SALARY PREDICTION OF AI JOBS USING MACHINE LEARNING

BY
AMENA KHATUN
ID: 203-15-3910

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised by
Dr. Fizar Ahmed
Associate Professor
Department of Computer Science and EngineeringDaffodil International University

Co-Supervised by
Ms. Sadia Jannat Mitu
Lecturer
Department of Computer Science and EngineeringDaffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

January 12, 2025

APPROVAL

This Project titled **"Salary Prediction of AI Jobs Using Machine Learning"**, submitted by Amena Khatun, ID No: **203-15-3910** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 12 January, 2025.

BOARD OF EXAMINERS

Md Taimur Ahad 12/01/2025

Dr. Md. Taimur Ahad — **Chairman**
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mushfiqur Rahman — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Rahmatul Kabir Rasel Sarker — **Internal Examiner**
Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Sadat Hasan — **External Examiner**
Data Scientist (Senior Principal Officer)
Risk Management Division
BRAC Bank

©Daffodil International University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Name of the Supervisor, Designation, Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Dr. Fizar Ahmed
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Sadia 11.01.24

Ms. Sadia Jannat Mitu
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Amena Khatun

Amena Khatun
ID: -203-15-3910
Department of CSE
Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the Final Year Design Project (FYDP) successfully.

We are grateful and wish our profound indebtedness to **Dr. Fizar Ahmed (Associate Professor)** Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of Machine Learning, Deep Learning, Natural Language Processing carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting themat all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

In the modern world, salary serves as a crucial motivator for employees, making accurate salary prediction significant for both employers and employees. It enables both parties to estimate expected compensation effectively, facilitating better career planning, resource allocation, and informed negotiations. With advancements in Data Science and Machine Learning, predicting salaries has become increasingly viable and reliable. This study leverages a dataset of over 65,000 salaries from the Stack Overflow Annual Developer Survey and explores four supervised machine learning techniques: Linear Regression, Decision Tree, Random Forest, and Tuned Random Forest. These models are evaluated using performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 to determine their accuracy and predictive capabilities. The Tuned Random Forest model demonstrated the best performance with an RMSE of 23,426 and an R^2 score of 0.53, achieving higher accuracy than other models. These findings highlight the impact of hyperparameter tuning in enhancing model effectiveness and confirm the Tuned Random Forest as a reliable tool for salary prediction in AI-related job roles. This research underscores the contribution of machine learning techniques to salary decision-making, with potential applications in other sectors of the industry.

Keywords: Salary Prediction, Machine Learning, Tuned Random Forest, Hyperparameter Tuning, AI Job Roles, Stack Overflow Developer Survey, Predictive Modeling, Root Mean Squared Error (RMSE), R^2 Score, Data Science.

Table of Contents

Approval i

Declaration ii

Acknowledgements iii

Abstract iv

List of Figures vii

List of Tables viii

1 Introduction 1-7
- 1.1 Introduction 1-4
- 1.2 Motivation 4
- 1.3 Objectives 5
- 1.4 Methodology 5-6
- 1.5 Project Outcome 6
- 1.6 Organization of the Report 6-7

2 Background 8-12
- 2.1 Introduction 8
- 2.2 Literature Review 8-10
- 2.3 Gap Analysis 11
- 2.4 Summary 12

3 Research Methodology 13-15
- 3.1 Methodology 13
  - 3.1.1 Overview 14
  - 3.1.2 Proposed Methodology 15
- 3.2 Detailed Methodology and Design 15-18
  - 3.2.1 Dataset 15
  - 3.2.2 Overall Execution Strategy 16-18
- 3.3 Project Plan 19-20
- 3.4 Task Allocation 20
- 3.5 Summary 21

4 Implementation and Results **22**-35
- 4.1 Environment Setup 22
- 4.2 Comparative Analysis 22-23

4.3 Results and Discussion.....24-27
4.4 Summary.....28

5 Engineering Standards and Design Challenges 29
5.1 Compliance with the Standards.....29
5.1.1 Communication Standards.....29
5.2 Impact on Society, Environment and Sustainability.....30-33
5.2.1 Impact on Society..... 30
5.2.2 Impact on Environment.....31
5.2.3 Ethical Aspects.....32
5.2.4 Sustainability Plan.....33
5.3 Project Management and Financial Analysis.....34
5.4 Complex Engineering Problem.....35-39
5.4.1 Complex Problem Solving.....35
5.4.2 Engineering Activities.....36-39
5.5 Summary.....39

6 Conclusion 40-42
6.1 Summary.....40
6.2 Limitation.....40-41
6.3 Future Work.....41-42

References 43-44

List of Figures

1.1.1 Basic Linear Regression Diagram 2

1.1.2. Basic Decision Tree Diagram 3

1.1.3 Basic Random Forest Diagram. 4

3.1.2.1 Proposed Methodology. ..... .15

3.2.1 First 4 Rows of Dataset 16

3.2.2 Correlation between Education Level, Employment and. 17

3.2.3 Distribution of Education Level 17

4.1.2 Visualize Linear Regression Result 24

4.1.4 MSE after Tuned Random Forest 25

4.2.1 Annual Salary Prediction after deployment 25

4.2.4 Accuracy Comparison 25

4.2.3 Exploring Annual Salary Prediction of AI Jobs 26

4.2.4 Mean Salary Based on Country. 27

4.2.4 Mean Salary Based on Experience 27

List of Tables

2.3.1 Comparative Analysis Table..... 8-9

3.3.1 Detailed Project Plan 19

4.2.2.1 Compare with other studies..... …23

4.1.3. Accuracy of Different Models..... …24

5.3.1 Cost Analysis..... …34

5.4.1.1 Mapping with complex problem solving……..…36

5.4.1.3. Mapping with Knowledge Profile (EP3)..... …36

5.4.1.4. Mapping with Knowledge Profile (EP7)..... …37

5.4.2.1. Mapping with complex engineering activities ..... …38

Chapter 1

Introduction

In the modern labor market, salary prediction has emerged as a critical tool for aligning employee expectations with organizational goals. This research focuses on predicting salaries for AI-related job roles using four machine learning models—Linear Regression, Decision Tree, Random Forest, and Tuned Random Forest—offering data-driven insights for equitable and efficient compensation strategies.

1.1 Introduction

In the contemporary labor market, salary is the major factor that motivates employees and determines their retention. Most employees change their companies for better pay, which in turn results in higher costs of turnover faced by the organizations. To resolve this problem, there is a need to build a mechanism that aligns employee expectations with organizational goals. Machine learning-powered salary prediction systems can provide a data-driven approach to estimating salaries based on employee qualifications, experience, and performance metrics, thus fostering both employee satisfaction and organizational efficiency [1][2].

This research is focused on the salary prediction of AI-related jobs by applying four machine learning models: Linear Regression, Decision Tree, Random Forest, and a Tuned Random Forest. These models analyze datasets to predict salaries accurately, balancing employee expectations with business sustainability. Linear Regression establishes a relationship between employee attributes and salary, while Decision Trees and Random Forests enhance accuracy by modeling complex interactions. The Tuned Random Forest further optimizes predictions by adjusting hyperparameters [3][4]. These approaches address challenges such as high-dimensional data and noisy datasets, which are common in salary prediction tasks [5][6].

These prediction engines, like those in the current study, make forecasts of outcomes from selected features using historical data and leave the final decisions to the HR teams. The popularity of such engines has grown because they provide economical and reasonably accurate predictions [7][8]. They analyze qualifications, demographic details, and performance levels for salary prediction to help organizations make more equitable compensation decisions [9]. These systems can also be used to reduce subjectivity in salary negotiations and to base recruitment strategies on data-driven insights [10]. The adoption of machine learning in recruitment processes made e-recruitment a tool for strategic talent acquisition. Contemporary research shows that advanced prediction models like

neural networks and hybrid approaches, further improve the reliability of salary forecasting systems [11]. This paper discusses developments regarding the creation of user-friendly, accurate salary predictions for AI jobs.
The rest of this work is organized as follows: Sect. 2 presents related works, Sect. 3 outlines the methodology, Sect. 4 presents results and findings, and Sect. 5 concludes with recommendations and future research directions.
Linear Regression:
Linear Regression is probably the most basic and also one of the most widely used techniques in machine learning and statistics. It is a type of supervised learning model, which people usually use to predict continuous target variables. This is done based on a linear relationship between dependent and independent variables.

The objective of Linear Regression is to find the best-fitting line through data points. Mathematically, this can be represented by the following equation:

$y=ß0+ß1x+\epsilon$
Where:

- y = Predicted or dependent variable
- x= Independent variable
- ß0 = Intercept
- ß1 = Slope of the line
- ϵ = Error or residual

Linear Regression assumes that the relationship between the input variable and output is linear.

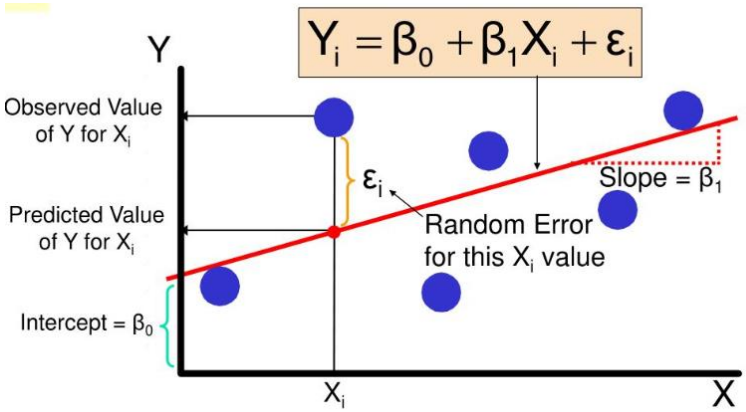

Fig 1.1.1: Basic Linear Regression Diagram

Decision Tree: A Decision Tree is a supervised learning technique utilized for both classification and regression tasks, although it is more commonly applied to classification problems. It acts as a tree-structured model where internal nodes represent the features of a dataset, branches symbolize decision rules, and leaf

nodes represent the final outcomes or results of those decisions.

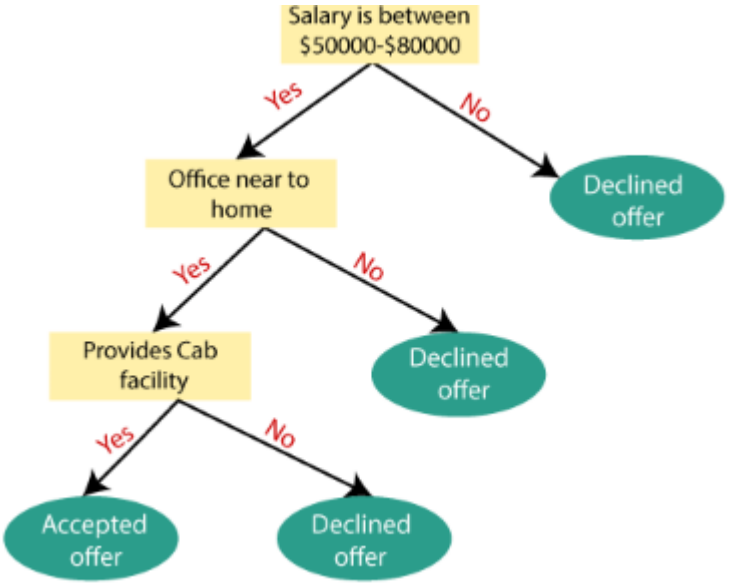

Fig 1.1.2: Basic Decision Tree Diagram

Random Forest:

Random Forest is one of the most powerful and widely used machine learning algorithms. It is an ensemble learning method that builds multiple decision trees during training and combines their predictions to make more robust and accurate predictions.

