Identifying, Measuring, and Addressing Algorithmic Bias in AI Admission Systems for Graduate Education

Ananya Prakash

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Mohammed S. Seyam, Chair
Dwayne C. Brown
Mohammed F. Farghally

April 14, 2025
Blacksburg, Virginia

Keywords: Graduate Education, Admissions, Computer Science, Machine Learning, Algorithmic Bias

Identifying, Measuring, and Addressing Algorithmic Bias in AI Admission Systems for Graduate Education

Ananya Prakash

(ABSTRACT)

The number of graduate students has been increasing rapidly to meet industry demands, with over 200% increase in competitive fields like computer science (CS) in the past decade. Several universities have adopted AI in their admissions processes for various tasks such as evaluating transcripts, extracting important information from essays, and scoring applications. While AI can greatly increase the efficiency of processing a large volume of applications, it is prone to data and algorithmic bias, which can lead to unfair outcomes for underprivileged subgroups among applicants. Recent changes in legislation such as the ban of affirmative action by the U.S. Supreme Court make it increasingly relevant to study the demographic composition of admitted students and ensure that we develop fair machine learning systems. We present a comprehensive two-phase methodology for detecting and mitigating algorithmic bias. Through analysis of graduate admissions data of the Computer Science department of a large R1 university, we found significant post-ban demographic shifts, including decreased applications from underrepresented groups and a 66% increase in applicants declining to report race post the affirmative action ban. Our preemptive bias detection phase includes exploratory data analysis, clustering, and subgroup discovery to identify both independent and intersectional sources of bias, revealing significant disparities based on citizenship status, gender and race. We then developed and evaluated a neural network model using fairness metrics, discovering substantial bias amplification for gender

and citizenship status. Our fairness evaluation and bias correction phase demonstrated that preprocessing and postprocessing mitigation techniques could significantly improve fairness metrics, though with varying effectiveness across different protected attributes. SHAP analysis confirmed that while academic metrics like GPA remained the strongest predictors, demographic features substantially influenced model decisions even after mitigation. This work provides a systematic framework for institutions seeking to implement fair AI admissions systems while navigating new legal constraints on affirmative action, emphasizing the importance of proactive bias detection and mitigation to maintain diversity in higher education.

Identifying, Measuring, and Addressing Algorithmic Bias in AI Admission Systems for Graduate Education

Ananya Prakash

(GENERAL AUDIENCE ABSTRACT)

This study examines how artificial intelligence (AI) systems used in graduate computer science admissions might perpetuate bias, especially following the recent ban on affirmative action. We developed a two-step approach to detect and mitigate bias in these systems. Our analysis revealed concerning trends: fewer applications from underrepresented groups and many more applicants choosing not to report their race after the ban. We discovered that admission prediction models significantly favored certain group,particularly US citizens, and sometimes amplified existing biases dramatically. While academic factors like GPA were the strongest predictors of admission, demographic characteristics still heavily influenced outcomes. We tested different bias correction techniques and found they could improve fairness, though no single approach worked perfectly for all groups. This research provides universities with practical tools to identify and reduce bias in AI admissions systems, helping maintain diversity in higher education while complying with new legal restrictions on affirmative action.

Acknowledgments

Several people have shaped my work and guided me throughout the journey of completing my thesis. With limited research experience and a lot of ambition, I took on this project with no dataset during a time of transition in the education landscape. I would first and foremost like to thank my advisor, Dr. Mohammed Seyam, whose guidance, knowledge and encouragement has helped me grow immensely as a researcher and an individual. Though my research lies in the complex interdisciplinary area of AI, education and society, Dr Seyam always helped me focus on the objective and navigate through the challenges associated with research in this field. Outside of being my research advisor, he also encouraged me to attend academic conferences where I could interact with other scholars and learn from their work. Admissions is a highly subjective topic and varies across universities and departments, and I would not have comprehended it without the various conversations I had with staff and faculty from the different universities. I'm grateful to Janice Austin, Kacy Lawrence, Catherine Cotrupi from the graduate school in Virginia Tech who shared their knowledge and expertise on the process of graduate admissions and made a significant difference in my research. I would also like to thank Danette Gomez from Radford University and Renee Cummings from University of Virginia who helped me understand the bottlenecks in admissions and how algorithms may be used in addressing these. I am grateful to have a wonderful thesis committee comprising of Dr Mohammed Seyam, Dr Chris Brown and Dr Mohammed Farghally, whose advice and feedback on my work was critical to improving my research. Finally, I would like to thank my family, friends and peers at Virginia Tech that have motivated me to continue my research and supported me through challenging times. Everything I have accomplished is a result of all the people in my life that have always encouraged me to reach for the stars.

Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence

AIF360 AI Fairness 360 Toolkit by IBM

CS Computer Science

EDA Exploratory data analysis

FIF Fairness Influence Function

FPR False Positive Rate

GPA Grade Point Average

HDBScan Hierarchical Density-based Spatial Clustering

LIME Local Interpretable Model-Agnostic Explanations

MDFA Multi-Differential Fairness Auditor

ML Machine Learning

MLP Multilayer Perceptron

MS Master of Science

NCES National Center for Education Statistics

NN Neural Network

Phd Doctor of Philosophy

SHAP Shapley Additive Explanations

SVM Support Vector Machine

TPR True Positive Rate

URM Under Represented Minorities

Chapter 1

Introduction

Over the last few decades, jobs in the technology industry have become far more competitive, with more students earning master's and doctorate level degrees for jobs motivated by nearly a 20% higher salary than bachelor's degree holders as per the U.S. Bureau of Labor Statistics [20]. According to the National Center for Education Statistics (NCES) [37], the number of graduates with a master's degree has grown from 14,990 in 2000 to 51,338 in 2019, a 242% increase over two decades. Similarly, the number of graduates with a doctorate has grown from 779 to 2790 in the same period, an increase of 258%. While this increase in pursuits of postgraduate degrees in the field reflects the rapid growth of the industry, universities still grapple with the task of evaluating increasingly large volumes of applications.

1.1 Demographic Disparities in Graduate CS Education

The rapid development of the technology industry led to an increased number of graduate degree holders yet the diversity among these graduates has not shown comparable growth. For instance, the male-to-female ratio among master's graduates has remained nearly constant in the last decade at 2:1 with 66% males and 33% females [37]. This supports Cuny and Aspray's observation that fewer women enroll in graduate computer science (CS) programs, with numbers dropping from master's to doctorate levels [14]. Despite the time gap between

the 2002 study [14] and the NCES report from 2023 [37], the findings align, emphasizing the persistence of the issue.

