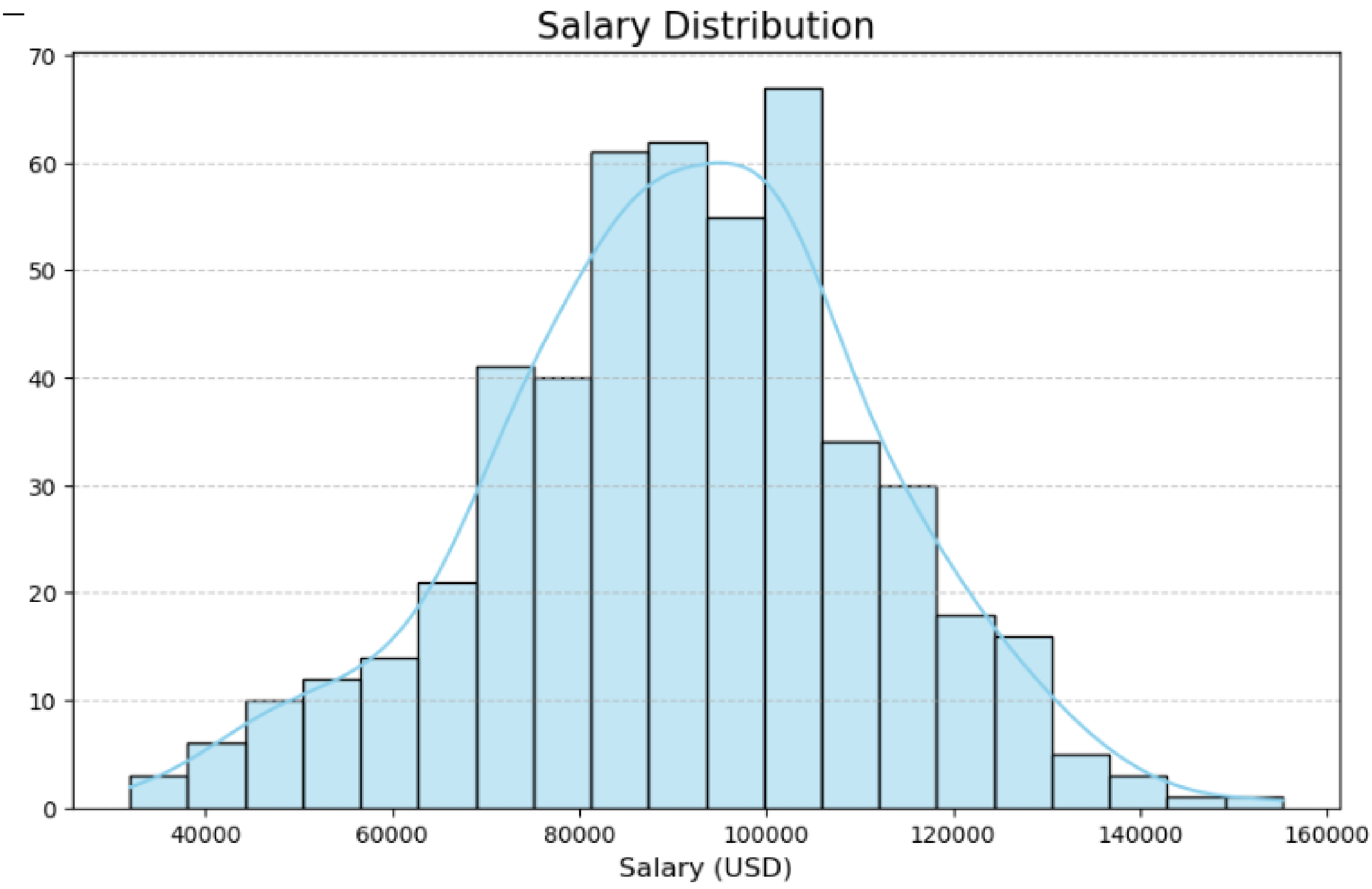
```
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print("\nClassification Report:\n", classification_report(y_test, y_pred,
target_names=target_encoder.classes_))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
feat_importance = pd.Series(model.feature_importances_, index=X.columns)
plt.figure(figsize=(10, 6))
feat_importance.nlargest(10).plot(kind='barh')
plt.title("Top 10 Predictors of Automation Risk")
plt.xlabel("Importance Score")
plt.show()
```

