DATS 6103: Final Project Topic proposal
Team 6
Fyrooz Anika Khan, Rasika Shrikant Nilatkar, Bairu Likhitha Reddy, Aniruddh Rajagopal
10/30/2024
Professor Ning Rui
Fall 2024

Data Science Job Salary Analysis Proposal

The field of data science is rapidly evolving, with the U.S. Bureau of Labor Statistics projecting significant job growth through 2029 (Ahmad & Hamid, 2023). However, challenges such as workforce fluctuations and the impact of technologies like generative AI are reshaping salary structures. To better understand these dynamics, this proposal aims to analyze data science salaries using a comprehensive dataset from Kaggle.

As more universities offer data science programs, competition intensifies, influencing salary expectations. The Kaggle Data Science Job Salaries Dataset, containing 6,600 observations and 11 key columns, provides a source of information to analyze these dynamics based on factors like job title, experience level, company size, and employment type. This dataset will enable a comprehensive analysis of salary dynamics within the data science field.

By addressing SMART questions below, the analysis aims to provide valuable insights for students and professionals navigating a competitive job market in the data science field.

SMART questions

1. What is the average salary of data scientists based on their level of experience in the United States?
2. Has the median salary for entry-level data science positions changed significantly from 2020 to 2024, and by what percentage?
3. What are the highest-paying job titles in the U.S. compared to other countries?

Dataset link

https://www.kaggle.com/datasets/sazidthe1/data-science-salaries

GitHub repo

https://github.com/aniruddhg43986683/Team_6_data_mining

Modelling Method:

For this project, we will apply classification techniques to develop a predictive model for salary estimation. Our approach will involve two primary classification algorithms: **logistic regression, support vector classifier (SVC)** or any other classification model. These algorithms will be selected due

to their effectiveness in handling classification tasks, even when dealing with complex datasets with multiple features.

We will start by preparing the dataset with essential preprocessing steps, including handling missing values, encoding categorical variables, and normalizing or scaling features as required. Following this, we will split the data into training and test sets to validate the model's generalizability. The chosen classification algorithms will be evaluated using metrics such as accuracy, precision, recall, and F1-score to ensure robust model performance. Based on these performance metrics, we will finalize the model that most accurately predicts the salary class of individuals based on the independent features in the dataset.

Reference

Ahmad, N., & Hamid, A. (2023). Will Data Science Outrun the Data Scientist? *Computer*, *56*(2), 121–128. Computer. https://doi.org/10.1109/MC.2022.3226929