A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Yumiko Maria Bejarano Azogue |
| *Student Number* | 2024144 |
| *Module Title* | Strategic Thinking |
| *Assessment Title* | Capstone Project Proposal |
| *Assessment Due Date* | 19th May 2024 23:59 |
| *Date of Submission* | 19.04.2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[Introduction 1](#_Toc162638295)

[Scope 1](#_Toc162638296)

[Problem definition 1](#_Toc162638297)

[Objectives 2](#_Toc162638298)

[General Objective: 2](#_Toc162638299)

[Specific Objectives: 2](#_Toc162638300)

[Data Sources (Technologies used) 2](#_Toc162638301)

[Libraries 2](#_Toc162638302)

[Accomplishment Data 2](#_Toc162638303)

[Attributes 2](#_Toc162638304)

[Multiple\_choice\_responses.csv 2](#_Toc162638305)

[Other\_text\_responses.csv 2](#_Toc162638306)

[Questions\_only.csv 3](#_Toc162638307)

[Survey\_schema.csv 3](#_Toc162638308)

[Data Analysis and Insights 3](#_Toc162638309)

[Gender Distribution Among Participants 3](#_Toc162638310)

[Age Distribution 3](#_Toc162638311)

[General Data Analysis 5](#_Toc162638312)

[Career and Education 5](#_Toc162638313)

[Salary 6](#_Toc162638314)

[Conclusion 7](#_Toc162638315)

[References 8](#_Toc162638316)

**Women workforce participation**

# CA1:

## Project proposal

### Introduction

This report focuses on analyzing workforce participation, with the aim of verifying the existence of gender disparities. We will examine the challenges and opportunities faced by individuals based on their gender in various job roles and positions.

Women face persistent gender barriers, the lack of role models and stereotypes that perpetuate gender inequality in this field, to quote the civil rights activist Marian Wright Edelman, “You can’t be what you can’t see”.

### Objectives

#### General Objective:

To investigate and understand the gender gap in the workforce, focusing on the participation of men and women in different job roles and positions.

#### Specific Objectives:

* Analyze the gender distribution in different job roles, including positions and departments.
* Investigate the relationship between age and workforce participation for men and women.
* Evaluate workforce performance and its relationship with gender.
* Explore the relationship between education and gender distribution in the workforce.

### Problem definition

The existence of gender disparities in the workforce can affect equity and equal opportunities. It is important to identify and understand these disparities to implement measures that promote gender equality and inclusion in the workplace.

## Scope and Methodology

### Scope

This project will analyze data related to job roles, gender, age, workforce performance, education, and other relevant factors. We will investigate patterns and trends in gender distribution across different job roles and departments, as well as their relationship with age, workforce performance, and educational level.

### Methodology

For this project, we will use the CRISP-DM methodology (Cross-Industry Standard Process for Data Mining):

Business Understanding:

* Identify business objectives.
* Determine data mining goals aligned with business objectives.
* Define success criteria.
* Develop a preliminary plan to achieve objectives.

Data Understanding:

* Gather relevant data sources.
* Explore data to understand its quality, content, and structure.
* Identify data issues, anomalies, and potential biases.
* Determine data requirements for modeling.

Data Preparation:

* Clean and preprocess data to handle missing values, outliers, and inconsistencies.
* Select relevant features for analysis.
* Transform and engineer features as needed for modeling.
* Split data into training and testing sets.

Modeling:

* Select appropriate modeling techniques based on data characteristics and objectives.
* Build initial models using selected techniques.
* Evaluate model performance using relevant metrics.
* Fine-tune models by adjusting parameters and algorithms.

Evaluation:

* Assess model performance against success criteria and business objectives.
* Validate models using cross-validation or holdout datasets.
* Interpret model results and identify areas for improvement.
* Document findings and recommendations for stakeholders.

### Ethical considerations

Identifying potential biases:

* Examine the data and models to identify any inherent biases.

Assessing potential impact:

* Evaluate how potential biases could affect different user groups or individuals impacted by the decisions made by the model. Consider whether bias could result in discrimination or injustice towards certain groups.

Mitigating identified biases:

* Implement strategies to mitigate the identified biases. This could include gathering more representative data, designing more equitable algorithms, and incorporating checks and procedures to prevent biased decisions.

