

*Project Report on*

**"DATA SCIENTIST’S SALARY PREDICTION"**

Submitted by

N.GREESHMA SREE BAI

20221CIT0097

5-CIT-02

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE and ENGINEERING

(INTERNET OF THINGS)

Submitted to

SK JAMIL AHMED

Presidency University, Bengaluru

# INDEX

|  |  |  |  |
| --- | --- | --- | --- |
| | **Sl. No.** | | --- | | **Content** | **Page Number** |
| 1 | Acknowledgement |  |
| 2 | Abstract |  |
| 3 | Introduction |  |
| 4 | Data Collection and Preprocessing |  |
| 5 | Data Description |  |
| 6 | Results and Conclusion |  |

# ACKNOWLEDGEMENT:

I would like to express my deepest gratitude to everyone who contributed to the successful completion of this Predicting Data Scientist Salaries Using Machine Learning project. I extend my heartfelt thanks to my mentor, **SK JAMIL AHMED**, whose expertise and guidance helped shape my understanding of machine learning techniques, specifically in the areas of natural language processing and text classification. Their valuable feedback throughout each stage of the project—from data preprocessing to model evaluation—was instrumental in enhancing the quality and accuracy of this work.

I am also grateful to **Presidency University** for providing the necessary resources, including computational infrastructure, access to datasets, and educational materials. These resources were essential for developing, testing, and refining the spam classification model. I would also like to acknowledge the creators and maintainers of the **data scientist salary.csv**, whose efforts have enabled researchers and students worldwide to work on real-world data, furthering research and innovation.

Lastly, I extend my appreciation to my friends, family, and colleagues for their constant encouragement and support throughout this journey. Their motivation and understanding were invaluable during challenging phases of the project, and their interest in my progress provided me with the drive to accomplish my goals.

# ABSTRACT:

Data science has emerged as one of the most sought-after career fields, characterized by competitive and varied compensation packages. This research aims to predict data scientist salaries using machine learning models, leveraging real-world datasets containing attributes such as job location, company size, industry sector, education requirements, and experience level. By identifying the factors influencing salary trends, the study seeks to provide actionable insights for professionals and organizations.

The methodology involves preprocessing data, performing exploratory data analysis (EDA), and employing regression models such as Linear Regression, Random Forest, and Gradient Boosting. Hyperparameter tuning and feature selection are applied to improve model accuracy and interpretability. Model evaluation is conducted using metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

The results demonstrate the effectiveness of machine learning in predicting salaries and reveal critical determinants such as geographical location, experience level, and industry type. This work contributes to career planning for data professionals and helps organizations optimize salary structures in a data-driven manner.

**Prediction for Data scientist’s salary prediction**

The demand for skilled data scientists has surged globally, driven by advancements in artificial intelligence, machine learning, and big data analytics. This study explores the application of machine learning models to predict salaries in the data science domain, leveraging publicly available datasets that include features such as job location, experience level, industry, company size, and skill requirements. By analyzing these variables, the study aims to identify key determinants of salary and provide accurate predictions.

Various machine learning techniques, including linear regression, random forests, and gradient boosting methods, are employed to model salary as a function of the input features. Feature engineering and hyperparameter optimization are performed to enhance model performance. Evaluation metrics such as mean absolute error (MAE) and R-squared are used to assess the accuracy of the predictions.

The findings provide insights into the factors influencing data scientist salaries and offer a tool for professionals and organizations to benchmark compensation expectations. This work underscores the value of data-driven approaches in decision-making for both job seekers and employers in the competitive tech industry.

# Chapter I - Introduction:

# 1.1 Background

# Data science has emerged as a high-demand field, offering competitive salaries that vary based on location, experience, industry, and skills. These variations make understanding and predicting salaries crucial for professionals and organizations alike. The dynamic nature of the job market adds complexity to forecasting compensation trends. Machine learning provides a robust method for analyzing historical data and identifying key salary determinants.

# 1.2 Problem Statement

# The aim of this study is to predict data scientist salaries using machine learning models, leveraging factors such as experience, location, and industry. It seeks to identify key determinants of salary trends and provide accurate forecasts.

# 1.3 Objectives

# 1*. Analyze Salary Determinants:* Identify key factors influencing data scientist salaries, such as location, experience, skills, and industry.

# 2. *Develop Prediction Models*: Build and evaluate machine learning models to accurately predict salaries based on relevant features.

# 3*. Provide Data-Driven Insights*: Offer actionable insights for professionals to make informed career decisions and for employers to structure competitive pay.

# 4. *Enhance Decision-Making*: Enable stakeholders to understand market trends and optimize resource allocation in talent acquisition and retention.

