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* | 8 November 2024, 11:59 PM |
| *Date of Submission* | 8 November 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

[Project proposal 1](#_Toc181720070)

[Introduction 1](#_Toc181720071)

[Objectives 1](#_Toc181720072)

[Problem definition 2](#_Toc181720073)

[Scope and Methodology 2](#_Toc181720074)

[Scope 2](#_Toc181720075)

[Methodology 2](#_Toc181720076)

[Ethical considerations 3](#_Toc181720077)

[Project management and planning 3](#_Toc181720078)

[Data Source Overview 3](#_Toc181720079)

[Project Plan 4](#_Toc181720080)

[Timeline 4](#_Toc181720081)

[Project Plan Timeline - Execution 6](#_Toc181720082)

[Data Exploratory Data Analysis (EDA) 7](#_Toc181720083)

[Data Description 7](#_Toc181720084)

[Summary Statistics 8](#_Toc181720085)

[Data Visualization 9](#_Toc181720086)

[Biases Analysis 12](#_Toc181720087)

[Pre-processing 13](#_Toc181720088)

[Data Cleaning 13](#_Toc181720089)

[Wage Gap 14](#_Toc181720090)

[Handling Outliers and Scaling 16](#_Toc181720091)

[Selecting the most relevant features for analysis 16](#_Toc181720092)

[Modeling and Evaluation 18](#_Toc181720093)

[Data Split 18](#_Toc181720094)

[Training and Test Data 18](#_Toc181720095)

[Cross-Validation 20](#_Toc181720096)

[K-Nearest Neighbors (KNN) 21](#_Toc181720097)

[Decision Tree (CART) 21](#_Toc181720098)

[Model performance correlation and comparative analysis 22](#_Toc181720099)

[Stability Analysis 22](#_Toc181720100)

[Confusion Matrices 22](#_Toc181720101)

[Class Performance Evaluation 22](#_Toc181720102)

[Visualization of Results 23](#_Toc181720103)

[Radar Charts 23](#_Toc181720104)

[Bar Charts 23](#_Toc181720105)

[Visualized Confusion Matrices 24](#_Toc181720106)

[Additional Cross-Validation 25](#_Toc181720107)

[Challenges and Strategies to Overcome Them in Gender Wage Gap Analysis 25](#_Toc181720108)

[Data Quality Issues 25](#_Toc181720109)

[Class Balancing 26](#_Toc181720110)

[High Dimensionality of Data 26](#_Toc181720111)

[Overfitting and Model Complexity 26](#_Toc181720112)

[Interpretation of Multiclass Results 26](#_Toc181720113)

[Conclusion 27](#_Toc181720114)

[Conclusions on the Results Found 27](#_Toc181720115)

[Recommendations 27](#_Toc181720116)

[General conclusion of the results found based on "The wage gap" 27](#_Toc181720117)

[Github 28](#_Toc181720118)

[References 29](#_Toc181720119)

**Gender Workforce Participation**

## Project proposal

### Introduction

This report focuses on analyzing participation in the workforce, with the aim of checking if there are gender disparities. We will look at the challenges and opportunities that people face based on their gender in different roles and job positions.

Women face persistent gender barriers, a lack of role models, and stereotypes that continue gender inequality. As civil rights activist Marian Wright Edelman said: "You cannot be what you cannot see."

### Objectives

#### General Objective:

Explore and analyze the gender gap in the workforce using the Eurostat dataset, starting with an exploratory data analysis (EDA), then develop and optimize a classification model to identify key patterns and predict gender disparity in employment accurately. Machine learning techniques and data visualization will be applied to interpret results and find the best solution, giving a clear understanding of the problem.

#### Specific Objectives:

* Perform exploratory data analysis (EDA) to explore gender gap characteristics in the workforce and understand the Eurostat data structure.
* Apply machine learning techniques and data visualization to find relevant patterns explaining the gender gap in employment.
* Develop a classification model to predict gender disparity in employment accurately, using Eurostat data.
* Compare and adjust at least three machine learning algorithms to choose the best-performing model.
* Optimize selected algorithms by tuning hyperparameters and using cross-validation to improve model accuracy.
* Visualize the final model results to make interpretation and understanding of the gender gap impact easier.

### 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.

#### Why fight the gender pay gap?

* Respect for rights: Respect for women's work rights is essential.
* Productivity: A competitive market needs motivated people who are paid fairly.
* Reducing poverty and inequality: Lower incomes for women affect their pensions and the economic situation of their children, especially single mothers.
* Social harmony: Equal pay encourages a fair division of care work, strengthening personal and work life and contributing to a more just society

## Scope and Methodology

### Scope

This project focuses on analyzing work participation from a gender perspective using data provided by Eurostat. The analysis includes identifying patterns and trends related to the gender gap in different work sectors and countries in Europe. It takes a detailed look at how factors like gender, age, and type of job affect work participation and salary differences. The project aims to implement predictive models to identify these disparities and propose recommendations based on the findings.

### 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 Possible Biases**: Check the data and models to find any biases. This includes seeing how the data was collected and making sure it comes from trusted sources, like Eurostat. Also, check if the data represents different gender groups and other demographics fairly, so no group is left out.

**Evaluating Potential Impact**: Check how possible biases might affect different groups or people who are impacted by the model's decisions. It is important to see if these biases could cause unfair treatment or discrimination against certain groups, especially those who are disadvantaged.

**Mitigating Identified Biases**: Understand the importance of reducing identified biases. During the project, strategies will be used to reduce these biases, like collecting more representative data and, if needed, looking for other reliable data sources.

## Project management and planning

### Data Source Overview

The dataset was obtained from Eurostat, specifically the data is broken down by demographic and social characteristics. The link to access the dataset is included below.

<https://ec.europa.eu/eurostat/databrowser/view/LFSI_EMP_A/default/table?lang=en>

