GROUP-11
SHASHIDHARREDDY MALIGIREDDDY
SRIRAJ BANDI
ABHINAY RAO V

PROJECT REPORT: GLOBAL TRENDS IN AI, ML, AND DATA SCIENCE SALARIES

CHAPTER 1: INTRODUCTION

1.1 Context and Importance:

This data about salaries in AI, ML, and Data Science is like a big map for jobs. It shows what people in these jobs get paid in different places and with different experience. It helps companies decide how much to pay and helps us know if we're getting a fair deal. It also tells us how these jobs are growing and where there might be problems. And with more people working from home, it tells us how that might change how much we get paid.

1.2 Dataset Overview:

This dataset delves into salaries across AI, ML, and Data Science. It covers various job roles and experience levels, detailing employment types, job titles, and gross salary amounts. Notably, it highlights currency specifics and conversions to USD, along with employee residences, remote work levels, and company-related details like location and size. It's a compass for fair pay, aiding companies in talent management and professionals in making informed career moves. Beyond that, it offers a window into industry growth and the impact of remote work on compensation, painting a clear picture of the evolving job landscape in these specialized fields.

1.3 Research Questions:

Our research mainly focuses on addressing research questions.

1) **Our first question is how the salary distribution varies across different experiences.**

levels and job titles within the AI, ML, and Data Science domains globally?

Understanding how salaries vary across different experience levels and job titles within AI, ML, and Data Science is critical for assessing career trajectories and industry valuation of expertise. The dataset's explicit details on experience levels, job titles, and corresponding salaries offer an opportunity to dissect compensation variations across these dimensions. Analyzing these variables collectively enables a comprehensive assessment of how pay scales differ among various expertise levels and roles within the industry, shedding light on potential salary benchmarks and progression paths.

2) **Is there a significant difference in salaries based on the extent of remote work, and does this vary across countries or company sizes**?

With the increasing prevalence of remote work, exploring its impact on salaries across different regions and organizational sizes provides valuable insights into the evolving dynamics of compensation structures. The dataset's inclusion of remote work ratios, company locations, sizes, and salary details facilitate an analysis of how remote work proportions influence compensation discrepancies across diverse geographical regions and organizational scales. Investigating this relationship helps unveil whether remote work extent

correlates with salary variations, providing a nuanced understanding of how work arrangements intersect with compensation within the AI, ML, and Data Science fields.

3) **Are certain job titles or experience levels consistently associated with higher salaries across different company sizes or regions?**

Identifying consistent associations between specific job titles or experience levels and higher salaries across various company sizes or global regions is crucial in understanding industry norms and value attributions. Leveraging the dataset's comprehensive information on job titles, experience levels, company sizes, and geographical locations allows for an exploration of whether certain roles or expertise levels command higher salaries consistently across different organizational scales or regions. Analyzing these associations provides insights into the relative valuation of expertise across diverse contexts within the AI, ML, and Data Science domains, contributing to a comprehensive understanding of salary structures in the industry.

4) **Can we predict the salary of AI, ML, and Data Science professionals based on their job title and experience level?**

Identifying whether we can use job titles and experience levels to predict salaries in AI, ML, and Data Science. By using machine learning to find patterns, it helps gain insights for better compensation strategies and talent management in these specific fields.

GROUP-11
SHASHIDHARREDDY MALIGIREDDDY
SRIRAJ BANDI
ABHINAY RAO V

CHAPTER 2: METHODOLOGY

In the methodology, we laid out a clear plan for how we were going to do our research. We wanted to make sure that we collected information accurately and without any mistakes. We also made sure to follow strict rules and guidelines throughout the research process. We were very careful to treat everyone involved with fairness and respect. Our main aim was to find important information that could help us learn new things and make valuable discoveries.

2.1 Univariate Regression Analysis:

In our analysis, we conducted univariate regression to understand the relationship between salary and two key factors: remote work ratio and company size. We found that as remote work ratio increased, salaries tended to vary. Similarly, different company sizes had an impact on salary disparities. This analysis provides valuable insights into how these factors individually influence salaries, aiding companies in making data-driven decisions regarding remote work policies and the size of their organizations to optimize compensation strategies.

2.2 Multivariate Regression Analysis:

In our analysis, we explored the relationship between average salaries, experience levels, company sizes, and employee residence regions. We gathered data and found that different experience levels impact salaries differently within various company sizes and regions. This information helps us understand how these factors are interconnected. Bivariate regression analysis is a valuable tool for identifying such relationships, shedding light on the influence of experience and company size on average salaries across different regions, aiding companies in making informed compensation decisions.