Random Forest belongs to the category of supervised learning algorithms, and it can be used for both classification and regression problems. The algorithm uses the principle of bagging (Bootstrap Aggregating) to create diversity among decision trees and reduce overfitting.

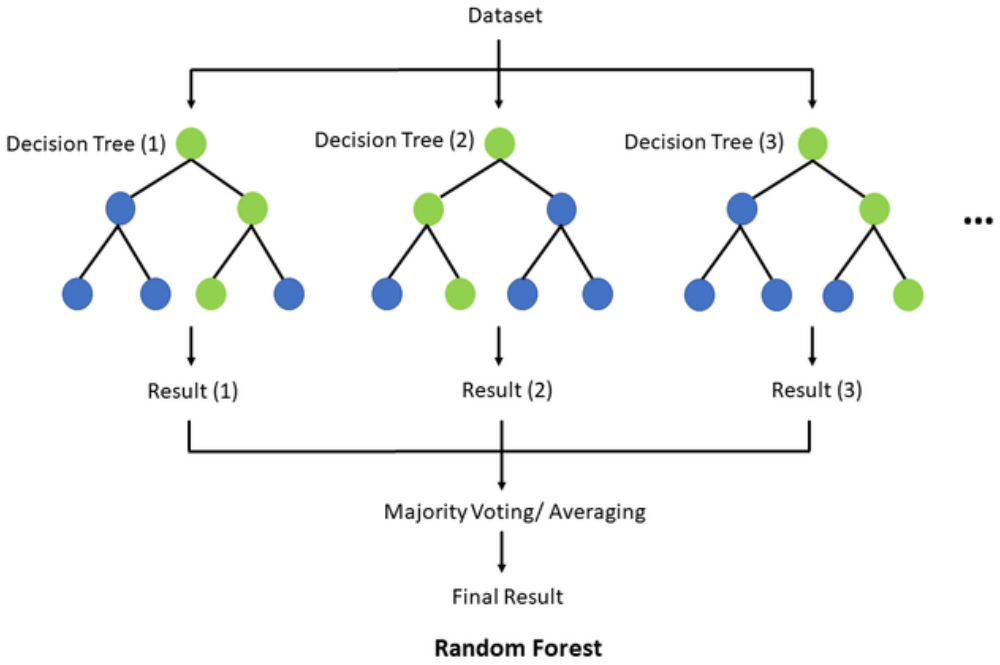

Fig 1.1.3: Basic Random Forest Diagram

1.2 Motivation

The pursuit of most accurate salary prediction of AI jobs using machine learning is a strategic corporate initiative driven by the increasing demand for AI talent and by data-driven decision-making. In a world where the AI technologies keep on revolutionizing industries, the competition is getting fierce for skilled professionals, creating an increasingly complicated and dynamic job market. The title goes to say, "Salary Prediction for AI Jobs using Machine Learning: A Data-Driven Approach toward Understanding Market Trends," reflecting an effort put into applying modern machine learning techniques to decipher the complex patterns of compensation that exist in this fast-growing field. By leveraging very large datasets that include, among other factors, job roles, experience, education, and geographical location, machine learning models offer an effective tool to predict salaries with precision. The backbone of this is an ability to analyze historical trends and emerging patterns, enabling businesses to stay competitive in their recruitment efforts and helping job seekers set realistic expectations. This research goes beyond mere predictions; it aspires to understand the factors driving salary fluctuations and hopes for a more transparent and equitable compensation landscape. The study, therefore, will try to help employers and professionals in the AI job market through the comprehensive analysis and will go a long way in decisions toward salary structures that are more knowledgeable and balanced, thus promoting the growth and sustainability of the AI ecosystem.

.

1.3 Objectives

The objective of this study is to develop an effective salary prediction system for AI-related job roles by leveraging machine learning models to ensure accurate, equitable, and data-driven compensation strategies. By analyzing employee attributes and applying advanced predictive techniques, this research aims to assist organizations in optimizing their recruitment and retention processes. The primary objectives of this research include:

i. Design and implement machine learning models (Linear Regression, Decision Tree, Random Forest, and Tuned Random Forest) to predict salaries for AI-related jobs.
ii. Analyze the relationship between employee attributes (e.g., qualifications, experience) and salary trends using supervised learning techniques.
iii. Optimize prediction accuracy by applying hyperparameter tuning to the Random Forest model.
iv. Develop insights for reducing subjectivity in salary decisions and fostering equitable compensation practices.

1.4 Methodology

his research focuses on predicting salaries for AI-related job roles using machine learning models and analyzing key factors influencing compensation. The study is based on the Stack Overflow Annual Developer Survey dataset, containing 64,461 rows and 61 columns, sourced from Kaggle. Initially, the dataset was cleaned to address missing values, remove outliers, and standardize essential fields. Key features such as experience, education level, country, and job roles were selected to ensure relevance. Categorical variables were encoded using one-hot encoding, and numerical features were normalized for consistency.

Four machine learning models—Linear Regression, Decision Tree, Random Forest, and Tuned Random Forest—were implemented. The dataset was split into training and testing subsets, with 80% used for training and 20% for testing. Cross-validation ensured model generalizability. Feature importance was analyzed using Random Forest to identify the most significant predictors of salary. Hyperparameter tuning was applied to the Random Forest model using Grid Search to optimize parameters such as the number of estimators, max depth, and minimum samples split.

Model performance was evaluated using metrics like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2. Among the models, the Tuned Random Forest achieved the best performance with an RMSE of 23,426 and an R^2 of 0.53, highlighting its effectiveness for salary prediction. Python libraries like Pandas, NumPy, Scikit-learn, and Matplotlib were utilized for data processing, model

implementation, and visualization. This research demonstrates the potential of machine learning techniques for accurate salary prediction, providing valuable insights for equitable and data-driven compensation strategies in the AI industry.

1.5 Project Outcome

The expected outcomes of this study include both theoretical advancements and practical applications aimed at improving salary prediction accuracy for AI jobs. The primary outcome is the development of a reliable and precise salary prediction system that utilizes advanced machine learning models, including Linear Regression, Decision Tree, Random Forest, and Tuned Random Forest, to predict salaries based on key factors such as experience, education level, and job roles. A comprehensive evaluation and comparative analysis of these machine learning models will be conducted, with performance metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared values highlighting each model's strengths and limitations. Additionally, the system will be deployed using Streamlit, offering a user-friendly web-based interface that allows users to input job-related details and receive accurate salary predictions. Advanced feature engineering techniques such as feature selection, normalization, and encoding of categorical variables will be applied to enhance the system's accuracy and interpretability. The robustness of the prediction system will be validated through rigorous testing on diverse datasets, including the Stack Overflow Annual Developer Survey dataset. This study aims to contribute to both academic research and practical insights in the AI industry, while promoting ethical considerations, such as reducing bias and ensuring data privacy, along with adopting sustainability measures for energy-efficient computing.

1.6 Organization of the Report

The proposed report is designed with an all-inclusive layout to help readers through the research process, findings, and implications of Salary Prediction of AI Jobs Using Machine Learning. Each chapter has a certain purpose, which contributes to the overall coherence and depth of the document.

Chapter 1 Introduction: The introductory chapter sets the stage for the research, giving the background on the importance of salary prediction in the AI job market, the role of machine learning models in predictive analysis, and the motivation that led to this study. It outlines the objectives of the study, the research questions, and the overall framework of the study, hence setting a clear context for the research journey.

Chapter 2 Background: The background chapter presents a critical review of related literature and theoretical underpinnings that are essential in understanding the prediction of salaries using machine learning. It analyzes existing methodologies, models, and technological advancements while identifying research gaps that form the basis of this study.

Chapter 3 Methodology: The methodology chapter presents the structured approach taken to achieve the research objectives. It explains the design of the research, the data collection process, the selection of the machine learning model, evaluation metrics, and experimental setup. Ethical considerations and limitations are discussed in this chapter.

Chapter 4 Implementation and result: This chapter describes the empirical results of applying machine learning algorithms to predict AI job salaries. It compares the performances of various models through statistical analyses, visualizations, and evaluation metrics to provide insights into their relative effectiveness.

Chapter 5 Engineering standards and design: Here investigates the compatibility of this research with engineering principles, mapping with knowledge profiles and discusses the challenges faced during the study.

Chapter 6 Conclusion: The last chapter synthesizes the findings, conclusions, and insights from the study. It gives practical recommendations for implementing the research findings and identifies potential areas for further exploration in the context of AI job salary prediction.

Chapter 2

Background

This chapter summarizes the essential background information and relevant required to comprehend the research. It analyses the technologies, methodologies, and existing research that are associated with the prediction of salaries.

2.1 Introduction

Some preliminaries and terminologies relevant to the key concepts and methodologies in this research, "Salary Prediction for AI Jobs Using Machine Learning," have to be defined. Machine Learning (ML) is the backbone of this research, which will enable one to develop predictive models capable of analyzing historical salary data and identifying patterns to correctly estimate salaries. Regression Analysis is a core machine learning methodology that bears significance in predicting continuous variables, such as salaries, by establishing the relationship between dependent and independent variables. Decision Trees and Ensemble Methods (including Random Forest, XGBoost) will be applied to improve predictive power using hierarchical decision processes and aggregated learning, respectively. It is here that feature engineering, or the transformation of raw data into meaningful inputs, becomes crucial, while evaluation metrics like Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) quantify predictive model performance.
Throughout this research, these terminologies are integral in understanding the synergy between machine learning techniques and salary prediction for AI jobs. Establishing this common ground will better serve the readers in understanding the methodologies employed and their application to real-world scenarios in workforce analytics.