Research shows that the diversity gap is further exacerbated for graduate degrees because of issues like lower GPAs or poor undergraduate experiences, low access to resources for standardized tests for underrepresented communities, and financial limitations [15]. Studies by the NCES also reveal that students of underrepresented minorities (URM) including Hispanic, Black, and American Indian or Alaskan Native students constitute a far lower percentage of graduate students compared to White and Asian students in science and engineering fields in the United States [22]. With the U.S. Supreme Court's decision to ban affirmative action in 2023 [54], a policy that earlier allowed universities to consider race as a criterion in the admissions process, there may be an increased threat to diversity, given its already deficit state in CS graduate education. With the minimal change in diversity over the last decade and the ban on affirmative action practices, it is imperative to find a solution to the challenge of diversity.

1.2 Optimizing Admissions Processing

Several large universities adopt a holistic review approach for admissions that is time-consuming and relies heavily on skilled human reviewers. The average time taken for each full review could vary between 10-30 minutes based on the skills of the reviewer [52]. A survey conducted by Intelligent in 2023, an education magazine [5], reported that 50% of 400 surveyed institutions already used Artificial Intelligence (AI) in their admissions process, and an additional 30% planned to do so in 2024. AI gives universities the advantage of increased efficiency, allowing them to focus their limited resources on other critical tasks like selecting students for financial aid and scholarships [13]. Therefore, it is essential to

innovate AI systems that assist in the admissions process and increase efficiency, but such systems could be susceptible to producing biased outcomes.

Despite AI's potential to optimize efficiency and reduce the workload of human reviewers, it also presents new risks. One such risk is bias, a phenomenon exhibited in AI systems that can amplify and perpetuate undesirable negative effects on individuals, organizations, and society [45]. The survey [5] also found that universities typically use AI to review letters of recommendation, transcripts, and essays and to communicate admission decisions to applicants. Though AI may be causing little to no harm in analyzing objective criteria like transcripts, previous studies have highlighted its ability to learn sensitive attributes like gender from letters of recommendation [55] and gender and household income from personal essays [7], which could potentially induce bias in the admissions process. To fully optimize the admissions process, machine learning systems may be employed to make final decisions on applications, as done by nearly 87% of the survey respondents [5]. However, institutions must pay careful attention to the details of how their model is trained, to reduce algorithmic bias in their system.

1.3 Research Objectives

While most large public universities already employ a rubric or some algorithms to make admissions decisions [13], machine learning models might learn unintended patterns from historical data that further perpetuate biases. This study aims to uncover the potential biases that a machine learning model trained on historical data may infer in its training phase, in the form of data bias from the historical data and the algorithmic bias from the model architecture and training parameters. The rapid adoption of machine learning for admission reviews reported in [5] highlights the importance of carefully analyzing what the

machine learning model may infer during training to prevent any unprecedented biases from being perpetuated by the model. We examine and evaluate the potential for such data biases, and experimentally determine the effectiveness of various bias mitigation approaches for data and algorithmic bias, by investigating the following research questions:

RQ 1: What are the independent features identified in the data that a machine learning model may infer bias from?

RQ 2: What are the intersectional sensitive attributes in the data that could induce bias in the model?

RQ 3: How can we evaluate and mitigate bias in a machine learning model for admission prediction?

RQ 4: How do we determine the underlying influence of demographic attributes on the model's decision-making?

RQ 5: How has the affirmative action ban impacted demographic diversity in graduate computer science education?

By studying these questions, this work aims to demonstrate how potential inferred data biases can be preemptively discovered before applying machine learning models to automate admission and how we can evaluate and mitigate bias using fairness toolkits and explainability methods.

1.4 Contributions of the Study

While prior works mentioned in 2.2 have experimented with ML models for admissions predictions, few have applied bias detection and mitigation approaches and evaluated their

effectiveness on the model. Some studies in the AI for admissions domain used explainability methods to understand model decision-making and assume bias, however, nearly all of them did not use fairness metrics to evaluate existing bias and assess the change in those metrics by applying mitigation techniques. This study presents a comprehensive two-phase methodology illustrated in Figures 3.1 and 3.2 to detect bias, apply mitigation strategies and evaluate their effectiveness. In Phase 1 (Preemptive Bias Detection), we begin with exploratory data analysis of the graduate admissions dataset, conducting detailed distribution and trend analysis (RQ5) while employing clustering techniques to identify potential demographic subsets. We then apply subgroup discovery to detect intersectional feature bias (RQ2) and random forest analysis to evaluate independent feature importance (RQ1). Phase 2 builds upon these findings through model development and fairness evaluation using the IBM AIF360 toolkit [8]. We implement both pre-processing and post-processing bias mitigation approaches (RQ3) while utilizing explainability tools to interpret model behavior and identify demographic features with the strongest influence on the model's decision-making (RQ4). This systematic approach allows us to both understand the potential biases that may arise from automating graduate computer science admissions with AI and develop effective strategies for mitigating algorithmic bias in the admissions process.

While we focus on bias in the context of graduate admissions for computer science, these methods can be extended to the admission processes of various programs and universities. By performing similar systematic bias discovery and mitigation methods, universities may develop machine learning solutions that improve the efficiency of admissions reviews, decrease the possibility of biased results, and encourage diversity in graduate computer science education. Thus, this study intends to present opportunities for universities to increase the efficiency of the admissions review process while minimizing potential harm to diversity emerging from the application of machine learning systems for decision-making in the

context of graduate admissions.

1.5 Thesis Roadmap

In chapter 2, we discuss the potential sources of bias in the admissions review pipeline, prior research in using AI for admissions, and explore in detail the various existing approaches for bias detection and correction. Chapter 4 presents the two-phased framework developed in this study for systematic detection of bias in historical admissions data and a fairness evaluation and mitigation pipeline. We also elaborate on some of the techniques used in the study such as subgroup discovery, fairness evaluation metrics and the bias mitigation approach. In chapter 5, we share our experiments and their implementation, including exploratory data analysis, model development, and fairness evaluation, the results of which are discussed in section 6. The discussion of our experiments and results are detailed in chapter 7, with some limitations and future work outlined in chapter 8. Finally, we conclude the findings of our study in chapter 9.

Chapter 2

Background and Related Work

To understand the existing methods and challenges associated with applying Machine Learning models in the context of admissions decision-making, we reviewed several prior works detailed below.

2.1 Admission Decision Pipeline and Sources of Bias

The admission process in U.S. universities typically requires applicants to submit various materials for evaluation such as transcripts, personal essays, multiple letters of recommendation, standardized test scores, and additional materials like extra-curricular certificates. When applying to graduate programs, applicants often submit additional essays that help the university understand their qualities, experience, and beliefs. These include essays on leadership, academic research, community service, and personal and professional ethics. Therefore, the data consists of numerical features such as standardized examination scores and Grade Point Averages (GPA), along with textual data from the essays and letters of recommendation. Applications also collect personal information including but not limited to the applicant's name, address, gender, and ethnicity. Among the range of application materials submitted, many of them could inform bias. Figure 2.1 details the potential stages in the admissions pipeline where bias could emerge and where AI is currently used as per the Intelligent survey [5].

