# EXPLORING THE GENDER WAGE GAP IN EUROPE


## Authors

Yumiko Maria Bejarano Azogue


## Project Description

This project studies the gender wage gap in Europe using data from Eurostat and machine learning techniques. It looks at gender, age, and job type to find inequalities and help reduce them. The project uses data analysis to give ideas about fairer work practices.

## Research Question

The following research questions guide the focus of this study:

- Q1: Can we accurately predict the gender wage gap using machine learning models?
- Q2: Which factors are most important in explaining the gender wage gap in Europe?
- Q3: Which machine learning model (KNN, Decision Tree, SVM) performs best in predicting gender disparities?
- Q4: How can the findings support new policies to promote gender equity in the workplace?

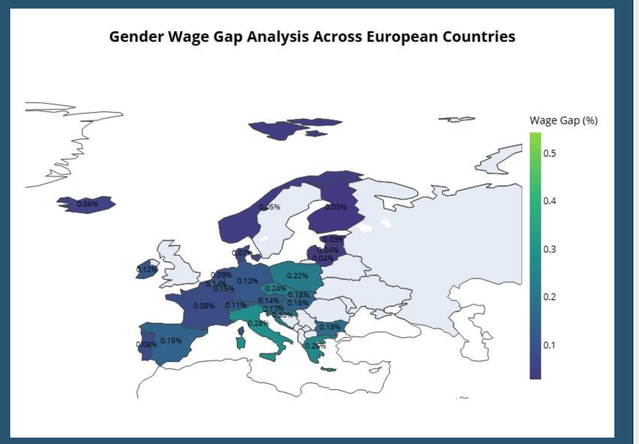



Figure : Gender Wage Gap Analysis Across European Countries *This map visualizes the gender wage gap percentages across various European countries. The darker areas indicate a higher wage gap, highlighting regions where income disparity between men and women is more pronounced. The data is based on the average difference in income for full-time workers of each gender.*

cct College Dublin
Computing • IT • Business

# 01 INTRODUCTION

The Capstone Project focuses on analyzing the Gender Wage Gap in Europe using Eurostat data and machine learning models to understand gender-based differences in employment. The study aims to identify patterns of wage disparities and provide recommendations for reducing inequalities. Three different machine learning models (K-Nearest Neighbors, Decision Tree, and Support Vector Machine) were used and compared. The findings of this project help in understanding which model performs best in predicting wage disparities and identifying factors that most impact gender wage inequality.

This study is important to inform policy changes that can promote fairer work environments.

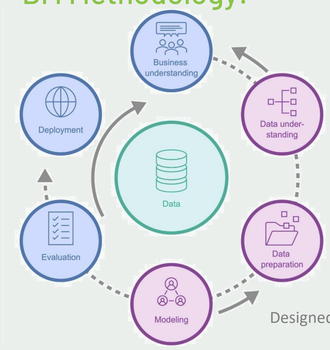
# 02 OBJECTIVES

For this project, the objectives are:

- Perform an Exploratory Data Analysis (EDA) to understand the gender wage gap.
- Use machine learning to find important patterns about the wage gap.
- Develop a model to predict gender differences in work.
- Compare and choose the best of three machine learning models (KNN, Decision Tree, SVM).
- Visualize the results to make them easy to understand.

# 03 METHODOLOGY

## CRISP-DM Methodology:




Designed by Fraunhofer IST

This project follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. This includes steps like data understanding, preparation, modeling, and evaluation.

The CRISP-DM method provides a structured approach to analyzing and modeling the wage gap, focusing on discovering key patterns and making reliable predictions.

## Ethical Considerations:

🔍 Identify potential biases: Review data for any biases that might affect results.

⚖️ Assess bias impact: Analyze how biases influence results and decision-making.

🛡️ Reduce biases: Use strategies to mitigate biases and ensure fair representation.

👥 Gender balance: Ensure equal representation of men and women in the dataset.

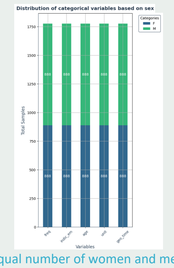



Figure: Equal number of women and men

## Requirements:

•Technologies: Jupyter Notebook, Python, Google Cloud Platform.
•Data: Dataset from Eurostat about work participation in Europe.
Assumptions:
•The dataset is big enough for the analysis.
•Information for each country is complete and correct.

# GENDER WAGE GAP

Designed by Freepik

## BUSINESS UNDERSTANDING

The gender wage gap is a significant challenge across Europe. This project aims to explore and analyze gender-based differences in employment and provide recommendations for improving wage equality.

By understanding the disparities in wages, policymakers can create solutions to promote fair pay practices and improve equality.

Using machine learning helps identify the key factors that influence the wage gap, providing insights for future policy adjustments.

## DATA UNDERSTANDING

We collected the data from Eurostat, which provides information about employment participation and wages across Europe. The dataset includes:

- Gender, Age, Employment Type, and Region.
- Years covered range from 2003 to 2023, allowing a detailed analysis of trends over time.