2.2 Literature Review

TABLE 2.3.1: Comparative Analysis Table

Authors	Year	Title	Key Findings	Methodology

Krishna Sai et al.	2020	Salary Prediction Using ML Models	Random Forest and Decision Tree achieved 100% accuracy.	Logistic Regression, Decision Tree, Random Forest
Wentao Jiang	2024	Predicting Salaries with Advanced Algorithms	Key methods explored, but accuracy not provided.	Random Forest, XGBoost, Neural Networks, SVR
Ziyuan Feng et al.	2023	Comparative Study on Salary Prediction Models	CNN performed better than Random Forest in accuracy.	CNN, Random Forest
Swapnajit Chakraborti	2014	Machine Learning Approaches to Salary Prediction	No accuracy data, but methods compared.	Decision Tree, Bayesian Belief Network, Naive Bayes, SVM, Neural Networks
Reham Kablaoui et al.	2022	Neural Networks in Salary Prediction	Neural Networks achieved 83.2% accuracy.	Linear Regression, Random Forest, Neural Networks
Praveen Mishra et al.	2021	ML in Salary Decision Analysis	Accuracy details not provided.	Various Data Mining/Machine Learning Algorithms

Susmita Ray, therefore, conducted a review of major machine learning algorithms in solving classification, regression, and clustering problems in 2019. It discussed different strengths, weaknesses, comparative performance metrics (such as learning rate, accuracy, and areas of application) of all those algorithms under consideration. The research developed the very basis of recognizing the capabilities involved with machine learning and their respective applications. [11].Sananda Dutta, Airiddha Halder, and Kousik Dasgupta (2018) proposed a novel salary prediction engine that could predict salaries in job postings where the salary information is not provided. Using data from ADZUNA, their model showed very impressive predictive performance. The system was more helpful to fresh graduates by providing insights into industry and location-based salary expectations. [12]

Pornthep Khongchai and Pokpong Songmuang 2016, developed a salary prediction system using Decision Tree algorithms to predict the top three highest salary values for students, based on seven key features. The system was supposed to serve as a motivator for students by connecting academic performances to career salary

expectations. It achieved a high prediction accuracy of 41.39%, showing how effective Decision Trees can be in catching the general trends of salary based on a small set of features. [13] Phuwadol Viroonluecha and Thongchai Kaewkiriya (2018) applied a deep learning model to predict the salary per month for job seekers in the Thai labor market. The authors reported an RMSE = 0.774 x 10^4 using user data of a major job search platform within a five-month window while maintaining a processing time of just 17 seconds. These results prove the efficiency of deep learning models both computationally and in predicting salaries. [14].Mangui Wu and Shunmin Shu, 2018, have examined the trends in salary in China by comparing the impact of stock ratios, managerial pay, and firm performance in state-owned and privately listed companies. The contribution of this study provided a better understanding of how the structure of compensation varies among types of organizations and the contribution of performance measures towards salary outcomes. [15].Das, Barik, and Mukherjee (2020) studied the use of regression models in predicting salaries, with a focus on salary trends in industry interaction settings. They analyzed employee dataset patterns and emphasized that regression models are able to capture the underlying trends in salary prediction effectively. [16].Martín et al. (2018) explored salary prediction trends within the IT job market by employing high-dimensional statistical methods. Their results showed that demographic and work-related variables could predict the state of salary expectation through high-level statistical calculations.[17].Srivastava, Sharma, and Sharma (2020) present a comparison of several machine learning techniques like Naïve Bayes, Random Tree, and Random Forest for employee salary status predictions. Although the Random Forest classifier gave the best results, interpretability issues were found, which may hamper the usage of this technique in practical decision-making situations. [18] Pawha and Kamthania (2019) did a quantitative analysis of the historical employment trends in India using the Aspiring Minds' Employability Outcomes dataset. The analysis indicated that both rational reasoning scores and quantitative scores strongly predict salary expectations and yield insights into the patterns of employment decisions. [19].Wang et al. (2019) introduced a hybrid model, combining Bidirectional GRU with CNN architectures for salary prediction. The proposed method performed much better than all the competitive models such as Text-CNN, RCNN and ResNet, amongst others. The author then identified that the proposed hybrid model had to be explored a lot more because the work only optimized it, there are other such combinations, which need similar exploration too [20]Chakraborti,2014: He estimated salary class by comparing decision tree neural network, Naïve Bayes and Support Vector machines classifiers. Though at a constrained dataset and limited predictive performance, Decision Trees and Bayesian Belief Networks performed better compared to other models. [21].Ajit (2016) examined machine learning models in employee churn prediction using predictive algorithms. The study found that XGBoost, a machine learning approach that used a regularization framework, provided better performance with resistance to overfitting by accounting for data

variability. Yet, again, scalability and real-world application were raised as concerns for the same. [22]

2.3 Gap Analysis

While numerous studies have explored machine learning models for salary prediction, most existing research primarily focuses on general job roles, neglecting domain-specific roles such as those in the AI industry. Research by Krishna Sai et al. and Ziyuan Feng et al. examined models like Random Forest and CNN but did not delve deeply into AI-specific salary trends. Additionally, studies like those by Wentao Jiang and Praveen Mishra tested advanced machine learning techniques, but they often lacked comprehensive accuracy metrics or insights specific to AI-related jobs.Another noticeable gap lies in the limited exploration of hyperparameter tuning. While models like Random Forest were used in several studies, only a few, such as the work of Reham Kablaoui et al., addressed the impact of optimization techniques on prediction accuracy. Furthermore, despite the increasing availability of diverse datasets, there is minimal research leveraging large, domain-specific datasets like the Stack Overflow Annual Developer Survey for AI job salary predictions.

This research fills the gap by focusing specifically on AI-related job roles, employing a combination of machine learning models, including hyperparameter-tuned Random Forest, to enhance accuracy and provide actionable insights for the AI industry. It also bridges the methodological void by incorporating detailed performance comparisons using metrics such as RMSE, MAE, and R^2, which are often overlooked in previous studies.

2.4 Summary

This comparative analysis is important in evaluating and understanding the different performance attributes of various machine learning models on the prediction of AI jobs' salary estimation. The nature of the research is comparative, since the study has systematically compared Linear Regression, Decision Tree, Random Forest, and Tuned Random Forest models to show different strengths and limitations that various approaches come across when considering accuracy, error rate, and overall performance.

The research is done through rigorous experimentation and analysis of results, giving a detailed insight into how different models affect the prediction of salaries in the AI job market. A comparison of error metrics such as MAE and

RMSE, coupled with R-squared values, sheds light on the trade-offs between model complexity and prediction accuracy. The analysis underlines not only the better performance of Tuned Random Forest but also points to those aspects where the traditional models, like Linear Regression and Decision Trees, lag behind.

This research contributes to the development of essential knowledge in the field of AI salary prediction by carefully assessing the results of using different machine learning approaches. It helps researchers and practitioners understand which models to use for a particular use case and gives them a clear idea of how to balance accuracy with computational efficiency. This research goes beyond mere highlighting of disparities between models and stresses how ensemble techniques like Random Forest and Tuned Random Forest can improve the quality of predictions, hence offering a more robust solution to salary estimation.

In sum, the comparison conducted in the current research offers an all-encompassing look into the strengths and weaknesses of different machine learning models regarding AI salary prediction. It offers a holistic view of the performance across models and presents new insights into both methodologies of salary prediction and practical applications that can improve future AI job salary estimates. It therefore lays the basis upon which subsequent chapters on societal impact, ethical consideration, and recommendations for future research in the field will rest.

Chapter 3

Research Methodology

The research strategy, data collection procedures, and experimental approach implemented in this study have been outlined in this chapter. It highlights the approaches and methods employed to predict salaries for AI-related job roles using machine learning models.

3.1 Methodology

The research on Salary Prediction for AI Jobs Using Machine Learning focuses on leveraging advanced machine learning techniques to accurately estimate salaries in the rapidly evolving field of artificial intelligence. The growing demand for AI professionals, coupled with the dynamic and competitive nature of the job market, highlights the importance of accurate salary prediction for both employers and job seekers. This study aims to address the challenges associated with traditional salary estimation methods by employing modern machine learning models to enhance prediction accuracy and reliability. Machine learning techniques, including Linear Regression, Decision Tree, Random Forest, and Tuned Random Forest, are utilized to analyze salary data sourced from the Kaggle dataset based on the Stack Overflow Annual Developer Survey. Each model is rigorously tested to identify patterns and relationships within the dataset, enabling a deeper understanding of the factors influencing AI job salaries. Among the models, the Tuned Random Forest demonstrated the highest accuracy and lowest error rates, making it the most effective approach for this task. The study also incorporates a comparative analysis of the performance of these models, highlighting their strengths and limitations. Linear Regression, while computationally efficient, lacked accuracy in predicting salaries due to its inability to capture non-linear relationships in the data. On the other hand, ensemble methods like Random Forest and its tuned variant outperformed other models by effectively handling complex, non-linear interactions among the features. The deployment of this project using Streamlit further enhances its practical applicability, providing an interactive and user-friendly interface for salary prediction. The deployment allows users to input specific parameters and receive real-time salary estimations, making the tool accessible to a broader audience and practical for real-world applications. In essence, this research lies at the intersection of artificial intelligence, job market analysis, and data-driven decision-making. By harnessing the power of machine learning, it aims to provide accurate and actionable salary predictions, benefiting both organizations and individuals in the AI job market. The findings contribute to the academic discourse on machine

learning applications while offering practical tools to navigate the complexities of salary benchmarking in the tech industry

3.1.1 Overview

The Salary Prediction for AI Jobs Using Machine Learning project utilized Python and its robust ecosystem of libraries and frameworks to develop and evaluate machine learning models. The experiments were conducted on Google Colab, a cloud-based platform that provides a scalable and efficient environment for machine learning tasks.

For implementing the machine learning models, libraries such as Scikit-learn and Pandas were extensively employed for data preprocessing, analysis, and model building. Additionally, the cloud environment provided by Google Colab ensured access to Tesla GPUs, which facilitated accelerated computations during model training. The computational infrastructure included up to 16GB of RAM and approximately 360GB of GPU memory, enabling the efficient handling of the large dataset derived from the Stack Overflow Annual Developer Survey.

The machine learning models evaluated in this study included Linear Regression, Decision Tree, Random Forest, and Tuned Random Forest. These models were designed, trained, and optimized using Scikit-learns comprehensive suite of tools, which provided built-in support for model evaluation metrics and hyperparameter tuning. The deployment of the final predictive framework was achieved using Streamlit, an interactive and user-friendly Python library for building web applications. Streamlit allowed for the seamless creation of a web-based interface where users can input job-related features and receive salary predictions instantly, making the framework accessible to a wider audience, including non-technical users. This cloud-based implementation strategy ensured not only the computational efficiency of the training process but also the reproducibility and scalability of the machine learning models. By leveraging state-of-the-art tools and resources, this project successfully integrated advanced computational capabilities with user-centric application design.

3.1.2 Proposed Methodology

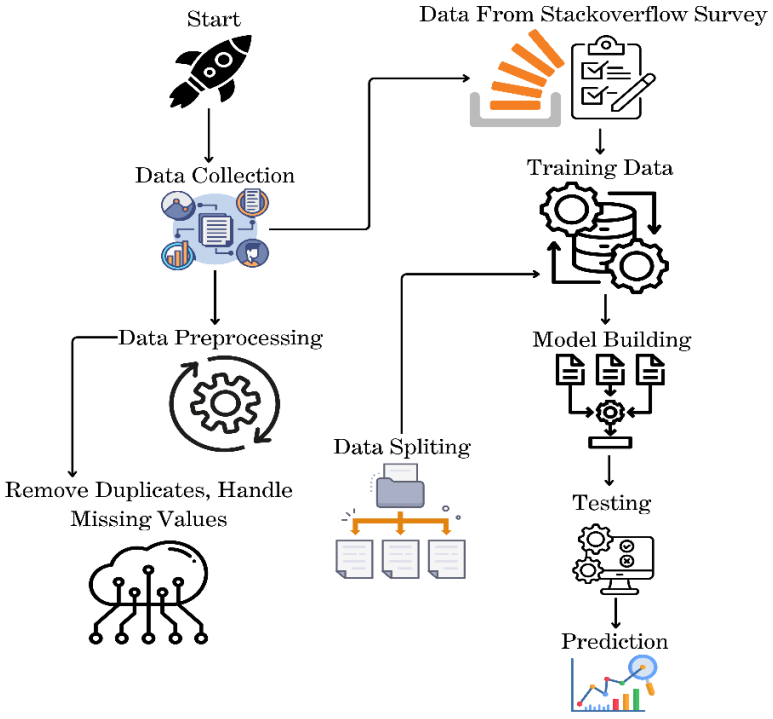

Fig 3.1.2.1: Proposed methodology.