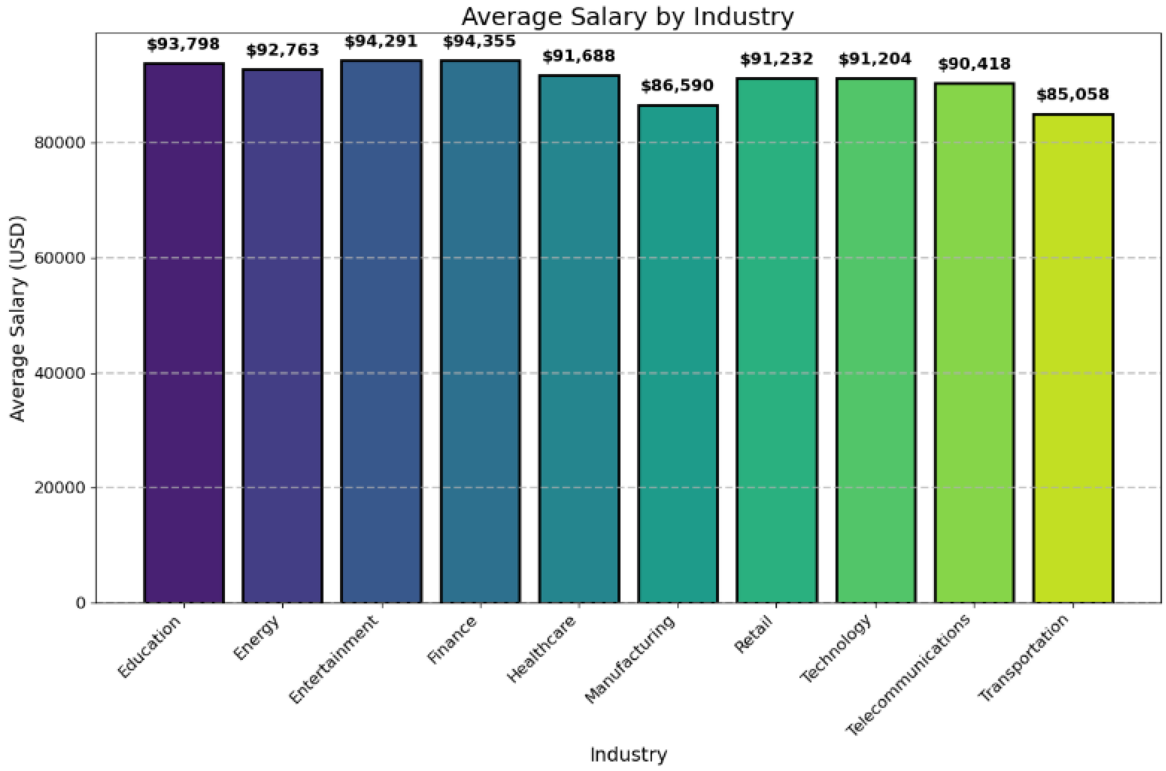


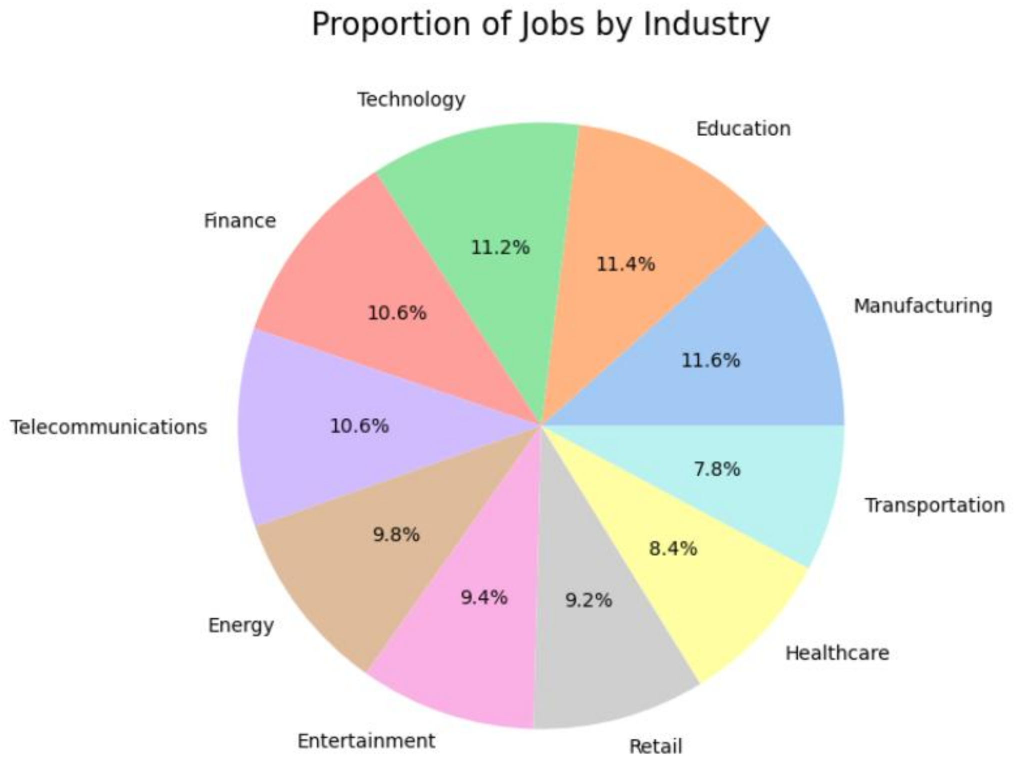
`

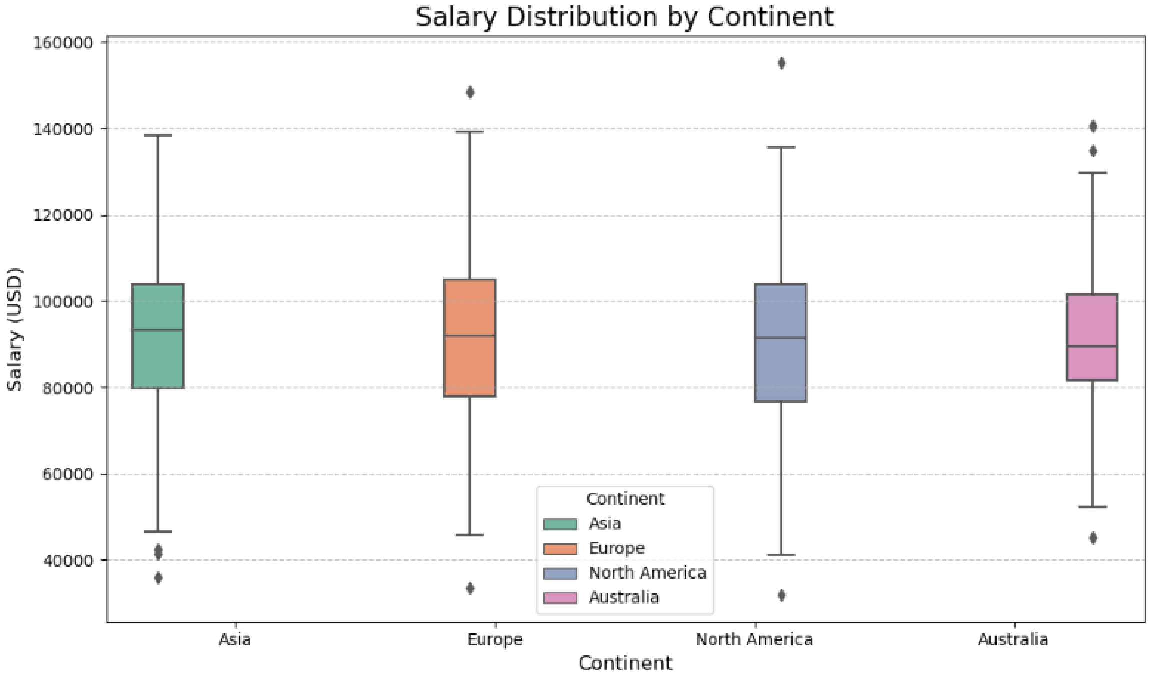