In the context of university admissions, features like gender and ethnicity are usually examined for bias, as done by Kahlor et al. [24]. Contrarily, several existing studies on machine learning systems for admissions also seem to exclude demographic features to remove bias [34, 41]. This could be due to these studies simulating the admission process in California, where Proposition 209[1] prohibits the use of gender and ethnicity for admission reviews. In the GRADE system [52], the authors note that gender and ethnicity were assigned zero weight when passed to the model as features, concluding that demographic features do not contribute to the model's decisions. Another study analyzing the effect of the test-optional policy on admission decisions [11] examined bias along the features of gender, ethnicity, and first-generation applicants.

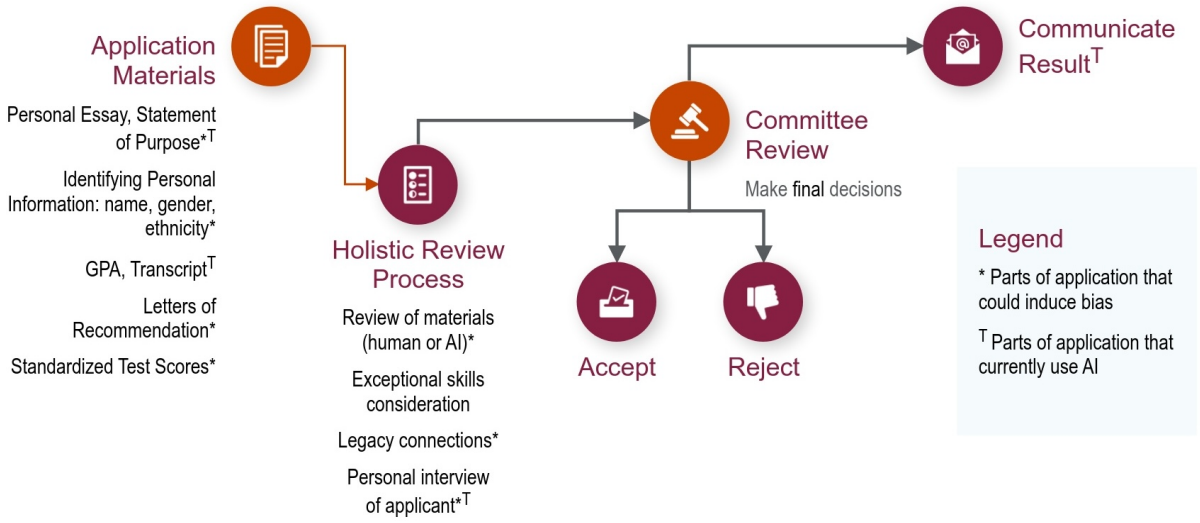


Figure 2.1: Admission Decision Pipeline

Though gender and ethnicity are the most commonly considered sensitive features when examining bias, other features such as nationality, first-generation students and median household income may also induce bias. Another occurrence of bias that is often neglected is intersectional feature bias of observations that belong to more than one protected group,

[1] California Proposition 209: https://lao.ca.gov/ballot/1996/prop209_11_1996.html

i.e., as a combination of multiple sensitive attributes [50]. This can be identified by scanning the dataset for subgroups with increased bias [23]. The holistic review process involves various application materials, with one of the primary mediums for students to share their narratives being essays. However, these essays often contain personal stories that may reveal demographic information related to an applicant as found in a study [7], creating a potential for bias. Similarly, sensitive attributes of the applicant may also be inadvertently extracted from letters of recommendation [55], proving to be another potential source of bias.

2.2 Prior Work on AI in Admissions

An important and necessary precursor to identifying and mitigating bias in the admission process is understanding the different ways in which AI has been previously applied to tackle admissions. While numerous studies have examined the feasibility of using AI in the admission process, there is limited structured work that compares and categorizes the various applications. To effectively describe the different applications, we classified these systems into two main categories as shown in Figure 2.2:

- AI-predicted decision-making: We refer to AI-predicted decision-making as the process where an AI algorithm is directly used to predict the admission decision for a given observation containing an applicant's information.

- AI-assisted decision-making: We define AI-assisted decision-making as the process where AI is used in the admission review pipeline to aid human evaluators. This involves making more information available for decision-makers, such as by validating essay scores or extracting and comparing transcript information.

Several studies have explored the domain of AI-predicted decision-making for admissions,