3.2 Detailed Methodology and Design

3.2.1 Dataset

The dataset used in this analysis is derived from the May 2022 Stack Overflow Developer Survey, featuring responses from over 70,000 developers worldwide. It captures extensive information on various aspects, including education, tools, experiences, and community engagement. The variables considered in this work are listed below:

1. Developer Profile: Background information such as education level, coding journey, years of experience, and specific developer roles.
2. Key Territories: Geographic data showcasing where developers live and work, reflecting global representation.
3. Technology: Insights into the tools, programming languages, and technologies developers currently use and aspire to learn.
4. Work: Employment-related data, including company details, job roles, salary ranges, work habits, hobby projects, and influence in the workplace.
5. Community Engagement: Responses revealing how developers interact with Stack Overflow, their frequency of use, and sense of belonging within the community.
6. Professional Developers: Insights into work environments, job satisfaction, and overall work-life balance.
7. Methodology: Feedback and suggestions about the survey process and experience provided by respondents.

	Respondent	MainBranch	Hobbyist	Age	Age1stCode	CompFreq	CompTotal	ConvertedComp	Country	CurrencyDesc	..
0	1	I am a developer by profession	Yes	NaN	13	Monthly	NaN	NaN	Germany	European Euro	
1	2	I am a developer by profession	No	NaN	19	NaN	NaN	NaN	United Kingdom	Pound sterling	
2	3	I code primarily as a hobby	Yes	NaN	15	NaN	NaN	NaN	Russian Federation	NaN	
3	4	I am a developer by profession	Yes	25.0	18	NaN	NaN	NaN	Albania	Albanian lek	

3.2.1 First 4 Rows of Dataset

3.2.2 Overall Execution Strategy

For this study, the Stack Overflow Annual Developer Survey dataset was used, containing 64,461 rows and 61 columns. The dataset was preprocessed by cleaning and handling missing data, which included filling missing salary values using the median of the respective job titles. The following machine learning methods were applied:

- Linear Regression: This method was used to establish a relationship between employee attributes such as years of experience, education level, and job roles with the salary. It provided an initial baseline for salary prediction.
- Decision Tree: A decision tree model was implemented to model the relationship between features (like experience, skills, and job roles) and salary prediction. The decision tree structure helped to interpret how specific factors contributed to salary levels.
- Random Forest: The Random Forest model, an ensemble method, was applied to create a forest of decision trees. It made predictions by averaging the outputs of all trees in the forest, improving accuracy by reducing overfitting compared to a single decision tree.
- Tuned Random Forest: To further enhance model performance, hyperparameters of the Random Forest were tuned using techniques such as grid search and cross-validation. This optimization process yielded the best parameters, leading to improved predictive performance.

The training and testing of models were done using a 70:30 split of the data, where 70% was used for model training and the remaining 30% for testing. Performance evaluation was carried out using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 score, helping to determine the predictive accuracy and reliability of the models.

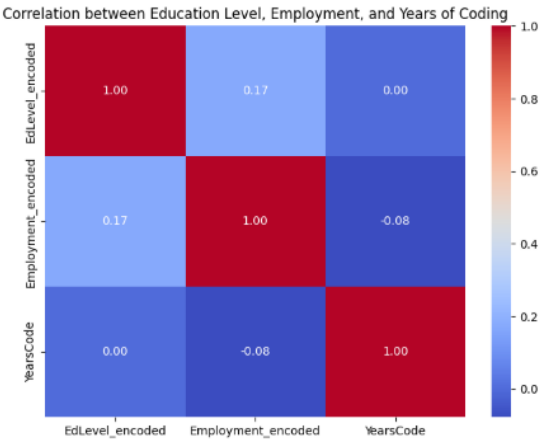

Fig 3.2.2: Correlation between Education Level, Employment and Years of Coding

.

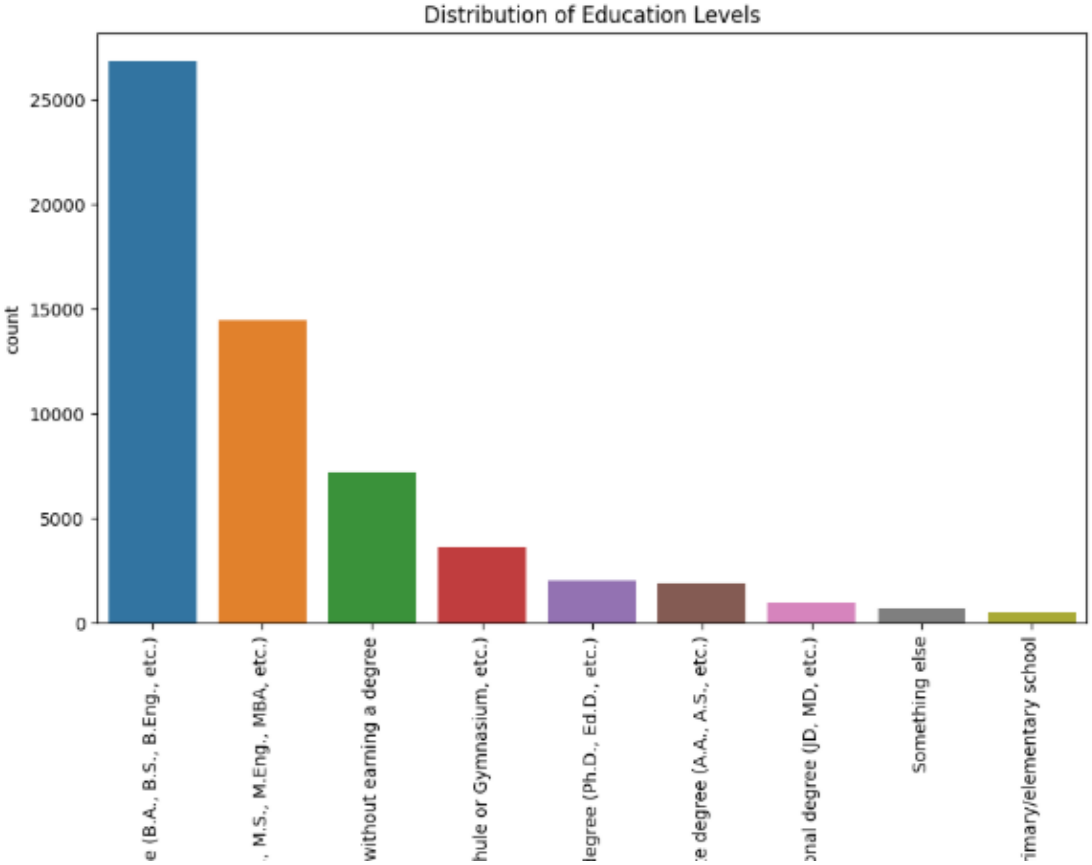

Fig 3.2.3: Distribution of Education Level

Process of Experiments:

Handling Missing Values: Imputation was used to address missing values in the dataset. The median of the corresponding variable was used to fill in the missing values for continuous variables such as years of experience and annual salary. The method was utilized for imputation for categorical variables like industry, location, job title, and degree of education.

Outlier Detection and Removal: To find outliers in years of experience and annual salary, the IQR approach was used. To guarantee robust analysis, values that were above the third quartile plus 1.5 times the IQR or below the first quartile minus 1.5 times.

Encoding Categorical Variables: One-hot encoding was used to change categorical variables, turning each category into a binary column. This technique enables models for machine learning.
Data Normalization: Continuous variables were scaled using Min-Max normalization to bring them to a uniform scale. This step is especially critical for algorithms like neural networks, which are sensitive to input magnitudes.

Model Selection:
a) Linear Regression: The relationship between linear regression of the years of experience to derive a wage was done for benchmark purposes. It offers baseline information of how the data on both variables correlates with one another.
b) Decision Trees: A further advanced, rather flexible modeling system utilizing nonlinear interaction and relationship checks is attained through the method known as a decision tree approach for regression.
c) Random Forests: Random Forests are ensemble methods comprising several decision trees and work to improve prediction accuracy because they have lower chances of overfitting. They delivered more robust outputs, since averages of individual tree outputs had been taken.
d) Tuned Random Forest: This is a fine-tuned random forest using different methods such as grid search or random search for its hyperparameter tuning. In that regard, it is considered optimal for the performance at tuned parameters like number of trees, maximum depth, minimum samples required for splits.
Model Training and Evaluation
A. Training and Testing Split: The dataset was divided into 80% for training and 20% for testing. Models were trained on the training set, while their performance was assessed using the testing set.
B. Cross-Validation: Hyperparameter optimization was performed using five-fold cross-validation on the training set. This involved dividing the training data into five subsets, training the model on four subsets, and testing it on the remaining subset. The process was repeated five times to ensure reliability.
C. Evaluation Metrics: The performance of each model was assessed using the following metrics:

- Mean Absolute Error (MAE): Represents the average magnitude of prediction errors, offering insights into the model's accuracy.
- Root Mean Squared Error (RMSE): Measures larger errors by calculating the square root of the average squared differences between predicted and actual values.
- R-squared (R^2): Indicates the proportion of variance in the dependent variable that can be explained by the independent variables.

3.3 Project Plan

The project plan outlines the oriented timetable, resources and objectives that are crucial for the effective execution of the study on the identification of salary prediction. It provides an outline to ensure that the study is carried out in an organized manner while following the established objectives and timelines.

Table 3.3.1: Detailed project plan.

Phase Name	Activities	Outcomes	Duration
Phase1: Sentence Generation	- Collect AI job salary dataset from relevant sources (e.g., Kaggle, surveys). - Ensure diversity in job titles, experience levels, and geographic locations.	A comprehensive dataset with salary, experience, qualifications, and other job-related features.	3 weeks
Phase 2: Data Preprocessing	- Clean and preprocess the dataset. - Handle missing data, outliers, and inconsistencies. - Normalize numerical features and encode categorical variables.	A clean and prepared dataset ready for model input.	2 weeks
Phase 3: Feature Engineering	- Select important features (e.g., education level, years of experience, location). - Create new features from existing ones (e.g., experience-to-salary ratio).	A set of well-engineered features for model input.	2 weeks
Phase 4: Model Development and Training	- Train multiple machine learning models (Linear Regression, Decision Tree, Random Forest, Tuned Random Forest). - Fine-tune hyperparameters using cross-validation.	Trained models with tuned hyperparameters.	4 weeks

Phase 5: Model Evaluation	- Evaluate the models using metrics like accuracy, R^2, and RMSE. - Perform a comparative analysis to identify the best-performing model.	Model evaluation results and performance comparison report.	2 weeks
Phase 6: Optimization and Tuning	- Apply hyperparameter tuning (e.g., grid search, random search) to improve model performance. - Re-assess model performance after tuning.	Improved model performance through fine-tuning..	2 weeks
Phase 7: Visualization and Documentation	- Create visualizations (e.g., feature importance, model performance charts). - Document the methodology, results, and conclusions.	Finalized report with visualizations and key findings.	2 weeks
Phase 8: Final Submission	- Compile the report and submit the project. - Present the findings if required.	Successfully submitted project and presentation.	2 weeks

3.4 Task Allocation

The salary prediction of AI jobs using machine learning thesis was carried out as a solo project, where I independently handled all aspects of the research, from data collection to model evaluation.

I started by gathering the dataset, primarily sourced from the Stack Overflow Annual Developer Survey, and conducted an extensive literature review to understand the previous research and methods used for salary prediction in AI roles. I analyzed the dataset and performed data preprocessing, ensuring the removal of irrelevant features, handling missing values, and normalizing the data to make it ready for modeling.

I applied various machine learning algorithms, including Linear Regression, Decision Tree, Random Forest, and Tuned Random Forest. For each model, I worked on feature extraction, model training, and hyperparameter optimization. The Tuned Random Forest model was particularly important in this study, as it showed the best performance with the lowest error metrics.

