National College of Ireland

Analyzing Gender Disparities in Corporate Leadership Using Machine Learning Techniques on Global Board Member Data

MSc Research Project
Data Analytics

Geerthana Moorthy
Student ID: x23267348

School of Computing
National College of Ireland

Supervisor: Aaloka Anant

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	GEERTHANA MOORTHY
Student ID:	X23267348
Programme:	M.SC Data Analytics **Year:** 2024-2025
Module:	Research Practicum
Supervisor: **Submission Due Date:**	11-08-2025
Project Title:	Analyzing Gender Disparities in Corporate Leadership Using Machine Learning Techniques on Global Board Member Data
Word Count:	8116 **Page Count** 19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Geerthana Moorthy

Date: 11-08-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Analyzing Gender Disparities in Corporate Leadership Using Machine Learning Techniques on Global Board Member Data

Geerthana Moorthy
X23267348

Abstract

Gender disparity in corporate leadership remains a significant and persistent issue in the global business landscape. This research addresses the challenge of analyzing the factors contributing to this disparity by developing a comprehensive methodological framework. Due to the scarcity of granular, publicly available data on corporate board members, this study pioneers an approach based on a large-scale synthetic dataset comprising 200,000 records for 20,000 companies. This dataset is engineered to reflect realistic global distributions and inherent biases related to industry, region, and company size. A machine learning pipeline is then constructed to predict the gender of a board member using a profile of individual and corporate attributes. The pipeline requires significant feature engineering, pre-processing and dimensionality reduction through Principal Component Analysis (PCA). Several classification models were developed, and a tuned XGBoost model emerged as the final model in the pipeline. The final model reports a test accuracy of 79.14% and an AUC of 0.7751. The evaluative process identifies a large performance disparity between the classes, specifically a low recall of 26.11% in the female category, indicating how the model was biased as a result of learning from imbalanced training data. The primary contribution of this study is to provide a replicable model to examine and study demographic disparities using synthetic data and machine learning, but also presents evidence of how algorithms can internalize and manifest existing systemic biases present in training data.

1 Introduction

The composition of corporate boards of directors is a vital component of contemporary corporate governance, with important implications for strategic direction, financial performance and social duty. A major theme for researchers, as well as government officials and the general public, for many decades has been the element of gender inequality in respect, not only to women's underrepresentation in senior leadership and board opportunities, but also men's over-representation in these spaces, within both domestic and international contexts. While the substantive progress made in many sectors is noteworthy, women globally continue to be significantly under-represented in senior leadership and board roles. This phenomenon is a greater issue beyond its social dimension; it affects organizational effectiveness and creates inequity (Gander and Sharafizad, 2025). Moreover, gender imbalance at the top has been shown to correlate with differences in corporate disclosure associated with transparency, risk management strategies, and innovative capacity (Afeltra, Alerasoul and Usman, 2022). Therefore, addressing gender inequality is increasingly being considered not just a moral imperative, but also strategic for developing resilient, innovative, and future forward-looking organizations that can manage the

challenges facing them in today's increasingly complex 21st century economy (Smith and Sinkford, 2022).

A primary barrier to understanding the confluence of many factors that contribute to this gender inequality, is the lack of meaningful, standardized and publicly accessible data. While aggregate statistics on board diversity are often published, they lack the granularity required for sophisticated, multivariate analysis. Detailed data profiling individual board members across a global spectrum of companies, industries, and regions is scarce and often proprietary (Behlau, Wobst and Lueg, 2024). This data deficit makes it exceptionally difficult to move beyond descriptive statistics and employ advanced analytical techniques. Such techniques are needed to uncover the subtle, interconnected factors—from personal attributes like age and education to corporate characteristics like industry and company size—that may predict an individual's presence and role on a corporate board.

The advent of machine learning and big data analytics offers a powerful new lens through which to examine such complex social phenomena (Li, 2025). These computational methods excel at identifying non-linear patterns and complex interactions within large, high-dimensional datasets that traditional statistical approaches might miss. By framing the problem of gender representation as a prediction task, it is possible to identify which features are most indicative of a board member's gender, thereby shedding light on the characteristics most commonly associated with male and female representation in leadership. The successful application of these techniques, however, is fundamentally contingent on the availability of a suitable, large-scale dataset.

This research directly confronts this data scarcity problem by first generating a large-scale, realistic synthetic dataset. This approach allows for the creation of a controlled yet complex environment in which to develop and rigorously test a machine learning framework. The motivation for this is twofold. First, it aims to create a tangible and replicable methodology that can be adapted and applied if or when real-world data of sufficient quality becomes available. Second, it utilizes the predictive model itself as an analytical tool. The way a model learns from intentionally biased data, the patterns it prioritizes, and the errors it makes can provide profound insights into the nature and structure of that bias, contributing to the critical field of algorithmic fairness and responsible data governance (Wu, 2023).

This study is therefore guided by the overarching goal of developing a computational framework to analyze gender disparities in corporate leadership. To achieve this, the following research questions and objectives have been established.

1.1 Research Questions

- What are the key individual and corporate features that are most predictive of a board member's gender in a simulated global corporate environment?
- To what degree of accuracy and fairness can machine learning models predict whether a board member is female, based on a comprehensive profile of personal, professional, and corporate attributes?
- Which factors exhibit the most significant influence on the predictions of the optimal machine learning model, and what does this imply about the systemic patterns of gender representation that the model has learned?

1.2 Research Objectives

- To design and implement a process for generating a large-scale synthetic dataset of global corporate board members, incorporating realistic, multi-faceted biases related to geography, industry, and other corporate characteristics.
- To develop a robust machine learning pipeline that includes data preprocessing, extensive feature engineering, dimensionality reduction, and model training to predict a board member's gender.
- To systematically train, evaluate, and compare the performance of multiple classification algorithms to identify the most effective model for the prediction task, with a focus on metrics suitable for imbalanced data.
- To conduct an in-depth analysis of the final, optimized model, including its performance on different subgroups, its classification errors, and the relative importance of predictive factors, to derive actionable insights into gender disparity patterns.

This thesis makes several key contributions to the existing scientific literature. Methodologically, it provides a novel and transparent framework for generating synthetic data for social science research, particularly in the domain of corporate governance where data access is a known and significant limitation (Behlau, Wobst and Lueg, 2024). Analytically, it demonstrates the application of a complete, modern machine learning workflow to the persistent problem of gender disparity, moving the analysis beyond traditional descriptive statistics. The findings offer empirical insights into how systemic biases are learned and reflected by predictive models, contributing to the growing conversation on algorithmic fairness and data governance (Wu, 2023; Sklavos et al., 2024). Finally, the entire framework, from data generation to a deployed web application, serves as a replicable and extensible blueprint for future research in this and related areas of social inquiry.