The gender wage gap measures the difference in earnings between men and women across different countries and years. This gap is calculated using the average wages for each gender. This graph compares the average gender wage gap between men and women in Europe, highlighting specific differences by country.

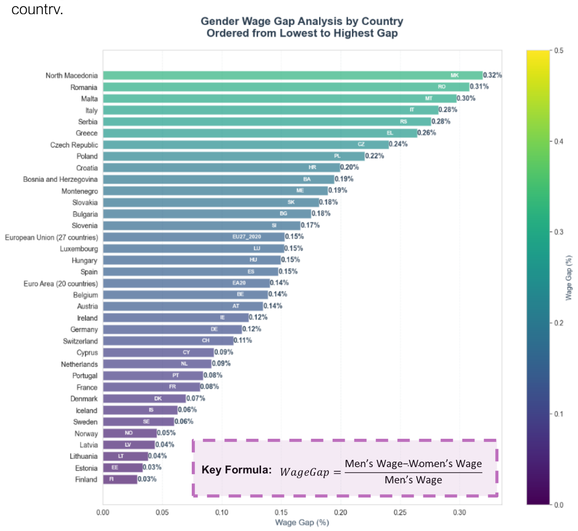



Figure: Gender Wage Gap Across European Countries

# 04 DATA PREPARATION

For data preparation, several important steps were taken:

- Data Cleaning: Missing values were handled using linear interpolation to ensure data quality.
- Feature Selection: Important features like age, employment type, and region were selected using correlation analysis.
- Balancing the Data: The dataset was balanced using SMOTE to handle gender class imbalance.

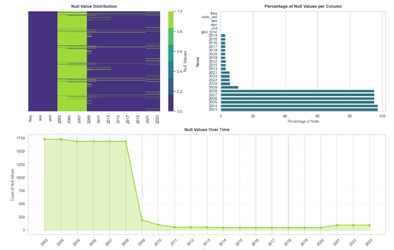



Figure: Analysis of Missing Values: Distribution, Percentage by Column, and Trend

# MODELLING TECHNIQUES

We used three different machine learning models to predict gender wage disparity:

| | |
|---|---|
| K-Nearest Neighbors (KNN) | • Best accuracy with k=4. Balanced between accuracy and generalization. |
| Decision Tree | • Good performance but risk of overfitting with deeper trees. |
| Support Vector Machine (SVM) | • Stable but lower performance compared to KNN and Decision Tree. |


Figure: Model Performance Comparison

A bar chart comparing the accuracy of each model, highlighting that KNN had the highest accuracy with a good balance among metrics.

In predictive modeling, K-Nearest Neighbors (KNN) with k=5 was the most successful, reaching an accuracy of 87.94% with well-balanced precision and recall. The Decision Tree model achieved an accuracy of 81.16% at a depth of 16, but showed problems with overfitting. These findings suggest KNN as a robust model for predicting wage gaps, with radar charts highlighting the comparative performance of each model.

.

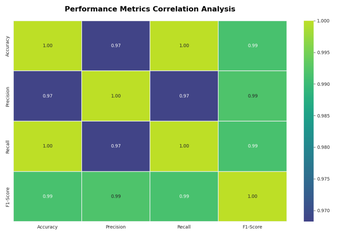



Figure: Correlation Analysis of Performance Metrics (Accuracy, Precision, Recall, F1-Score)

# 05 CONCLUSION

The results from the analysis provide insights into the gender wage gap:

- Q1: The KNN model can accurately predict the wage gap, making it a useful tool for analyzing wage inequalities.
- Q2: Factors such as employment type, age, and region were found to be important in explaining the wage gap.
- Q3: KNN was the best-performing model, providing a good balance of accuracy and generalization.
- Q4: Findings can support the creation of fair wage policies to reduce gender inequalities in the workplace. These results support the implementation of policies aimed at reducing wage inequality, and organizations can use these predictions to identify key areas for improvement.

## References


- Eurostat Employment and activity by sex and age - annual data. Available at: https://ec.europa.eu/eurostat/databrowser/view/LFSI_EMP_A (Accessed: 10 November 2024).
- UN Women (no date) Gender pay gap: Causes, figures, and why it should be fought. Available at: https://lac.unwomen.org/es/que-hacemos/empoderamiento-economico/epic/que-es-la-brecha-salarial (Accessed: 10 November 2024).
- Chapman, P. et al. (2000) Step-by-step data mining guide. Available at: https://www.the-modeling-agency.com/crisp-dm.pdf (Accessed: 10 November 2024).
- Hotz, N. (2024). What Is CRISP DM? [online] Data Science Project Management. Available at: https://www.datascience-pm.com/crisp-dm-2/
- VanderPlas, J. (no date) Handling Missing Data | Python Data Science Handbook. Available at: https://jakevdp.github.io/PythonDataScienceHandbook/03.04-missing-values.html (Accessed: 10 November 2024).