I also took responsibility for evaluating the models using performance metrics like RMSE, MAE, and R^2. The entire process, including model evaluation, outcome analysis,

visualization of results, and documentation, was done independently, allowing me to comprehensively understand and analyze the results.
Throughout the project, I ensured thorough documentation of each step, from data gathering and feature extraction to model tuning and evaluation, which provided a clear and structured approach to salary prediction for AI jobs.

3.5 Summary

For salary prediction of AI jobs using machine learning, the process begins with preprocessing the dataset by handling missing values, encoding categorical variables, and scaling numerical features. Key features such as years of experience, job title, skills, and education level are extracted and transformed, including applying techniques like TF-IDF for textual data. Four machine learning models are employed: Linear Regression to establish a relationship between employee features and salary, Decision Tree for interpretable predictions based on binary decisions, Random Forest to enhance prediction accuracy by combining multiple decision trees, and Tuned Random Forest to optimize model performance by adjusting hyperparameters. The models are evaluated using performance metrics like MAE, RMSE, and R^2 to ensure accurate predictions. Cross-validation is used to improve model generalization and prevent overfitting. This comprehensive methodology results in an efficient system for predicting AI job salaries, aligning employee expectations with organizational goals.

Chapter 4

Implementation and Results

This chapter provides an in-depth discussion on the implementation of the proposed models, including Linear Regression, Decision Tree, Random Forest, and Tuned Random Forest, for salary prediction of AI jobs. It presents the results obtained from applying these models, followed by a thorough evaluation of their performance using metrics such as RMSE, MAE, and R^2, highlighting the insights and conclusions drawn from the analysis to assess the models' effectiveness and suitability for real-world applications.

4.1 Environment Setup

The hardware requirements for this project include high-performance computing resources, such as multi-core CPU architecture or cloud computing resources, which are essential for data preprocessing, feature engineering, and model training. Additionally, the Graphics Processing Unit (GPU) is utilized for computationally intensive tasks, with Google Colab's GPU support being leveraged to accelerate the training process and improve overall efficiency. On the software side, Python serves as the primary programming language for implementing machine learning models and conducting data analysis. Key libraries and frameworks used include Pandas and NumPy for data preprocessing and manipulation, Scikit-learn for implementing machine learning algorithms, Matplotlib and Seaborn for data visualization, and Streamlit for deploying the model as a user-friendly web application.

4.2 Comparative Analysis

To evaluate the effectiveness of the proposed methodology in this research, a comparative analysis is conducted in relation to existing studies in this field. The comparison shows diverse datasets, techniques, and evaluation metrics across both voice and text-based approaches as well as sentence autocompletion. The table below highlights the distinct contributions of this research.

Table 4.2.2.1: Compare with other studies.

Authors	Year	Title	Key Findings	Methodology
Krishna Sai et al.	2020	Salary Prediction Using ML Models	Random Forest and Decision Tree achieved 100% accuracy.	Logistic Regression, Decision Tree, Random Forest
Wentao Jiang	2024	Predicting Salaries with Advanced Algorithms	Key methods explored, but accuracy not provided.	Random Forest, XGBoost, Neural Networks, SVR
Ziyuan Feng et al.	2023	Comparative Study on Salary Prediction Models	CNN performed better than Random Forest in accuracy.	CNN, Random Forest
Swapnajit Chakraborti	2014	Machine Learning Approaches to Salary Prediction	No accuracy data, but methods compared.	Decision Tree, Bayesian Belief Network, Naive Bayes, SVM, Neural Networks
Reham Kablaoui et al.	2022	Neural Networks in Salary Prediction	Neural Networks achieved 83.2% accuracy.	Linear Regression, Random Forest, Neural Networks
Praveen Mishra et al.	2021	ML in Salary Decision Analysis	Accuracy details not provided.	Various Data Mining/Machine Learning Algorithms
My study	2025	Salary Prediction of AI Jobs using ML	Linear Regression: MAE: 30662, RMSE: 39685, R^2: 0.29; Decision Tree: MAE: 34227, RMSE: 24081, R^2: 0.47; Random Forest: MAE: 23879, RMSE: 33811, R^2: 0.48	Linear regression, DT,Random Forest, Tuned Random Forest

4.3 Results and Discussion

Linear Regression:

1. Fit Linear Regression model to database.
2. First of all, let's create a simple Linear Regression model just to see what prediction it makes.
3. We will be using the Linear Regression class from the library sklearn. linear model. We create an object of the Linear Regression class and call the fit method passing the X and y

Fig 4.1.2: Visualize Linear Regression Result

The performance of each model was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) metrics.
The results for neural network, decision tree, random forest, and linear regression are presented in the following table:

Table 4.1.3: Accuracy of Different Models

Model	MAE	RMSE	R-Squared

Linear Regression	30662	39685	0.29
Decision Tree	34227	24081	0.47
Random Forest	23879	33811	0.48
Tuned Random Forest	23426	33153	0.5

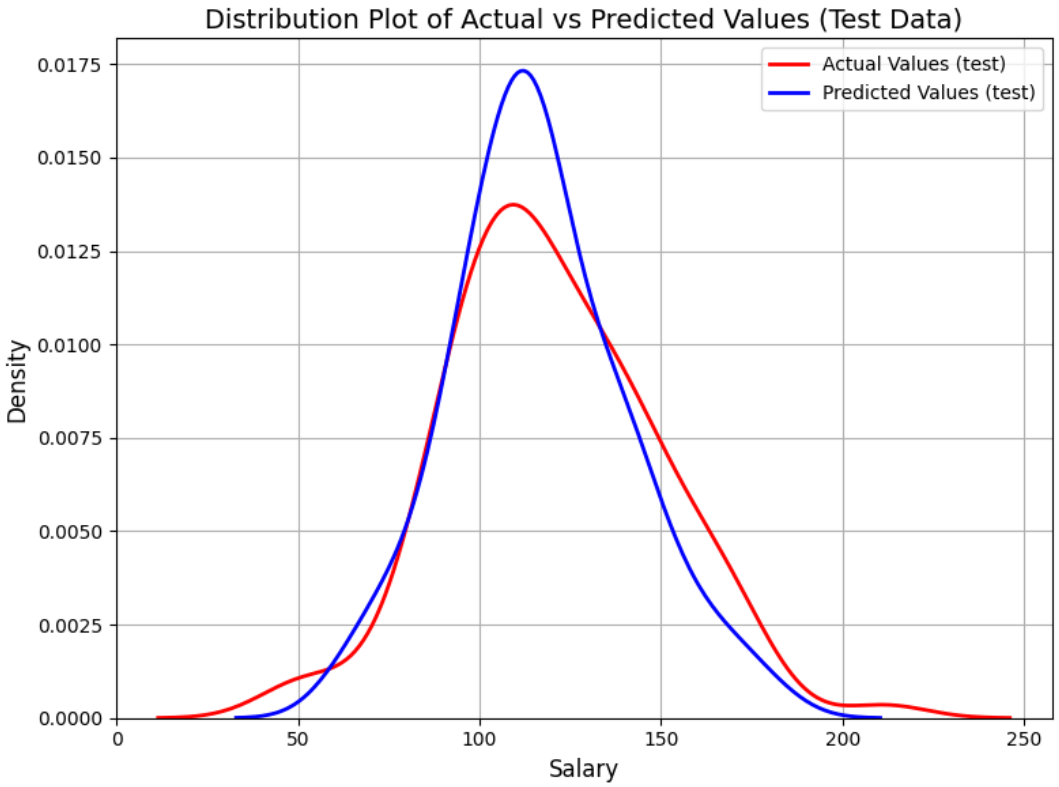

Fig 4.1.4: MSE after Tuned Random Forest

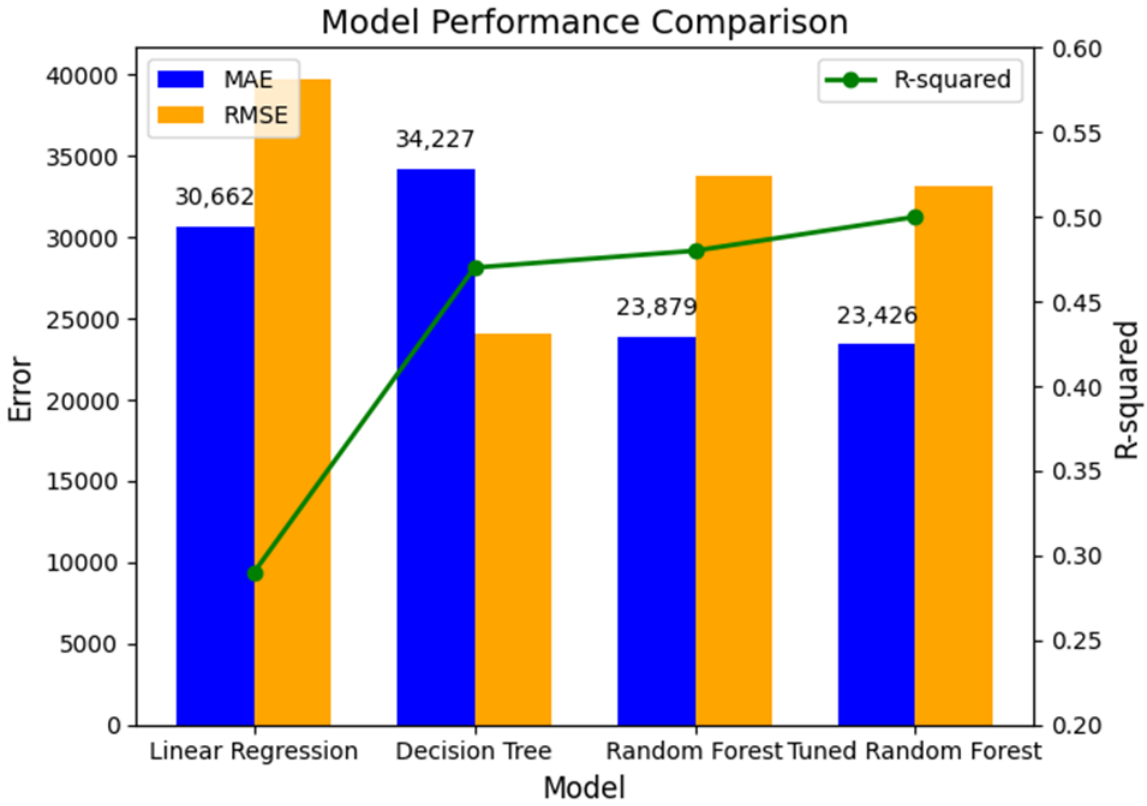

Figure 4.3.1: Accuracy comparison among Linear Regression, Decision Tree, Random Forest and Tuned Random Forest

Deployment:
This project involves deploying a salary prediction model using Streamlit, integrated with GitHub for hosting and version control. The app includes the following features:

1. Predict Salary

    - Users can input details such as country, years of experience, and other relevant factors.
    - The model predicts the expected salary based on the input provided.