The remainder of this report is organized to guide the reader through this research journey. Chapter Two provides a critical review of the related academic literature. Chapter Three details the research methodology, offering a comprehensive description of the synthetic data generation process and the design of the machine learning pipeline. Chapter Four presents the design specification of the system, including the overall architecture. Chapter Five describes the implementation details. Chapter Six presents a thorough evaluation of the experimental results. Finally, Chapter Seven concludes the report, summarizing the research, discussing its limitations, and proposing meaningful directions for future work.

2 Related Work

This chapter provides a critical review of the academic literature consulted for the analysis of gender disparity in corporate leadership using machine learning. Subsequently, it attempts to position the present study within this scholarly landscape, identifying the specific contributions of the previous work while pinpointing the particular areas of knowledge this research seeks to fulfill. The thematic strands underlying this review comprise the primary literature about the role and composition of corporate boards, a large body of work pertaining to gender fairness in leadership, nascent applications of machine learning to corporate and ESG contexts, and long-standing methodological difficulties about data and measurement in

diversity research. It is through the synthesis of these very streams that the contribution and innovations of this thesis will be substantiated.

2.1 Board Composition and Corporate Governance

The board of directors is at the top of corporate decision-making, and its composition has long been amongst the subjects of intensive academic reflection. Research in the field builds strong associations between board characteristics and various organizational outcomes. Earlier investigations primarily focused on structural variables, for instance, board size, proportion of independent directors, duality of CEO and chairperson role, seeking to correlate these factors with firm performance. More recently, the focus has broadened to include demographic diversity as a key element of effective governance. An analysis of the systematic literature network made by Afeltra, Alerasoul, and Usman (2022) reveals an increasing number of works relating board characteristics to corporate social reporting, with high diversity and independence of boards often being associated with enhanced transparency. A similar bibliometric analysis by de Enrique Arnau and Pinillos-Costa (2024) shows that the structure of boards increasingly is seen as a pivotal point in the context of managing complex business transformations.

The concept of board diversity has become more nuanced to understand the antecedents of top management team and board gender diversity, Yao (2023) completed a comprehensive review of the gender diversity literature, and notes that both external drivers (regulatory pressure, investor demands) and internal drivers (corporate culture, nomination procedures) contribute to gender diversity on boards and TMTs. Taken together, this research stream suggests that board composition is a consequence of many intersecting forces and not arbitrary. However, one significant limitation across this area is the reliance on Google Scholar or a similar approach to query search terms and a traditional quantitative statistical perspective to obtain measurements such as regression analysis. Although these are valid approaches, they may not capture the complexities and non-linear interactions that multiple attributes characterizing a board member may allow. One particular challenge identified within the critical review undertaken by Behlau, Wobst, and Lueg (2024) is that of data quality, with the inconsistencies and incompleteness of public data sources being a strong pitfall in diversity research. This problem of data scarcity is an important argument in favor of using the synthetic data in this thesis.

2.2 Gender Equity in Leadership and Business

A vast body of literature underlay for the problem of gender disparity. The consistent theme in this body of research has revealed that systematic barriers exist. Such barriers inhibit women's access at the highest levels of corporate power. For example, Smith and Sinkford (2022) discuss such obstacles in the way of women accessing leadership positions in global health, an area that employs a majority of females but where the management is male-dominated, a pattern found in many other sectors as well. Gander and Sharafizad's systematic literature review (2025) on advancing gender equity in senior leadership synthesizes evidence on effective interventions, such as formal mentorship programs and transparent succession planning, among other things. According to Górska and Burlakova (2025), such enablers of

gender inequality usually include enduring stereotypes and work-life balance as some of the main issues women face while pursuing business leadership roles.

This movement in research has undoubtedly become specialized- into the interaction of gender and specific business functions within particular industries. Mehnaz and Yang (2025), for example, explore gender diversity in accounting research, concluding that although women are slowly increasing in numbers, their presence is absent in top journals and their positions in academia mirror the corporate "pipeline" problem-their underrepresentation in top positions. A systematic review of gender differences in accessing entrepreneurial equity funding by Koziol, Schmitz, and Bort (2025) highlights that finance is a field in which female ventures experience more scrutiny and receive less funding compared to those run by men. A similar finding of gender gap prevalence is reported by Prusak and Wacławek (2023) in the quickly expanding FinTech sector. The methodologies within this area are widespread-ranging from applied qualitative case studies to large-scale surveys. Though these methods afford in-depth understanding, a scoping review by Belingheri et al. (2021), covering twenty years of research on gender equality, identified a clear need for new analytical tools capable of handling the increasing complexity of available data. This thesis provides that by presenting a machine learning-based framework as one such advanced tool.
Improperly paraphrasing sentences can prevent students from being aware of future-business research trends. The problem is that more than 70 percent of research on gender differences is based on variables such as education, age, and major income sources. There are many problems with those numbers. First, they adapt to students or to part-time employees who work full time in a business, but get no compensation for that work. The percentage of part-time students is about 51. Then there are high research costs because as many as 58 percent of all part-time students do not receive any financial aid.

2.3 Machine Learning in Corporate and ESG Research

Currently, AI and ML applications are gaining traction as a new area of inquiry into issues concerning corporate governance, finance, and ESG criteria. These techniques hold great promise for assisting in novel insights generation through large and complex datasets. For example, Kalusivalingam et al. (2022) advocate for the use of NLP and ML algorithms for good corporate governance by automating the analysis of reports and legal documentation. Sklavos et al. (2024) discuss the concept of AI-based governance to digitalize firms' leadership and HR management and suggest that AI can be beneficial for more objective decision-making.
In a comparative study pertinent to this thesis's methodology, Xue, Ong, and Demir (2024) apply various machine learning models to assess the effects of CEO and chairman characteristics on green innovation. Such works present the ability of ML to forecast important factors in a complex organizational setting. A review by Li (2025) on Big Data and ML in ESG research particularly highlights their usefulness for sifting through huge quantities of unstructured data to create ESG scores and risk identification. Even meta-analysis has benefited; Off and Baltes (2021) used an ML-based literature review to analyze the context of women in entrepreneurship, showcasing ML's utility for finding patterns within the body of research itself.

Critics have raised several issues against AI. Deepak et al. (2025) review the role of women in AI-enabled management, raising questions around whether AI systems risk perpetuating existing biases if not deliberately designed. Doo and McGinty (2023) emphasize the significance of representation—both in data and the workforce—in building equity for AI systems. In an extensive bibliometric analysis, Wu (2023) documents the exponential increase in algorithmic discrimination literature, progressively raising critical concerns on how data governance and algorithmic design can impact human rights. This body of literature underscores the need for not just the realization of predictive models but also an evaluation of their fairness and possibly harmful impacts, which is a central theme within the evaluation chapter of this thesis.