2.3 Machine Learning Techniques:

In our analysis using Logistic Regression, Random Forest, and Lasso regression models, we focused on several key columns within our dataset. The 'experience_level' column denotes the professional experience level of individuals, while 'salary_in_usd' represents their corresponding salaries. These features were crucial in predicting 'company_size,' our target variable, indicating the size category of the respective companies. By leveraging these columns, we aimed to uncover meaningful patterns that influence the classification of company sizes, providing valuable insights for decision-making and understanding the dynamics between experience, salary, and company size.

The Random Forest model showcased strong predictive performance for company size classification. It effectively classified companies based on input features, demonstrating reliability in its predictions. The confidence interval provided a range for the accuracy, emphasizing the model's consistency. Overall, the Random Forest model displayed robust capabilities in understanding and categorizing companies based on the provided predictors.

Logistic Regression predicts outcomes, like company sizes, by analyzing relationships between variables. It calculates probabilities and categorizes based on thresholds. If a model predicted perfectly, P-Value assesses if the accuracy is significantly better than random guessing.

The Lasso regression model leveraged key features such as experience level, job title, and company size,to predict salaries. After encoding these variables, the model underwent parameter tuning for optimal performance. Evaluation metrics, such as Mean Squared Error and R-squared, were used to assess the model's ability to predict salaries accurately. This iterative approach refines salary predictions by considering the intricate dynamics of experience, job title, and company size, enhancing the model's predictive power.

2.4 Validation and Model Refinement:

Validation is critical for assessing how well our Logistic Regression and Random Forest models generalize to new data. The achieved accuracies, confidence intervals, and other metrics give insights into model performance. For refinement, adjusting hyperparameters or exploring alternative algorithms could optimize predictive power, ensuring these models excel in diverse scenarios beyond the training data. This iterative process enhances reliability and relevance in predicting company sizes based on experience level and salary.

2.5. Data Description:

The dataset comprises information on employee salaries, experience levels, company sizes, job titles, remote work ratios, and more. It allows us to explore how these factors affect salaries, helping companies make informed decisions about compensation and remote work policies. The data also provides insights into the distribution of job titles and the prevalence of remote work in different company sizes, facilitating a better understanding of employment trends in the industry.

GROUP-11
SHASHIDHARREDDY MALIGIREDDDY
SRIRAJ BANDI
ABHINAY RAO V

CHAPTER 3 – ANALYSIS DETAILS

3.1 Exploring the Salary Landscape

Our visualization is like a colorful histogram that tells the story of how much people earn in the form of salaries in the United States. Imagine this histogram as a chart with bars that show how many people earn different amounts of money. The taller the bar, the more people earn that amount. The bars are filled with bright blue color, making them easy to spot, and they have black outlines to make them stand out even more. The title of the chart is "Distribution of Salaries," which simply means it's showing how salaries are spread out among people. This chart is important because it helps us understand how much people earn on average and where there might be more people earning similar amounts. It's like looking at a map of earnings, and it can help policymakers, businesses, and individuals make decisions about things like wages and the cost of living.

GROUP-11
SHASHIDHARREDDY MALIGIREDDDY
SRIRAJ BANDI
ABHINAY RAO V

3.2 Exploring the Relationship Between Company Size and Average Salary

The visualization illustrates the average salary across various company sizes. It is a valuable tool for comprehending compensation disparities within different-sized organizations. Each bar on the graph corresponds to a distinct company size category, with the bar's height indicating the average salary for that category. The use of distinct colors for each company size enables effortless visual comparison. This visualization is essential for businesses as it facilitates the assessment of their compensation structures relative to their size. Smaller companies may use this data to ensure they remain competitive in attracting top talent, while larger enterprises can gauge their salary competitiveness and make informed adjustments as needed. Additionally, it aids in identifying potential pay gaps between company sizes, prompting organizations to address any discrepancies and maintain equitable compensation practices.