# CA2:

## Project management and planning

### Data Source Overview

The dataset was obtained from Kaggle and focuses on incomes for various job titles by gender. Below is the link to access the dataset.

<https://www.kaggle.com/datasets/nilimajauhari/glassdoor-analyze-gender-pay-gap>

Data has been released under a CC 2.0 license:  <https://creativecommons.org/licenses/by/2.0/>

### Project Plan

Business Understanding:

* Identify project objectives and key research questions.
* Establish success criteria and define the project scope.

Data Understanding:

* Select and acquire the dataset on women in the technology industry.
* Explore the data to understand its quality, content, and structure.
* Identify potential data issues and biases.

Data Preparation:

* Clean and preprocess the data to handle missing values, outliers, and errors.
* Select relevant features and transform them as necessary for analysis.
* Split the data into training and testing sets.

Exploratory Data Analysis (EDA):

* Conduct exploratory data analysis using visualization techniques and descriptive statistics.
* Identify patterns, trends, and potential insights related to women's participation in the technology industry.

Machine Learning Implementation:

* Select at least one machine learning algorithm to apply to the dataset.
* Train the model using the training dataset and evaluate its performance.
* Fine-tune and optimize the model as necessary.

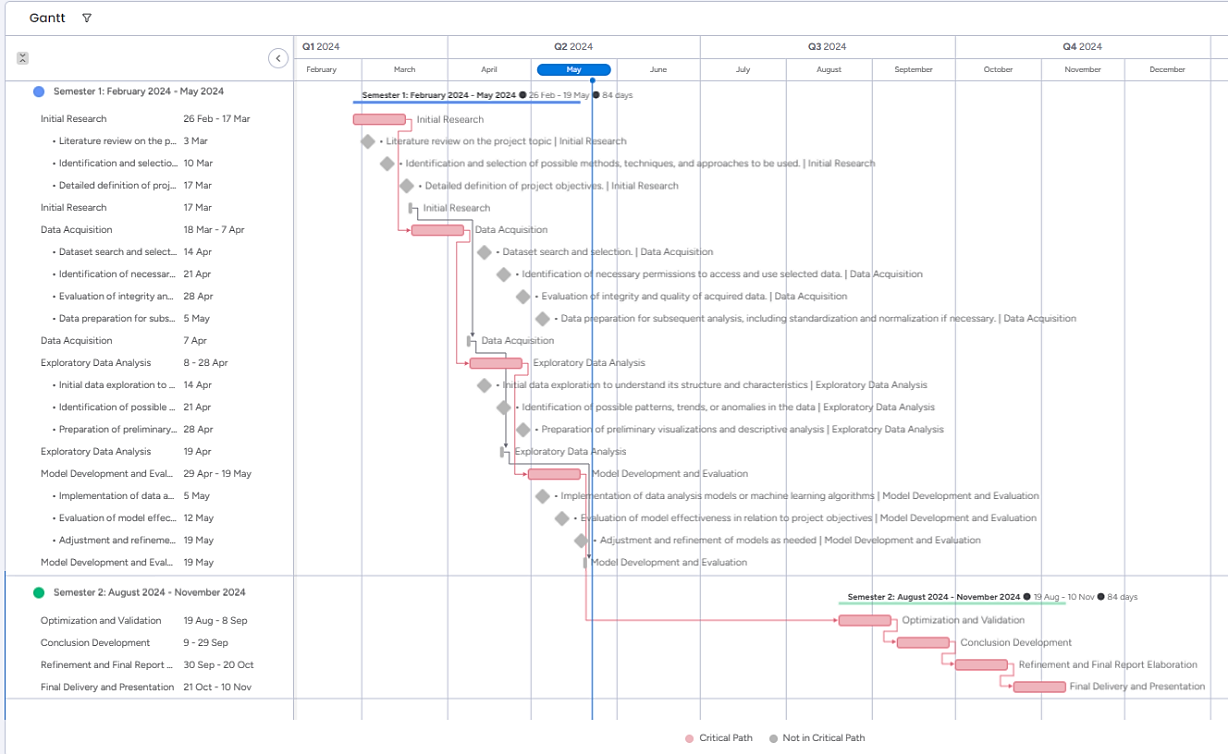
Evaluation:

* Evaluate the model's performance using appropriate metrics such as accuracy, recall, F1-scor.
* Validate the model using a separate test dataset or cross-validation techniques.