2.4 Synthesis and Identification of the Research Gap

Where the intersection of these research themes reveals a significant gap that this thesis attempts to fill is the challenge to apply advanced computational methods to a problem area suffering from scarcity of data and issues of measurement. The literature on corporate governance and gender equity has, on the whole, demonstrated how important board diversity is, but it has often found itself constrained by limitations of data, as emphatically argued by Behlau, Wobst, and Lueg (2024). The very problem of data scarcity militates against the application of machine-learning models that have been established, by the very field in which the disputes of computational business analysis compel us to see them as powerful. Therefore, there is a methodological void. This thesis resides squarely at that intersection, accepting the data limits as its starting point and then proposing a remedy: construction of a high-fidelity synthetic data set, to which would be applied the advanced predictive methods captured in the works of Li (2025) and Xue, Ong, and Demir (2024). Next, it used those advanced methods with the explicit goal of identifying potential algorithmic bias, making fairness assessment of the models a central concern of the study, as has recently been raised by Wu (2023). Providing a complete methodological landscape that the literature currently lacks, this study establishes a pathway for rigorously studying the drivers of gender disparity in a controlled computational environment.

3 Research Methodology

This chapter comprehensively presents a systematic and meticulous understanding of the research process exercised in this work. The methodology is open, replicable, and considerably robust, targeting the research issues put forward in Chapter One. Overall, the approach is quantitative-computational and divided into distinct phases: synthetic data generation, data preparation and exploratory analysis, feature engineering, and development of a predictive modeling pipeline. Each phase is designed to build on a previous phase and hence allow for coherent and logical research work flow.

3.1 Overall Research Design

The multi-phase, model-based design is used in this investigation. It is designed exactly to tackle the main limitation of missing real-world data, a well-known challenge in corporate governance research (Behlau, Wobst and Lueg, 2024). The main thrust of the approach constructs an artificial reality-synthetic dataset - that emulates known patterns of gender

imbalance, and allows the experimentation and development of a machine learning framework in this artificial environment. This would ensure a level of control and transparency that could not be achieved with the incomplete or proprietary real-world data. The process of the research was organized in the following broad phases:

1. **Synthetic Data Generation:** A Python-based process to create a large-scale, feature-rich dataset of corporate board members.
2. **Data Preprocessing and Exploratory Data Analysis (EDA):** Cleaning, preparation, and visualizations of the generated data in order to understand its structure, distributions, and inherent relationships.
3. **Feature Engineering:** The addition of some new informational features from the available data into the existing data, to improve the predictive performance of models.
4. **Modeling Pipeline Construction:** Designing a systematic workflow for feature scaling, encoding, dimensionality reduction, model training, and evaluation.

3.2 Phase One: Synthetic Data Generation

The backbone of this study is a synthetically created dataset. The goal is not to create random data but actually programmatically creating a dataset that has the complex multilayered biases found in the real world. The data generation process was invoked through Python, making use of standard libraries including Pandas, NumPy, and Faker.

The process started by expressing the core defining parameters of the dataset, setting an aspirational target of generating board member records for 20,000 unique companies, giving a total of 200,000 entries. Foundational probability distributions for various attributes were then established based on general observations from the literature. This includes basic baseline probabilities for gender, educational levels, company size, and industry sectors.

The sophistication of the script lies in its systematic injection of realistic biases. This was done through several programmed mechanisms:

- **Industry-Specific Gender Bias:** A dictionary was created to match gender probabilities to the applicable industry. For instance, the industry 'Energy' was meant to have higher male representation (85%) than an 'Healthcare' (65%).
- **Regional Gender Bias:** Again, a similar dictionary was maintained for shaping the different degrees of gender diversity across North America, Europe, Asia Pacific, and Emerging Markets.
- **Dynamic Probability Calculation:** A function was created that mingles these biases, also includes an adjustment for company size, with larger companies modeled to have slightly better gender diversity. This ensured that the gender of a generated board member was a function of their company's specific context.

Realistic attributes were also generated using probabilistic functions. Tenure for a member is modeled as a function of age, role, and gender to reflect historical patterns, assigning a bit shorter mean tenure for women. The size of a company's board was determined by its size and industry. The final build-up then happened in that the 20,000 entry companies were first generated, the 200,000 board members were generated and associated with the companies, and lastly, each of their individual attributes was calculated using these context-dependent models. The final output was a single CSV file with 29 columns, including identifiers,

individual demographics, professional attributes, company-level attributes, and several calculated diversity scores.

3.3 Phase Two: Data Modeling and Analysis

At this point, the generated dataset would need to be subjected to a strict analytic workflow in a Jupyter Notebook. The first step is cleaning and preparing the data. For example, the Board_Join_Date column was converted into a datetime object. Data consistency checks were made, and an integer transformation was performed on boolean-like columns. The most important step was transforming the Education_Level feature into an ordinal feature, Education_Rank, which would capture its inherent order.

Following all this, a full visual presentation of the data was accomplished through an Exploratory Data Analysis (EDA) exercise, which validated all patterns embedded during the data generation process. This understanding is crucial for the entire data landscape, i.e., prior to modeling. Libraries like Seaborn Plotly created a whole suite of visualizations within the EDA.

- **Distribution Analysis:** The male-female distribution distributions across the Industry_Sector, Board_Role, Education_Level, and, Company_Size were visualized through bar charts.
- **Continuous Variable Analysis:** The box plot and histogram deal with the study of Age distributions as per gender.
- **Correlation Analysis:** The heatmap of the correlation matrix was depicting that of the salient numerical features, where potential multicollinearity can be identified and relationships between variables better understood.
- **Geographic Analysis:** Box plot comparison of Female_Board_Percent by world Regions.

This EDA confirmed the synthetic dataset to reflect the biases as intended and hence laid a very strong foundation for steps that follow in modeling.

3.4 Phase Three: Feature Engineering

New engineered features from existing data were created to enhance model performance. The explicit signals were to be created better for the model to learn from. Different kinds of features were developed. Role-based features were generated by making binary flags for specific powerful roles like Is_CEO and Is_Chairperson. Demographic grouping comprised binning continuous variables like Age and Tenure into categorical groups. The individual was then positioned according to their board's average with respect to Age_vs_Board_Avg. Finally, the ten industry sectors were consolidated into broader categories like 'Technology' and 'Finance' to create an Industry_Group feature, which can help model learning more general patterns. Feature engineering forms an important part in making raw data possible for machine learning algorithms.

3.5 Phase Four: Modeling Pipeline Development

Finalizing an appropriate methodology would entail the building and implementation of a machine learning pipeline. At first, the target variable was defined. The Gender column was then converted to a binary target Gender_Binary, where 'Female' is coded 1 and all other

categories ('Male', 'Non-binary') as 0. This framed out the problem as a classification exercise of Female vs Non-female.

The dataset was then divided in the following way: 75% was set as training data, while the remaining 25% was used as the testing data. In both the sets, stratification on the target variable preserved the class distribution for their reliable evaluation on an imbalanced dataset. Subsequently, a systematic preprocessing pipeline was constructed using Scikit-learn's Pipeline and ColumnTransformer. This is a best practice, making consistency in and leakage prevention from data very efficient. The pipeline was configured to apply StandardScaler to all numerical and binary features, and all categorical features would be handled by OneHotEncoder.