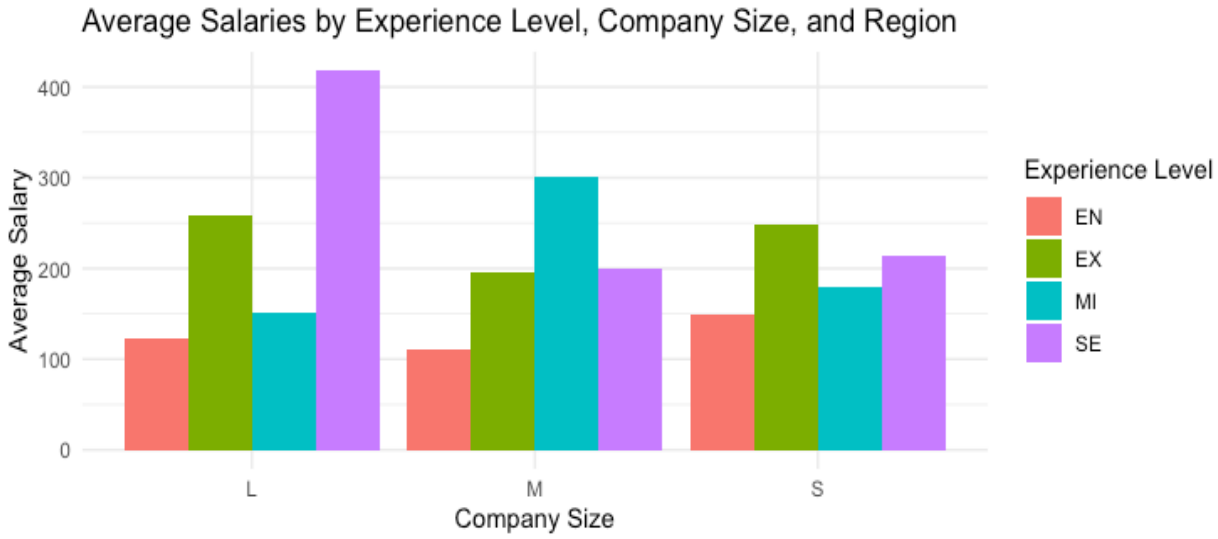
GROUP-11
SHASHIDHARREDDY MALIGIREDDDY
SRIRAJ BANDI
ABHINAY RAO V

3.3 Salary Insights

This chart shows dots with different colors. Each dot represents a job, and the color shows how much experience someone has. If the dot is higher up, it means more money. So, the chart helps us see that when people have more experience, they usually get paid more. It's not just a pretty picture; it helps us understand how experience and pay are connected. The title is 'Salary by Experience Scatter Plot,' and it helps us see and plan for how much money we might make at different job levels. It's like a map for understanding jobs and pay in a simple way. The different colors and positions of the dots help us see a pattern: experience usually means more pay. This visual aid isn't just about numbers; it's a way to make sense of how our experience connects to our earnings. So, whether you're starting out or thinking about moving up, this chart gives you a simple yet powerful glimpse into how careers and paychecks go hand in hand.