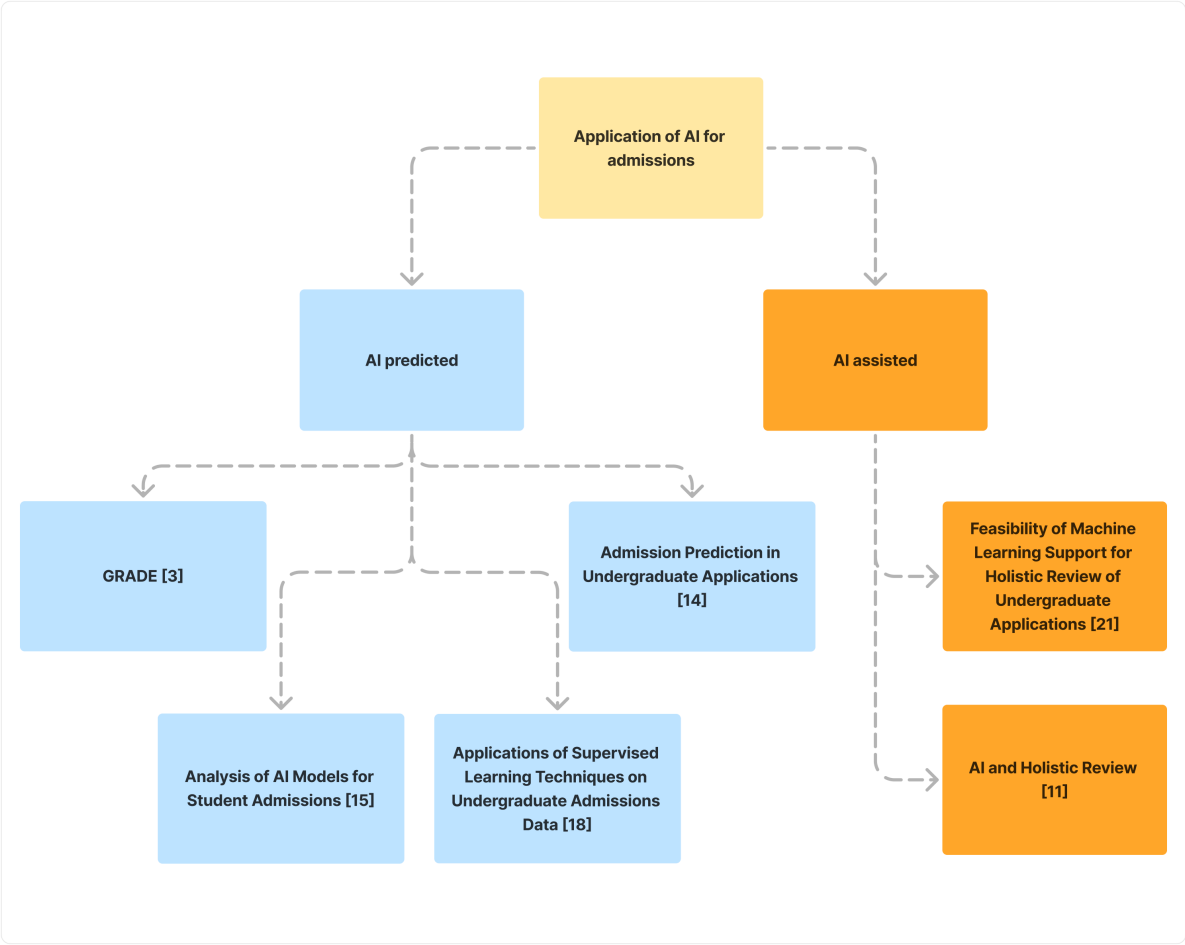


Figure 2.2: Prior research on AI for admission review

such as the GRADE system [52], deep learning algorithms to predict undergraduate admissions [41], and a Multilayer Perceptron(MLP) and Support Vector Machine(SVM) implementation [29]. Another study [11] developed an AI model to predict admission decisions with the test-optional policy. [41] and [29] developed a classifier trained on historical data to determine admission offers and used feature selection to identify feature contributions to the classifier's output. Both these works attempted to identify which features contribute more to the decisions, with the first study directly making inferences from learned weights of the classifier and the second study using LIME [42].

While most researchers argue that using AI for admission prediction potentially reduces personal biases from human evaluators' decisions, training on historical data could also produce biased outcomes and lead to a vicious feedback loop between the data and algorithm [30]. A few studies have also applied AI to create systems that aid decision-making in admissions. In an experimental study to increase the efficiency of the holistic review process while sensitizing reviewers, Alvero et al. [7] used AI to extract hidden sensitive attributes in personal essays. Another interesting use case of AI to support holistic review is its application in validating review scores assigned by application reviewers [34]. Both these studies demonstrate the possibility of AI systems to aid admissions decision-making but require further research and experimentation to prove their effectiveness. Therefore, AI can be applied to directly make admissions decisions or as a supporting tool to aid human reviewers in decision-making. Both these applications offer benefits but must be implemented carefully to ensure that they do not deliver biased outputs.

2.3 Tackling Bias in Machine Learning Systems

There have been increasing efforts in academia and industry towards tackling bias in the recent past. While previous works discuss the mitigation of bias briefly along the lines of pre-processing data, selecting fairness optimized models and post-processing decisions to achieve parity [18] and suggest the use of tools to balance the trade-off between accuracy and model fairness [30], there is a lack of comprehensive documentation of the various mitigation approaches. While bias is a largely complex problem and has multiple ways of emerging in data and algorithms, we broadly categorize the attempts to tackle bias as

- Bias detection and;
- Bias correction.

We define bias detection as the use of various algorithms and techniques to identify the presence of bias or attributes that induce bias in decision-making systems. We then define bias correction as the phenomenon of utilizing algorithmic or systematic techniques to alleviate or eliminate bias exhibited in decision making.

2.3.1 Bias Detection

A crucial step to tackling bias is its detection in a given model. Bias detection strategies can vary based on the type of bias and can also differ for each type of data considered. In a study that aimed to reduce bias in order to produce similar decision outcomes for people with dissimilar demographic features, Neda suggested determining features with privileged values that cause inadvertent bias [35]. In another study, Neda also argues that bias in algorithms for holistic review could be controlled by performing feature analysis and advocates using explainability tools to understand the decision-making process [34]. The different approaches to identifying bias revolve around isolating bias-inducing features based on importance or correlation, and the use of explainability tools to identify features that influence the decision outcome. Ghosh et al. [19] and Wamburu et al. [49] have also proposed alternate methods which involve identifying a subset of features that may be causing bias.

Explainability tools to detect bias

A popular approach to discovering bias in machine learning algorithms is the use of explainability tools. For example, [41] used the Local Interpretable Model-Agnostic Explanations (LIME) method [42] to understand which features contributed with greater magnitude either positively or negatively to the admissions decision of undergraduate applicants in California. While developing a bias mitigation pipeline, Kaya et al. [46] used the Shapley Additive Explanations tool (SHAP) [28], to compute feature importance for sensitive attributes to

infer bias in the decision-making algorithms. Both these tools provide insights into feature influence on the model outcomes and can be used to interpret the significance placed on demographic and non-demographic features, which can provide insights on model bias.

Feature subsets and intersectional bias

Bias need not be specifically attributed only to independent sensitive features but may be induced by a combination or subset of features. [49] proposed MDScan- a method to discover bias patterns in data without assuming predetermined attributes such as race or gender to bias inducing. When evaluated on the Stanford Open Police Search dataset, they found that Black and Hispanic males of specific age groups were searched much more frequently than other subgroups, indicating the presence of bias over the attribute race. Race was not presumed to cause bias, yet it appeared in many highly deviating subsets, showing its significant impact on police search bias. Intersectional feature bias can be addressed through subgroup discovery, a statistical method that identifies population segments with disproportionate probabilities of specific target outcomes [23].

Similar to the concept [49] that a subset of features may induce or reduce bias, [19] also studied the effect on bias by individual features and intersecting features through bias decomposition using the Fairness Influence Function (FIF). They proposed the fairXplainer algorithm, which captures FIFs for individual and intersectional features, and experimentally showed that it outperforms methods like SHAP [28]. This study revealed that a subset of features, or intersectional features, have significant impacts on bias which are not apparent when these features are considered individually.

Proxy features causing bias

Bias is often attributed to sensitive features like race and gender, but may also be present in proxy features or strongly correlated features. Gitiaux and Rangwala [21] demonstrated

experimentally how a discrimination assessment approach termed Multi-Differential Fairness Auditor (MDFA) is effective in identifying characteristics of minority groups susceptible to discrimination in black-box classification models. The auditor balances feature distribution across sensitive attributes and extracts correlations between the sensitive attributes and classifier outcome. When evaluated on the COMPAS dataset, the study found that a subgroup of African American individuals having minimal criminal record were three times more likely to be classified as high risk of recidivism. Though the classifiers were adjusted for aggregate fairness measures, the MDFA approach revealed discrimination against smaller sub-populations across the Adult, German and Crime datasets.