The last step in data preparation was dimensionality reduction. The high-dimensionality feature space was created after the ones creating one-hot encoding. To manage this, PCA was applied to the preprocessed data. The number of components was chosen to retain 95% of the explained variance in the training set to achieve a balance between keeping information and dimension reduction. This resulted in a final set of 33 principal components. This reduced-dimensionality dataset was then used to train and evaluate the suite of machine learning models described in the evaluation chapter. This thorough and multi-phase methodology ensures that what is done in this study is not merely an application of a model but rather considered research from beginning to end.

4 Design Specification

This chapter outlines the architectural design and technical specification of the solution developed for this research. It describes the overall system architecture, from data generation to model deployment, and provides a specification for the core algorithm chosen for the final model. The design prioritizes modularity, reproducibility, and a clear separation of concerns between different functional layers of the system.

4.1 System Architecture

The system is designed as a modular, end-to-end pipeline that encompasses data creation, model training, and predictive deployment. This architecture ensures that each stage of the process is a self-contained unit. The architecture, depicted in Figure 1, can be broken down into three primary layers.

- **Data Layer:** This is the foundational layer responsible for creating the dataset. Its core component is the Synthetic Data Generation Script (datageneration.ipynb). This script programmatically generates the dataset.csv file, which serves as the static, canonical source of data for the entire project, ensuring reproducibility.
- **Modeling Layer:** This is the core analytical engine of the project, implemented in the Modeling and Training Notebook (modelling.ipynb). This notebook orchestrates the entire machine learning workflow, including data ingestion, preprocessing, dimensionality reduction with PCA, model training, evaluation, and hyperparameter tuning. The final output of this layer is a set of serialized model artifacts (.pkl files) that encapsulate the learned knowledge from the training process.

- **Application Layer:** This layer provides a practical interface to the trained model, exposing its predictive capabilities via a REST API. Its core component is a Flask Web Application (app.py). On startup, the application loads the saved model artifacts. It then exposes a /predict endpoint that accepts new data in JSON format, processes it through the saved prediction pipeline, and returns a JSON response containing the prediction and confidence scores.

This layered design ensures a logical flow of data and processes, from raw data creation to actionable, real-time predictions.

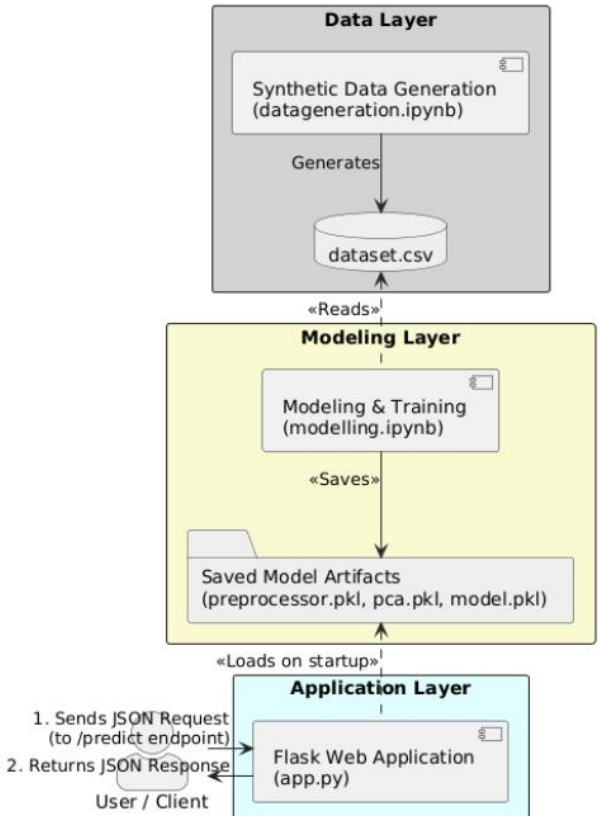

Figure 1: The end-to-end system architecture, illustrating the flow from synthetic data generation in the Data Layer, through model training and artifact creation in the Modeling Layer, to real-time inference via a web API in the Application Layer.

4.2 Algorithm Specification: XGBoost

The final model selected for this study is based on the XGBoost (eXtreme Gradient Boosting) algorithm, a highly efficient and effective implementation of the gradient boosting framework. Gradient Boosting is an ensemble learning technique that builds a strong predictive model by sequentially adding weak learner models, typically decision trees. Each new tree is trained to correct the errors made by the ensemble of all previous trees.

XGBoost enhances this concept with several key innovations. It takes a more regularized model formalism, which utilizes both L1 and L2 regularization to penalize complexity and overfitting. This becomes paramount with complicated, high-dimensional data. For optimization, XGBoost takes advantage of both the first-order and second-order derivatives of the loss function, thereby producing more accurate approximations and faster convergence time. In addition, the algorithm is highly engineered for, and optimized for, performance,

with capabilities like parallel processing, cache-aware access, out-of-core computation, which makes XGBoost exceptionally fast, and highly scalable. Finally, XGBoost possesses a built-in routine for missing values, which learns a default direction for missing values at each split. Collectively, these aspects made XGBoost the favorable option for this project based on its performance during the preliminary analysis.

5 Implementation

This chapter describes the operationalization of the research project, describing the development tools, technologies, and essential software components that were developed. The description gives a high-level overview of the development journey and the artifacts that were produced, but does not include raw code listings.

5.1 Tools and Technologies

The project was operationalized fly using open-source software, allowing access and reproducibility. The technology stack was chosen based on its popularity and creditability in the community of research data science.

- **Primary Programming Language:** Python (version 3.x) was utilized for all scripting and application development.
- **Core Data Science Libraries:** The project utilized, and depended primarily on Pandas for data manipulation, NumPy for numerical analysis, and Scikit-learn for a variety of usages including preprocessing, dimensionality reduction, model evaluation, and hyper-parameter optimization. Since the final classification model was XGBoost based, the XGBoost library was used (a separate library, not part of scikit-learn).
- **Data Visualization Libraries:** Matplotlib and Seaborn were used for creating static plots during exploratory analysis, while Plotly was used for generating interactive and publication-quality visualizations for the final report.
- **Development Environment:** All data generation and modeling tasks were conducted in Jupyter Notebooks hosted on the Google Colab platform, chosen for its ease of use and access to computational resources.
- **Web Application Framework:** Flask was used to build the lightweight and flexible REST API for model deployment.

5.2 Key Implemented Components

The implementation resulted in a set of distinct, functional software components that correspond to the architectural layers described in the design specification.

- **Synthetic Data Generation Script (datageneration.ipynb):** A standalone Jupyter Notebook containing the Python code responsible for creating the project's foundational dataset, dataset/dataset.csv.
- **Modeling and Analysis Notebook (modelling.ipynb):** A comprehensive Jupyter Notebook that encapsulates the entire machine learning workflow, from data loading and cleaning to final model analysis.