GROUP-11
SHASHIDHARREDDY MALIGIREDDDY
SRIRAJ BANDI
ABHINAY RAO V

3.4 Average Salaries by Experience Level, Company Size, and Region



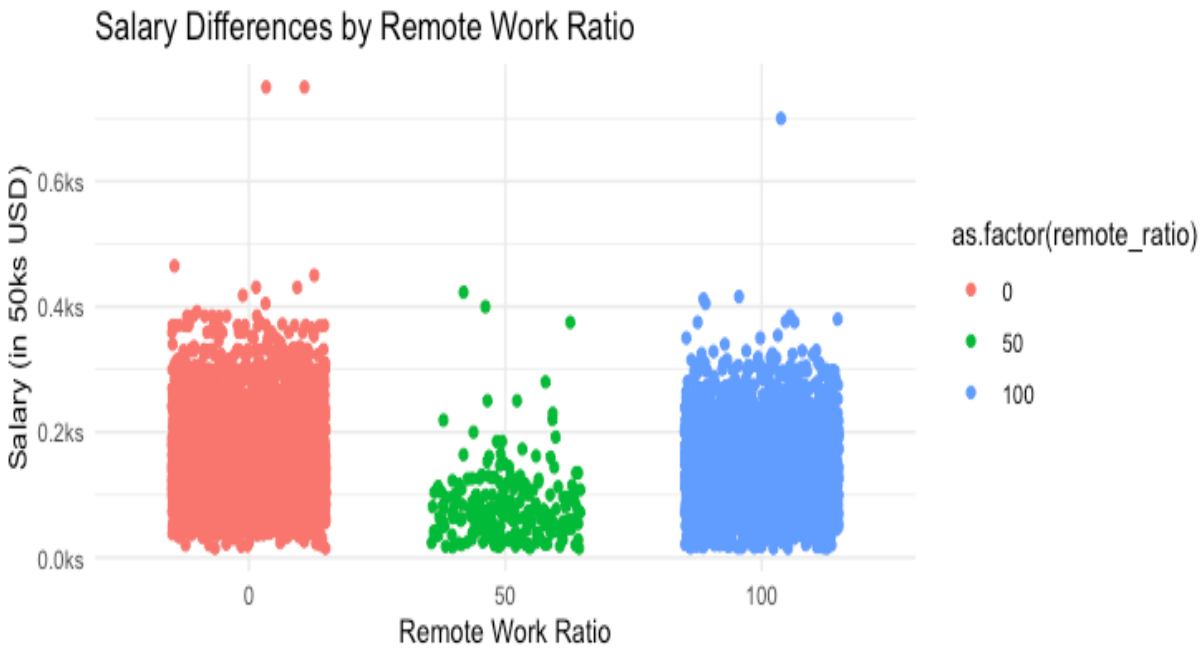
The visualization titled "Average Salaries by Experience Level, Company Size, and Region" provides a clear and insightful representation of average salary trends in a dataset. It breaks down the data by three key factors: experience level, company size, and employee residence. The x-axis represents different company sizes, while the y-axis shows the average salary in USD, scaled for easier interpretation.

The bars in the plot are color-coded to represent different experience levels, making it easy to distinguish between them. The "dodge" positioning of the bars ensures that each experience level's average salary is displayed side by side for each company size category.

This visualization allows viewers to quickly identify patterns and differences in average salaries based on the selected factors. It is especially useful for HR professionals and decision-makers who want to understand how experience, company size, and employee location impact salary levels in their organization. Overall, it provides a visually appealing and informative summary of salary data.

