

Exploring Machine Learning and AI Basics:

| Salary Prediction System

BY Deepthi CH[S190218]





TABLE OF **CONTENTS**

- Problem Statement
- Abstract
- Introduction
- Existing works
- proposed method
- Experimental setup
- AI Basics
- conclusions
- Future scope and references



| Problem statement

Problem Statement:

The objective of this summer internship project is to develop a machine learning model capable of predicting salaries based on various factors. The primary challenge lies in identifying the key features that influence salary predictions and building a robust model that can handle real-world data variability and noise.

Learning the basics of Artificial Intelligence (AI) is crucial for this project. AI encompasses a wide range of techniques and algorithms that enable computers to perform tasks that typically require human intelligence. Understanding AI basics provides a solid foundation for developing effective machine learning models and addressing the challenges associated with salary prediction.



| ABSTRACT

This documentation presents the work completed during the summer internship focused on salary prediction using machine learning techniques. The project involved comprehensive data collection, preprocessing, implementation of various regression models, and rigorous evaluation of model performance. The primary goal was to develop a robust machine learning model capable of accurately predicting salaries based on factors such as job title, company size, sector, and location. During the internship, I also delved into the basics of Artificial Intelligence (AI), gaining insights into fundamental concepts, supervised and unsupervised learning algorithms, data preprocessing techniques, and model evaluation metrics. This foundational understanding was crucial in addressing the challenges of data variability and noise in real-world datasets. The findings demonstrate the effectiveness of machine learning in predictive analytics, highlighting the challenges encountered, such as handling missing data and selecting relevant features, and the solutions implemented.

Keywords : ML Techniques, Artificial Intelligence, Regression models , Evaluation Metrics



| INTRODUCTION

1) MOTIVATION FOR THE WORK:

Accurate salary prediction is crucial for both companies and job seekers. Companies use salary predictions for budgeting and recruitment strategies, while job seekers use them for career planning and salary negotiations. This makes me to Select this project In internship .



2) REAL WORLD APPLICATIONS:

A) salary predictions for budgeting

B) Salary prediction System job seekers



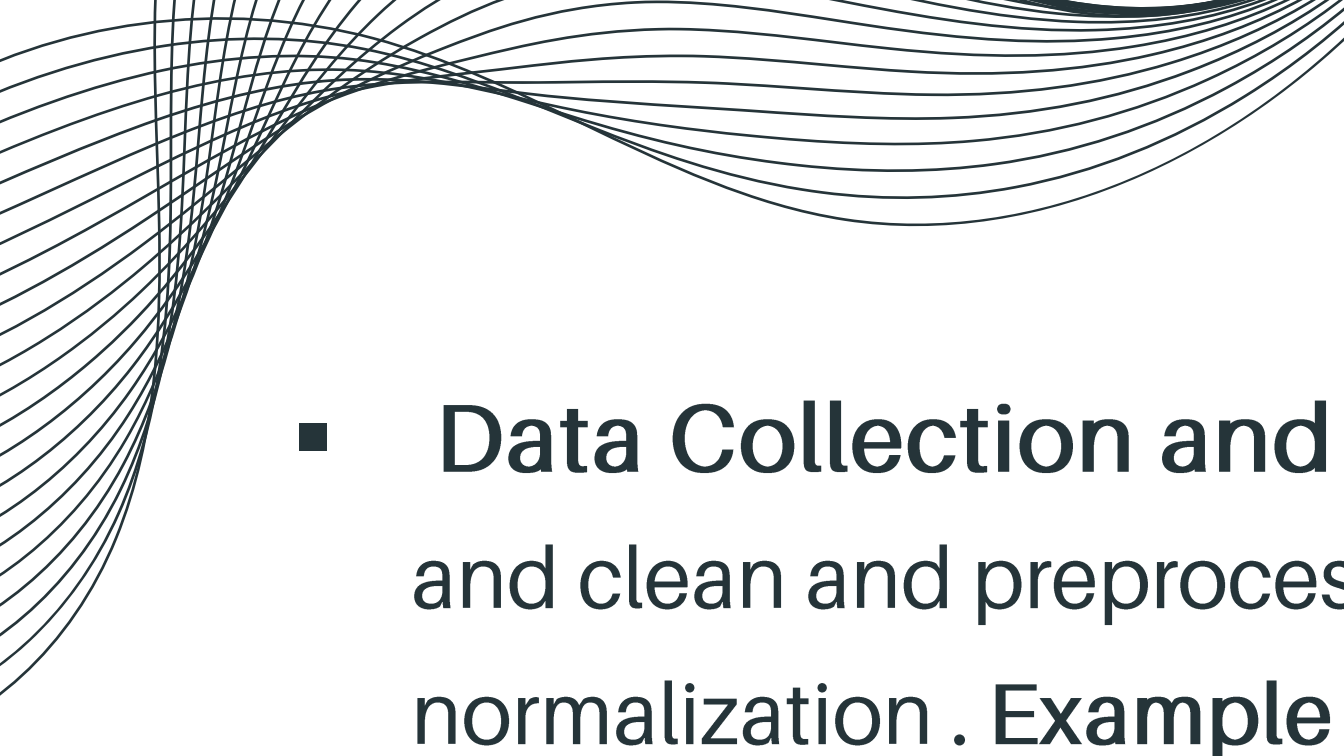
| ***EXISTING WORKS***

Numerous studies have explored salary prediction using machine learning. Most approaches leverage regression techniques to model the relationship between salary and various predictors such as education, experience, and job location. Commonly used models include linear regression, decision trees, and support vector machines. Previous research indicates that data preprocessing and feature selection significantly impact model performance. This project builds on these foundations, aiming to enhance prediction accuracy through comprehensive data preprocessing and model optimization.



| PROPOSED SYSTEM

The rapid advancement of technology has made data-driven decision-making an integral part of various industries. Predictive analytics, a subset of machine learning, is increasingly used to forecast outcomes based on historical data. This project aims to predict salaries using machine learning models. The project spans data collection, preprocessing, model implementation, and evaluation phases. By exploring different regression models, the project aims to identify the most accurate method for salary prediction.

- 
- **Data Collection and Preprocessing:** Gather salary data from reliable sources and clean and preprocess the data to handle missing values, outliers, and normalization . **Example dataset fields:** Job Title, Salary Estimate, Rating, Headquarters, Size, Sector, Revenue, Min Salary, Max Salary, Avg Salary, Same State, Age.
 - **Feature Selection:** Identify and select key features influencing salary based on domain knowledge and statistical analysis.
 - **Model Implementation:** Implement various regression models including linear regression, decision trees, and support vector machines.
 - **Model Evaluation:** Evaluate model performance using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).
 - **Optimization:** Fine-tune model parameters to enhance prediction accuracy.



ALGORITHM:


STEP 1 : Start

STEP 2 : Collection of data from CSV file downloaded from Kaggle.

STEP 3 : Access data using Pandas library

STEP 4 : Data cleaning using pandas

STEP 5 : Modification of raw data into useful metrics by using Feature Engineering



STEP 6 : Visualizing data frame columns to find key columns that is useful for better prediction System EX: Heat maps

STEP 7 : Make new Data Frame by deleting Unwanted columns

STEP 8 : Building regression models Using linear regression and Random Forest Regression model from sklearn.ensemble and sklearn.linear_regression libraries using split test dataset