- **Saved Model Artifacts (models/ directory):** A directory created to store the serialized objects produced by the modeling notebook, allowing the trained intelligence to be decoupled from the training environment. This includes the saved preprocessor (preprocessor.pkl), the PCA transformer (pca_transformer.pkl), the final model (final_model.pkl), and files containing performance metrics and model information.
- **Web Application (app.py):** A Python script that uses the Flask framework to create a simple web server. It implements the application layer of the system, loading the saved models and providing API endpoints for health checks, schema information, and, most importantly, real-time predictions.

6 Evaluation

This chapter presents a comprehensive and critical evaluation of the machine learning models developed within the scope of this research. The central purpose of this evaluation is to rigorously assess the models' capacity to predict a board member's gender based on the synthetically generated profile of attributes, to meticulously analyze their performance characteristics using appropriate metrics, and to interpret the findings in the broader context of the research questions and the problem of gender disparity. The evaluation is systematically structured into a series of well-defined experiments, each designed to answer a specific question about the models' performance, culminating in a detailed discussion that synthesizes the results and their implications.

6.1 Evaluation Metrics

Appropriate metrics need to be evaluated first when any models project in this case, as this data set is imbalanced, and specific attention is required in this case. Because the target variable, Gender_Binary, is heavily skewed, it will be about 22.4% of 'female' (positive) class samples. In such a situation, an understanding derived on the basis of only one metric like accuracy can be very misleading. To add a more refined and broader understanding of model performance, some complementary metrics were used for the assessment:

- **Accuracy:** The number of correct predictions as a percentage of all predictions. It will be a fairly global measure of the overall performance, although it is not the metric that should be driving selection in this case.
- **Precision:** The number of positive predictions that were true positive. Checks reliability of predicting it positive.
- **Recall (Sensitivity):** The proportion of true positives that were correctly detected. Checks the ability of the model to find all positive instances.
- **F1-Score:** The harmonic mean of Precision and Recall. This is a very important metrics for imbalanced datasets since it provides a single score that balances the trade-off between precision and recall.
- **ROC AUC:** An aggregate measure of the model's ability to discriminate between positive and negative classes over all possible classification thresholds.

6.2 Experiment One: Baseline Model Comparison

The first of these tests was on the creation of a profile. The objective is to find out what is the average performance value and the best algorithm candidate for further optimization. This includes training and testing on a selected six well-known classification algorithms. The same

PCA-derived training data and the same held-out test data were used for all trained models. The models were the Logistic Regression, Random Forest, XGBoost, Gradient Boosting, K-nearest Neighbors (KNN), and standard Decision Tree.

The comparative performance of these models is detailed in Table 1, which provides the precise metric scores and training times. To better visualize the trade-offs between the models, these key performance metrics are also presented in Figure 2.

Table 1: Performance Comparison of Baseline Classification Models

Model	**CV Accuracy**	**Test Accuracy**	**Test Precision**	**Test Recall**	**Test F1-Score**	**Test AUC**	**Training Time (s)**
XGBoost	0.7869	0.7892	0.5652	0.2611	**0.3572**	0.7703	8.53
Logistic Regression	0.7932	0.7932	0.5913	0.2540	0.3553	0.7781	2.40
Decision Tree	0.7018	0.6991	0.3392	0.3599	0.3492	0.5785	17.25
K-Nearest Neighbors	0.7599	0.7569	0.4346	0.2777	0.3388	0.6708	27.09
Gradient Boosting	0.7933	0.7930	**0.6005**	0.2306	0.3332	**0.7810**	327.38
Random Forest	0.7846	0.7830	0.5425	0.2106	0.3035	0.7507	215.47

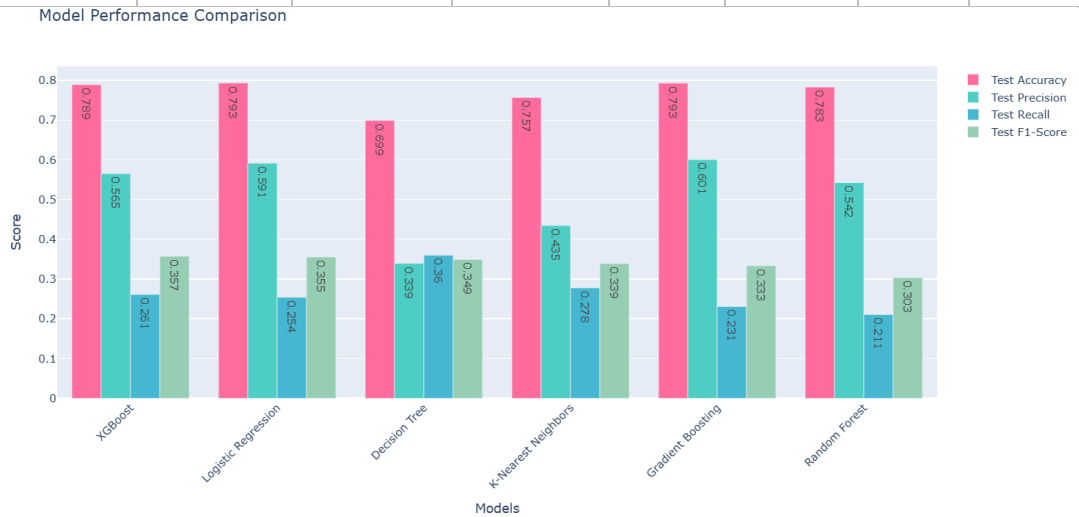

Figure 2: Visual Comparison of Model Performance Metrics (Test Accuracy, Precision, Recall, and F1-Score) for the six baseline classification models. The chart highlights the trade-offs between different performance aspects for each algorithm.

An analysis of these results reveals several key insights. As illustrated in Figure 2 and detailed in Table 1, the XGBoost model emerged as the top performer based on the F1-Score (0.3572), which is the most critical metric for this imbalanced problem. While other models, notably Gradient Boosting and Logistic Regression, achieved slightly higher accuracy or AUC scores, they did so at the expense of recall. The Gradient Boosting model, for instance, had the highest precision (0.6005) but one of the lowest recall scores (0.2306), indicating that while its positive predictions were reliable, it was extremely conservative and failed to identify a large proportion of the actual female board members.

Furthermore, the more complex ensemble models, Random Forest and Gradient Boosting, were computationally very expensive, with training times of several minutes. In contrast, XGBoost offered a superior balance between precision and recall while also being exceptionally efficient to train. Based on its superior F1-Score and its computational efficiency, XGBoost was unequivocally selected as the best model to carry forward for hyperparameter tuning.