In cases where demographic data is not directly available, Wang and Singh extract hidden sensitive feature proxies to approximate demographic features in training data and propose fixing methods to reduce the dependency of the prediction algorithm on the sensitive feature proxies [51]. They evaluated this approach by experimenting with two Twitter related datasets, one on gun ownership and the other on COVID-19 vaccination, and found that with their standard fixing approach, the classifiers had an improved fairness score with a slight performance trade off.

2.3.2 Bias Correction

Once bias has been detected and its source has been identified, there are various approaches to mitigate the effects of bias on decision outcomes. These range from simply excluding bias-inducing features from model training to muting and re-balancing feature weights to reduce the effects of bias. While there are limited attempts to correct bias in the context of university admissions, we broaden the scope of academic works to those that could potentially be applied effectively to this context.

Exclusion of sensitive features

In the context of university admissions, a popular method has been the exclusion of sensitive features such as gender and ethnicity. This could be due to the belief that exclusion of sensitive attributes leads to a more unbiased evaluation, and could also be attributed to laws such as the California Proposition 209, which prohibits the use of race, ethnicity, or sex as criteria for selection in public education. For example, in their interpretable deep learning approach to undergraduate admission prediction, [41] eliminated all demographic features and focused solely on academic features to evaluate student applications. While this seems like a reasonable approach and aligns with the California Proposition 209, research shows that there may still be intersectional features or proxy features causing bias.

Muting of sensitive features

[46] proposed a bias-mitigation approach – ProxyMute for the context of job applications, that detects proxy features of a given sensitive attribute using global feature attribution-based explainability. Sensitive feature contributions to decision outcomes are identified using SHAP [28], and then passed through a bias-optimization module that selects the most influential features to mute. This method was experimentally proven to be effective in reducing bias while maintaining model performance on the datasets FairCVdb [38] and ChaLearn LAP-FI [16].

Tackling bias through NLP

Another source of bias in admissions data is found to be in the essays written by applicants [7]. Textual data is likely to contain patterns from which sensitive demographic information can be inferred and could influence decision-making involuntarily. A potential approach to tackling this would be through debiasing of word embeddings, a popular method in the NLP

domain. Bolukbasi et al. [10] experimented with two types of debiasing with respect to gender on the W2Vnews dataset, hard-debiasing and soft-debiasing, both of which apply some transformations to neutralize the embeddings. They found that hard-debiasing is effective in reducing all forms of gender bias, while soft de-biasing is useful in specific settings. A similar approach could be taken with embeddings extracted from applicant essays, along various sensitive attributes such as gender, income, age and ethnicity.

Fairness Evaluation Tools

Fairness tools are also widely used to detect bias in decision-making where machine learning models are involved. These tools provide an overall fairness auditing suite, including the capability to generate bias reports for features used in the model. Most of these tools are offered by pioneers in the AI field, including IBM, Google, and LinkedIn[26, 30]. Some of these tools are described below.

- **AIF360:** This is IBM's solution for fairness assessment and mitigation. It is open source and integrated with Python, and includes comprehensive fairness metrics such as equalized odds, statistical parity and disparate impact. Additionally, it offers built-in bias mitigation functions including group fairness optimization, adversarial de-biasing and several others. It is model-agnostic and allows users the flexibility to integrate custom algorithms [6, 8].

- **LiFT:** This is LinkedIn's offering to address bias and fairness in large-scale ML models. Among its features offered, a significant advantage is that it is offered as a library with Apache Spark, making it efficient for use in big data applications. Similar to AIF360, this also supports an array of fairness metrics including statistical parity and equalized odds. It provides APIs that can be integrated into any stage of an ML pipeline and provides post-processing mitigation strategies along with ad-hoc fairness analysis [47].

- **WhatiF:** This toolkit offered by Google addresses ML model evaluation challenges such as interpretation difficulty and insufficient diverse testing. It provides an interactive interface that allows users to modify inputs and observe resulting changes in outcome through visualizations and feature importance analysis. This toolkit is friendly for users from a non-technical background as well and encourages transparency in machine learning [53].

While these are some of the popular tools available, there are also other tools such as Microsoft's FairLearn [9] and Aequitas [43]. All of these tools are compatible with specific libraries and models and present their own unique advantages and disadvantages.

Chapter 3

Proposed Framework

This work aims to demonstrate the systematic analysis of bias that can be performed by universities adopting machine learning models to automate their admissions review process. To conduct the various bias detection and mitigation experiments methodically, we have developed a two-phased study comprising of the following:

- **Phase 1:** Preemptive Bias Detection
- **Phase 2:** Fairness Evaluation and Bias Correction

Sections 3.1 and 3.2 will describe the details of each phase and explain the various methods involved. Since bias and fairness is a largely complex topic and involves extensive analysis, we have sectioned our work into the two phases to aid future researchers and university technology teams in understanding and applying our methodology for bias detection and correction of their admissions prediction machine learning models.

3.1 Phase 1: Preemptive Bias Detection

We introduce a pipeline to systematically detect potentially inferred biases in the data shown in Figure 3.1. Once the data is preprocessed, exploratory data analysis (EDA) is performed to reveal distributions and trends. This is crucial to understanding relationships between the features and the decision variable, as well as other patterns in the data that may be useful

to analyze our results. As part of the EDA, we also perform clustering to identify potential subsets in the data to further extract patterns in subgroups of the dataset that may lead to data biases. The second step involves searching for intersectional feature bias through subgroup discovery. Subgroup discovery is used to identify a combination of features that could have a higher tendency toward a particular decision outcome [50]. Step 3 includes training a machine learning model to extract the feature importance of the dataset features and subsequently identify demographic features that the model considers important but may not be typically prioritized by an admissions review committee.