Conclusion and Presentation:

* Interpret the results of the analysis and discuss implications for women's participation in the technology industry.
* Summarize key findings and provide recommendations for future research or actions.
* Prepare a detailed report and an effective presentation of the project results.

### Project Plan Timeline





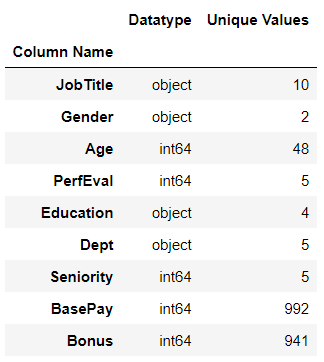


## Data Exploratory Data Analysis (EDA)

### Data Description

Dataset Size: The dataset comprises 1000 rows and 9 columns. It consists of information about employees, encompassing details such as salary, gender, position, location, education level, and work experience. Among these, there are 4 categorical variables (Gender, JobTitle, Education, Dept) and 5 numeric variables (Age, PerfEval, Seniority, BasePay, Bonus).

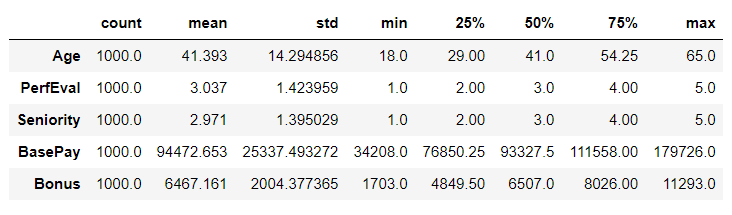
Structure and Variables: Upon examination, the dataset revealed a structured layout consisting of 9 variables. Among these variables, 4 are categorical (Gender, JobTitle, Education, Dept), and 5 are numeric (Age, PerfEval, Seniority, BasePay, Bonus). These variables were further analyzed to gain insights into their distributions and relationships within the dataset.



### Summary Statistics:

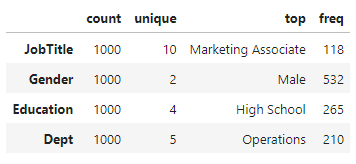
Descriptive statistics for numerical variables in the dataset:

* Age: The average age of individuals is approximately 41 years, with a standard deviation of around 14 years. Age ranges from 18 to 65 years.
* PerfEval: The average performance evaluation score is about 3.04, with scores ranging from 1 to 5.
* Seniority: On average, individuals have a seniority of around 2.97 years, with a range from 1 to 5 years. Half of the individuals have seniority equal to or less than 3 years.
* BasePay: The average base salary is approximately $94,472.65, with salaries ranging from $34,208 to $179,726.
* Bonus: The average bonus is about $6,467.16, with bonuses ranging from $1,703 to $11,293.
* The minimum recorded bonus is $1,703, and the maximum is $11,293.



Frequencies of the categorical variables in our dataset:

* JobTitle: There are 10 job titles in the dataset, with "Marketing Associate" being the most common, appearing 118 times.
* Gender: There are two categories: "Male" and "Female," with "Male" being the most frequent, totaling 532 records.
* Education: There are four educational levels, with "High School" being the most common, with 265 records.
* Dept: There are five departments, with "Operations" being the most common, with 210 records.



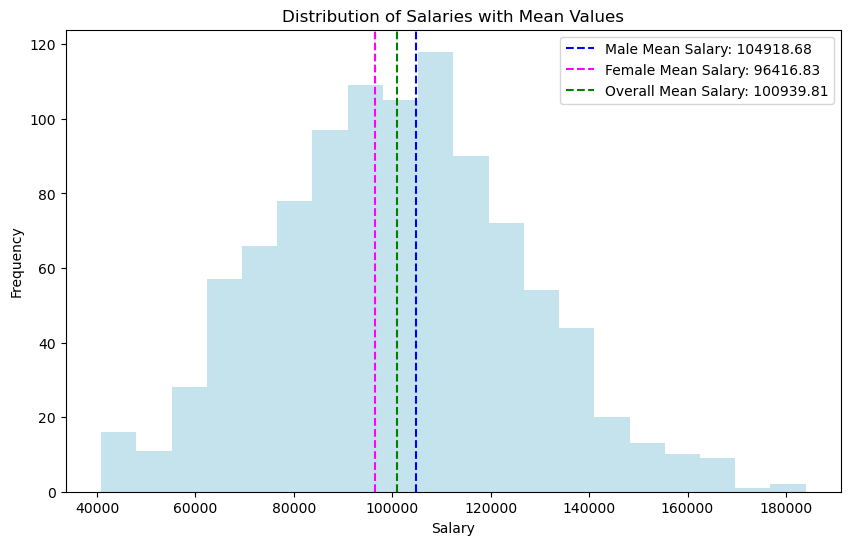
#### Salary:

The salary is calculated as sum of the base salary and the yearly bonus.