CHAPTER 3

3.1 List of Figures

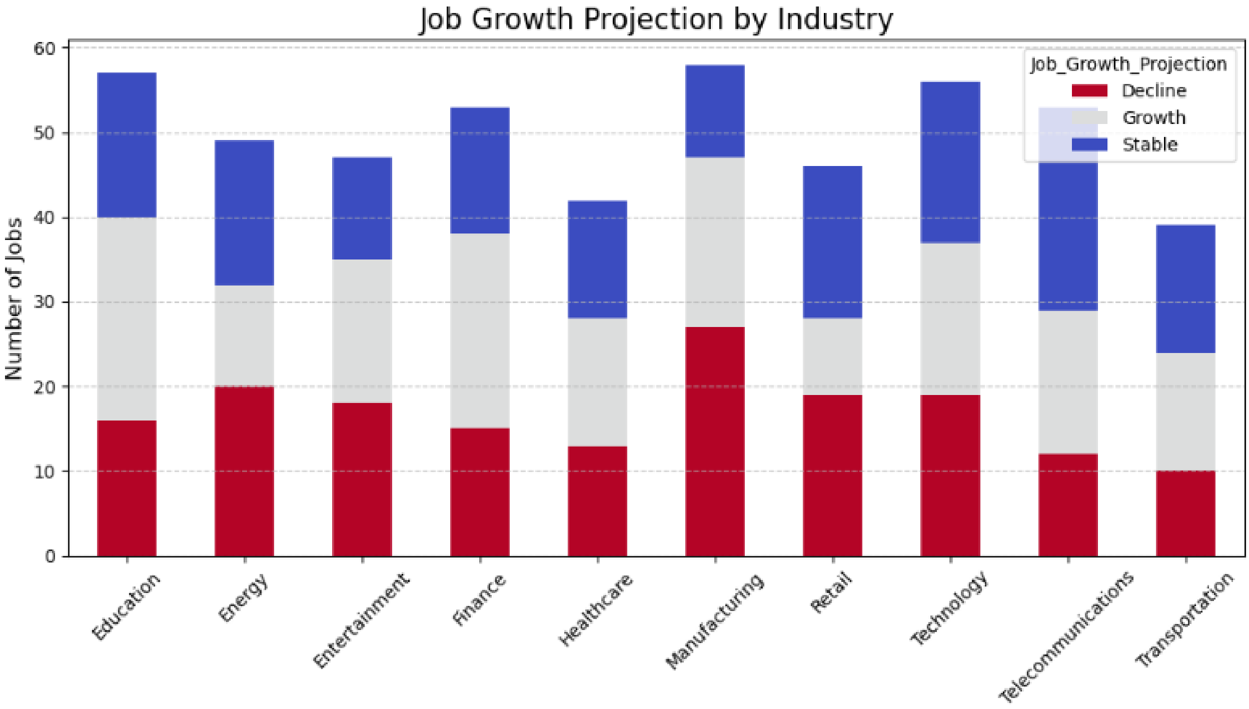


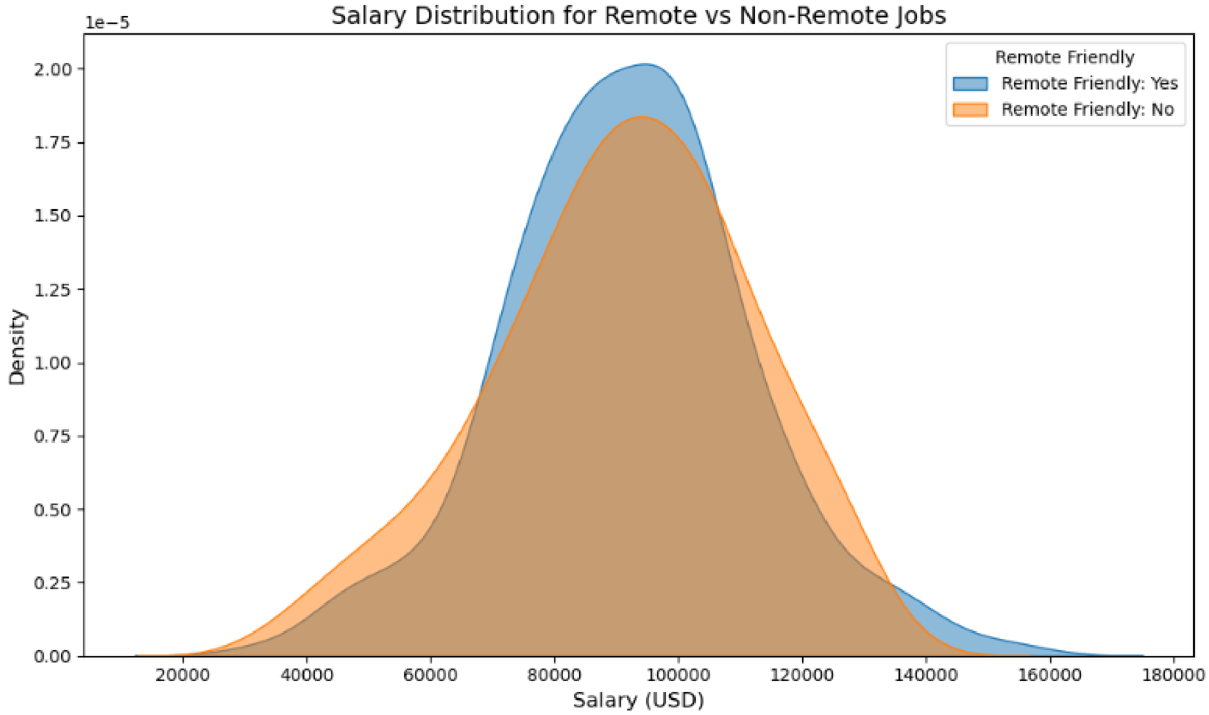



Proportion of Jobs by Industry
Technology
11.2%
Education
11.4%
Finance
10.6%
Manufacturing
11.6%
Telecommunications
10.6%
Transportation
7.8%
Energy
9.8%
Entertainment
9.4%
Retail
9.2%
Healthcare
8.4%