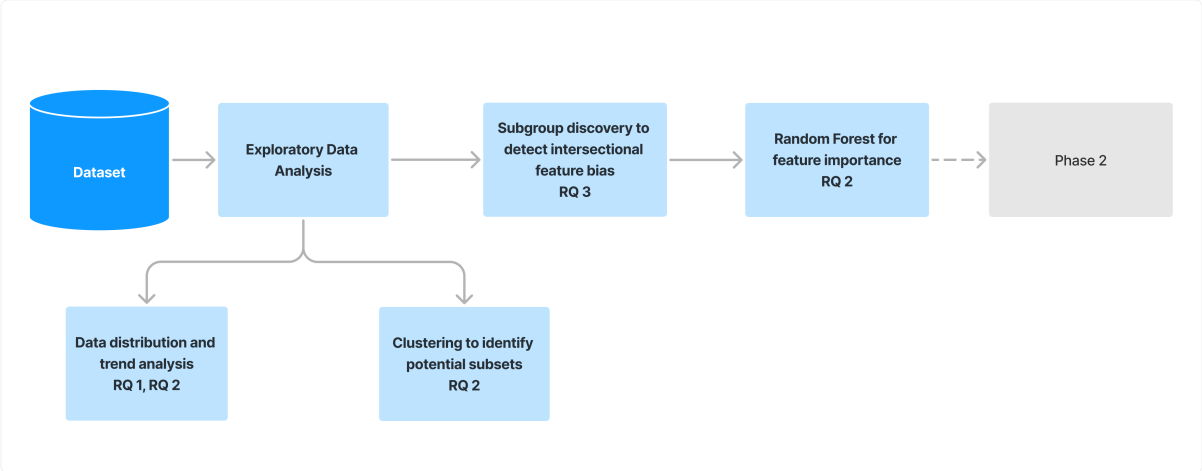


Figure 3.1: Phase 1: Preemptive Bias Detection

This framework would help ascertain how bias might be introduced into machine learning models, by comprehending the data, feature linkages to the admission decision, intersectional feature bias, and the early effects of the affirmative action ban. Using this framework, researchers and universities may analyze the induction of bias to a model and potentially rectify it to mitigate bias propagation, while still increasing the efficiency of admissions committees by using AI for admissions. Figure 3.1 displays the proposed framework and relates it to corresponding research questions in this study.

3.2 Phase 2: Fairness Evaluation and Bias Correction

While 3.1 focused on a priori detection of inferred biases from the dataset, phase two involves applying a machine learning model for admissions prediction and evaluating its fairness. The first stage of the pipeline is feature engineering, where features are transformed to reduce dimensionality and make them suitable for the machine learning model and further analysis of bias. The next stage of the pipeline involves the development of a sequential neural network model as described in 5.3.2. To interpret the outcomes of the model's predictions, we use explainability tools from 2.3.1. These provide an insight on model decision-making factors such as feature importance, the values of each feature that may positively or negatively influence model decisions, as well as the magnitude of influence each feature may have on the model outcomes. In this study, we focus on determining explanations at a global level to

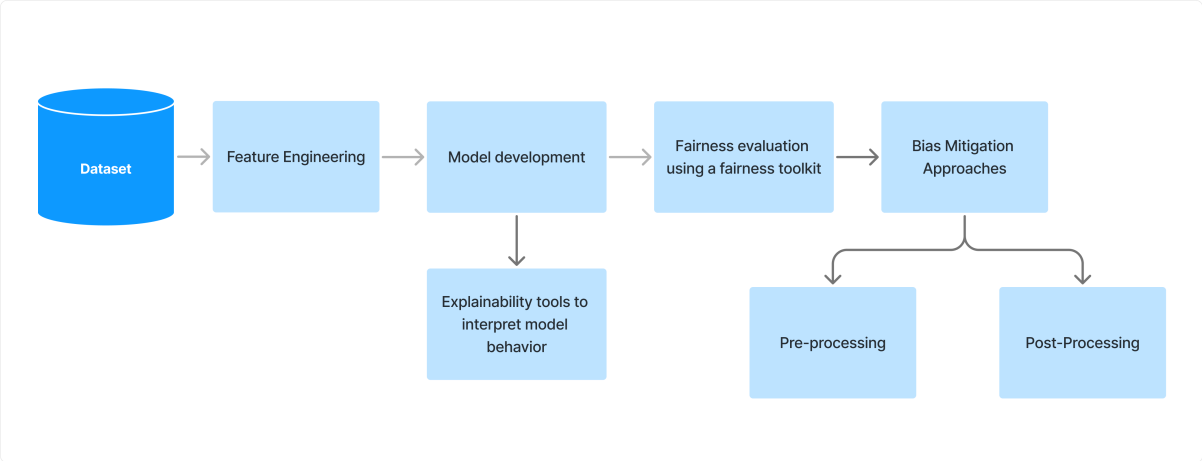


Figure 3.2: Phase 2: Fairness Evaluation and Bias Correction

obtain a macroscopic view of model behavior, rather than for each individual observation, although this capability is provided by most tools. This is followed by fairness evaluation using a fairness evaluation toolkit from section 2.3.2.

Chapter 4

Methodology

This study aims to uncover the potential biases that a machine learning model trained on historical admissions data may learn in its training phase and demonstrate how these biases can be identified and mitigated using data analysis and fairness tools. As established in the introduction, while admissions decisions are made holistically through different rubrics specific to each university, they are often made by human reviewers, who not only evaluate objectively but also look for qualities that indicate student success in the university. When a machine learning model trains on historical data, it can potentially learn patterns in admissions outcomes as rules for decision-making, regardless of whether the admissions committee intended for these to be the rules. We refer to this as inferred bias, wherein a machine learning model develops biases in its algorithm while training based on inferences from the biased data. The inferred bias can be a consequence of various types of biases from the data, such as representation bias, historical bias, human bias, independent bias and intersectional bias. When the model trains on this data, it is likely to infer bias from the dataset and develop further algorithmic biases that may emerge from the design choices made while training the machine learning algorithm, such as the selection of regularization methods and optimizers [30].

4.1 Subgroup Discovery

Many have studied the bias-contributions of individual features in the dataset in the context of university admissions [7, 11, 24, 29, 41, 55]. While these are useful in determining which features may lead to a biased outcome, they do not consider all possible combinations of features that may lead to a bias, also known as intersectional feature bias. As argued by Wamburu et al. in [49], systematic scanning [36] without presupposing bias-inducing features may reveal subsets of features that contribute to bias that are difficult to discover otherwise. One way to tackle intersectional feature bias is through subgroup discovery, which is a statistical approach to extract subgroups that have an increased likelihood of achieving a particular target outcome [23]. Subgroup discovery identifies distinct subsets of data where the target variable behaves differently from the overall population. These are represented as condition-target pairs where the condition could be a specific combination of feature values such as

$$Gender = female \ AND \ Age > 35 \tag{4.1}$$

which may have a higher probability of the admission decision being 'reject' in the dataset. We adopt this approach to identify intersectional feature configurations in the dataset that exhibit statistically significant deviations in their predictive patterns towards the target variable which is the admission decision. The objective of using subgroup discovery in this study is to reveal underlying recurring patterns in the dataset that may be potential sources of algorithmic bias inapparent when examining these features independently.