### 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.

### Timeline

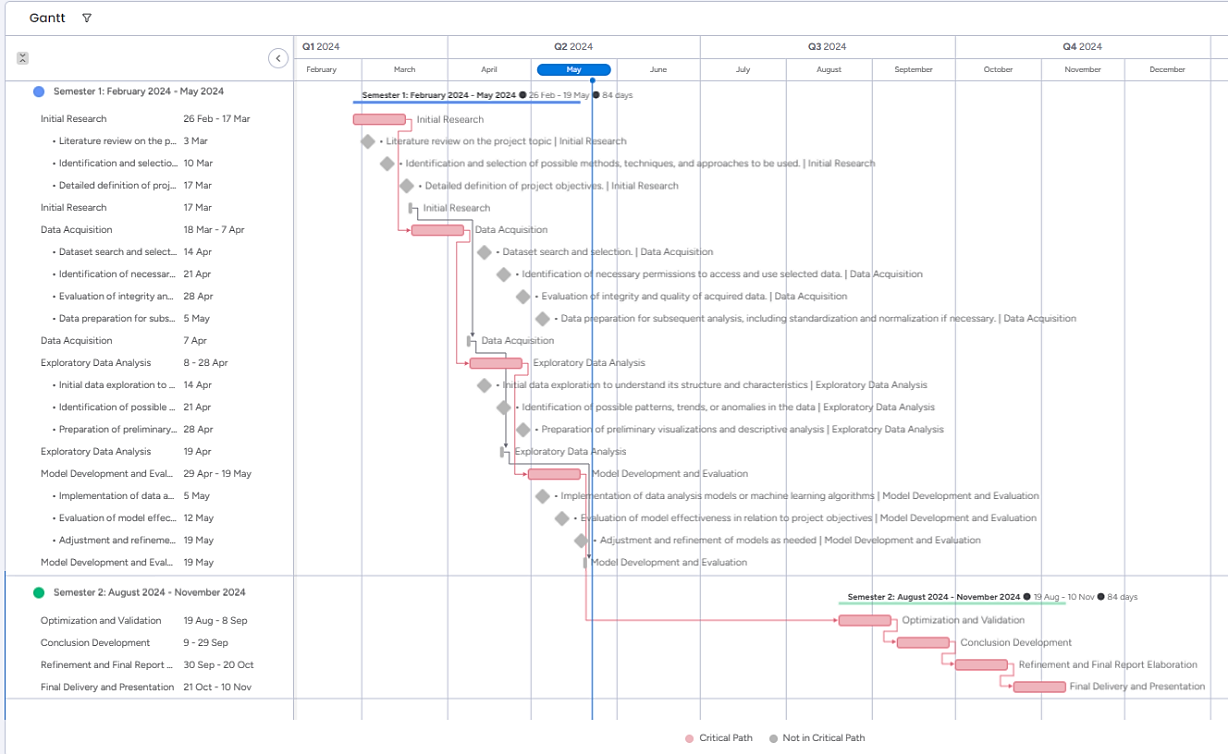
Semester 1: February 2024 - May 2024

* Initial Research (Week 1 - Week 4):
  + Literature review on the project topic.
  + Identification and selection of possible methods, techniques, and approaches to be used.
  + Detailed definition of project objectives.
* Data Acquisition (Week 5 - Week 10):
  + Dataset search and selection.
  + Identification of necessary permissions to access and use selected data.
  + Evaluation of integrity and quality of acquired data.
  + Data preparation for subsequent analysis, including standardization and normalization if necessary.
* Exploratory Data Analysis (Week 11 - Week 16):
  + Initial data exploration to understand its structure and characteristics.
  + Identification of possible patterns, trends, or anomalies in the data.
  + Preparation of preliminary visualizations and descriptive analysis.
* Model Development and Evaluation (Week 17 - Week 22):
  + Implementation of data analysis models or machine learning algorithms.
  + Evaluation of model effectiveness in relation to project objectives.
  + Adjustment and refinement of models as needed.

Semester 2: August 2024 - December 2024

* Optimization and Validation (Week 23 - Week 28):
  + Further optimization of selected models.
  + Cross-validation and robustness testing of models.
  + Evaluation of accuracy and reliability of obtained results.
* Conclusion Development (Week 29 - Week 32):
  + Interpretation of results obtained from data analysis.
  + Formulation of conclusions and recommendations based on project findings.
  + Preparation of preliminary reports and presentation of results.
* Refinement and Final Report Elaboration (Week 33 - Week 36):
  + Review and refinement of the final project report.
  + Incorporation of feedback and suggestions received during preliminary presentation.
  + Preparation of final presentations and additional materials for delivery.
* Final Delivery and Presentation (Week 37 - Week 40):
  + Final project presentation to evaluation committee.
  + Delivery of final project report and other related materials.
  + Closure and completion of all project-related activities.

### Project Plan Timeline - Execution



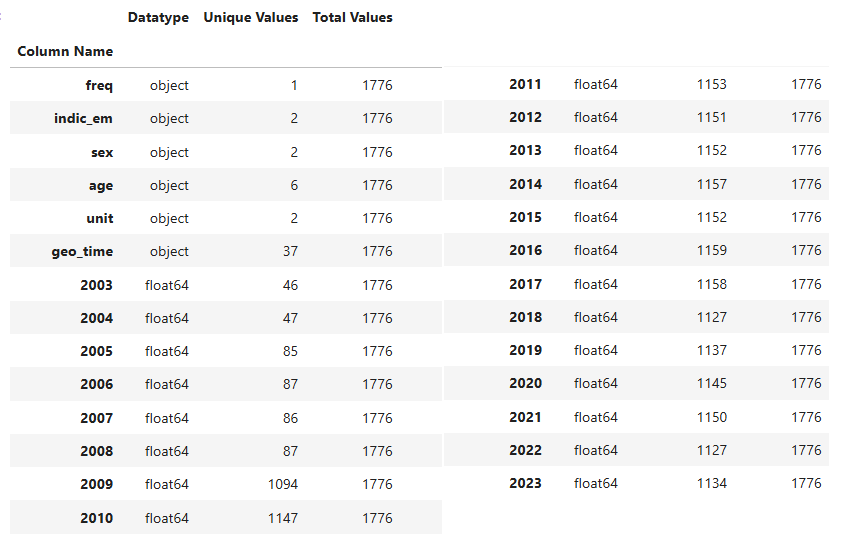




## Data Exploratory Data Analysis (EDA)

### Data Description

The exploratory data analysis (EDA) focuses on understanding the dataset's characteristics and identifying patterns relevant to studying the gender gap in work participation. The dataset consists of 1776 rows and 27 columns, including 6 categorical variables and 21 numerical variables representing employee information such as time frequency, employment type, gender, age, unit of measure, location, and salary.



Data Dictionary

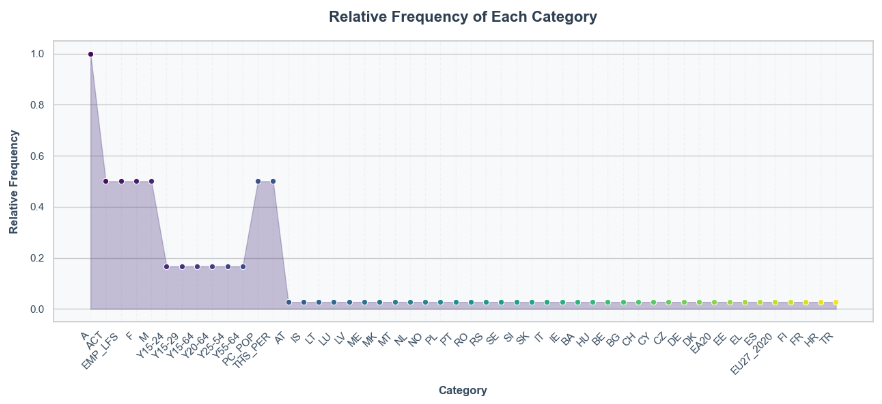
Below is a data dictionary that describes the main variables used in the analysis:

* Time Frequency: All data belongs to category "A", which means there is uniformity in the time frequency of observations.
* Employment Indicator (ACT, EMP\_LFS): Employment indicator, where "ACT" means economic activity and "EMP\_LFS" refers to employment based on the Labor Force Survey.
* Gender: Distribution between men and women, each representing 50% of the data.
* Age Range (Y15-24, Y15-29, Y15-64, etc.): Age range of employees, evenly distributed across the indicated categories.
* Unit of Measurement (PC\_POP, THS\_PER): "PC\_POP" means percentage of the population, while "THS\_PER" refers to thousands of people.
* Geographic Area/Period: Includes various countries and regions, each with 48 observations.

### Summary Statistics

#### Categorical variables

* Time Frequency: Uniform across all observations.
* Employment Indicator: Two categories, "ACT" and "EMP\_LFS," both representing 50%.
* Gender: Equal representation between men and women (50% each).
* Age Range: Categories such as Y15-24, Y15-29, and Y15-64 are equally represented (16.67% each).
* Unit of Measurement: Both "PC\_POP" and "THS\_PER" have equal representation.
* Geographic Area/Period: Various countries and regions are represented with 48 observations each.