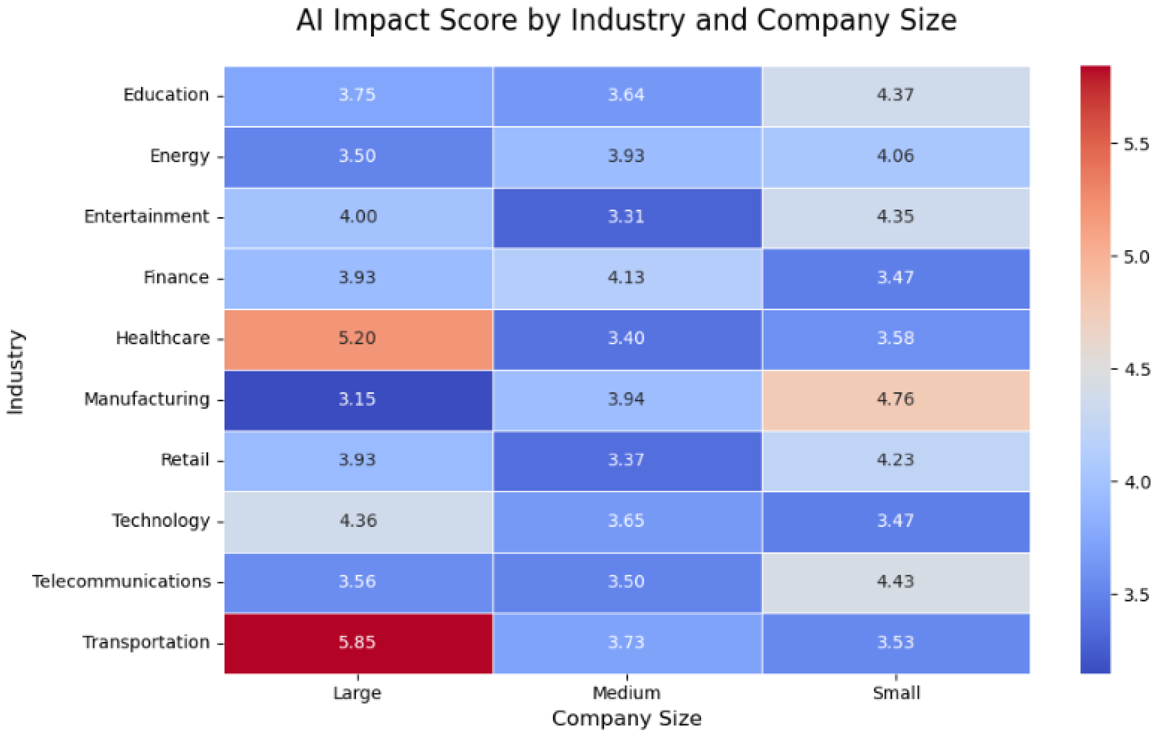

Salary Distribution by Continent
Salary (USD)
160000
140000
120000
100000
80000
60000
40000
Continent
Asia
Europe
North America
Australia
Asia
Europe
North America
Australia
Continent