### Data Visualization:

Distribution Plots: Distribution plots were used to visualize salary distribution and other numeric variables.

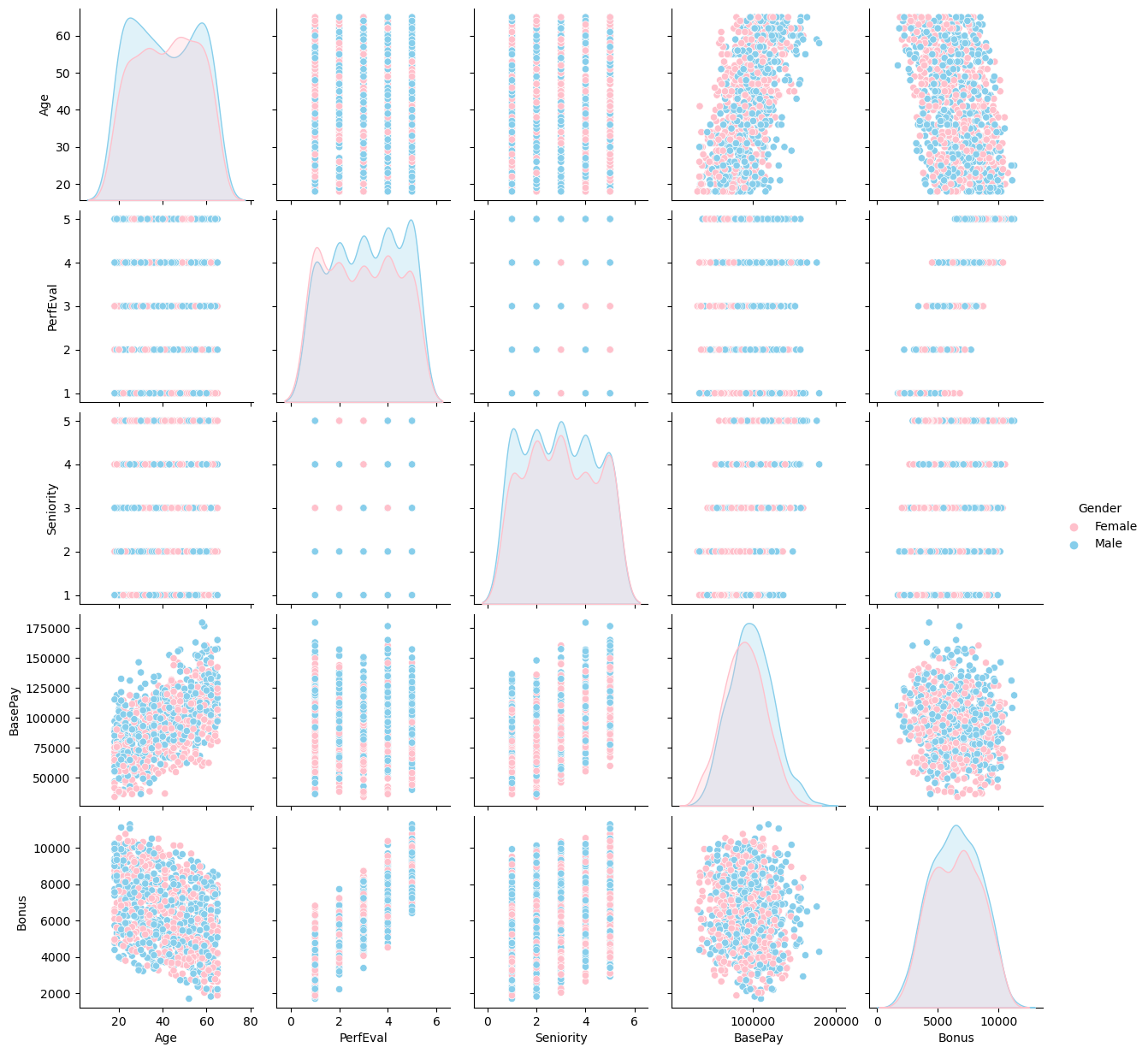


In the histograms, we observe the distribution of salaries for all individuals in the dataset, alongside their respective mean values for both men and women.