The categorical variables have a balanced distribution, which helps in analyzing employment and workforce in different demographic contexts.

#### Numeric variables

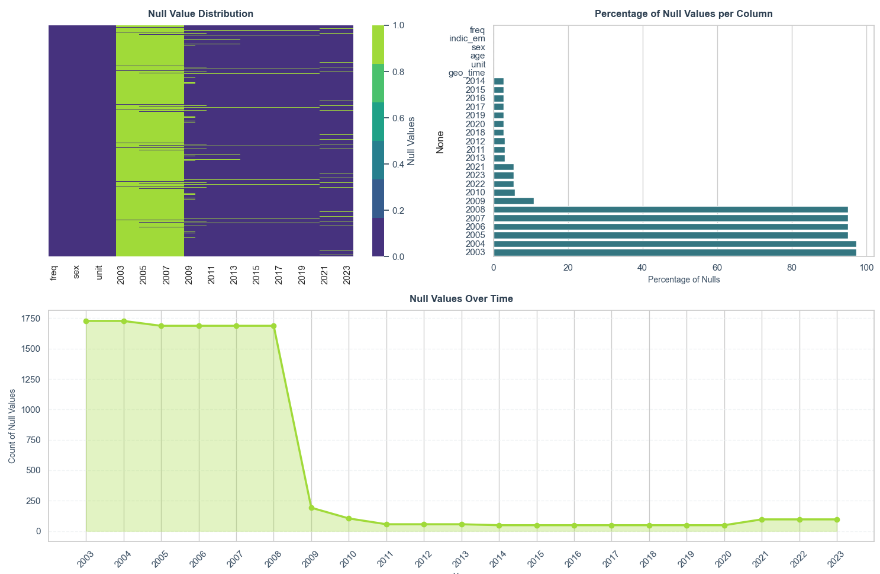
* **Data Count**: The number of records varies by year. From 2003 to 2008, there are fewer records (48-88), while from 2009 onwards, it increases to about 1728.
* **Min and Max Values**: Minimum values range from 4.0 to 27.7, while maximum values have grown from 14,619 in 2003 to 113,314 in 2023.
* **Mean and Median**: The mean stabilizes between 2100 and 2300 from 2009. The median also stabilizes between 80 and 86 from 2009.
* **Mode**: Varies between 35.3 and 81.7 before 2009, and stabilizes between 54 and 86 since then.
* **Range and Standard Deviation**: Both have increased, showing greater data spread in recent years.



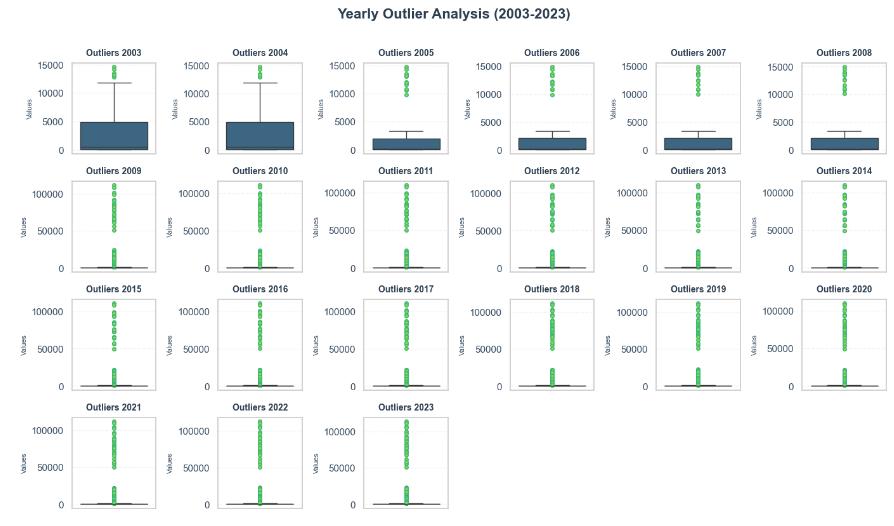
In summary, there has been an increase in both the amount and variability of data, with more stability in some central metrics but greater spread in recent data

### Data Visualization

Missing Data: There are periods with significant missing values, which can affect accuracy and representativeness.

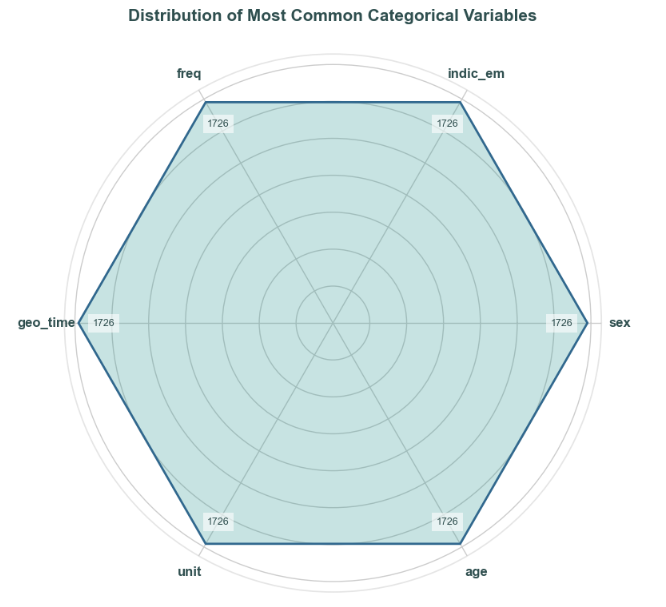
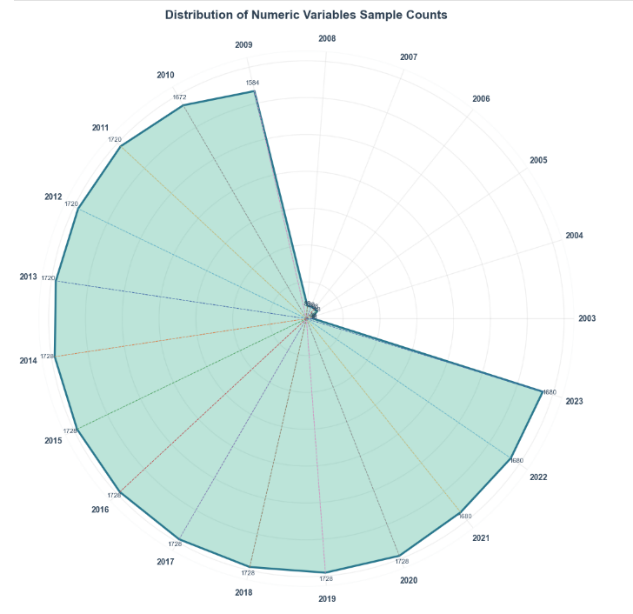


Boxplots: In 2003-2004, there is more spread with several outliers. From 2005, data concentration improves, though some outliers remain.



