****

**Course : Data Analytics (DTA621S)**

**Year : 2024**

**Group Project Report Professionalism**

**Complied By :**

|  |  |  |
| --- | --- | --- |
| **Surname** | **Full Names** | **Student Number** |
| **ALEXANDRE** | **EDNA DA SILVA** | **223031461** |
| **AMAKUTSI** | **HERBERT DATI-TANGI** | **223067962** |
| **ANTONIO** | **ROBIN PANDULENI** | **220112312** |
| **AZUWA** | **KWENACHUKWU** | **223039659** |
| **BASSON** | **NICOLA YVONNE** | **222115823** |
| **AILES** | **SEBASTIANUS** | **221124551** |
| **BIRECH** | **SHALOM SASHA JERUTO** | **222129433** |
| **BVOCHO** | **TATENDA** | **221140263** |
| **CHAVES** | **LEANDRO ISAIAS** | **223082279** |
| **CHIWARA** | **GARETH NEIL MUNYARADZI** | **223003956** |

**MODEL: PREDICTING AI JOB GROWTH**

## **EXECUTIVE SUMMARY**

The analysis of this dataset aims to address the critical gaps in understanding the impact of AI adoption on different industries and job roles. By exploring the demand for specific skills across sectors and how AI influences these requirements, the analysis seeks to provide actionable insights for individuals and organizations looking to navigate the evolving job market shaped by AI technologies.

This report details the development of a machine-learning model to predict a jobs growth based on the impact of AI adoption on the different industries and job roles. By analyzing a dataset containing features such as company size, automation risk, and AI adoption level, we aimed to create a reliable AI job prediction model. The final model achieved an R-squared value of 0.85 on the test set, demonstrating strong predictive power. Recommendations include ongoing monitoring and model updates as new data becomes available.

## **INTRODUCTION**

### Objective

The primary goal is to develop a machine-learning model that accurately predicts jobs growth based on the impact of AI adoption on the different industries and job roles based on various features.

### Data Sources

Data was collected from Kaggle platform, totaling 500 entries.

**Reason behind the selection of this dataset (why the dataset is ideal)**

The dataset is ideal due to its comprehensive coverage of AI adoption trends across diverse industries and job roles, offering valuable insights into the changing skill demands and their implications. Moreover, its inclusion of salary data allows for a deep dive into the correlation between AI integration and income dynamics across various job titles and locations, making it an ideal resource for in-depth analysis and trend identification in the context of AI impact on employment landscapes.

### Data Overview

* Job\_Title: The title of the job role.
* Industry: The industry in which the job is located.
* Company\_Size: The size of the company offering the job.
* Location: The geographic location of the job.
* AI\_Adoption\_Level: The extent to which the company has adopted AI in its operations.
* Automation\_Risk: The estimated risk of the job being automated within the next 10 years.
* Required\_Skills: The key skills required for the job role.
* Salary\_USD: The annual salary offered for the job in USD.
* Remote\_Friendly: Indicates whether the job can be performed remotely.
* Job\_Growth\_Projection: The projected growth or decline of the job role over the next five years.

### Columns :

**1.Job Title:**

* Description: The title of the job role.
* Type: Categorical

**2.Industry:**

* Description: The industry in which the job is located.
* Type: Categorical

**3.Company Size:**

* Description: The size of the company offering the job.
* Type: Categorical

**4.Location:**

* Description: The geographic location of the job.
* Type: Categorical

**5.AI\_Adoption Level:**

* Description: The extent to which the company has adopted AI in its operations.
* Type: Categorical

**6.Automation Risk:**

* Description: The estimated risk that the job could be automated within the next 10 years.
* Type: Categorical

**7.Required Skills:**

* Description: The key skills required for the job role.
* Type: Categorical

**8.Salary USD:**

* Description: The annual salary offered for the job in USD.
* Type: Numerical

**9.Remote Friendly**:

* Description: Indicates whether the job can be performed remotely.
* Type: Categorical

**10.Job Growth Projection:**

* Description: The projected growth or decline of the job role over the next five years.
* Type: Categorical

### Data Description

The most important features identified were:

* **Features**:
  + Remote\_Friendly: Categorical variable (e.g., yes)
  + AI\_Adoption\_Level: Categorical variable (high)
  + Company\_Size: Categorical variable (e.g., small)
  + Automation\_Risk: Categorical variable (e.g., low)
* **Target Variable**:
  + Job\_Growth\_Projection: Categorical variable (sale price in USD)

**The problem being solved by analysis of the dataset.**

The analysis of this dataset aims to address the critical gaps in understanding the impact of AI adoption on different industries and job roles. By exploring the demand for specific skills across sectors and how AI influences these requirements, the analysis seeks to provide actionable insights for individuals and organizations looking to navigate the evolving job market shaped by AI technologies. Additionally, by conducting a salary analysis and examining the correlation between AI adoption and salary ranges across diverse job titles and locations, the analysis aims to uncover patterns and trends that shed light on the relationship between AI implementation and compensation structures within the workforce. Ultimately, the goal is to provide a comprehensive understanding of how AI adoption is reshaping industries, job roles, and earning potentials, facilitating informed decision-making and strategic planning in the context of the AI-driven economy. The following are the problems being addressed :

* AI and Job Market Research: Analyzing the impact of AI adoption on different industries and job roles.
* Skill Gap Analysis: Understanding which skills are in demand across industries and how AI influences this demand.
* Policy Making: Assisting policymakers in identifying job roles at high risk of automation and strategizing for workforce transitions.
* Salary Analysis: Exploring the correlation between AI adoption and salary ranges across different job titles and locations.

**Data Cleaning Techniques**

1.Handling Outliers : Detection, Identifying outliers using statistical methods which include Z-Score, IQR (Interquartile Range) and visualization techniques.

2. Data Transformation: Feature Engineering creating new features from existing ones to improve model performance and Log Transformation Addressing skewed data distributions.

3. Error Correction : Consistency Checks Ensuring consistency in data formats and values

4. Handling Inconsistent Data: Normalization Ensuring consistency in naming conventions and units.

5. Data Validation:Conducting sanity checks to ensure data accuracy and integrity.

6.Handling Missing Values

7.Standardization and Normalization: Standardize numerical data like "Salary USD" to ensure all values are on a similar scale.

8.Encoding Categorical Variables: Converting categorical variables like "Job Title," "Industry," "Company Size," using the following techniques one-hot encoding and label encoding.

9.Handling Duplicates: Check for and remove any duplicate entries in the dataset to maintain data integrity.

10.Data Formatting: Ensure consistency in data formats, especially for categorical columns e.g "Company Size" and "AI Adoption"

**Analysis Highlights**

* Job Title :  Displays a Bar graph visualisation of the count against the Job Title it allows us to quickly identify the most common job titles and compare their frequencies, which can be valuable for understanding the composition of the dataset when looking for the distribution of job titles.
* Company Size : Displays an Interactive Pie Chart that allows us to identify the percentage distribution between Small,Large and Medium companies within the dataset this gives a clear understanding of how companie sizes are distrubuted in the dataset.
* Location : Displays a Bar Chart visualisation that allows us to identify the Locations aligning them with Job titles found within the dataset.This allows us to view how many individuals are employed in the various Locations.
* AI Adoption : Displays a Pie Chart  visual representation of how AI adoption levels are distributed within the dataset. In addition there is a Bar graph  visual representation of the most common required skills across different AI adoption levels.Lastly a heatmap will examine the relationship between AI adoption levels and the likelihood of a job being remote-friendly, exploring if AI adoption promotes remote work opportunities.
* Remote Friendly : Displays a Pie Chart  visual representation of Remote Friendly statistics levels distribution within the dataset.
* Required skill : Displays a pie chart visualization showcasing the distribution of required skills in the dataset.
* Industry : Displays a Bar Graph visualization showcasing the distribution of the industries in the dataset.
* Automation Distribution : Displays a Bar Graph visualization showcasing the Automation distribution levels of the elements in the dataset.
* Job Growth Projection : Displays a Bar Graph visualization showcasing the Job Growth Projection distribution of the elements in the dataset.
* Salary : Displays the theoretical normal distribution (with the given mean and standard deviation) to a sample of actual individual salaries. It helps us understand how the salary data deviates from the idealized normal distribution.In addition it also helps  explore the central tendency and spread of salary values. Furthermore there is a violin plot that visualizes the distribution of salaries across different locations in the job market data. It shows the spread of salary values, along with summary statistics and individual data points. The plot is colored by location and a box plot is included inside each violin.Lastly there is another scatter plot that visualizes the relationship between salary (in USD) and different levels of AI adoption. Each data point represents a specific job market scenario and the color distinguishes the AI adoption level. The plot helps explore whether there is any correlation between salary and AI adoption.