Mean Salary:

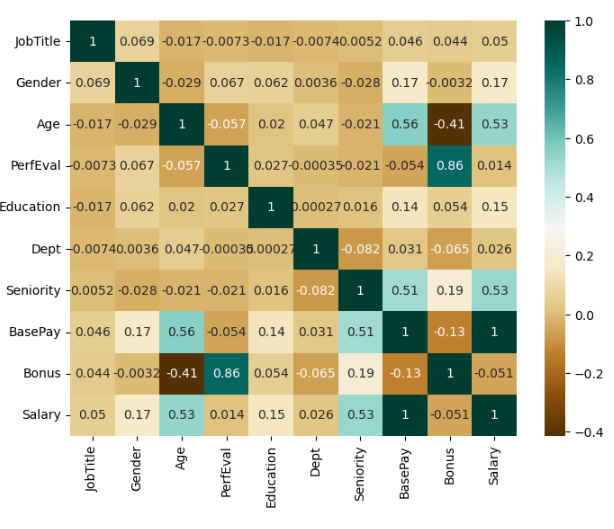
* For the overall dataset: $100,939.814
* For men: $104,918.68
* For women: $96,416.83

These results provide a comprehensive view of the salary distribution within the dataset. We can observe that the mean salary for men is higher than that for women, indicating a potential gender disparity. This trend is consistent with the mean salary reflecting a higher average salary for men compared to women.

Pair Plots: Pair plots visually explore relationships between variables, using color to distinguish between categorical and numeric variables. This helps identify patterns, including those related to the gender pay gap, by showing how salary varies across different groups like men and women. 

### Model Performance:

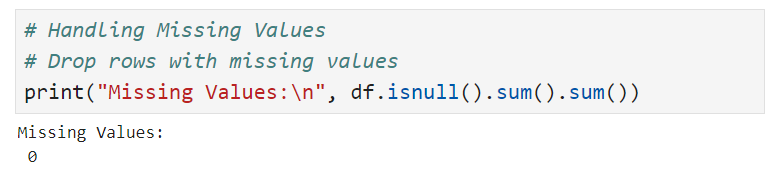
Correlation Analysis: A correlation matrix was calculated to understand relationships between variables like gender, job position, and salary. It revealed moderate positive correlations between age, seniority, and salary. Additionally, significant correlations were found between performance evaluation, base pay, and bonus. Visualizing these correlations with a heatmap highlighted strong relationships, such as those between salary and age, and salary and base pay.



### Pre-processing

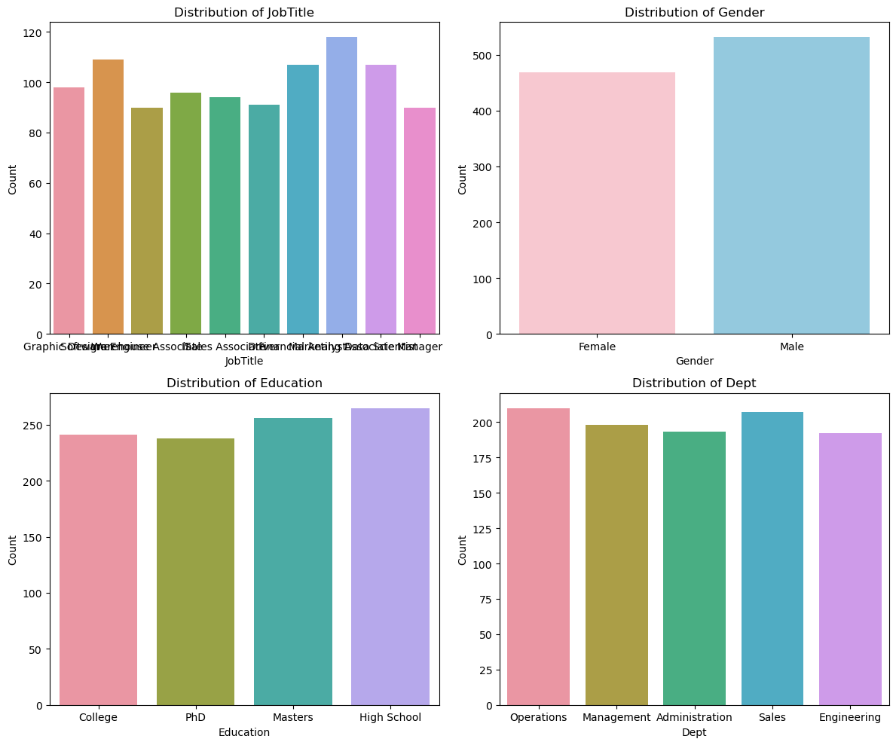
#### Handling Missing Values:

**No missing values** were detected in any of the variables in the dataset, as confirmed during the data exploration phase. Therefore, no correction or removal of missing values is required.



##### Encoding Categorical Variables:

To facilitate the analysis of the dataset, categorical variables were encoded into a numerical format. In our dataset, we identified four categorical variables: "Job Title," "Gender," "Education," and "Department".



The label encoding method was used to transform these variables into numerical values. The results of this encoding are shown at the end, allowing for more effective use of the data in subsequent analyses.

##### Scaling Numeric Features:

Scaling was applied to numeric features to ensure all variables were on the same scale and contributed equitably to the analysis.

##### Outlier Detection and Removal:

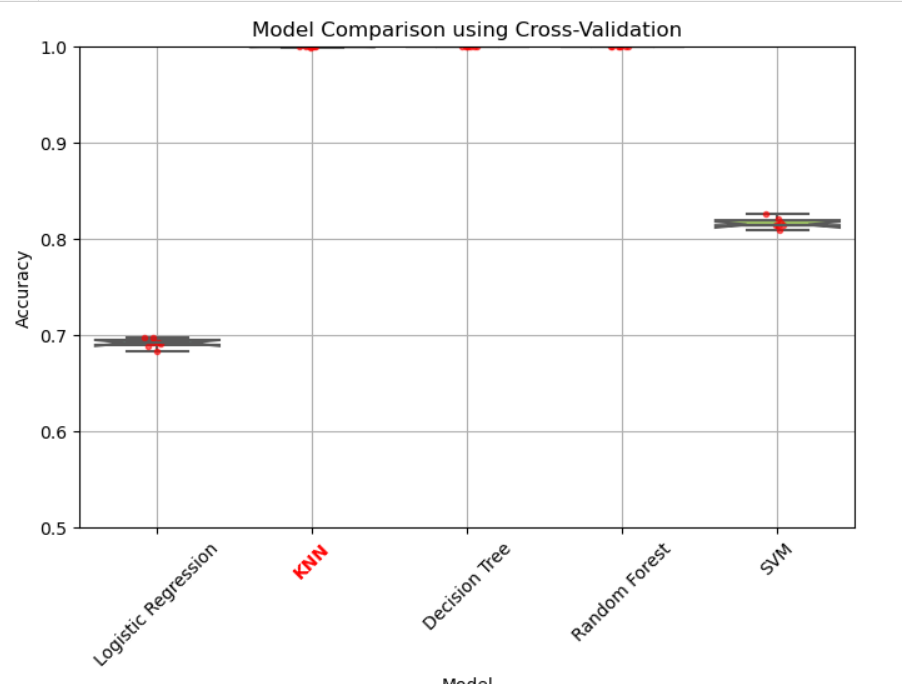
Outliers were identified and removed to mitigate potential negative effects on the pay gap analysis.

## Implementation and Evaluation of Machine Learning Models

### Classification Models:

A list of classification machine learning models, including Logistic Regression, K Nearest Neighbors (KNN), Decision Tree, Random Forest, and Support Vector Machines (SVM), were evaluated to predict female workforce participation. This variety of models is chosen to compare and select the most suitable one for the dataset at hand.

Each model is evaluated using ***stratified cross-validation*** with 5 folds. Cross-validation provides more accurate estimates of model performance by training and testing the model on multiple data subsets. The average accuracy and standard deviation of each model during cross-validation are printed to assess their stability and performance. Finally, the performance comparison of the algorithms is visualized using a box plot.



Finally, the performance comparison of the algorithms is visualized using a box plot. This graph allows identifying differences in the distribution of performance among the evaluated models, aiding in selecting the most suitable model for the classification problem.