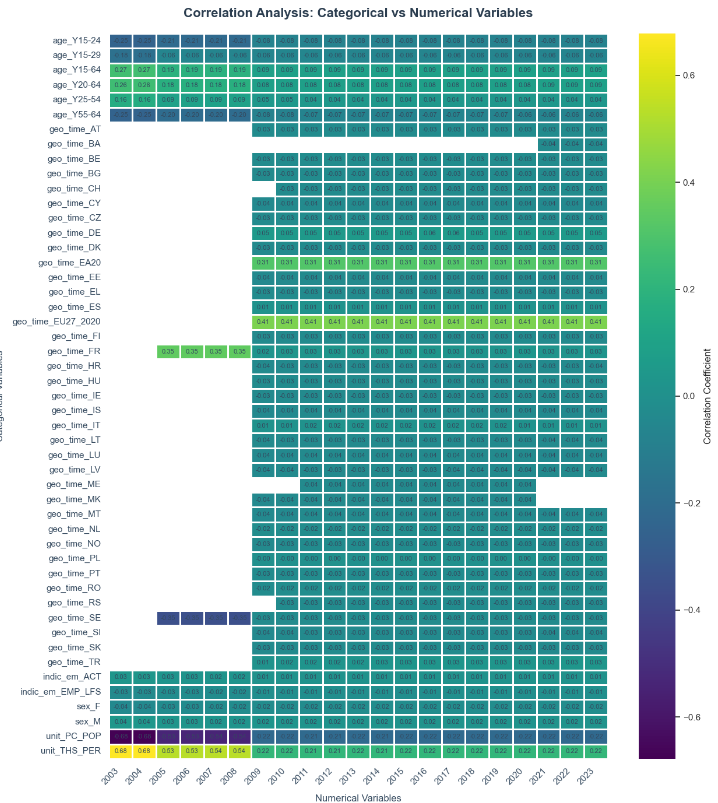
#### Sample Distribution and Annual Variability:

Each variable has 1776 samples, ensuring uniform representation and avoiding biases. The number of samples varies greatly between 2003 and 2023. From 2003 to 2008, there are fewer samples, while from 2009 onwards, there is an increase. The largest discrepancy occurred in 2009 and 2010.

#### Heatmap

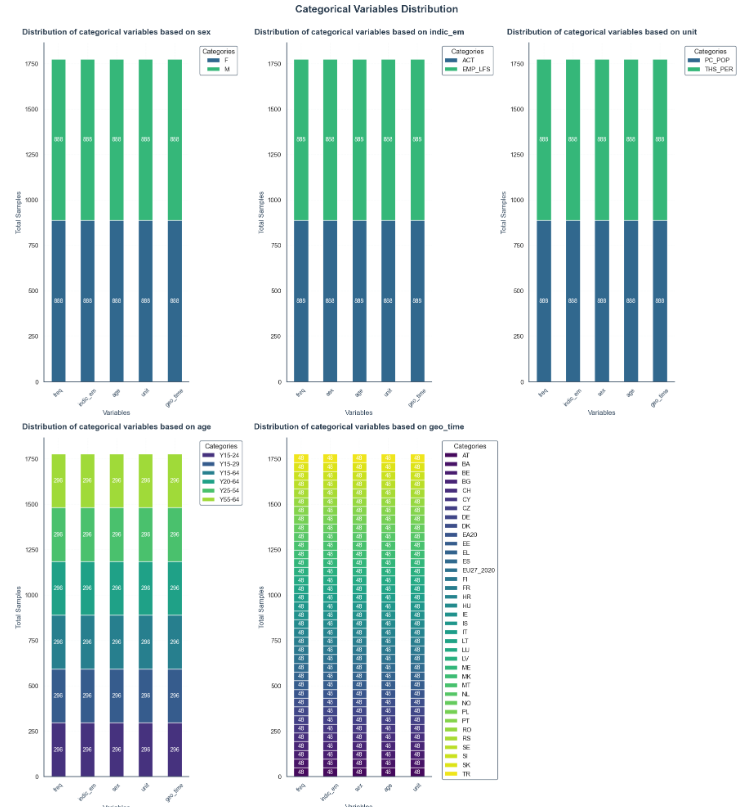
It visualized correlations between variables, helping to identify relationships that could affect model performance.



### Biases Analysis

Although the sample distribution across variables is balanced, potential biases may exist within categories, such as gender. For example, if there is more representation of males than females, this could indicate a bias that affects the generalizability of the results.

Overall, the exploratory data analysis provides a good perspective on the evolution of metrics over time, highlighting both stability and variability in the data.



## Pre-processing

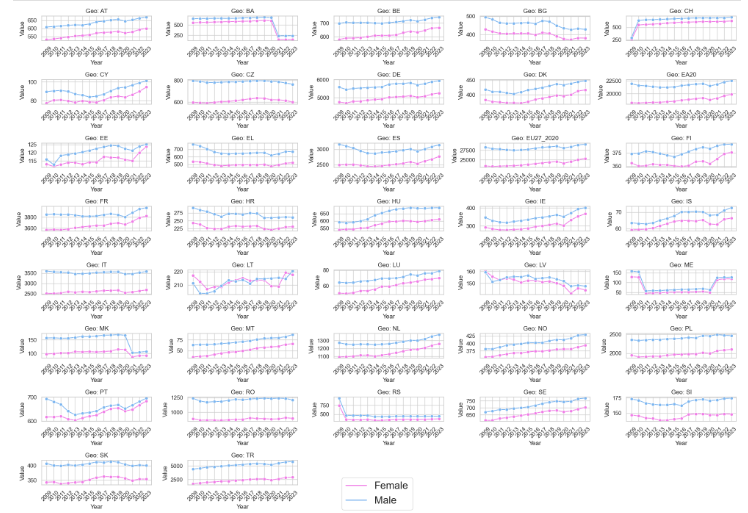
### Data Cleaning

Columns with less than 85% complete data were removed, while remaining missing values were handled using linear interpolation. This approach helped maintain dataset integrity while minimizing information loss.



### Gender Pay Gap Calculation

To calculate the wage gap, we first separated the data by gender for each region. In the graphs, a clear trend is seen: in most countries and years, women's values (pink line) are generally lower than men's (blue line). This shows a possible gap in employment or income where women are at a disadvantage.

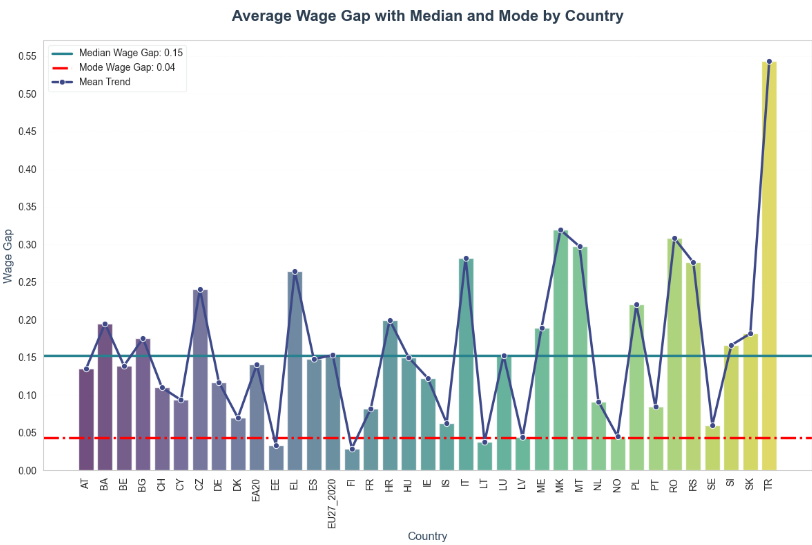


The gender pay gap is calculated year by year, and then averaged for each row. This average is added as a new column called Gender Pay Gap in our database.