4.2 Fairness Evaluation Metrics

While there are numerous fairness metrics to measure different types of fairness outcomes, we primarily use four metrics in this study:

- **Disparate Impact:** Disparate impact follows the legal notion that the ratio of selection rate for privileged and underprivileged groups must be high, such that both underprivileged and privileged groups have a similar likelihood of receiving a positive outcome, or admit [39].

$$P[\hat{Y} = 1|S \neq 1] \leq P[\hat{Y} = 1|S = 1] \geq 1 - \varepsilon, \tag{1}$$

  where S represents the protected attribute and S=1 is the privileged group, and S≠ 1 is the unprivileged group. $\hat{Y} = 1$ indicates that the prediction is positive.

- **Statistical Parity Difference:** Statistical parity suggests that the likelihood of receiving a positive outcome should be equal for all observations regardless of their attributes being protected [26, 30].

$$P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1) \tag{4.2}$$

  where Y is the predictor and A is a protected attribute. Statistical parity difference would measure the probability of favorable outcome for unprivileged instances - probability of favorable outcome for privileged instances, i.e., the difference in selection rates between underprivileged and privileged groups. This is useful to identify whether certain subgroups have a lower admission selection rate compared to more privileged subgroups.

- **Equal Opportunity:** This fairness metric suggests that different demographic subgroups should have a similar chance at receiving a positive outcome regardless of their protected attributes [26]. This approach focuses on the true positive rates (TPR) solely, since we are trying to ensure that qualified applicants from all demographic

groups have an equal chance at being accepted into the university.

$$P(\hat{Y} = 1|S = 0, Y = 1) = P(\hat{Y} = 1|S = 1, Y = 1) \tag{4.3}$$

where Y is the predictor and S is a demographic subgroup.

- **Average Odds:** The average odds difference can be described as the difference in false positive rate (FPR) and true positive rates (TPR) between underprivileged and privileged groups [6].

$$\frac{(FPR_{D=\text{unprivileged}} - FPR_{D=\text{privileged}}) + (TPR_{D=\text{unprivileged}} - TPR_{D=\text{privileged}})}{2} \tag{4.4}$$

average odds difference helps us measure the error in false admissions and false rejections and ideally should be close to zero to ensure no difference in error rates based on protected attributes.

Based on the protected attributes identified as biased, bias mitigation approaches are implemented to effectively reduce the model's bias detailed in 4.3.

4.3 Bias Mitigation Approach

When considering bias in algorithmic decision-making applications, we must acknowledge that the bias can emerge from different stages in the ML pipeline, such as the dataset, the design of the model, feedback loops and user interaction [30]. The authors of [30] also categorize the emergence of bias in algorithmic decision-making into three different processes, data to algorithm, algorithm to user and user to data. In the context of university admissions, we may potentially find bias in all three of these categories as described below:

- **Data to algorithm:** This could occur in the form of representation bias, if our dataset contains underrepresented subgroups that may have received a particular admission decision outcome.

- **Algorithm to user:** This could occur in the form of algorithmic bias, when the design of our machine learning model induces newly added bias to the outcome.

- **User to data:** This often exists in socio-technical systems, such as university admissions and is a result of human biases which may have influenced the decision-making for the candidates that are the observations recorded in our dataset.

In this study, we focus on the data to algorithm and algorithm to user biases, as these may be mitigated using algorithmic approaches. The data bias can be tackled by using a pre-processing strategy, wherein we pre-process the dataset prior to training the machine learning model to remove any underlying biases [30]. This can be done by Reweighing, which works on the principle of assigning different weights to different subgroups in the training data to equalize their representation in each admission decision pool (i.e, accept or reject) [25]. For each configuration of protection attribute values such as gender and age, and target variable value which is admission decision in our case, a weight is calculated as follows:

$$W(attr, target) = \frac{P(attr) \times P(target)}{P(attr, target)} \tag{4.5}$$

This reweighing ensures that the model learns fairly from the data without actually modifying the outcomes of observations or upsampling or downsampling the data. However, it bears the disadvantage that for each sensitive attribute that the dataset is debiased towards, the model needs to be retrained form scratch, which can become computationally expensive for large datasets.

Algorithmic bias can be tackled using post-processing, where the algorithm output is modified achieve calibrated equalized odds [18]. Equalized odds are defined in [30] as the scenario in which the true positive rate and false positive rate is equal for both protected and unprotected subgroups. The calibrated equalized odds approach, adopted from [40] is a post-processing strategy that systematically identifies probabilities for output labels that is optimized to achieve equalized odds. Since this is a post-processing algorithm, it is advantageous in the fact that it is very computationally efficient since there is no retraining required.

This study uses a combination of the pre-processing and post-processing bias mitigation strategies to tackle bias on multiple stages in the machine learning automated solution, so that both data and algorithmic bias may be reduced effectively.

Chapter 5

Experiments

5.1 Dataset

The dataset consists of 14850 observations with 11 features of graduate applications to a competitive Computer Science department of a large public research university in the United States. These are from applications to three distinct graduate-level programs: a research-focused Master of Science (MS) program, an industry-focused Master of Engineering program (MEng), and a Doctor of Philosophy (PhD) program. The data spans a period of ten years from 2014 to 2024 for the MS and PhD programs but was only available for 2 years for the MEng program. To maintain uniformity in results, we dropped all observations of the MEng program. Our final dataset consists of the 9315 observations of the MS and PhD programs available from 2014-2024.

The features used from the dataset can be categorized into demographic and non demographic features as shown in Table 5.1.

Though application data consists of a mixture of test scores, GPA, essays, personal statements, and research experience for graduate applications, the scope of this study is limited to non-essay data. The data was preprocessed to impute missing values with mean for numerical features and median for categorical features. Optional fields such as 'Birth Nation', 'Race', 'Current or Former Military', and 'First Generation' were imputed with placeholder

Table 5.1: Demographic and non-demographic features in the dataset

Demographic	Non-Demographic
Person Age	GPA
Gender	IELTS
Birth Nation	TOEFL
Citizenship (Primary)	GRE
Hispanic	Decision
Race	
Current or Former Military	
First Generation	

values such as 'Unknown' and -1 to indicate that they were opted out. Since the applicants had undergraduate education from various countries, the GPA scores were rescaled to the standard U.S. 4.0 scale for uniformity. The GRE, TOEFL, and IELTS scores were aggregated from the initial dataset, which contained separate features for each subsection of these standardized tests. Unlike the centralized application reviews done for undergraduate applications, the graduate admissions process for the data-providing institution consists of decentralized application reviews from the CS department where reviewers examine the applicant information along with essays and supporting documents submitted, to determine admission decisions. The university uses volunteer readers to score essays, and the average of multiple reviews is considered for evaluation along with the remaining application details.

5.2 Phase 1

5.2.1 Exploratory Data Analysis

Data Distributions

The first stage of EDA involves examining the distributions of various demographic features shown in 5.1 and their contribution to the subset of accepted and rejected applications.

