**Proposal**

**DATA SALARY PREDICTIVE MODEL**  
**Joseph Choi | Jenny Overby | Daniel Meier | Serge Nane**  
**DSC450 Applied Data Science  
Spring 2024**

# Domain

This project falls under the jobs and salaries domain, as our focus is to develop a predictive model that forecasts salary trends for data-related roles based on the updated 2024 from [The Global AI, ML, Data Science Salary Index for 2024 | ai-jobs.net](https://ai-jobs.net/salaries/2024/).

We have compiled ten references to help us make sense of the domain and the data we’ll be working with. These references were selected to understand the best methodologies for building our predictive salary model. They were also chosen as they cover salary trends, the data on the job market, and economic factors impacting these positions.

1. [A Data-Driven Approach to Salary Prediction: | by Amar saish | Medium](https://medium.com/@mummineniamar/a-data-driven-approach-to-salary-prediction-158f6cb8f121): Discusses the use of Linear and Tree Regressions to predict salaries using age, experience, and other factors.
2. [(PDF) Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits -A Literature Review (researchgate.net)](https://www.researchgate.net/publication/362280362_Salary_Prediction_in_Data_Science_Field_Using_Specialized_Skills_and_Job_Benefits_-A_Literature_Review): Reviews new and existing techniques to build advanced salary predictions in Data Science, focusing on specialized skills and job benefits.
3. [Employee Salaries Analysis and Prediction with Machine Learning | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/9943146): Explores various regression models to predict salaries based on influencing factors and evaluates the use of R2 and RMSE metrics.
4. [Predict Data Science Salaries with Data Science | by Junting (JT) Lai | Towards Data Science](https://towardsdatascience.com/the-us-data-science-job-market-in-2020-463520a9d5a): Uses EDA and machine learning algorithms to analyze the US data science job market and salaries in 2020.
5. [Job Salary Prediction with NLP, Machine Learning and Deep Learning | by Bonnie Ma | Towards Data Science](https://towardsdatascience.com/job-salary-prediction-with-nlp-machine-learning-and-deep-learning-b87a96336b08): Analyzes 10,000 plus job postings from Indeed to build a model for job salaries using machine learning and deep learning techniques.
6. [Machine Learning Models for Salary Prediction Dataset using Python | Semantic Scholar](https://www.semanticscholar.org/paper/Machine-Learning-Models-for-Salary-Prediction-using-Kablaoui-Salman/5a213573154231fe97113f1c41bc3651a3eea409): Highlights the result of applying three supervised machine learning techniques (linear regression, random forest, and neural networks) to a U.S. salary dataset.
7. [The Recession’s Impact on Analytics and Data Science (mit.edu)](https://sloanreview.mit.edu/article/the-recessions-impact-on-analytics-and-data-science/): Provides information regarding the potential decline in demand for data scientists and analytics due to economic disruptions.
8. [Will Recession Impact Data Science and Analytics? | Analytics Insight](https://www.analyticsinsight.net/will-recession-impact-data-science-and-analytics/): Looks into the challenges and potential declines in the demand for analytics and data science during economic recessions.
9. [Key Data And Analytics Trends To Watch In 2023 (forbes.com)](https://www.forbes.com/sites/forbestechcouncil/2022/12/15/key-data-and-analytics-trends-to-watch-in-2023/?sh=1b65271629b0): Summarizes trends in data and analytics in 2023.
10. [Data Scientist Job Market 2024: Analysis, Trends, Opportunities | 365 Data Science](https://365datascience.com/career-advice/data-scientist-job-market/): Examines the current trends in the data job market, focusing on the increasing demand for data roles and the evolving job requirements heading into 2024.

# About the Dataset

## Description

This dataset contains 2024 job and salary data in the data industry, sourced from ai-jobs.net. It provides information about salary distributions across roles and demographics, making it ideal for a detailed analysis to understand major salary influencing factors and develop a predictive salary model.

## Data Fields

* **work\_year**: The year the data was recorded
* **job\_title**: Specifies the job role
* **job\_category**: Classification of the job role
* **salary\_currency**: The currency in which the salary is paid
* **salary**: Annual gross salary in the local currency
* **salary\_in\_usd**: Annual gross salary converted to USD
* **employee\_residence**: Residence of the employee
* **experience\_level**: Professional experience level of the employee
* **employment\_type**: Type of employment
* **work\_setting**: Work environment
* **company\_location**: The location of the company
* **company\_size**: Size of the company

## Dataset Location

The dataset can be located and downloaded in the link: [Jobs and Salaries in Data field 2024 (kaggle.com)](https://www.kaggle.com/datasets/murilozangari/jobs-and-salaries-in-data-field-2024)

# Research Questions

For students about to graduate with a bachelor's degree in data science, understanding salary trends in the data industry is essential for making informed decisions about job opportunities and negotiating salaries effectively. It also allows them to set realistic salary expectations when applying to roles.

To this end, our project plans to address the following research questions:

1. What are the key factors influencing salary levels in data-related roles?
2. Can we develop a predictive model that accurately forecasts salaries for data-related roles based on historical data?

Our 2024 data science and analytics job market analysis will provide insights into salary distributions and influencing factors, supporting those looking to break into the data industry.

# Method

The plan is to approach this project by following the outlined process workflow below:

1. **Data Wrangling**: During this phase, we will perform our initial data cleaning/transforming tasks to identify and address any inconsistencies, missing values, or outliers present in the dataset. The objective is to ensure that the raw data is ready for the data science phase of the project.
2. **Data Science**: In this phase, we plan to conduct exploratory data analysis (EDA) to understand the characteristics of each input feature and our target variable. Our goal with EDA is to extract insights to help us develop an approach to our feature selection and feature engineering process. Once the dataset is ready for model training, we'll experiment with five models (Logistics, Ridge, Lasso, Random Forest, KNN), running three iterations each. We'll document and evaluate performance metrics for each iteration, selecting the best-performing model as our final choice. We'll make intentional adjustments throughout each iteration based on insights from EDA and the previous iterations.
3. **Data Visualization**: During this phase, we will create plots to effectively communicate our extracted insights to address our research questions. This involves providing visuals to support our findings on key factors impacting salary and display a clear view of our model performance.

# Potential Issues

After the initial review of our methods and the dataset, we do not anticipate many issues will arise when we start executing each phase of our project. However, the project’s success relies on effective data cleaning and model fine-tuning for optimal performance. If the data quality or our modeling approaches fall short, we could encounter problems that affect our output. We plan to work closely together to come up with mitigation strategies to address these issues and prevent delays.

# Conclusion

Bringing everything together, our project proposal centers on understanding the current job market for all data-related roles, focusing on salary trends. Through carefully executed data wrangling, analysis, science, and visualization, our team’s main objective is to identify key factors influencing salary levels and create a predictive model based on 2024 data. We are eager to get started and look forward to sharing our insights and final model at the end of the project.