STEP 9 : Finding Error Value by using mean absolute error from sklearn.metrics

STEP 10 : Find Accuracy of the model by testing with test data set



| EXPERIMENTAL SETUP

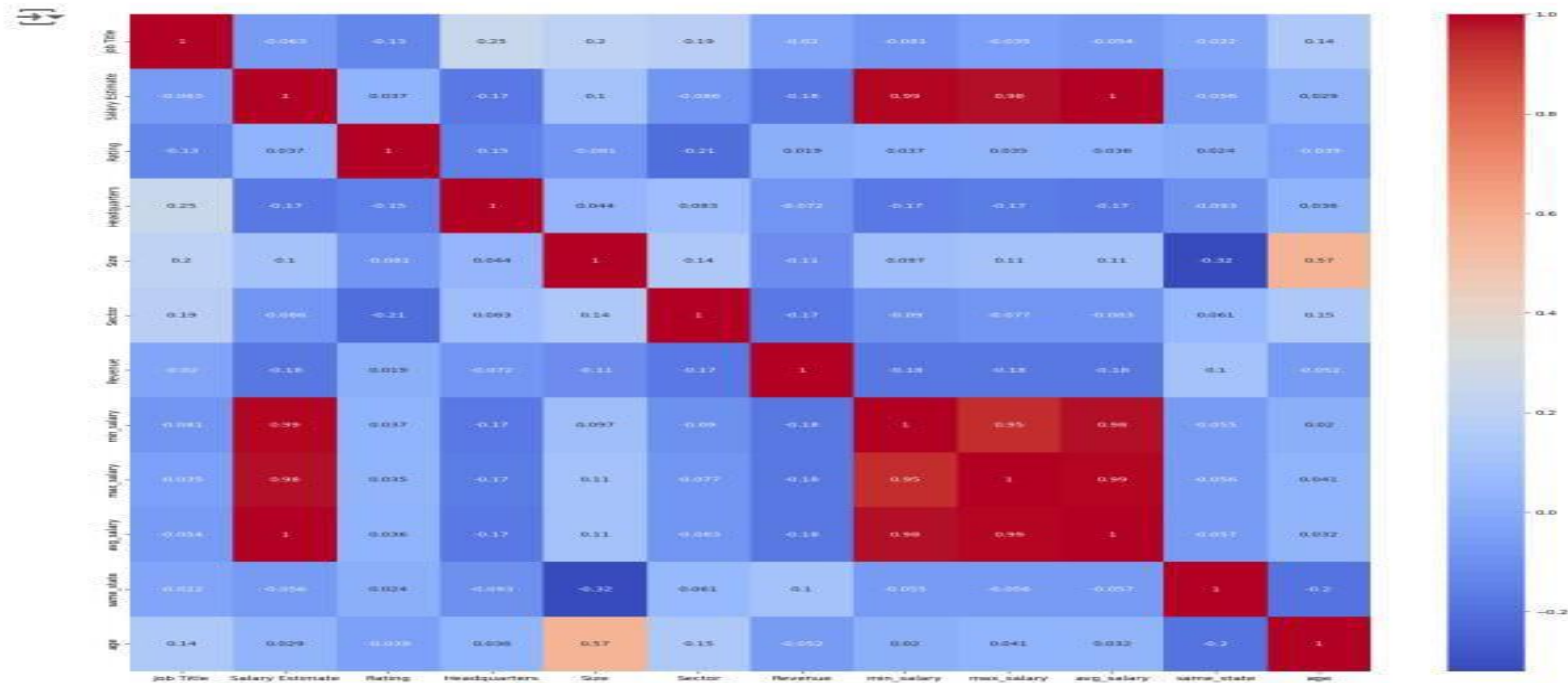
Libraries Used

- 1) Pandas
- 2) Seaborn
- 3) sklearn
- 4) matplotlib

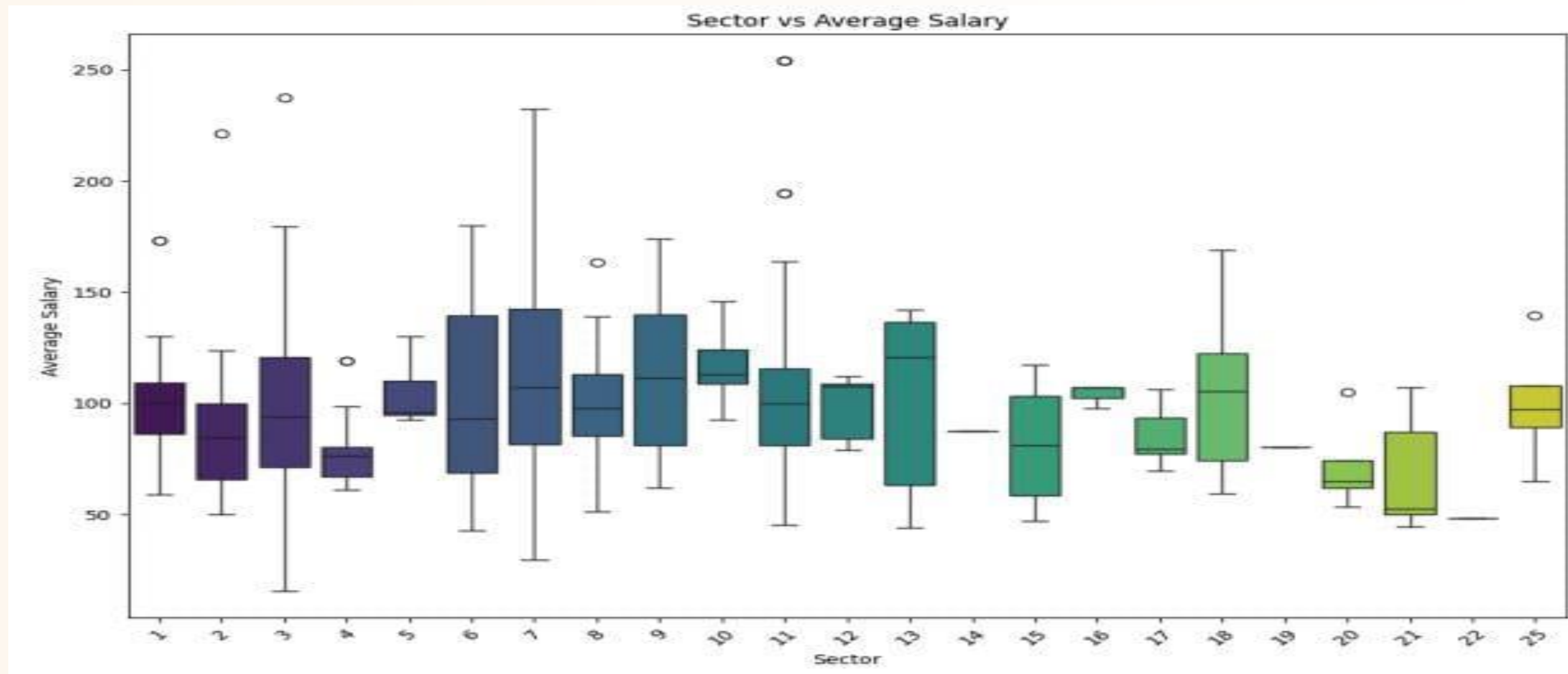
DataSets

- 1) salary_data_cleaned dataset

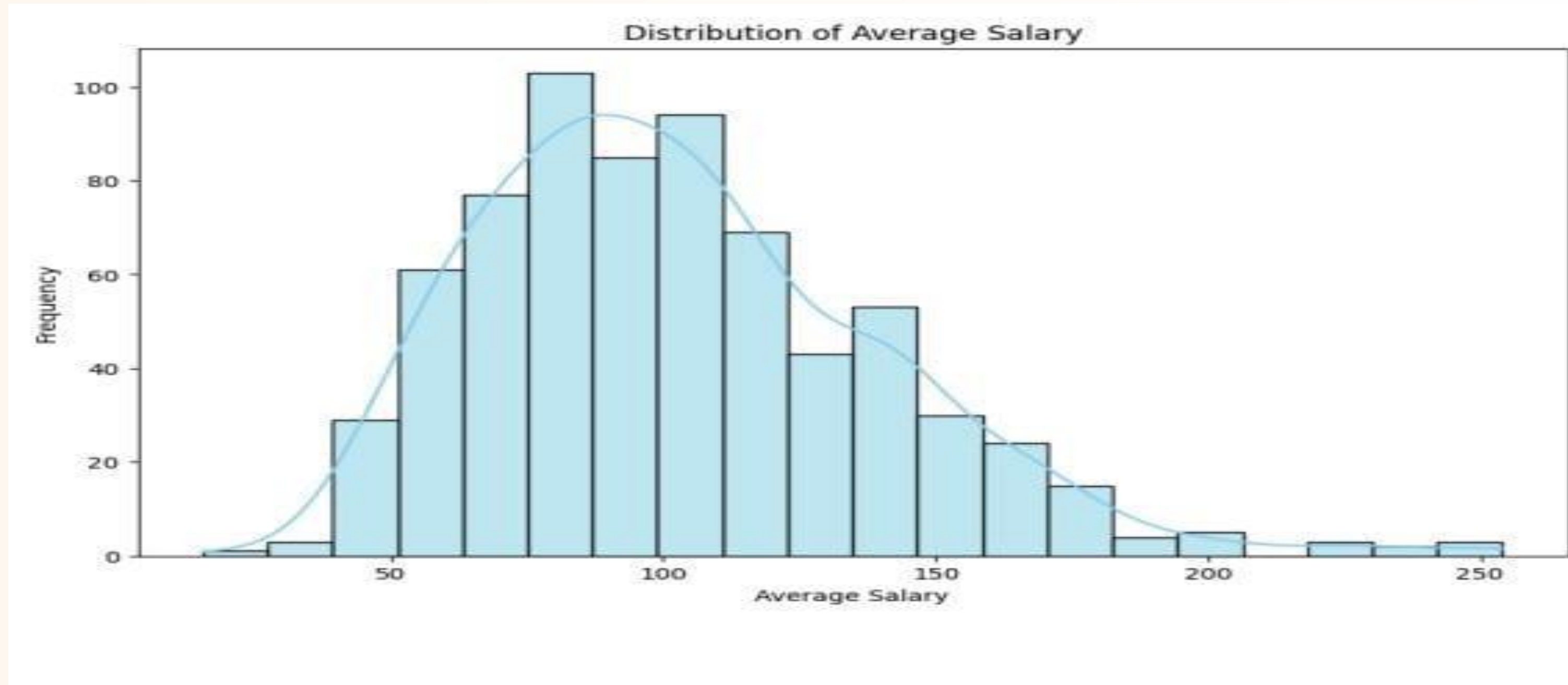
```
# correlation matrix with heatmap
data.corr()
plt.figure(figsize=(20,20))
sns.heatmap(data.corr(), annot=True, cmap='coolwarm')
plt.show()
```



Heat map for identifying correlation between all numerical Variables



Box Plot: Compares the distribution of avg_salary across different Sector categories using a box plot.



Histogram: Visualizes the distribution of avg_salary using a histogram with kernel density estimation .



| AI Basics

1) Introduction to AI, Real time applications, Importance of Artificial Intelligence

2) gadget learning has the following algorithms:

- *Supervised learning*
- *Unsupervised learning*