The formula used to calculate the gender pay gap is:

#### Average Wage Gap with Median and Mode by Country:

The graph shows the average, median, and mode of the wage gap for each country. These measures together provide a comprehensive view of income disparities between men and women. The results indicate a significant pay gap in most countries, emphasizing the need for stronger policies to address this inequality.



### Encoding and Scaling:

Label encoding was used to convert categorical variables into numerical values. Outliers were removed using the interquartile range (IQR), and numerical variables were scaled to have a mean of 0 and a standard deviation of 1, ensuring standardized data for modeling.



### Selecting the most relevant features for analysis

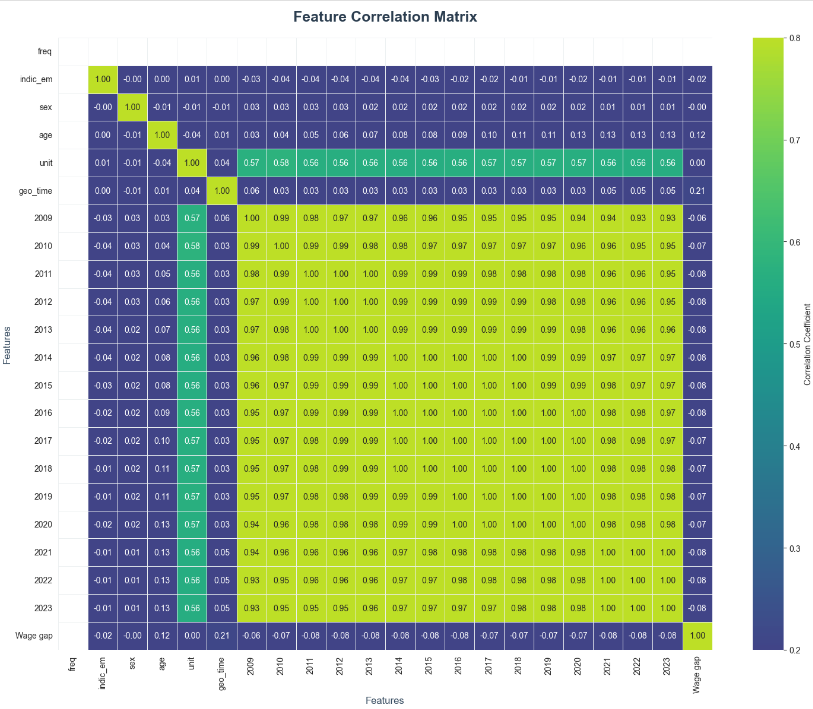
Three feature selection methods were used:

1. Correlation: A correlation matrix was calculated to understand relationships between variables. The correlation of each feature with the target (Pay Gap) was visualized.
2. Univariate Selection (SelectKBest): This method selected the top 10 features based on statistical tests (ANOVA F-value).
3. Lasso Regularization: Lasso regression penalized features with coefficients close to zero, selecting only those with significant influence.
4. Recursive Feature Elimination (RFE): RFE used Lasso as a base model to recursively eliminate less important features, resulting in the top 10 features.

The results from the three methods were combined to identify the most important features. It was recommended to prioritize features selected by Lasso and RFE for predicting the gender pay gap.

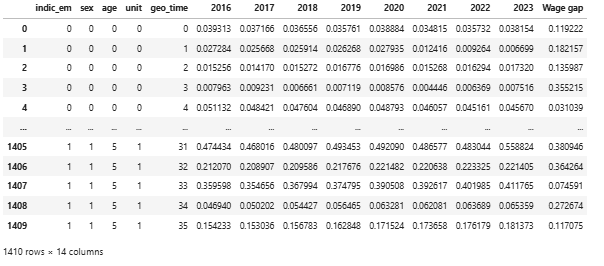
#### Correlation Matrix

The matrix shows the relationship between different variables in the dataset, including the target variable "Gender Pay Gap." The columns for the years (2009-2023) show high correlation among themselves. The correlations of categorical variables with the years are low, indicating no strong relationship between these variables and employment data over the years. The pay gap also has low correlations, suggesting it depends on a more complex combination of factors.



#### Final Feature Selection

The exploratory data analysis provided insights into how metrics evolved over time, highlighting both stability and variability. Identifying patterns and trends and selecting relevant features ensured a cleaner dataset suitable for modeling.



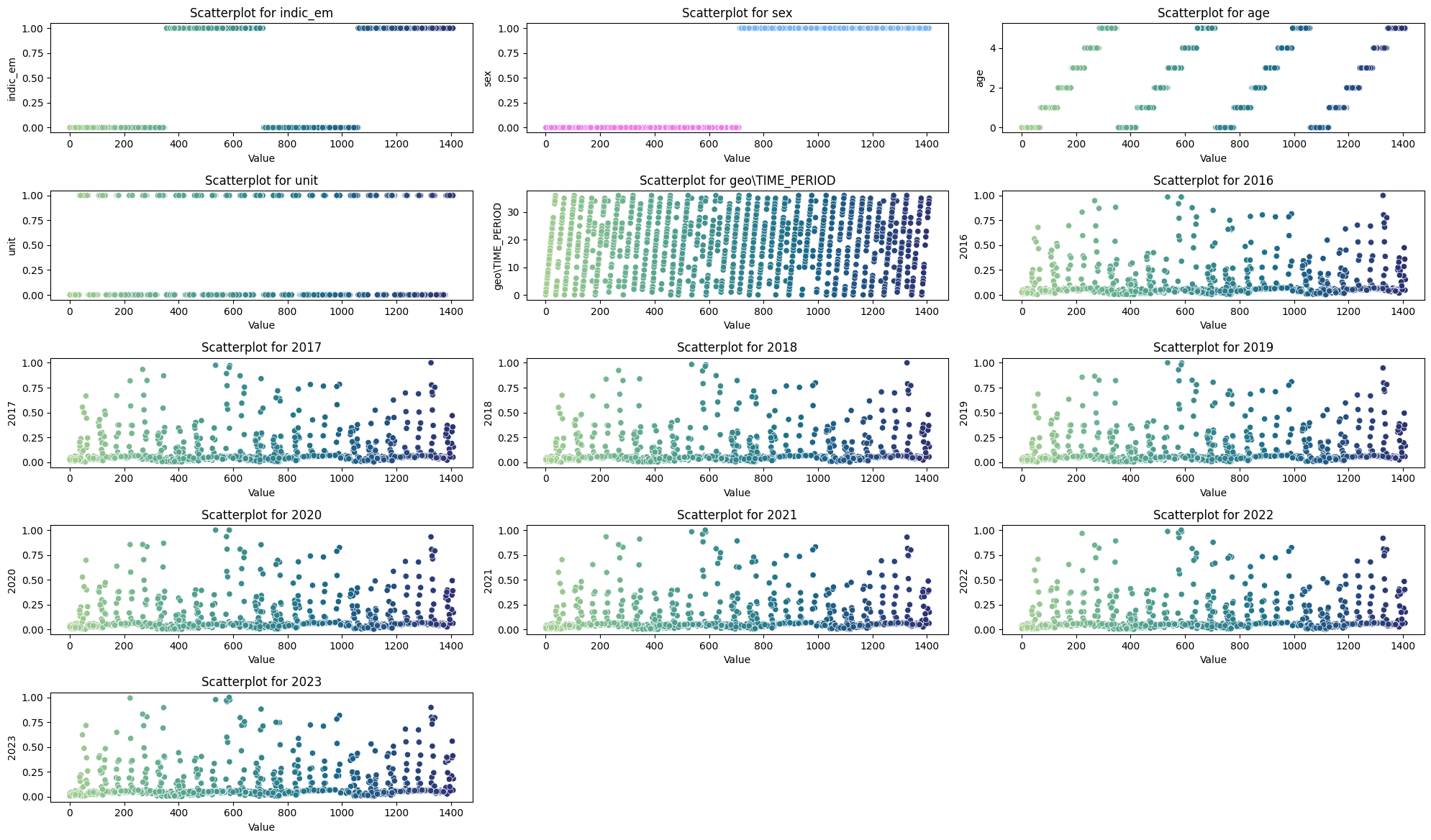
## Modeling and Evaluation

### Data Split

The data was split into training and test sets with a ratio of 70% and 30%, respectively. This split was done in a stratified way based on the salary gap levels to ensure that all categories were represented fairly in both sets. Stratification allowed the models to train and evaluate with results that were comparable and reliable.