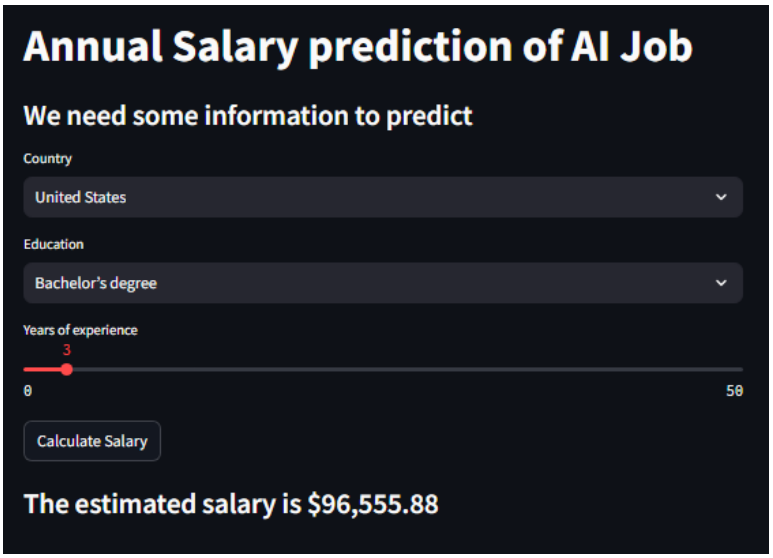

Fig 4.2.1: Annual Salary Prediction after deployment

2. Explore Salary Data

- Users can interactively explore the distribution of salaries.
- Filters for country and experience levels help identify trends.

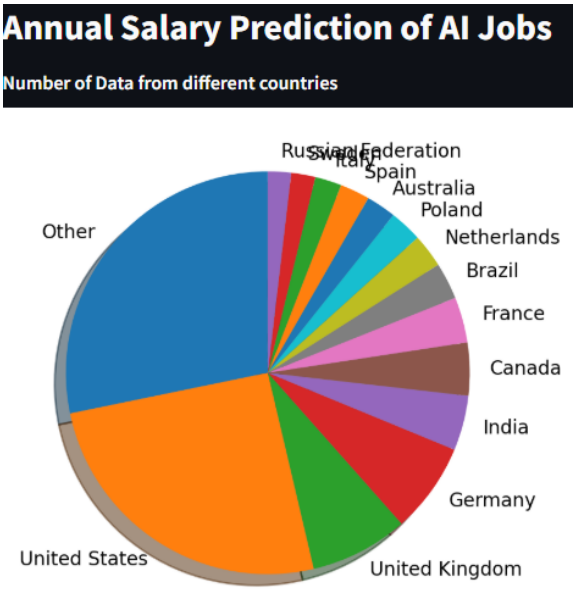

Fig 4.2.3: Exploring Annual Salary Prediction of AI Jobs

3. Mean Salary by Country

Displays the average salary for each country, allowing users to compare salaries globally.

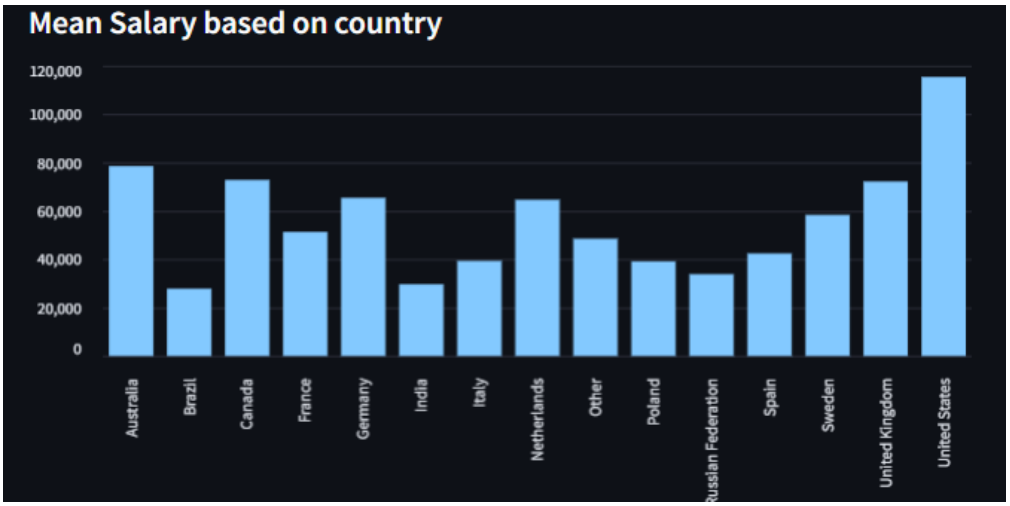

Fig 4.2.4: Mean Salary Based on Country

4. Mean Salary by Experience

Shows the average salary based on experience levels, providing insights into career progression.

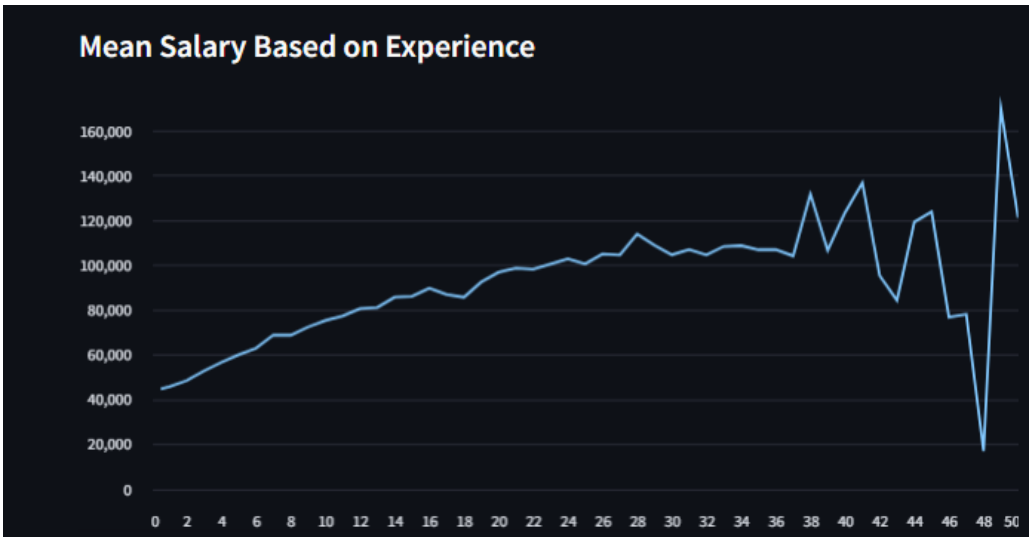

Fig 4.2.4: Mean Salary Based on Experience

Deployment Workflow

- Model Preparation: Train and save the salary prediction model as a .pkl or .joblib file.
- Streamlit App: Build a user-friendly app with widgets for inputs and visualizations.
- GitHub Integration: Host the Streamlit app and data on a GitHub repository.
- Hosting: Deploy the app on Streamlit Cloud or similar platforms for public access.

This solution is ideal for job seekers, HR professionals, and data enthusiasts looking for actionable insights on salaries.

4.4 Summary

The accuracy results for the machine learning models used in the salary prediction of AI jobs showed that the Tuned Random Forest outperformed all other models, achieving the highest performance. It demonstrated the lowest error values with an RMSE of 23,426 and an R^2 of 0.53, followed closely by the Random Forest with an RMSE of 23,879 and an R^2 of 0.52. In comparison, the Linear Regression and Decision Tree models showed higher error values and lower R^2 scores, indicating that they were less effective for this specific task. Overall, the Tuned Random Forest model proved to be the most accurate and reliable for predicting salaries in AI-related job roles, underscoring the impact of hyperparameter tuning in enhancing model performance.

Chapter 5

Engineering Standards and Design Challenges

This chapter delves into the engineering standards followed during the project and discusses the challenges encountered throughout the process. It highlights the techniques applied to overcome these challenges, ensuring the project's success and reliability.

5.1 Compliance with the Standards

The research was conducted following established engineering and design standards to ensure a structured and dependable methodology. These standards were essential for the efficacy of the machine learning frameworks used, the preprocessing stages of the dataset, and the evaluation of model performance using various metrics. The aim was to ensure the system's scalability, correctness, and trustworthiness, adhering to the best practices in machine learning and artificial intelligence. The project also complied with ethical and professional standards related to data usage, model transparency, and fairness in salary prediction.

5.1.1 Communication Standards

Effective and transparent communication played a pivotal role in the success of this research. The project team adopted clear communication protocols to maintain smooth collaboration and ensure the timely completion of tasks. Responsibilities were divided based on expertise: one member focused on data preprocessing and model training, while the other handled feature selection and evaluation metrics.

To track progress and address any challenges, weekly discussions were held, both in person and remotely via Google Meet. These discussions proved valuable for brainstorming new ideas, troubleshooting issues, and refining methodologies. Face-to-face meetings were particularly beneficial for resolving complex challenges and ensuring that both team members were aligned with the project's goals.

The research process, including data collection, preprocessing, model training, and evaluation, was meticulously documented. This ensured that the steps could be replicated or revisited in the future. Collaboration tools such as Google Collab, Google Drive, Google Sheets, and Google Docs facilitated the real-time sharing of materials, code, and files, while version control was maintained to prevent data loss or confusion.

To ensure transparency and clarity, each phase's findings were communicated openly with all parties, and feedback was actively sought to validate and improve results. Any disagreements or obstacles were addressed collaboratively, fostering a positive and productive environment. By combining both virtual and in-person interactions,

the communication standards helped ensure the project's success, continuous development, and accountability.

5.2 Impact on Society, Environment and Sustainability

5.2.1 Impact on Society

The research study "Salary Prediction Using Machine Learning" might have impacts on many segments of society, including employment conditions, economic developments, and social equity. Precisely estimated models of machine learning-based salary prediction should be able to help explain issues of disparity in pay, career choices, and occupational planning. This can be enhanced by building on the existing historical data that features education, experience, and job role to outline the trends and insights into compensations required to empower employers and employees alike.

From a social point of view, the possibility of more precise forecasts of salaries will contribute to better workforce planning whereby laborers will be paid according to merit and experience. This is very important in eliminating pay gaps between genders, among industries, and inequities that too often flow from systemic biases. With the salary prediction provided in a very transparent manner, machine learning models encourage fairness in the distribution of wages, foster workplace diversity, and contribute toward goals of economic equality.

Democratizing access to information about salaries can have further implications for job seekers based on this research. Predictive models can give a better view of salary expectations with regard to particular positions and branches, which may affect the choice of profession, additional education, or negotiation strategies. Thus, it allows individuals to make better decisions and, therefore, avoid cases of underpayment, enhancing job satisfaction overall. Moreover, it will provide a lead to educational and training institutions on how curricula should be oriented according to the demand in the labor market to increase the employability of graduates.

The social effect extends to organizations as well. Employers can make use of salary prediction models to ensure that their remuneration practices are competitive and equitable. By adopting these models, businesses are able to attract top talent and retain skilled employees to contribute to productivity and innovation in different sectors. These kinds of models may also assist an organization in keeping its compensation strategy in line with industry standards and assure compliance with labor laws, besides helping to maintain a good corporate image.

The societal impact of the research will be manifold in its influence on both the micro

and macro levels of economic well-being. Given that machine learning models bring objectivity to salary-related predictions, they can make labor markets fairer, facilitate social mobility, and foster economic growth. In other words, salary prediction research does not constitute a technical development in nature but can redefine the employment perspective of the future so that compensation is fair, transparent, and truly reflective of the worth employees bring into an organization.

5.2.2 Impact on Environment

The research on AI Job Salary Prediction Using Machine Learning aims to enhance the salary estimation accuracy in the AI sector by using advanced models of machine learning. Therefore, in this context, most of the project focuses on enhancing the prediction capability of the proposed model. Thus, the consideration of environmental impact becomes relevant only regarding computational processes consisting of training and deploying machine learning models.