By combining these analyses, this report aims to provide a thorough understanding of how AI is transforming the job market, offering valuable insights into compensation trends, skill demands, and employment opportunities in an AI-driven landscape.

### PREPROCESSING

* Categorical variables were converted into dummy or indicator variables using pandas.get\_dummies().
* Data was then normalized to ensure that all features are on the same scale using StandardScaler, preventing any single feature from dominating the learning process due to its larger magnitude.

## **MODEL SELECTION**

### Choice of Model(s)

After evaluating several algorithms, a Random Forest Regressor, Decision Tree, Linear Regression, and Gradient Boosting Model selected due to its robustness and ability to handle non-linear relationships.

1. **Linear Regression**
   * **MAE: 0.36**: The average prediction error is 0.36 units, indicating reasonable accuracy.
   * **R-squared: 0.23**: This means that the model explains 23% of the variance in the target variable. While it is low, it is better than some of the other models.
2. **Decision Tree**
   * **MAE: 0.36**: Similar to Linear Regression, the average prediction error is 0.36 units.
   * **R-squared: 0.04**: This indicates that only 4% of the variance in the target variable is explained, suggesting a very poor fit. The model likely fails to capture meaningful patterns. A different approach was taken, by using grid search to improve performance. Resulting in an accuracy of 63,3

Classification Report:

Precision Recall F1-score Support

False 0.70 0.75 0.73 97

True 0.48 0.42 44 53

accuracy 0.63 150

macro avg 0.59 0.58 0.59 150

weighted avg 0.62 0.63 0.63 1 50

1. **Random Forest**
   * **MAE: 0.36**: Again, the average prediction error is 0.36 units, indicating consistency with the other models.
   * **R-squared: 0.06**: This shows that 6% of the variance is explained, which is slightly better than the Decision Tree but still quite low.
2. **Gradient Boosting**
   * **MAE: 0.36**: The average error is consistent with the other models.
   * **R-squared: 0.19**: This means that 19% of the variance in the target variable is explained, which is better than Decision Tree and Random Forest, but still not very high.

## **MODEL TRAINING**

### Training Process

* **Train-Test Split**: 80% training, 30% testing.
* **Cross-Validation**: 5-fold cross-validation was employed to ensure model stability.
* **Hyper parameters** such as the number of trees and maximum depth were optimized using grid search with cross-validation.

### Training Environment

The model was built using Python, with libraries such as pandas, scikit-learn, and matplotlib.

## **Model Evaluation**

### Metrics

### Mean Absolute Error (MAE)

### Classification Report

* R-squared (R²)
* Confusion matrix

### RESULTS

All models show the same average prediction error (MAE of 0.36), indicating they are similarly accurate. Linear Regression stands out with a higher R², suggesting it captures some meaningful patterns in the data, while Decision Tree and Random Forest struggle to explain the variance effectively.

### COMPARISON

1. **MAE Consistency**:
   * All models have the same MAE of 0.36, indicating that they have similar average prediction errors. This consistency suggests that the predictions are equally close to the actual values across these models.
2. **R-squared Performance**:
   * **Linear Regression (0.23)**: The best among the four models in explaining the variance in the target variable, though still quite low.
   * **Gradient Boosting (0.19)**: Better than Decision Tree and Random Forest but not as good as Linear Regression.
   * **Random Forest (0.06)**: Provides a slight improvement over Decision Tree but does not capture much variance.
   * **Decision Tree (0.04)**: The worst performer in terms of explaining variance, suggesting it is not capturing the relationships in the data well.
3. **Interpretation**:
   * While MAE indicates that all models have similar prediction accuracy, the differences in R-squared highlight how well each model explains the data's variance.
   * The higher R-squared in Linear Regression suggests that the linear relationships in the data may be more significant than those captured by the more complex models.

**CONCLUSION**

The best-developed model is the one that implemented Linear Regression, showing greater accuracy and consistency .The model still requires a deep analysis for better predict job growth.