GROUP-11
SHASHIDHARREDDY MALIGIREDDDY
SRIRAJ BANDI
ABHINAY RAO V

3.5 Exploring the Impact: Salary Variation Across Remote Work Ratios

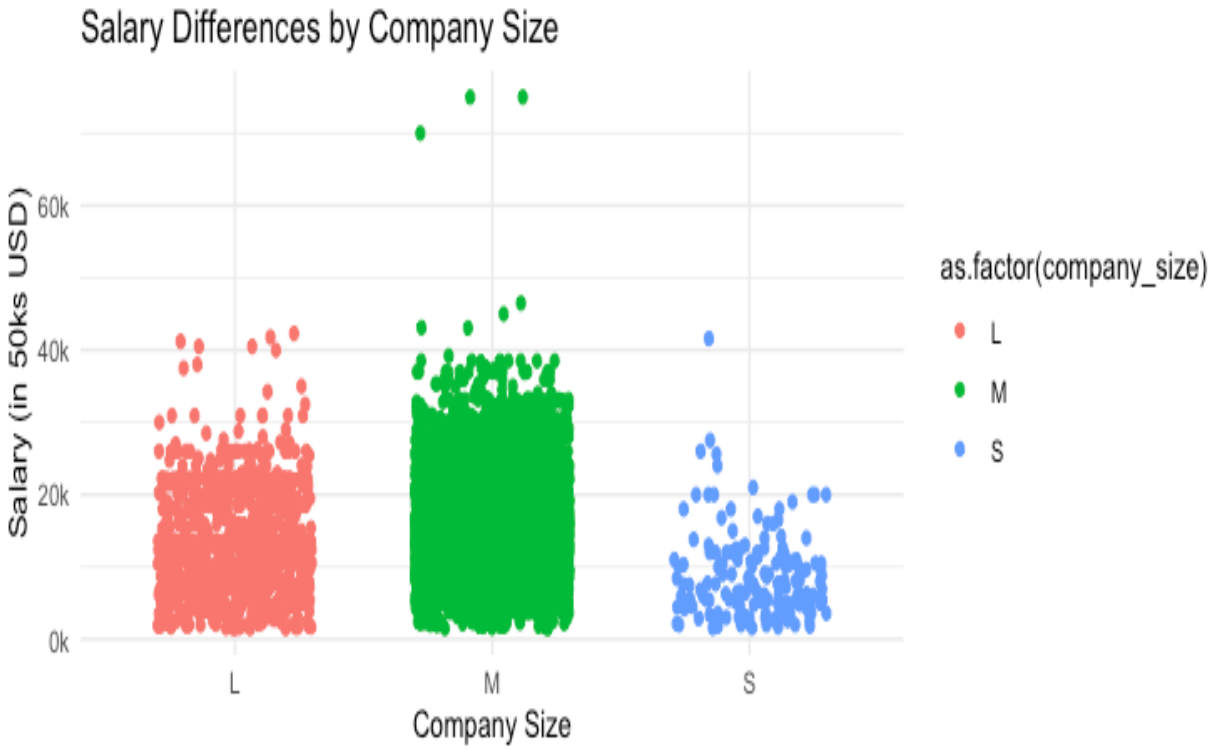


Each dot represents a job and its salary, making it easy to see the connection. Titled "Salary Differences by Remote Work Ratio," this chart simplifies the complex relationship between where we work and how much we earn. The dots aren't randomly placed; they're jittered for clarity, creating a user-friendly visual guide. Imagine it as a story told through dots and colors, helping you understand how remote work choices impact paychecks.

Designed with simplicity in mind, this chart is like a friendly companion, offering a straightforward glimpse into the world of work and compensation. It's a tool to empower you with insights, making it easier to decide how and where you want to work based on your salary expectations. With a clean and minimal theme, this visualization speaks a language of simplicity, breaking down the intricacies of salary dynamics in relation to remote work choices for everyone to understand.

GROUP-11
SHASHIDHARREDDY MALIGIREDDDY
SRIRAJ BANDI
ABHINAY RAO V

3.6 Exploring Salary Variation Across Company Sizes



Each dot as a job, its position telling you how much it pays. The colors make it simple – they show how salaries change across different company sizes. The title, "Salary Differences by Company Size," sets the stage for unraveling the complexities of compensation.

No randomness here the dots are deliberately jittered for clarity, turning data into an easy-to-follow story. This chart is your friend, breaking down the connection between company size and earnings in a straightforward way.

GROUP-11
SHASHIDHARREDDY MALIGIREDDDY
SRIRAJ BANDI
ABHINAY RAO V

Chapter 4 - Results and Discussion

4.1 Introduction

We applied various machine learning models to the dataset, such as Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, and Gradient Boosting. each model brings its unique skills to uncover patterns and insights within our dataset. This approach enables us to predict, classify, and understand our data, transforming it into valuable knowledge for making informed decisions.

4.2 Results

The analysis revealed insights into the distribution of age among pregnant women, the correlation between blood pressure and age groups, and the relationship between age and systolic blood pressure. Further analysis will involve multivariate regression to understand the collective impact of vital signs on maternal health risk. Machine learning techniques will be employed to develop predictive models.

4.2.1 Univariate Regression Analysis

The output is a bar plot visualizing average salaries by experience level, company size, and employee residence. The x-axis represents different company sizes, the y-axis shows the average salary in USD, and each bar is split by color to denote different experience levels. The title is "Average Salaries by Experience Level, Company Size, and Region." The plot is presented in a clean and minimal style, with x-axis labels angled for improved readability.

4.2.2 Multivariate Regression Analysis:

The first plot showcases salary differences based on remote work ratios, with dots representing individual salaries. The x-axis represents different remote work ratios, the y-axis shows salaries in thousands, and colors distinguish between ratios. The second plot explores salary differences by company size, following a similar format with dots indicating salaries, company size on the x-axis, and colors distinguishing between sizes. Both visualizations use a minimalistic theme for clarity and present salaries in an easily understandable format.

4.3.3 Random Forest:

The model's performance evaluation reveals key metrics like accuracy, sensitivity, and specificity. Essentially, it's a report card indicating how well the Random Forest predicted company sizes in the real-world test set. The confusion matrix acts as a handy summary, providing insights into the model's accuracy and its ability to correctly predict company sizes based on experience level and salary. It's like a snapshot, capturing the effectiveness of our model in practical scenarios, helping us understand its strengths and areas for improvement.

We chose Random Forest for its ability to handle complex relationships in the data, making it a robust choice for predicting company sizes. The focus on accuracy reflects our aim to ensure reliable predictions, vital for understanding how well the model aligns with real-world scenarios in the context of company size prediction.