The models used in this paper, such as Random Forest and Tuned Random Forest, are notorious for their computational intensity, especially when the training has to be performed on big datasets, such as that obtained from the Stack Overflow Annual Developer Survey. Large computing jobs and datasets require the use of complex algorithms that call for expensive hardware with high-performance-capability Graphics Processing Units (GPUs). Computing tasks that require so much energy result in large carbon footprints.

The environmental concerns arise from the high energy demand in both training and optimization phases of machine learning models. While AI and machine learning applications are becoming increasingly widespread, so is the growing demand for computational power and, by extension, an increased environmental footprint. This becomes particularly relevant in large-scale deployments where resources need to be constantly spent to maintain the accuracy and performance of a model. This should not, however, blind us from the fact that there is ongoing work to meet these challenges. Advances in cloud computing and energy-efficient hardware have thus given rise to more sustainable approaches to training AI models. Furthermore, greater adoptions of green computing practices and the use of renewable energy sources by data centers should go a long way in considerably mitigating the environmental impact. Besides, researchers and practitioners try to optimize algorithms so that the consumption of energy is less without compromising performance, which will make AI development eco-friendly.

While the direct value of the project is in the precise predictions of AI job salaries, there is a need to acknowledge the environmental impacts that such extensive deployment of advanced machine learning models could entail. The research field is

trending toward greener practices, and sustained attention will make AI advances, including supporting salary prediction tools, socially responsible and environmentally sensitive.

5.2.3 Ethical Aspects

Ethical issues in the research of Salary Prediction for AI Jobs Using Machine Learning have been taken into consideration to make sure that machine learning model deployment is responsible and fair. Core ethical issues in this project include the use of data from the Stack Overflow Annual Developer Survey. This dataset contains sensitive information about job salaries, demographic data, and professional experiences; hence, it is important that this data is treated with due care and respect for privacy. First and foremost, data privacy is one of the main ethical considerations in any machine learning project, and this research is no exception. Although this dataset is publicly available, anonymization and aggregation must be performed where possible to avoid identifying individual participants. One must make sure that the data protection laws and guidelines, such as the GDPR, are followed, especially when sensitive information is being used or processed. Another important ethical consideration in this study is the fairness of predictive models. If not designed with care, machine learning algorithms can easily embed biases from the data to which they are exposed; for example, gender, age, or racial biases. In the context of predicting salaries, it is necessary to assess whether these models could inadvertently reproduce historical salary disparity or discriminatory practices. The research underlines the importance of transparency in the model selection process and develops methods that reduce biases to ensure that AI job salary predictions are fair and equitable across different demographic groups.

Besides, transparency of the research process is a significant constituent of ethical accountability. The development and deployment of machine learning models should be well-documented to let other researchers scrutinize and reproduce the findings. This openness will help build trust in the methodology and make sure that the performance and assumptions of the model are clearly communicated for responsible knowledge dissemination. Finally, the societal implications of salary predictions in the AI job market must be considered. Salary estimations could influence hiring practices, career decisions, and wage negotiations. Thus, it is crucial to ensure that the predictions do not unfairly impact individuals or perpetuate inequality. Developers and stakeholders must be conscious of these potential consequences and work to mitigate any adverse effects. In a nutshell, the ethical consideration of the project Salary Prediction for AI Jobs Using Machine Learning will involve giving due attention to data privacy, fairness, transparency of the predictions, and ramifications on society as a whole. By addressing these concerns, the project ensures that the outcomes are not only technically sound but also ethically responsible and aligned with societal values.

5.2.4 Sustainability Plan

The sustainability plan for the 'Salary Prediction for AI jobs using Machine Learning' will move toward the least environmental impacts, keeping the model as efficient and accessible for such a long period of time. This project recognizes that there is a further opportunity to optimize the usage of computational resources to bring down the carbon footprint from training models and deploying them. The design of the system will use energy-efficient algorithms that make sure the computing power utilization is optimized during the training process for reduced energy consumption. Besides, the cloud computing solution will be explored in order to offload the intense processing to more energy-efficient data centers, reducing the local use of hardware and making use of infrastructure that uses renewable sources of energy. By leveraging cloud platforms with green computing certifications, the project puts itself in a position to keep up with sustainability concerns related to data management and storage.

It also points out that continuous optimization is needed for machine learning models in the sustainability plan. Other techniques for reducing computational complexity will be discovered as time goes on; if found, the project shall move to more efficient methods and update the models so performance and resource consumption remain in balance. Besides, periodic assessments of the environmental impact of the project will be carried out for further improvement, with the intention of keeping sustainable practices at the front line in the development of the project.

In summary, this research's sustainability plan relies on the basis of energy-efficient computing, eco-friendly cloud computing usage, and continuous reevaluation for sustainability. Green computing principles will be integrated into every step of the AI job salary prediction framework to ensure not only a positive contribution to society with accurate salary estimates but also to keep the ecological footprint as low as possible in view of global sustainability goals.

5.3 Project Management and Financial Analysis

Provide a cost analysis in terms of budget required and revenue model. In the case of budget,you must show an alternate budget and rationales.

Table 5.3.1: Cost analysis.

Name	Estimated Cost
GPU (GeForce RTX 3050)	30,500 BDT
Processor (intel i5 12th gen)	16,000 BDT
RAM (32 GB)	15,800 BDT
Storage (1 TB)	16,000 BDT
Total Hardware cost	78,300 BDT
Data collection cost	12,000 BDT
Development and research cost	30,000 BDT
Web-based application	45,000 BDT
Total cost	168,000 BDT

5.4 Complex Engineering Problem

5.4.1 Complex Problem Solving

In this section, provide a mapping with problem solving categories. For each mapping add subsections to put rationale (Use Table 5.1). For P1, you need to put another mapping with Knowledge profile and rational thereof.

Table 5.4.1.1: Mapping with complex problem solving.

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
✓		✓				✓

Mapping with Knowledge Profile for EP1

Table 5.4.1.2: Mapping with knowledge Profile (EP1).

K1 Natural Sciences	K2 Mathematics	K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K7 Comprehension	K8 Research Literature
		✓	✓	✓	✓		✓

K3(Engineering Fundamentals): This research incorporates the fundamental principles of data science, machine learning, and statistical modeling to predict AI job salaries. The study applies core concepts such as regression analysis, classification techniques, and performance metrics (RMSE, MAE, R^2) to analyze the factors influencing salaries in the AI industry and make accurate predictions based on historical data.

K4(Specialist Knowledge): Specialized knowledge in feature engineering and machine learning models such as Linear Regression, Decision Trees, Random Forests, and Tuned Random Forest is utilized in this research. These models are applied to extract meaningful insights from the dataset and make precise salary predictions for AI-related job roles. Advanced tuning techniques enhance model performance, ensuring that predictions align closely with actual salary values.

K5(Engineering Design): The design of the predictive models for AI salary forecasting is based on understanding the relationship between various employee attributes (such as experience, skills, and education) and salary levels. Workflows for data preprocessing, feature extraction, model training, and evaluation are carefully structured to create reliable and accurate prediction systems that balance both employee expectations and business needs.

K6(Engineering Practice): The practical application of machine learning pipelines is demonstrated through the use of tools such as scikit-learn, Keras, and TensorFlow for model training and evaluation. These tools facilitate the implementation of algorithms and allow for the deployment of predictive models in real-world scenarios, ensuring scalability, efficiency, and real-time applicability of salary predictions.

K8(Research Literature): This research critically analyzes existing studies on salary prediction, exploring how machine learning models have been applied in various sectors to predict compensation. It also reviews prior work on the application of regression models and ensemble methods (like Random Forests) for salary prediction, providing a comparative foundation for assessing the effectiveness of the models used in this study.

Mapping with Knowledge Profile for EP3

Table 5.4.1.3: Mapping with knowledge Profile (EP3).

K1 Natural Sciences	K2 Mathematics	K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K7 Comprehension	K8 Research Literature
	✓	✓	✓	✓			✓

K2 (Mathematics): Mathematical techniques such as feature extraction and performance evaluation are employed for this project. Techniques like mean squared error (MSE) and root mean squared error (RMSE) are used to assess model performance. Statistical principles guide the implementation of machine learning models like Linear Regression (LR), Decision Tree, Random Forest, and Tuned Random Forest. These techniques help to calculate important metrics like R^2, mean absolute error (MAE), and optimize hyperparameters for better predictive accuracy.

K3(Engineering Fundamentals): The principles of machine learning and deep learning are utilized for feature extraction and model optimization. For example, in this context, machine learning models such as Linear Regression, Decision Tree, Random Forest, and Tuned Random Forest are applied to predict salaries based on data like years of experience, skills, and education. The data preprocessing step, such as handling missing values and normalizing features, is a fundamental part of this process.

K4(Specialist Knowledge): Specialist knowledge is applied in handling complex datasets related to salary prediction. Expertise in Decision Tree, Random Forest, and Tuned Random Forest models helps in improving the accuracy of predictions. Understanding how these models process features such as experience, education, and technical skills enhances the ability to build accurate salary prediction models for AI jobs. Hyperparameter tuning techniques are applied to optimize the Random Forest model for better performance.

K5(Engineering Design): The engineering design involves the creation of workflows for data preprocessing, model training, and performance evaluation. Key steps in the process include feature extraction, model selection (LR, Decision Tree, Random Forest, Tuned Random Forest), hyperparameter tuning, and comparison of models' performance metrics like RMSE, MAE, and R^2. This structured workflow ensures a systematic approach to building the salary prediction model and allows for comparison between the models to select the best-performing one.

K6(Engineering Practice): The practical implementation involves the development of data analysis pipelines using tools like scikit-learn, Keras, and TensorFlow. These tools are used to implement and train the Linear Regression, Decision Tree, Random Forest, and Tuned Random Forest models. Visualizing results using libraries like matplotlib and seaborn aids in the analysis of the model's performance. Additionally, the pipeline involves saving trained models, visualizations, and performance metrics for future reference or deployment. Development of data analysis pipelines with the help of scikit-learn, keras, and TensorFlow. Visualize the findings and save necessary files for future use.

K8(Research Literature): Previous research on salary prediction models and machine learning techniques is reviewed to inform model selection and improvement. The results of similar studies in the fields of machine learning and AI job salary prediction are analyzed to identify gaps and areas for improvement in your models. Research literature guides the hyperparameter tuning of models like Random Forest and helps in understanding the significance of various features. This research process ensures the development of robust models and provides a benchmark for evaluating performance metrics such as R^2 and RMSE.

Mapping with Knowledge Profile for EP7

Table 5.4.1.4: Mapping with knowledge Profile (EP7).

K1 Natural Sciences	K2 Mathematics	K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K7 Comprehension	K8 Research Literature
✓	✓	✓	✓	✓	✓	✓	✓

K1 (Natural Sciences): The breakdown of salary data is achieved using advanced statistical and machine learning techniques. These methods help predict AI-related

salaries by analyzing relationships between job roles, qualifications, and salary trends.

K2 (Mathematics): Statistical techniques, such as AUC-ROC and TF-IDF, are applied to extract features from datasets. Models like linear regression, decision tree, random forest, and tuned random forest use activation functions like ReLU and SoftMax for evaluation.

K3 (Engineering Fundamentals): Natural language processing (NLP) and machine learning methods are used for analyzing AI salary prediction data. These techniques help optimize the connection between various components in the data preprocessing, feature extraction, and model training process.