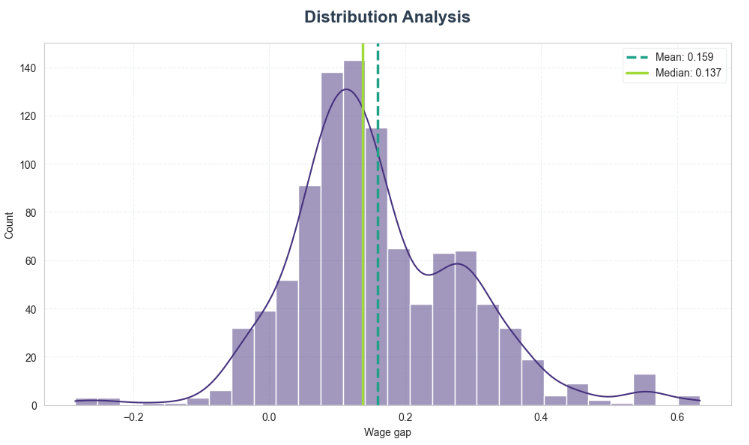
### Training and Test Data

Training Data (X\_train)

* indic\_em: Balanced distribution between men and women (0.4954).
* gender: Almost equal representation between men and women (0.4964).
* age: Data concentrated in the younger age group (mean of 2.38).
* unit: Balanced distribution across different units of measure.
* geo\_time: Great diversity in represented geopolitical entities.

Training Data (y\_train - Gender Pay Gap)

* Number of observations: 987.
* Average: 15.9%, suggesting that, on average, women earn 15.9% less than men.
* Standard deviation: 0.127, indicating moderate variability in the pay gap.
* Outliers: -28.5% (cases where women earn more) and 63.4% (cases of high inequality).

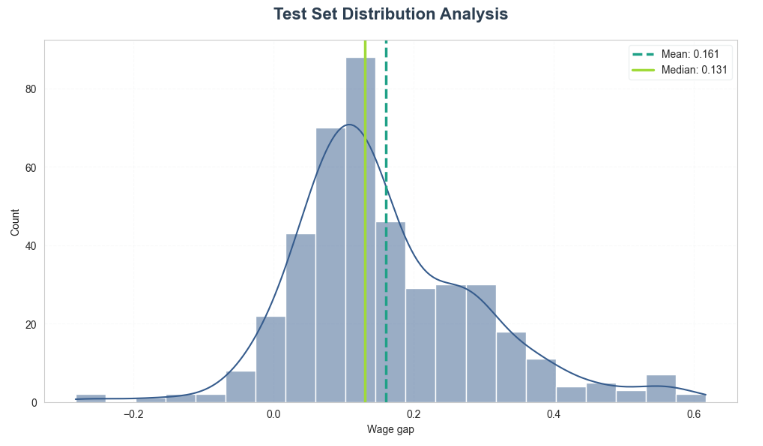


Test Data (X\_test)

* indic\_em: Slight bias towards employment indicators (0.5177).
* gender: Balanced distribution between men and women (0.4941).
* age: Data concentrated in the middle age group (2.57).
* unit: Predominance of the PC\_POP measure.
* Geo\_time: Good geographic representation of EU countries.

Test Data (y\_test - Gender Pay Gap)

* Number of observations: 423.
* Average: 16.0%, indicating that, on average, women earn 16.0% less than men.
* Standard deviation: 0.136, showing moderate variability.
* Outliers: -24.2% (cases where women earn more) and 62.2% (cases of high inequality).



In conclusion, the data show that the gender pay gap remains a significant issue, with prevalent wage inequality in most analyzed countries.

### Cross-Validation

We employed a 5-fold cross-validation strategy to rigorously evaluate the stability and reliability of our models. In this approach, the dataset was divided into five equal parts, and the model was trained and validated five times. In each iteration, one of the five parts was used as the validation set, while the remaining four parts were used for training. This ensured that every data point was used for validation exactly once and for training four times, providing a comprehensive evaluation of the model's performance.

Cross-validation was crucial in avoiding overfitting by ensuring that each subset of data was used for both training and validation at some point. By testing the model on different subsets of the data, we could assess how well it generalized to unseen data. This process also helped fine-tune the model hyperparameters, such as n\_neighbors in KNN or max\_depth in Decision Tree, to maximize the accuracy, precision, recall, and overall robustness of the models. By leveraging cross-validation, we were able to ensure that our models were neither too simplistic (underfitting) nor too complex (overfitting), striking the right balance for optimal predictive performance. This method was key to ensuring that the models did not overfit to specific patterns in the training set, providing robust results.

### K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) model was implemented with different values of k to evaluate its performance in predicting the salary gap. The evaluation metrics for different values of k are presented below:

* **Accuracy**: The overall accuracy of the model showed that the optimal value of k is 5, with an accuracy of 86.76%. As the value of k increased beyond this point, the accuracy tended to decrease, indicating that the model's ability to make accurate predictions was negatively impacted by the use of more neighbors, which introduced noise.
* **Precision and Recall**: The precision and recall metrics also peaked at k = 5. The best performance in terms of precision (86.30%) and recall (86.10%) indicated that the model was effective in correctly identifying positive cases while maintaining a relatively low false positive rate.
* **F1-Score**: The F1-score, which combines precision and recall, also showed the best performance for k = 5, with a value of 0.8676. This reinforces that a moderate number of neighbors resulted in a more effective model for this task.

The results from the KNN model indicate that for the analyzed dataset, the optimal value of k is 5, which allows the model to effectively capture the complex relationships between the variables involved in the wage gap.

### Decision Tree (CART)

The CART Decision Tree model was evaluated across different tree depths, and performance metrics were collected to analyze its ability to predict the salary gap. The results of the evaluation metrics for different tree depths are presented below:

* **Accuracy**: The overall accuracy of the model showed an increasing trend with increasing tree depth, reaching a peak accuracy of 85.82% at a maximum depth of 7. This indicated that while deeper trees improved accuracy, they also increased the risk of overfitting. The cross-validation accuracy of 78.92% was lower, reflecting this risk.
* **Precision and Recall**: Both precision and recall peaked at a depth of 7, with precision reaching 85.40% and recall at 85.00%. These metrics suggested that the model was moderately effective at identifying wage gap instances but struggled to generalize beyond the training set.
* **F1-Score**: The F1-score, which combines precision and recall, was also highest at a depth of 7, with a value of 0.8582. This indicated that while increasing the model complexity improved performance, the gains were limited compared to the risk of overfitting.