6.3 Experiment Two: Hyperparameter Tuning of XGBoost

Having identified XGBoost as the ideal baseline model, the second experiment focused on optimization and hyperparameter tuning in order to maximize performance. The goal of hyperparameter tuning is to identify a combination of model parameters which will yield the best possible result on a given dataset. To accomplish this, the tuning process used the RandomizedSearchCV function from Scikit-learn to tune hyperparameters over a predefined grid of distributions for the parameters, using the F1-score as the metric to optimize. The tuning process identified the best parameter set and the model was retrained. The tuned model's final performance on the holdout test set included an accuracy score of 0.7914 (true positive rate: 1.0000, false positive rate: 0.5208), an F1-Score of 0.3596 (previous value was 0.3572), and an AUC of 0.7751 (previous value was 0.7690). Upon analyzing the results, it was clear that hyperparameter tuning yielded a small but measurable improvement to the F1-Score, from 0.3572 to 0.3596. Though the improvement seemed slight, this process was an important validation tool to demonstrate that the baseline model was already functioning near optimal capacity, and to provide confidence that the final model was indeed well-calibrated. This tuned XGBoost model was thus classified as the final_model for the next stages of in-depth analysis.

6.4 Experiment Three: In-Depth Analysis of the Final Model

To gain a deeper understanding of the final model's behavior, its strengths, and its critical weaknesses, a more detailed analysis of its predictions on the test set was conducted. To visualize the model's classification performance in detail, a confusion matrix was generated, as shown in Figure 3. This matrix provides a clear breakdown of the model's correct and incorrect predictions for each class.

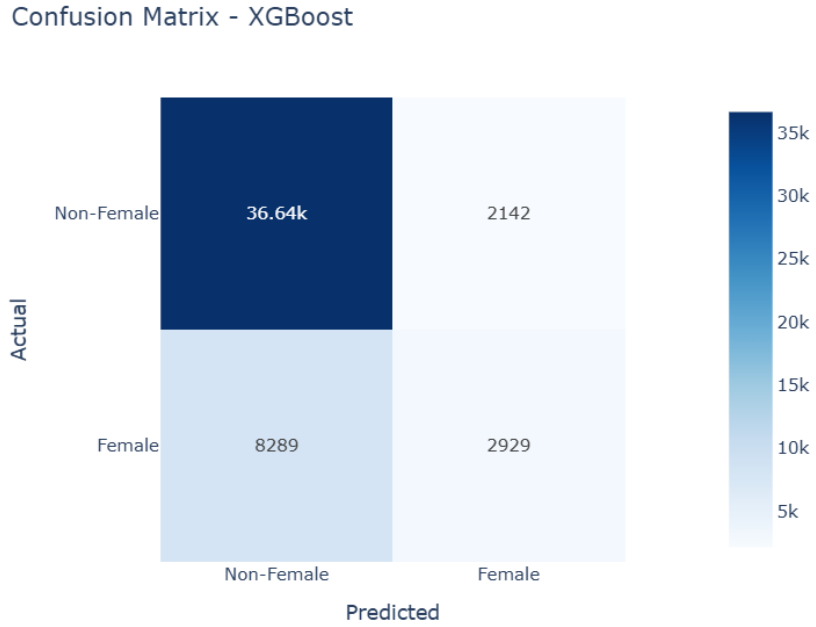

Figure 3: Confusion Matrix for the Final Tuned XGBoost Model on the test set. The matrix shows the counts of True Negatives (TN), False Positives (FP), False Negatives (FN), and True Positives (TP).

The matrix clearly shows that the model is very effective at identifying the majority class, with 36,640 True Negatives. However, the most striking feature of the matrix is the high count of False Negatives (8,289). This means the model failed to identify nearly three-quarters (74%) of the actual female board members present in the test set. This catastrophic failure is quantified by the extremely low recall of 0.26 for the 'Female' class. This result

unequivocally indicates that the model has learned a powerful bias towards the majority class. It has effectively learned the "rule" from the data that an arbitrary board member is highly likely to be non-female, and therefore it requires an overwhelming amount of evidence to predict 'Female', defaulting to 'Non-Female' in any case of ambiguity.
The model's precision for the 'Female' class, at 0.58, is more moderate. This means that when the model does venture a 'Female' prediction, it is correct 58% of the time. While this is significantly better than the baseline random chance of 22.4%, it also means that its positive predictions are incorrect 42% of the time. To complement the confusion matrix, an ROC curve was plotted to assess the model's overall discriminative ability across different thresholds, as seen in Figure 4.

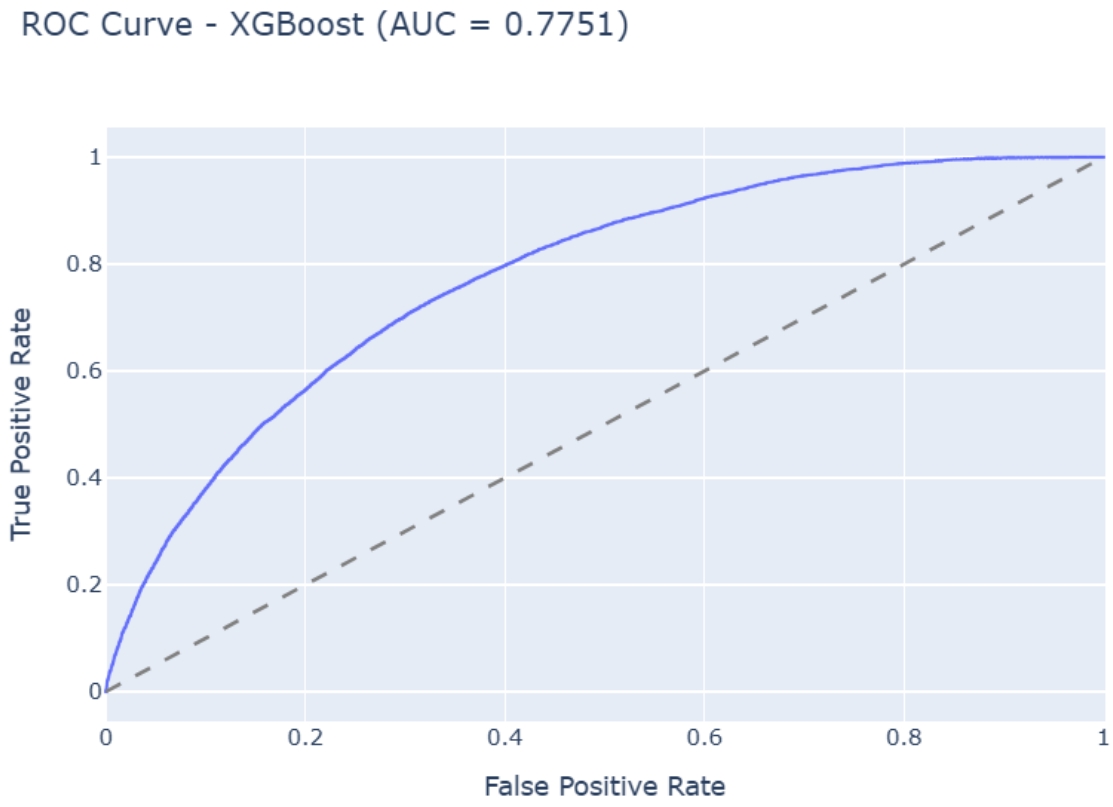

Figure 4: Receiver Operating Characteristic (ROC) Curve for the final XGBoost model. The Area Under the Curve (AUC) of 0.7751 indicates a moderate ability to distinguish between the 'Female' and 'Non-Female' classes.
The AUC score of 0.7751 suggests that the model has a fair to good ability to differentiate between the two classes, significantly better than random chance (an AUC of 0.5, represented by the dashed line). However, the shape of the curve, which rises steeply at first but then flattens out, reinforces the findings from the confusion matrix: the model can achieve a reasonable true positive rate only at the cost of a significant false positive rate.