K4 (Specialist Knowledge): Advanced tasks include classification, feature extraction, and model implementation for salary prediction. Machine learning models such as linear regression, decision tree, random forest, and tuned random forest are applied for accurate predictions.

K5 (Engineering Design): Procedures for preprocessing, training, and deployment are developed to ensure model scalability. This ensures that the predictions are reliable and consistent for various AI job roles.
K6 (Engineering Practice): Effective data processing and pipeline management are implemented using tools like NumPy, TensorFlow, Keras, and scikit-learn. These tools ensure the optimization of models and high-quality performance metrics in salary prediction.

K7 (Comprehension): Ethical and sociological considerations are addressed to ensure fairness and transparency in salary predictions. The project maintains unbiased algorithms for equitable outcomes in AI job salary predictions.

K8 (Research Literature): A review of over twenty research studies helps identify best practices for salary prediction models. The project incorporates lessons learned to improve models and manage complex interdependencies effectively.

5.4.2 Engineering Activities

Table 5.4.2.1 Mapping with complex engineering activities.

EA1 Range of resources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓	✓	✓	✓	✓

EA1 (Range of resources): This study applies a wide range of resources, including machine learning and deep learning frameworks like scikit-learn, Keras, and TensorFlow. These resources are utilized to build and evaluate predictive models such as Linear Regression, Decision Tree, Random Forest, and Tuned Random Forest for accurate AI job salary predictions..

EA2 (Level of interaction): The research addresses challenges in integrating data preprocessing, feature extraction, and model optimization for accurate salary prediction. It combines multiple models to balance high performance with real-world applicability, ensuring practical use in predicting AI job salaries.

EA3 (Innovation): This project innovatively applies machine learning models to predict AI job salaries, leveraging advanced techniques like Random Forest and Tuned Random Forest. By fine-tuning these models, the research improves the accuracy and reliability of salary predictions in the AI industry.

EA4 (Consequences for society and the environment): The research contributes to society by providing organizations with reliable salary predictions, promoting fair compensation and helping employees make informed career decisions. It enhances industry practices by improving salary transparency and reducing pay inequality in AI roles.

EA5 (Familiarity): The project demonstrates familiarity with engineering principles applied to machine learning for solving complex salary prediction tasks. By combining traditional and advanced machine learning models, it creates a data-driven approach to predicting AI salaries across various roles.

5.5 Summary

This project applies machine learning to predict AI job salaries based on factors such as experience, location, skills, and education. The study focuses on feature engineering, model optimization, and hyperparameter tuning with algorithms like Linear Regression, Decision Tree, Random Forest, and Tuned Random Forest.
The research highlights the use of advanced ensemble models and optimization techniques to enhance prediction performance. By comparing various machine learning approaches, the project identifies insights into the relationship between skill sets, experience, and compensation. This contributes to informed decision-making for both job seekers and recruiters, aligning AI-driven insights with practical workforce trends.

Chapter 6

Conclusion

This project applies machine learning to predict AI job salaries based on factors such as experience, location, skills, and education. The study focuses on feature engineering, model optimization, and hyperparameter tuning with algorithms like Linear Regression, Decision Tree, Random Forest, and Tuned Random Forest.
The research highlights the use of advanced ensemble models and optimization techniques to enhance prediction performance. By comparing various machine learning approaches, the project identifies insights into the relationship between skill sets, experience, and compensation. This contributes to informed decision-making for both job seekers and recruiters, aligning AI-driven insights with practical workforce trends.

6.1 Summary

The objective of this project is to study the techniques for machine learning in predicting salary for AI-related job roles using the Stack Overflow Annual Developer Survey dataset sourced from Kaggle. A performance comparison is done for the Linear Regression, Decision Tree, Random Forest, and Tuned Random Forest models based on performance metrics: RMSE, MAE, and R^2. It follows that the tuned random forest had the best score, since it presented the least amount of error, as noted: RMSE was 23,426 and R^2 was 0.53, while random forest is very close to that result with an RMSE of 23,879 and R^2 of 0.52. Meanwhile, linear regression and the decision tree represented higher errors and low values of R^2. These findings suggest that the model Tuned Random Forest is effective for salary predictions in AI jobs, proving that optimization significantly enhances the performance of models. The research underpins the contribution of machine learning models toward salary decision-making in the AI industry, with further scope for the adoption of these techniques in other sectors.

6.2 Limitation

The following limitations were identified and considered for future improvements:

i. Data Limitations: The study relied on the Stack Overflow Annual Developer Survey dataset, which may not fully represent the diversity of AI-related job roles across different industries and geographic regions. The dataset may also have biases due to self-reporting by respondents, which could affect the generalization of the findings.
ii. Model Generalization: While the machine learning models demonstrated good performance on the available dataset, their effectiveness may vary when applied to different AI job datasets or real-world scenarios. Models may struggle with unseen data or different salary distribution patterns, requiring further testing on more diverse data.
iii. Feature Selection: Despite the extensive feature selection process, some factors that influence salary (such as company culture, job satisfaction, or individual negotiation skills) may have been overlooked. Incorporating additional features could improve model accuracy.
iv. Technological Requirements: The models used in this study require significant computational resources, including high-performance hardware and software environments. This could limit their accessibility to organizations with fewer resources, making the practical implementation more challenging.
v. Ethical Concerns: The use of salary data raises potential concerns regarding data privacy, fairness, and bias. Ensuring that the models are fair and transparent in predicting salaries without reinforcing existing inequalities is crucial. Additionally, the handling of personal and sensitive data should adhere to ethical guidelines and privacy regulations.

6.3 Future Work

This present study of salary prediction using machine learning on AI jobs, however, contributes to a wide perspective and creates several avenues for further research. One such direction to these future studies could be the application of other machine learning methods, such as neural networks or ensemble methods, in pursuit of more accurate predictions with nonlinear relationships between complex datasets. Although models such as Random Forest, Decision Tree (DT), Tuned Random Forest, and Linear Regression have shown their value, future studies might want to examine other deep learning models or ensemble approaches that merge the outcomes of several techniques for better predictability.

Moreover, future studies could be directed to incorporate other features, such as location, company size, or AI job skill requirements, in order to increase the accuracy

of salary predictions. Future work may also investigate whether transfer learning or the use of pre-trained models will be helpful in enhancing model performance, especially in conditions of a small or unbalanced dataset.

In this study, given the performance of Tuned Random Forest, further investigation into hyperparameter tuning and advanced optimization techniques may improve the predictive capability of this method even more. Further efficiency in the implementation of algorithms of Random Forest and Decision Trees is also worth exploring in order to reduce computational resources and time during the training process, especially on bigger datasets. Cloud-based solutions for distributed computing frameworks might provide scalable solutions for handling increasingly complex data.

These implications for further study on the prediction of AI jobs salaries are to explore new techniques such as neural networks or hybrid models, use more features, improve the performance of Random Forest, DT, Tuned Random Forest, and Linear Regression, and computational efficiency. This may provide significant improvements in the accuracy, adaptability, and real-world applicability of salary prediction models in the AI job market.

References

[1] Das, S., R. Barik, and A. Mukherjee, Salary Prediction Using Regression Techniques. Proceedings of Industry Interactive Innovations in Science, Engineering & Technology (I3SET2K19), 2020.
[2] Dutta, S., A. Halder, and K. Dasgupta, Design of a Novel Prediction Engine for Predicting Suitable Salary for a Job. 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), IEEE.
[3] Martín, I., et al., Salary Prediction in the IT Job Market with Few High-Dimensional Samples: A Spanish Case Study. International Journal of Computational Intelligence Systems, 2018, 11(1): pp. 1192-1209.
[4] Srivastava, S., D. Sharma, and P. Sharma, Comparing Various Machine Learning Techniques for Predicting the Salary Status. 2020, EasyChair.
[5] Khongchai, P. and P. Songmuang, Random Forest for Salary Prediction System to Improve Students' Motivation. 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), IEEE.
[6] Khongchai, P. and P. Songmuang, Implementation of Salary Prediction System to Improve Student Motivation Using Data Mining Techniques. 2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS), IEEE.
[7] Pawha, A., and D. Kamthania, Quantitative Analysis of Historical Data for Prediction of Job Salary in India—A Case Study. Journal of Statistics and Management Systems, 2019, 22(2): pp. 187-198.
[8] Wang, Z., S. Sugaya, and D.P. Nguyen, Salary Prediction Using Bidirectional-GRU-CNN Model. Assoc. Nat. Lang. Process, 2019.
[9] Chakraborti, S., A Comparative Study of Performances of Various Classification Algorithms for Predicting Salary Classes of Employees. Volume 5, 2014, pp. 1964-1972.
[10] Ajit, P., Prediction of Employee Turnover in Organizations Using Machine Learning Algorithms. Algorithms, 2016, 4(5): p. C5.
[11] Susmita Ray, "A Quick Review of Machine Learning Algorithms," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT Con), India, 14th -16th Feb 2019.
[12] Sananda Dutta, Airiddha Halder, and Kousik Dasgupta, "Design of a novel Prediction Engine for predicting suitable salary for a job," 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN).
[13] Pornthep Khongchai and Pokpong Songmuang, "Improving Students' Motivation to Study using Salary Prediction System," 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE).
[14] Phuwadol Viroonluecha and Thongchai Kaewkiriya, "Salary Predictor System for Thailand Labour Workforce using Deep Learning," The 18th International Symposium on Communications and Information Technologies (ISCIT 2018).
[15] Mangui Wu and Shunmin Shu, "Top Management Salary, Stock Ratio and Firm Performance: A Comparative Study of State-owned and Private Listed Companies in China."
[16] Das, S., Barik, R., and Mukherjee, A., "Salary prediction using regression techniques,"

Proceedings of Industry Interactive Innovations in Science, Engineering & Technology (I3SET2K19), 2020.
[17] Martín, I., et al., "Salary Prediction in the IT Job Market with Few High-Dimensional Samples: A Spanish Case Study," International Journal of Computational Intelligence Systems, 2018, 11(1): p. 1192-1209.
[18] Srivastava, S., Sharma, D., and Sharma, P., "Comparing various Machine Learning Techniques for Predicting the Salary Status," 2020, EasyChair.
[19] Pawha, A., and Kamthania, D., "Quantitative analysis of historical data for prediction of job salary in India-A case study," Journal of Statistics and Management Systems, 2019, 22(2): p. 187-198.
[20] Wang, Z., Sugaya, S., and D.P. Nguyen, "Salary Prediction using Bidirectional-GRU-CNN Model," Assoc. Nat. Lang. Process, 2019.
[21] Chakraborti, S., "A Comparative Study of Performances of Various Classification Algorithms for Predicting Salary Classes of Employees," 2014, vol. 5: p. 1964-1972.
[22] Ajit, P., "Prediction of employee turnover in organizations using machine learning algorithms," 2016, 4(5): p. C5.

Salary Prediction of AI Jobs Using Machine Learning

ORIGINALITY REPORT

22%	18%	13%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ijisrt.com Internet Source	3%
2	Submitted to Daffodil International University Student Paper	2%
3	ijrpr.com Internet Source	1%
4	ijasret.com Internet Source	1%
5	www.mdpi.com Internet Source	1%
6	Submitted to Liverpool John Moores University Student Paper	1%
7	Submitted to United International University Student Paper	1%
8	Sukhpreet Kaur, Sushil Kamboj, Manish Kumar, Arvind Dagur, Dhirendra Kumar Shukla. "Computational Methods in Science and Technology", CRC Press, 2024 Publication	1%