The results from the CART Decision Tree model suggest that although deeper trees improved performance metrics, the model's ability to generalize remained relatively limited compared to KNN, indicating that the Decision Tree might not be the most effective technique for this dataset without further refinement.

## Model performance correlation and comparative analysis

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Cross-Validation Score | Test Accuracy | Difference |
| KNN | 84.39% | 86.76% | +2.37% |
| Decision Tree | 78.92% | 85.82% | +6.89% |
| SVM | 70.82% | 75.18% | +4.36% |

**Analysis**: KNN demonstrated the most consistent performance, with minimal variance between training and testing phases, indicating strong generalizability. The Decision Tree, although achieving high accuracy, exhibited signs of overfitting due to a larger discrepancy between cross-validation and test scores. SVM, while stable, had lower performance, indicating difficulty in learning complex patterns present in the data.

### Stability Analysis

* **KNN**: The model exhibited the highest stability, with a small gap between cross-validation and test accuracies, indicating that it effectively generalized to new data.
* **Decision Tree**: The Decision Tree displayed good test performance but higher variability, which suggests overfitting and lower generalizability.
* **SVM**: SVM showed moderate stability but lagged behind in terms of overall accuracy. The model would benefit from improved tuning and class balancing techniques to achieve better results.

### Confusion Matrices

Confusion matrices were generated for each model to evaluate classification accuracy across the wage gap categories (low, medium, high):

* **KNN**: Demonstrated balanced classification across all categories, with relatively few misclassifications.
* **Decision Tree**: Performed well overall but showed a higher rate of misclassification in the medium wage gap category, indicating potential overfitting.
* **SVM**: Struggled particularly with the medium wage gap category, with a higher number of false positives and false negatives, suggesting that the model had difficulty with overlapping features.

### Class Performance Evaluation

**Category Analysis**

The evaluation metrics—accuracy, recall, and F1-score—were analyzed for each wage gap category:

* **Low Wage Gap**: All models performed well in identifying instances in the low wage gap category, with KNN achieving the highest accuracy.
* **Medium Wage Gap**: This category was the most challenging, with both Decision Tree and SVM showing higher misclassification rates. KNN was more effective in handling this category, though improvements are needed.
* **High Wage Gap**: KNN and Decision Tree performed comparably well in this category, while SVM struggled due to overlapping features.

**Metrics by Class**

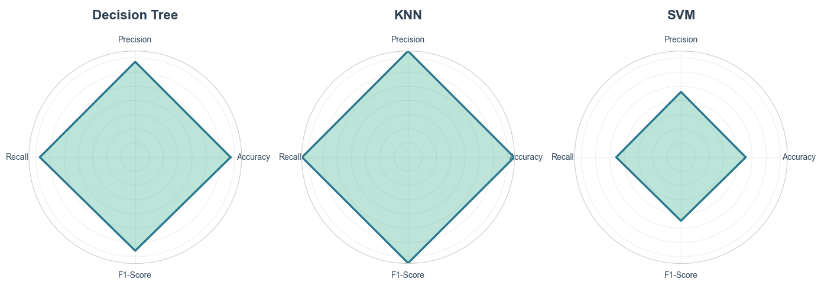
* **Accuracy**: KNN outperformed the other models in all categories, minimizing false positives and ensuring accurate predictions.
* **Recall**: Decision Tree had a high recall, particularly in the low and high wage gap categories, indicating that it was successful in identifying most positive instances.
* **F1-Score**: SVM had lower F1-scores across all categories, reflecting its difficulty in maintaining a balance between precision and recall, especially in the medium category.

## Visualization of Results

### Radar Charts

Radar charts were used to compare the performance metrics of each model (accuracy, recall, and F1-score). The radar charts clearly illustrated that:

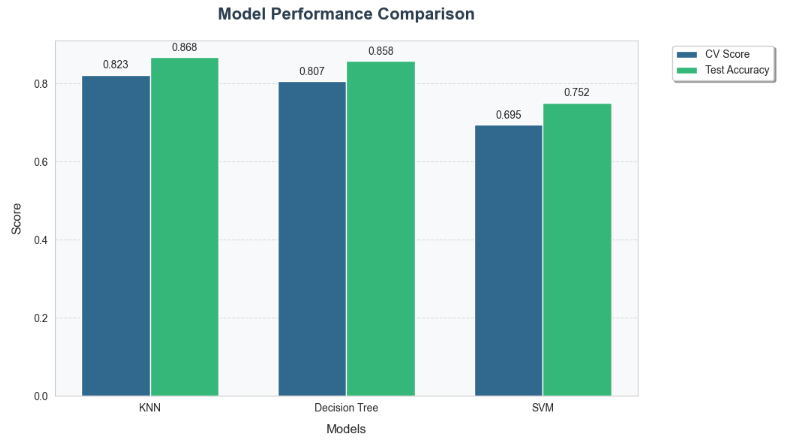
* **KNN**: Had a well-rounded performance, with all metrics showing high values, indicating balanced and effective classification.
* **Decision Tree**: Showed high recall, especially for the low and high wage gap categories, but slightly lower precision.
* **SVM**: Formed the smallest radar chart, indicating weaker performance in all metrics.



### Bar Charts

Bar charts provided a comparison of the average metrics for each model, showing:

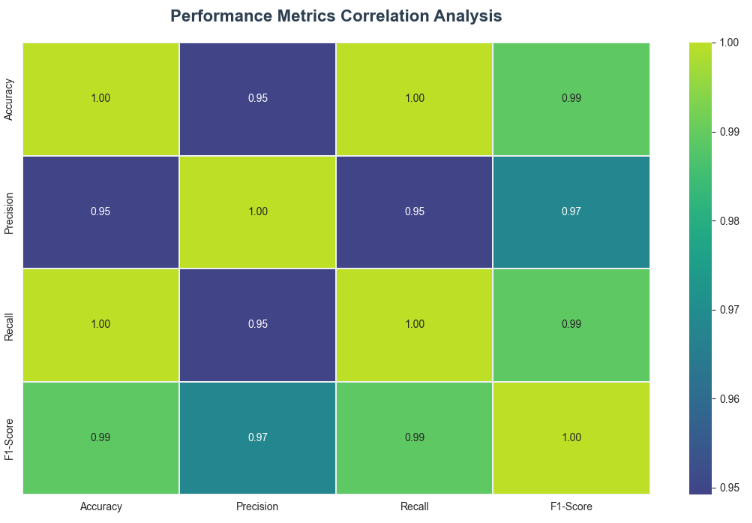
* **KNN**: Outperformed the others in terms of accuracy and overall stability.
* **Decision Tree**: Displayed competitive performance, particularly in recall, but with slightly lower precision.
* **SVM**: Lagged behind in all metrics, indicating the need for further improvement.



### Visualized Confusion Matrices

Confusion matrices for each model were visualized to analyze the distribution of true and false classifications across the categories:

* **KNN**: Showed balanced and effective classification, with minimal bias.
* **Decision Tree**: Revealed some concentration of errors in the medium wage gap category, suggesting a need to refine the model.
* **SVM**: Displayed significant misclassification in the medium category, highlighting areas for improvement.



### Additional Cross-Validation

An additional cross-validation was conducted to further validate the robustness of the models. The results confirmed that:

* **KNN** maintained consistent performance across different folds, indicating high reliability.
* **Decision Tree** showed variability, which confirmed its tendency to overfit, especially at higher depths.
* **SVM** demonstrated moderate consistency but generally underperformed compared to KNN and Decision Tree.