6.5 Experiment Four: Feature Importance Analysis

The final experiment was an analysis of the feature importances from the final XGBoost model. It is important to note that because the model was trained on the data after the PCA transformation, the feature importances relate to these abstract principal components, not directly to the original business-level features. The relative importance of the top 15 PCA components is presented in Figure 5.

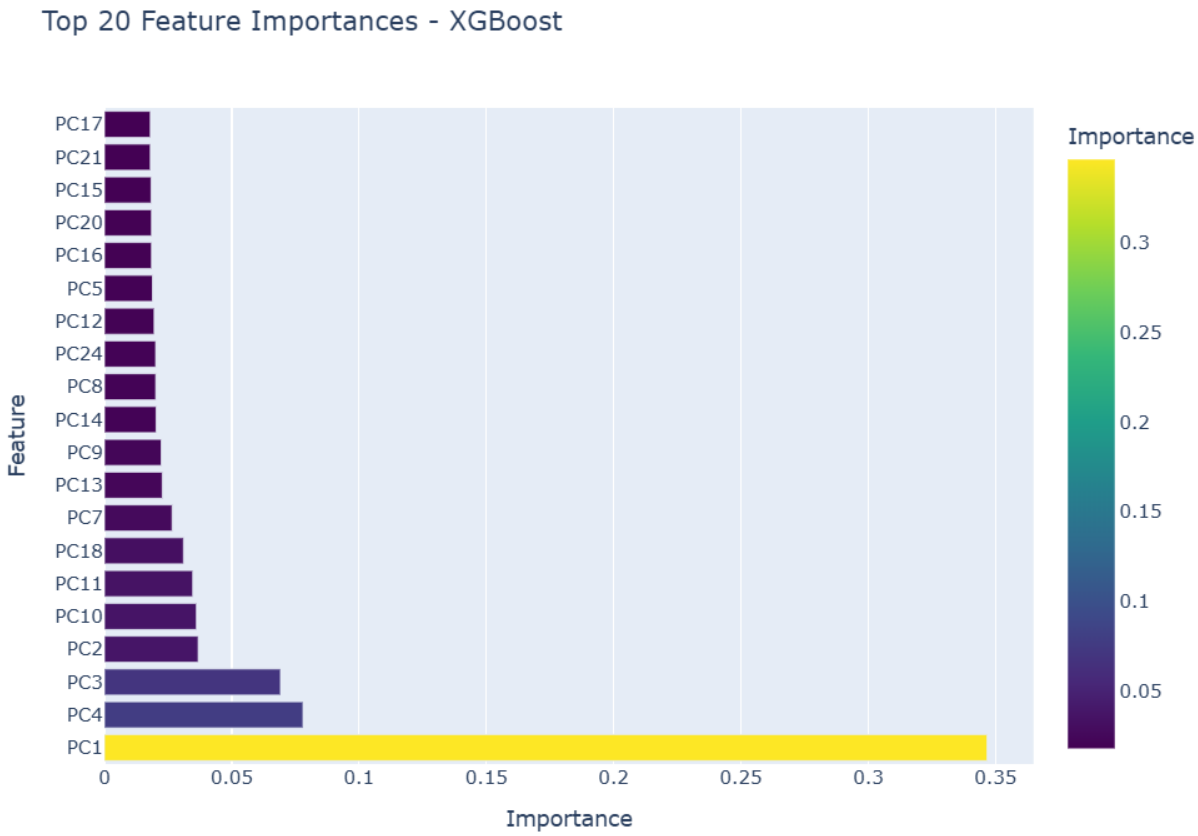

Figure 5: Relative Feature Importance of the Top 15 PCA Components for the final XGBoost model. The chart shows the normalized importance score for each principal component, highlighting their contribution to the model's predictions.

The most prominent result from this analysis is the overwhelming dominance of the first principal component (PC1), which accounts for over a third of the model's decision-making weight by itself. This suggests that the primary axis of variation within the preprocessed data—the direction in which the data is most spread out—is highly predictive of a board member's gender. While the abstract nature of PCA components makes direct interpretation challenging, it is highly probable that PC1 is a composite feature that is heavily influenced by the original features most directly related to gender diversity itself, such as the pre-calculated Female_Board_Percentage and the regional and industry-level biases that were explicitly hard-coded into the data. The subsequent components have a much smaller, though still significant, impact. The key limitation of this analysis is this lack of direct interpretability due to the PCA transformation.

6.6 Discussion

The comprehensive evaluation process yields several critical and multi-layered insights. On one level, the study successfully demonstrated that it is possible to develop a machine learning model that can predict a board member's gender with a respectable overall accuracy of nearly 80%. This confirms that the synthetically generated features contain a strong, coherent, and learnable signal that is systematically correlated with gender. On a deeper and more important level, however, the model's performance is profoundly flawed, and these flaws reveal more about the nature of the data and the challenges of algorithmic fairness than they do about creating a perfect classifier.

The most significant and sobering finding from the entire evaluation is the stark and dramatic performance gap between the majority ('Non-Female') and minority ('Female') classes. The model's abysmal recall of just 26% for the female class is a textbook and powerful illustration

of a classifier learning and amplifying the biases present in imbalanced data. The model effectively learned the statistical reality from the training data: that most board members are not female. It then operationalized this learning into a predictive strategy that is highly hesitant to ever predict the minority class, leading to a massive number of false negatives. This empirical result provides a concrete and quantitative validation for the serious concerns raised in the academic literature about the potential for AI systems, if not carefully designed and audited, to perpetuate and even exacerbate existing societal biases and inequities (Wu, 2023; Deepak et al., 2025).

The model's specific errors are also highly informative. The 2,142 false positives represent non-female individuals whose profiles, according to the model, are so similar to the typical female profile in the dataset that they are misclassified. A deeper analysis of these cases could reveal what characteristics define a "female-like" profile in the model's view. Conversely, and more critically, the 8,289 false negatives represent the female board members whose profiles are more aligned with the dominant non-female "archetype." These are the women who are, in a sense, "invisible" to the model because their features do not provide a strong enough signal to overcome the model's inherent bias towards the majority class.

In conclusion, the evaluation phase of this research has been highly successful, not because it produced a perfectly accurate model, but because it produced a model whose imperfections are profoundly instructive. The evaluation clearly demonstrates that while standard machine learning techniques can be very effective at identifying the complex patterns related to gender disparity, a naive application of these powerful tools on biased data will inevitably lead to a biased model. The true value of this evaluation is not in the model's raw predictive accuracy, but in its ability to serve as a high-fidelity mirror to the systemic imbalances that were embedded within the data, providing a stark, quantitative, and unambiguous illustration of the profound challenges of achieving fair and equitable gender representation in corporate leadership.