SUPERVISED LEARNING:

This is a list of predictions. These predictions are unrelated variables. This learning algorithm's goal is to make predictions based on this set of independent variables.

Variability in result predictability. This is a conditional variable. A function is formed from a group of independent variations that aids in the delivery of our intended output inputs.

The machine is constantly trained in order to attain a particular level of accuracy in our training data. Linear regression, hindsight, KNN decision tree, random forest, and other guided readings are examples.

UNSUPERVISED LEARNING:

No specific aim or outcome can be estimated or predicted with this approach. It is used to join many groups, which are then divided into various groups in order to interfere. K-Means are another example of unregulated learning.



| CONCLUSIONS

The project successfully demonstrated the application of machine learning techniques in salary prediction. The linear regression model, after fine-tuning, provided the most accurate predictions among the tested models. The project highlighted the importance of data preprocessing and feature selection in building robust predictive models. The challenges encountered, including handling missing data and selecting relevant features, were effectively addressed, leading to improved model performance. Additionally, the internship included learning AI basics, which provided a foundational understanding crucial for the project.



| FUTURE SCOPE

Future work can explore the integration of additional features such as industry trends and economic indicators to enhance prediction accuracy. Implementing advanced models such as neural networks and ensemble methods could further improve performance. Additionally, deploying the model as a web application could provide real-time salary predictions for users.



| REFERENCES

<https://ieeexplore.ieee.org/document/9943146>

NoteBook Link:

https://colab.research.google.com/drive/1mRAiVI3GS1OtXB0-xNNdfQZMXS_9XQMz?usp=drive_link

https://ijirt.org/master/publishedpaper/IJIRT151548_PAPER.pdf



THANK YOU