### Development K Nearest Neighbors (KNN) - model with the best accuracy score

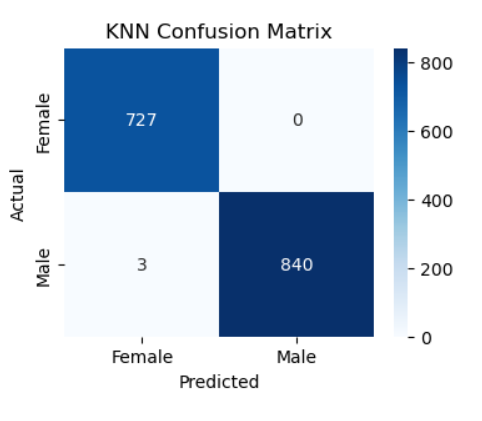
The K Nearest Neighbors (KNN) model was chosen for classification based on comparing different models. Selecting the right model is crucial for ensuring optimal performance of the classification system.

The chosen model is trained using the entire training dataset. This step is critical for the model to learn the relationships between predictor variables and the target variable. Training the model is essential for its predictive capability. Additionally, an analysis is conducted to determine the importance of features in the model.

**Prediction and Model Evaluation:** The trained model is used to make predictions on the validation dataset. Subsequently, its performance is evaluated using metrics such as accuracy and the confusion matrix.

### Visualization of the Confusion Matrix:

The confusion matrix is visualized using a heatmap. This graphical representation helps understand the relationships between the model's predictions and the actual classes.



## Conclusion

Summary of analysis conclusions:

**Data Description:** The dataset contains information about salaries for different job titles, categorized by gender. It includes 1000 entries with features like Job Title, Gender, Age, Performance Evaluation, Education, Department, Seniority, Base Salary, and Bonus.

Various statistical analyses were conducted, including mean, median, mode, range, standard deviation, and others.

* **Summary Statistics:** Descriptive statistics reveal that the average age of individuals is around 41 years, with a standard deviation of approximately 14 years. The average performance evaluation score is about 3.04, and the average seniority is around 2.97 years. The average base salary is approximately $94,472.65, with average bonuses around $6,467.16.
* **Categorical Variables Frequency:** There are 10 job titles, with "Marketing Associate" being the most common, and two gender categories, with "Male" being the most frequent. Most individuals have high school education, and the most common department is "Operations".
* **Data Pre-processing:** No missing values were found in any of the variables. Categorical variables were encoded into numerical format to simplify analysis.
* **Visual Data Analysis:** Salary distribution was observed in histograms, showing mean, median, and mode for both men and women. The results suggest possible gender salary disparities.
* **Salary Range:** Salary ranges for men and women in each department were examined, revealing potential salary differences across departments.
* **Standard Deviation:** The standard deviation of salaries indicates greater variability for men than for women.

Regarding categorical variables, there are 10 job titles, with "Marketing Associate" being the most common, and two gender categories, with "Male" being the most frequent.

**Data Preprocessing:** During preprocessing, no missing values were found in any of the variables. Categorical variables were encoded into numerical format to facilitate analysis. Outliers were also detected and removed in the "BasePay" and "Salary" columns.

**Exploratory Data Analysis (EDA):** Exploratory Data Analysis (EDA) was employed to understand the distribution and characteristics of the data. A trend towards a gender salary disparity was observed, evidenced by differences in the median, mode, and mean of salaries. Salary ranges for men and women in each department were examined, revealing potential salary differences across departments.

**General Conclusions:** Overall, the statistical analyses reveal detailed information about the data, from basic description to pattern and anomaly detection. Model comparison highlights the effectiveness of the KNN model in this specific context. Visualization of the confusion matrix provides a clearer understanding of the model's performance in data classification

# References

Mooney, P. (2019). 2019 Kaggle Machine Learning & Data Science Survey. Kaggle. URL: <https://www.kaggle.com/competitions/kaggle-survey-2019>.

Engelbert, C. 2016, 'Starting A New Conversation for Women: [Deloitte CEO Catherine Engelbert on Gender Equality', TIME](https://time.com/4587281/catherine-engelbert-gender-equality/).

OpenAI. (2024). ChatGPT [Artificial Intelligence model]. Conversation with ChatGPT on March 29, 2024, at 10:00 a.m. <https://chat.openai.com/>