7 Conclusion and Future Work

Thus, this last chapter summarizes the insights it has gathered through conducting research, revisits the research questions and objectives, discusses the findings and their importance, recognizes the limitations of the study, and indicates good future research perspectives.

7.1 Conclusion

This study is aimed at creating a computational framework for studying the gender gap in corporate leadership, an area that has been hobbled from deep analysis due to the unavailability of public data at a granularity level. This objective was well achieved, making a realistic synthetic data set, developing a robust machine learning pipeline, being able to test different predictive models, and analyzing the final model in order to extract insights. The project proved that using a machine-learning model, it is possible to predict a board member's gender with a 79.14% accuracy level depending on the person's professional and corporate profile. The major one indicates, however, substantial bias on the part of the model, evidenced by a very low recall of 26.11% for the female class, which indicates that this measure is almost always failing to detect female members on boards.

Implications from the study therefore include both methodological and practical ones. Methodologically, it establishes a repeatable argument in the application of synthetic data for the examination of the difficult social science questions arising from the absence of real-world data. For practitioners, this boils down to a forceful, quantitative demonstration of how

systemic bias creeps into data-driven systems. In fact, it warns against applying models trained on biased data without critical evaluation in the context of fairness.

Limitations of the Study

There are several limitations that must be acknowledged in this research.

- **Synthetic Data:** The main limitation is that this research uses synthetic data which, while it was thoroughly designed is still a simulation and cannot capture all the complexities of the real-world.
- **Interpretability:** While PCA was an effective methodology for dimensionality reduction in this research, it came at the cost of directly interpreting features.
- **Binary Classification:** The model was simplified to a binary "Female" vs. "Non-Female" classification task due to the small sample size of the "Non-binary" category.

7.2 Future Work

This research opens up several promising avenues for future work.

- **Validation with Real-World Data:** The most immediate next step would be to apply the developed framework to a real-world, albeit smaller or more localized, dataset to validate the findings.
- **Advanced Techniques for Imbalance:** To address the poor recall, future work should explore techniques specifically designed for imbalanced classification, such as SMOTE for oversampling or cost-sensitive learning.
- **Enhancing Interpretability:** A crucial extension would be to focus on model explainability by training models without PCA and applying eXplainable AI (XAI) frameworks like SHAP or LIME to get precise, feature-level explanations.
- **Expanding the Data Simulation:** The data generation script itself could be enhanced to include more dynamic factors, such as simulating career trajectories, board interlocks, or the impact of national diversity quotas over time.

References

[1] Afeltra, G., Alerasoul, A. and Usman, B., 2022. Board of directors and corporate social reporting: A systematic literature network analysis. Accounting in Europe, 19(1), pp.48-77.

[2] Behlau, H., Wobst, J. and Lueg, R., 2024. Measuring board diversity: A systematic literature review of data sources, constructs, pitfalls, and suggestions for future research. Corporate social responsibility and environmental management, 31(2), pp.977-992.

[3] Belingheri, P., Chiarello, F., Fronzetti Colladon, A. and Rovelli, P., 2021. Twenty years of gender equality research: A scoping review based on a new semantic indicator. Plos one, 16(9), p.e0256474.

[4] de Enrique Arnau, L. and Pinillos-Costa, M.J., 2024. Board of directors and business transformation: a bibliometric analysis. European Journal of Management and Business Economics, 33(2), pp.212-236.

[5] Deepak, D., Shenbagavalli, T., Rahul, V. and Manjunath, B., 2025. Navigating the Future on a Comprehensive Review of Women in AI-Powered Management. Public Sector and Workforce Management in the Digital Age, pp.81-96.

[6] Doo, F.X. and McGinty, G.B., 2023. Building diversity, equity, and inclusion within radiology artificial intelligence: representation matters, from data to the workforce. Journal of the American College of Radiology, 20(9), pp.852-856.

[7] Gander, M. and Sharafizad, F., 2025. Progressing gender equity in senior leadership: a systematic literature review. Gender in Management: An International Journal, 40(2), pp.352-369.
[8] Górska, M. and Burlakova, I., 2025. THE ROLE OF WOMEN'S LEADERSHIP IN BUSINESS: CHALLENGES AND PROSPECTS. Economics, Finance and Management Review, (1 (21)), pp.116-129.
[9] Kalusivalingam, A.K., Sharma, A., Patel, N. and Singh, V., 2022. Enhancing corporate governance and compliance through ai: Implementing natural language processing and machine learning algorithms. International Journal of AI and ML, 3(9).
[10] Koziol, K., Schmitz, M. and Bort, S., 2025. Gender differences in entrepreneurial equity financing—a systematic literature review. Small Business Economics, pp.1-56.
[11] Li, K., 2025. Big Data and Machine Learning in ESG Research. Asia-Pacific Journal of Financial Studies, 54(1), pp.6-21.
[12] Mehnaz, L. and Yang, C., 2025. Women in accounting research: a review of gender diversity, equity and inclusion. Meditari Accountancy Research, 33(7), pp.30-59.
[13] Off, R. and Baltes, G.H., 2021, June. Glimpse on Women in Entrepreneurial Context A Machine Learning based Literature Review. In 2021 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC) (pp. 1-9). IEEE.
[14] Paspuel, J.H., Vera-Ortega, R., Plaza, A.M. and Correa, B.H.O., 2025. Global Gender Inequality in Industry: A Systematic Review with Neutrosophic Analysis. Neutrosophic Sets and Systems, 81, pp.79-88.
[15] Prusak, B. and Wacławek, Ł., 2023. Gender Disparity in the FinTech Sector: Systematic Literature Review. Digital Transformation and the Economics of Banking, pp.99-112.
[16] Sklavos, G., Theodossiou, G., Papanikolaou, Z., Karelakis, C. and Ragazou, K., 2024. Environmental, social, and governance-based artificial intelligence governance: Digitalizing firms' leadership and human resources management. Sustainability, 16(16), p.7154.
[17] Smith, S.G. and Sinkford, J.C., 2022. Gender equality in the 21st century: Overcoming barriers to women's leadership in global health. Journal of Dental Education, 86(9), pp.1144-1173.
[18] Wu, Y., 2023. Data Governance and Human Rights: An Algorithm Discrimination Literature Review and Bibliometric Analysis. Journal of Humanities, Arts and Social Science, 7(1).
[19] Xue, R., Ong, T.S. and Demir, E., 2024. Do CEO and chairman characteristics affect green innovation? Evidence from a comparative analysis of machine learning models. Environment, Development and Sustainability, pp.1-28.
[20] Yao, T., 2023. Antecedents of top management team and board gender diversity: A review and an agenda for research. Corporate Governance: An International Review, 31(1).