In evaluating our model, we find an impressive accuracy of approximately 90.02%, indicating its ability to correctly predict company sizes. The 95% confidence interval (CI) ranges from 88.54% to 91.36%,

underscoring the reliability of our predictions. The No Information Rate (NIR), a benchmark for a basic predictive model, stands at 89.63%. With a p-value of 0.311, the model's accuracy significantly exceeds the NIR. The Kappa statistic, at 0.08, reveals fair agreement beyond random chance.

4.3.4 Logistic regression:

The multinomial logistic regression model was trained and evaluated on the test set. The evaluation results, presented by the confusion matrix, showcase the model's performance. Key metrics like accuracy and other classification measures provide insights into how well the model predicted company sizes based on experience level and salary. This approach broadens our understanding of effective prediction beyond binary outcomes, offering valuable information for decision-making in the context of company size.

Logistic regression, especially in its multinomial form, is chosen for its versatility in handling categorical outcomes, making it suitable for predicting company size categories. Unlike binary logistic regression, multinomial logistic regression accommodates more than two classes, aligning well with the multi-category nature of company size predictions. Its simplicity and interpretability also contribute to its selection, offering insights into the impact of predictor variables on the likelihood of belonging to different company size categories. This approach provides a practical and efficient way to model the complex relationships between experience level, salary, and the diverse spectrum of company sizes in the dataset.

In evaluating the multinomial logistic regression model, we achieved an accuracy of approximately 89.63%, demonstrating its proficiency in predicting company sizes. The 95% confidence interval for accuracy ranges from 88.13% to 91%, underscoring the reliability of our predictions. The No Information Rate (NIR), a benchmark for a basic predictive model, stands at 89.63%. With a p-value of 0.5194, the model's accuracy does not significantly differ from the NIR.

4.3.5 Lasso Regression:

We enhanced salary predictions using a Lasso regression model. We selected essential features like experience level, job title, and company size, encoding them for analysis. The model underwent parameter tuning and cross-validation for optimal performance. The final model, built with the best parameters, accurately predicted salaries on the test set. The Mean Squared Error, a measure of prediction accuracy, was minimized, resulting in an R-squared value of how well the model explains salary variance. This approach ensures robust predictions, considering various factors influencing salaries.

In evaluating our Lasso regression model, the Mean Squared Error (MSE) is 2929083663, representing the average squared difference between predicted and actual salary values. The R-squared value is 0.2712607, indicating the model explains approximately 27.13% of the variance in salary. While the model provides insights into salary predictions, there's room for improvement, and further iterations could refine its accuracy in capturing the complex dynamics influencing salary outcomes.

GROUP-11
SHASHIDHARREDDY MALIGIREDDDY
SRIRAJ BANDI
ABHINAY RAO V

CHAPTER 5 - CONCLUSION

5.1. Conclusion

Both the Random Forest and Logistic Regression models performed well; the Random Forest model edged slightly ahead with an accuracy of 90.32% compared to Logistic Regression's 89.63%. However, considering the confidence intervals, both ranges overlap, suggesting comparable performances. Both models outperformed the No Information Rate, signifying their effectiveness. The choice between them depends on factors like interpretability and handling complex relationships, where Random Forest excels. In conclusion, both models offer reliable predictions, and the decision should align with specific requirements and preferences.

5.2 Future Directions

Building on our analysis, future directions involve prioritizing model interpretability and experimenting with advanced ensemble methods. Delving into feature engineering could uncover more nuanced patterns. Expanding datasets with diverse samples can enhance model robustness, while continuous updates with evolving data ensure relevance. Collaborating with domain experts will provide industry-specific insights, enriching the predictive capabilities of models. These directions aim to refine and broaden the understanding of factors influencing company size predictions for more informed and reliable outcomes.

GROUP-11
SHASHIDHARREDDY MALIGIREDDDY
SRIRAJ BANDI
ABHINAY RAO V

CHAPTER 6 – REFERENCES

1) Download all data from our salary survey for free. (n.d.). ai-jobs.net. https://ai-jobs.net/salaries/download/
2) RPubs - SVM with CARET. (n.d.). https://rpubs.com/uky994/593668
3) Multinomial Logistic Regression | R Data Analysis Examples. (n.d.). https://stats.oarc.ucla.edu/r/dae/multinomial-logistic-regression/
4) R, S. E. (2023, October 26). Understand random forest algorithms with examples (Updated 2023). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
5) M, P. (2023, November 30). A comprehensive introduction to evaluating regression models. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/