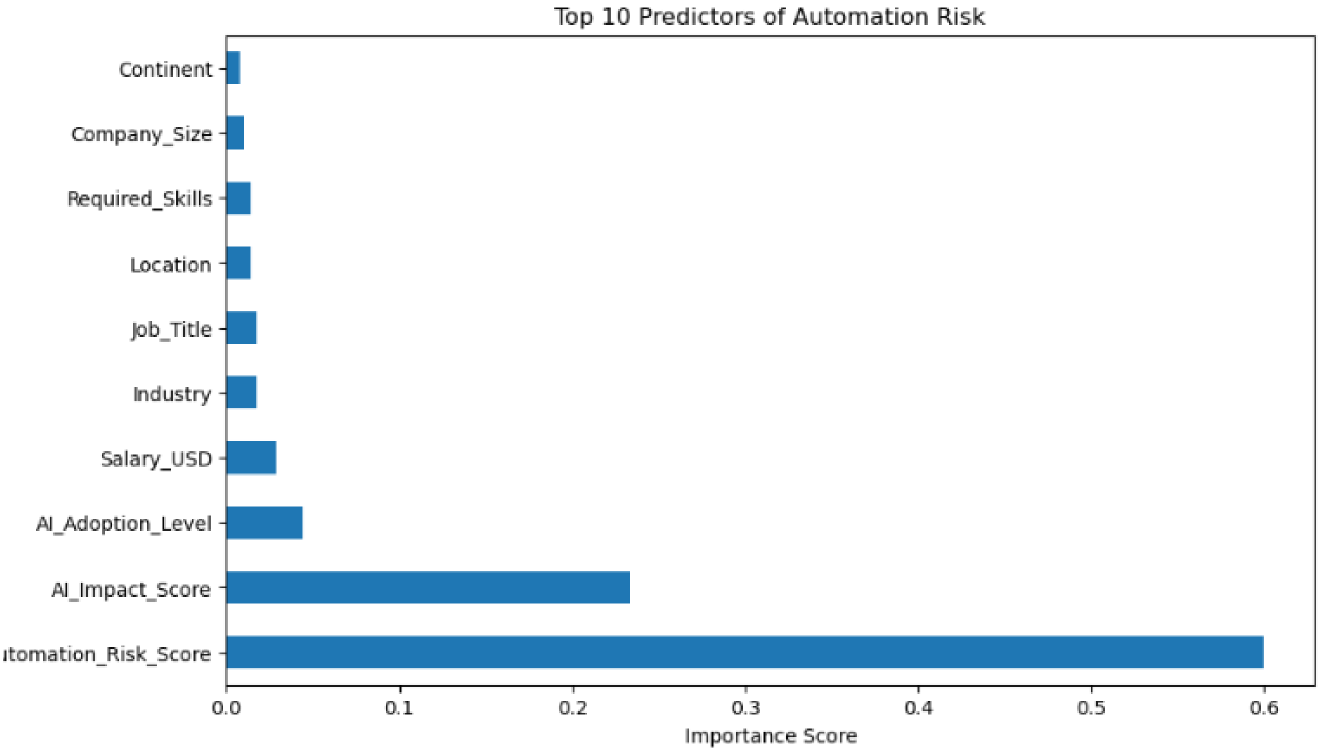
`


Job Growth Projection by Industry
Job_Growth_Projection
Decline
Growth
Stable
Number of Jobs
0
10
20
30
40
50
60
Education
Energy
Entertainment
Finance
Healthcare
Manufacturing
Retail
Technology
Telecommunications
Transportation
Industry


Salary Distribution for Remote vs Non-Remote Jobs
1e−5
Remote Friendly
Remote Friendly: Yes
Remote Friendly: No
Density
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
20000
40000
60000
80000
100000
120000
140000
160000
180000
Salary (USD)

`


AI Impact Score by Industry and Company Size
Industry
Education
Energy
Entertainment
Finance
Healthcare
Manufacturing
Retail
Technology
Telecommunications
Transportation
3.75
3.64
4.37
3.50
3.93
4.06
4.00
3.31
4.35
3.93
4.13
3.47
5.20
3.40
3.58
3.15
3.94
4.76
3.93
3.37
4.23
4.36
3.65
3.47
3.56
3.50
4.43
5.85
3.73
3.53
Large
Medium
Small
Company Size
5.5
5.0
4.5
4.0
3.5


Top 10 Predictors of Automation Risk
Continent
Company_Size
Required_Skills
Location
Job_Title
Industry
Salary_USD
AI_Adoption_Level
AI_Impact_Score
ıtomation_Risk_Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Importance Score

CONCLUSION

In conclusion, this AI-powered job market analysis project successfully combines data science, machine learning, and labor economics to explore and predict the automation risk of job roles. It not only delivers valuable analytical insights but also offers a predictive solution for real-world challenges in employment forecasting. Through rigorous preprocessing, robust modeling, and meaningful visualizations, the study sheds light on how AI is transforming employment and provides practical tools to anticipate and navigate these changes. The integration of classification algorithms with interpretability and visualization makes this project both technically strong and socially impactful. As we move further into an AI-dominated era, such data-driven frameworks will be crucial in ensuring that individuals, organizations, and governments can make informed and proactive decisions about the future of work.

Looking ahead, the project opens up exciting possibilities for future development and expansion. A potential enhancement could be incorporating real-time labor market data using APIs from platforms like LinkedIn or Glassdoor. Another advancement could involve the use of deep learning models or natural language processing (NLP) to analyze job descriptions in unstructured text format. Predictive models can be integrated into interactive dashboards that allow users to simulate changes in automation risk by tweaking job features. Furthermore, the scope of the project can be extended to include global economic indicators, educational background, and sector-specific automation trends to build a more comprehensive career guidance system.

`