# 1.4 Significance of the Study

Data Scientist's Salary prediction provides numerous benefits, including:

* **Career Planning**: Helps data scientists make informed decisions about salary expectations and skill development.
* **Competitive Pay**: Assists employers in creating attractive and fair compensation packages.
* **Market Transparency**: Promotes understanding of salary trends and fosters equity in compensation..

# Chapter II – Data Description:

**2.1 Data Collection**

The dataset used for this project is the **data scientist’s salary prediction dataset**, data for predicting data scientist salaries can be collected from job listings on platforms like LinkedIn and Glassdoor, which provide details on salaries, required skills, and experience levels. Geographical location and company industry/size data are also essential for understanding regional and sectoral salary variations.

* 1. **Dataset Overview**
* **Total Records**: The dataset contains thousands of data scientist’s salary entries.
* **Features**: Includes job title, location, company size, industry, skills, and experience level.
* **Target Variable**: Salary data (base and total compensation).
* **Additional Data**: Education, certifications, and benefits for enhanced predictions.

**2.3 Data Preprocessing**

To prepare the data scientist’s salary text data for machine learning, several preprocessing steps were performed:

# 1. Load Dataset

# Python code

# data = pd.read\_csv('data\_scientist\_salaries.csv')

# Description: Reads the dataset from a CSV file into a Pandas DataFrame, making it easy to manipulate and analyze.

# 2. Handle Missing Values

# Python code

# data.fillna(method='ffill', inplace=True)

# Description: Fills missing values using forward-fill, which propagates the last valid observation forward. Ensures no empty cells disrupt further analysis. Alternative methods include mean or median imputation.

# 3. Encode Categorical Variables

# Python code

# label\_encoders = {}

# categorical\_cols = ['location', 'company\_size', 'industry', 'education\_level']

# for col in categorical\_cols:

# label\_encoders[col] = LabelEncoder()

# data[col] = label\_encoders[col].fit\_transform(data[col])

# Description: Converts non-numeric categorical variables (e.g., location, company\_size) into numerical codes using LabelEncoder. This makes them compatible with machine learning algorithms.

# 4. Feature Engineering

# python code

# data['city'] = data['location'].apply(lambda x: x.split(',')[0])

# Description: Extracts the city name from the location field, assuming it is stored in "City, State" format. Adds a new feature that might better capture salary variation.

# 5. Drop Irrelevant Columns

# Python code

# data.drop(['job\_id', 'posting\_date'], axis=1, inplace=True)

# Description: Removes columns like job\_id and posting\_date that do not provide meaningful information for predicting salaries. Reduces noise and focuses the dataset on relevant features.

# Result and Conclusion:

# 3.1 Split Data into Features and Target

#### Python code

#### **X = data.drop(['salary'], axis=1)**

#### **y = data['salary']**

#### Description: Separates the dataset into features (X) and the target variable (y). Here, salary is the variable to predict, and all other columns are predictors.

#### 3.2. Train-Test Split

#### Python code

#### **X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)**

# Description: Divides the data into training (80%) and testing (20%) subsets to train the model on one set and evaluate its performance on another.

#### 3.3Scale Numerical Features

#### pythoncode

#### **scaler = StandardScaler()**

#### **X\_train\_scaled = scaler.fit\_transform(X\_train)**

#### **X\_test\_scaled = scaler.transform(X\_test)**

# Description: Standardizes numerical features to have a mean of 0 and a standard deviation of 1.

#### **3.3 Conclusion**

#### In this study, we applied machine learning techniques to predict data scientist salaries based on factors such as location, experience, education, and industry. By preprocessing the data, including handling missing values, encoding categorical variables, and scaling numerical features, we ensured the dataset was ready for accurate predictions. The predictive model developed can assist both data professionals and employers in understanding salary trends and making informed decisions. As the demand for data scientists continues to grow, this model can provide valuable insights into compensation expectations, guiding career planning and talent acquisition strategies. Further refinement of the model, incorporating more advanced algorithms and additional data sources, could further improve prediction accuracy and adaptability to market changes.

#### **3.4 Future Work**

For further improvements, future work could involve:

1.**Informed Career Decisions**: Data scientists can leverage salary predictions to make strategic career choices, focusing on high-demand skills and industries with better compensation opportunities.

2.**Attracting Top Talent**: Employers can use salary prediction models to design competitive compensation packages, helping them attract and retain skilled data science professionals.

3.**Market Insights**: Provides valuable insights into emerging trends and salary variations across regions, industries, and experience levels, aiding businesses in staying competitive.

4.**Fair Compensation Practices**: Helps in ensuring equitable salary structures by highlighting discrepancies in compensation, promoting fair pay practices across the data science field.