## Challenges and Strategies to Overcome Them in Gender Wage Gap Analysis

During the analysis of the gender wage gap in the workforce using classification models like KNN, Decision Tree, and SVM, several key challenges emerged that affected the performance of the models and the interpretation of the results. Below are the main challenges and the strategies used to solve them:

### Data Quality Issues

One of the main challenges during the analysis was the quality of the dataset. Missing values, inconsistent data entries, and outliers caused significant problems. Several columns in the dataset had missing values, which affected the quality of the analysis. To solve this, we used imputation techniques, such as linear interpolation, to fill in the missing values. In more critical cases, rows or columns with too many missing values were removed. Solving these issues was important to ensure that the models trained with a reliable and consistent dataset, improving the robustness of the results.

### Class Balancing

The dataset had an imbalance between different wage gap categories, with more instances in the low wage gap category compared to medium and high. This imbalance made the models biased towards the majority class, affecting the accuracy and reliability of predictions for minority classes. After performing a detailed analysis to categorize the data, we found that the classes were indeed imbalanced in the training and test sets. To solve this, we used techniques like SMOTE (Synthetic Minority Over-sampling Technique) to oversample the minority classes and create a balanced training set. This allowed the classification model to learn more effectively, avoiding bias towards the majority classes and improving its ability to classify all categories correctly.

### High Dimensionality of Data

Using multiple years and features in the analysis increased the dimensionality, which raised the risk of overfitting in some models. Dimensionality was reduced by removing irrelevant variables and applying techniques like data normalization. In addition, optimizing hyperparameters, such as the number of neighbors in KNN or the depth of the Decision Tree, was essential to maximize performance. Cross-validation was used to adjust the hyperparameters of each model, testing different configurations until the best ones were found. This process ensured that the chosen model achieved a balance between accuracy and generalization capability.

### Overfitting and Model Complexity

Another major challenge was the risk of overfitting, especially with models like Decision Tree, which tend to fit too closely to the training data. Overfitting was reduced by using cross-validation to adjust the model hyperparameters and by applying pruning techniques to limit the complexity of the tree. In addition, regularization techniques were applied to SVM to prevent it from fitting the training data too closely, which would lead to poor generalization on unseen data. These strategies helped the models achieve a balance between capturing complex patterns and avoiding overfitting.

### Interpretation of Multiclass Results

Evaluating results in a multiclass scenario made interpretation more difficult, as traditional metrics like accuracy were not enough to reflect model performance. Interpreting the results of machine learning models, especially in the context of wage gap analysis, was a challenge. While models like Decision Tree provided some level of interpretability, others like SVM and KNN were harder to interpret. To solve this, we used multiple evaluation metrics, including precision, recall, F1-score, and multiclass ROC curves, providing a more complete view of performance. We also used detailed visualizations like decision paths, radar charts, and confusion matrices to identify patterns and better understand the impact of variables on predictions, making model behavior easier to understand.

These strategies helped overcome the main technical challenges, improving both the accuracy of the models and the interpretation of the results in the analysis of the gender wage gap in the workforce.

## Conclusion

### Conclusions on the Results Found

The analysis shows that the KNN model is the most effective at predicting the wage gap in the analyzed dataset. Its ability to correctly classify instances, along with a remarkable balance between precision and recall, makes it a valuable tool to address the gender pay gap issue.

### Recommendations

Based on the findings of the analysis, the following recommendations are made:

Detailed analysis of pay gaps: Since the exploratory analysis of the data revealed significant differences in gender pay in certain countries, it is recommended to conduct a more detailed analysis to identify the underlying causes of these disparities. This may involve investigating factors such as compensation structure, promotion policies and distribution of roles within companies in the different countries.

Review of policies and procedures: Based on the identified pay gaps, it is suggested to review the company's policies and procedures related to compensation and promotion. It is important to ensure that these processes are transparent, equitable and free of gender bias. In addition, the implementation of specific measures to address any identified pay disparities should be considered.

Awareness of gender equity: Since the analysis highlighted the importance of raising awareness about gender equity in the workplace, it is recommended to implement awareness programs targeting all levels of organizations. These programs may include training sessions on unconscious bias, diversity and inclusion, as well as promoting an inclusive work environment where all voices are valued.

Continuous monitoring and impact evaluation: To ensure that the measures taken are effective in reducing gender pay gaps, it is recommended to establish a system of continuous monitoring and impact evaluation. This involves continuing to collect gender-disaggregated salary data and conducting regular analyses to assess progress towards pay equity. The results of these evaluations should serve as a basis for making further adjustments to companies’ policies and practices.

### General conclusion of the results found based on "The wage gap"

The results obtained through the analysis and application of classification models have provided deep insight into the gender pay gap in the workforce. The KNN model, selected as the best in this study, showed solid performance, evidenced by an AUC ranging from 0.8 to 1.0 for various classes. This suggests that the model has an acceptable ability to discriminate between different groups of workers based on their salary.

Across the metrics evaluated

* Precision and Recall: A reasonable balance was observed between the model's ability to correctly identify employees affected by the pay gap and its false positive rate. This indicates that the model is effective in detecting relevant patterns that contribute to wage inequality, although the need to improve recall is identified to ensure that all positive cases are detected.
* F1-Score: The results suggest that although the model has a moderate performance, there are opportunities to further optimize it and improve fairness in classification, especially in less represented groups.

These findings underscore the complexity of the gender pay gap, which is not only manifested in terms of wage disparities between men and women, but is also influenced by demographic and employment factors. The identification of key patterns in the gender pay gap provides a solid basis for policymaking to address these disparities, offering practical recommendations for policymakers and organizations seeking to promote equal pay.

As the model continues to be refined and optimized, these insights are expected to be valuable in identifying strategies to help close the gender pay gap, fostering a more fair and equitable work environment.

## Github

<https://github.com/CCT-Dublin/capstone-project-feb-2024-ft-YumikoBejarano>

# References

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). Step-by-step data mining guide. [online] Available at: https://www.the-modeling-agency.com/crisp-dm.pdf.

Chatgpt.com. (2024). Available at: https://chatgpt.com/share/681cfb33-4726-43ca-8d24-3a2431901e84 Get instant answers, find inspiration, learn something new.

Kazil, J. and Jarmul, K. (2016). Data Wrangling with Python: Tips and Tools to Make Your Life Easier. [online] Google Books. ‘O’Reilly Media, Inc.’ Available at: https://books.google.ie/books?id=XmeDCwAAQBAJ&printsec=copyright&redir\_esc=y#v=onepage&q&f=false [Accessed 22 May 2024].

Müller, A.C. and Guido, S. (2017). Introduction to machine learning with Python : a guide for data scientists. Beijing: O’reilly.

Peng, R.D. and Matsui, E. (2015). The art of data science : a guide for anyone who works with data. Victoria, British Columbia: Lulu.com.

Python, R. (n.d.). Using k-Nearest Neighbors (kNN) in Python – Real Python. [online] realpython.com. Available at: https://realpython.com/courses/knn-python/ [Accessed 22 May 2024].

VanderPlas, J. (n.d.). Handling Missing Data | Python Data Science Handbook. [online] jakevdp.github.io. Available at: https://jakevdp.github.io/PythonDataScienceHandbook/03.04-missing-values.html.