Since this work is focused on determining bias on applying a machine learning model to historical data, we thoroughly examine the distributions of various features categorized as demographic in Table 5.1. The dataset consists of applications to two graduate programs in the department of computer science, with 6317 observations belonging to the MS degree and 2998 observations belonging to the PhD degree. Some of the demographic features like 'Current or Former Military' and 'First Generation' only had a small subset of values filled in, where 8.6% of applications reported as first-generation students and 0.6% of applications reported having military experience. This minority makes it challenging to include these features in different analyses since a majority of the values were unfilled and had to be imputed. We plotted several plots to examine the acceptance rate, distribution of race of domestic and internal applicants, distribution of race of admitted applicants to each program, gender distribution and age distribution. We also tried to analyze the various countries from which the dataset contained applications. These distributions also helped us evaluate the preliminary effects of the affirmative action ban.

Clustering

A common unsupervised learning approach for identifying subgroups in a dataset is clustering. In this study, we used clustering to identify underlying patterns among different subgroups that may lead to a positive or negative decision on an application. Since we are experimenting with admissions data, the two classes of data would be those with the decision 'admit' and those with the decision 'reject'. However, clustering into these two classes would not render much insight into the contributions of different features towards the admission decision. Therefore, we assumed an unknown number of clusters in the data and performed Hierarchical Density-based Spatial Clustering (HDBScan) [12] to segregate the observations into subsets for each cluster. We chose the density-based clustering approach as it is robust to noisy data since it detects dense clusters and categorizes the remaining observations as

noise [3]. It can also tackle high-dimensional data, unlike other clustering algorithms like KMeans, especially when there is no predefined number of clusters.

5.2.2 Subgroup Discovery for Intersectional Bias

As discussed in 4.1, subgroup discovery enables us to identify underlying subgroups of feature intersections in the data that may receive differential treatment from the model due to their statistical relationship with a particular target outcome. In essence, some feature intersections may be present in the dataset such that they have a higher probability of having the target value 'accept', while some other subgroups may have a higher probability of the target value 'reject'. We applied the subgroup discovery algorithm using the Pysubgroup library [27] to our dataset with the application decision set as the target variable. We limited our search space to only include features defined as demographic in Table 5.1. However, we had to exclude the features 'Current or Former Military' and 'First Generation' since these did not have a sufficient number of filled values and distorted the results.

5.2.3 Feature Importance

A straightforward explainability method to understand the perception of training features by machine learning models is the visualization of the feature importance. We trained a Random Forest classifier on our training set which contained 84% of the original dataset filtered by application year from 2014-2023. This was done to simulate the practical development of the model in an applied sense, where historical data would be used as training data to predict future outcomes.

5.3 Phase 2

5.3.1 Feature Engineering for ML Model

Originally, the dataset consisted of the features described in 5.1 of which Gender, Birth Nation, Citizenship, Hispanic were used for the exploratory data analysis in 5.2.1. While this provided valuable insights, it was not directly usable to train the machine learning model since a) we needed to encode categorical features, which would increase the dimensionality of the model drastically due to the number of unique countries present in the Birth Nation and Citizenship features, and b) analyzing the model's predictions trained on each country as a feature for bias is challenging, since we cannot identify underprivileged and privileged values easily.

To combat these issues and draw meaningful insights, we dropped the Birth Nation feature and extracted citizenship information from the citizenship feature as a boolean feature titled 'US Citizen', which would differentiate domestic and international applicants. All the categorical features were one hot encoded and all the numerical features were standardized.

	Hispanic	IELTS	GRE total	TOEFL IBT total	GPA_scaled	US Citizen	Gender Code	Age Quartile	Race Code	Decision
Hispanic	1.000000	-0.055906	0.068697	0.037185	0.025516	0.055705	0.012778	0.072265	-0.068269	0.017074
IELTS	-0.055906	1.000000	-0.183893	-0.277743	0.000800	-0.088607	0.029675	-0.214673	0.110626	-0.114128
GRE total	0.068697	-0.183893	1.000000	0.386284	-0.044614	-0.104294	-0.023050	0.371339	-0.186735	-0.043837
TOEFL IBT total	0.037185	-0.277743	0.386284	1.000000	-0.030574	-0.311306	-0.003960	0.158077	-0.137258	-0.133845
GPA_scaled	0.025516	0.000800	-0.044614	-0.030574	1.000000	0.063469	0.022135	-0.079506	-0.026651	0.124916
US Citizen	0.055705	-0.088607	-0.104294	-0.311306	0.063469	1.000000	-0.029632	-0.050152	0.093249	0.306518
Gender Code	0.012778	0.029675	-0.023050	-0.003960	0.022135	-0.029632	1.000000	-0.014307	-0.030602	0.016273
Age Quartile	0.072265	-0.214673	0.371339	0.158077	-0.079506	-0.050152	-0.014307	1.000000	-0.196457	-0.025520
Race Code	-0.068269	0.110626	-0.186735	-0.137258	-0.026651	0.093249	-0.030602	-0.196457	1.000000	0.008921
Decision	0.017074	-0.114128	-0.043837	-0.133845	0.124916	0.306518	0.016273	-0.025520	0.008921	1.000000

Figure 5.1: Correlation Heatmap of Model Input Features

The final feature correlation map is shown in Figure 5.1. We observe low correlations between most features and the target variable 'Decision', with 'US Citizen' having the highest positive

correlation at 0.3 with the decision feature. This indicates that there is a weak underlying relationship with the attribute 'US Citizen' and 'Decision' as found in the analyses from 5.2.2.

5.3.2 Model Development

The objective of this work is to train a machine learning model to demonstrate how bias can be systematically evaluated and mitigated for the purpose of admissions reviews. Since the task focuses on identifying if a given observation shall receive an admit or a reject, this is a binary classification task with the model output being 0 for reject and 1 for admit.

Layer (type)	Output Shape	Param #
dense_64 (Dense)	(None, 128)	5,248
batch_normalization_38 (BatchNormalization)	(None, 128)	512
dropout_46 (Dropout)	(None, 128)	0
dense_65 (Dense)	(None, 64)	8,256
batch_normalization_39 (BatchNormalization)	(None, 64)	256
dropout_47 (Dropout)	(None, 64)	0
dense_66 (Dense)	(None, 32)	2,080
batch_normalization_40 (BatchNormalization)	(None, 32)	128
dropout_48 (Dropout)	(None, 32)	0
dense_67 (Dense)	(None, 1)	33

Total params: 48,645 (190.02 KB)
Trainable params: 16,065 (62.75 KB)
Non-trainable params: 448 (1.75 KB)
Optimizer params: 32,132 (125.52 KB)

Figure 5.2: Neural Network Model Architecture

Model Architecture

The neural network model architecture is shown in Figure 5.2.The model contains four fully connected layers alternating with batch normalization and dropout layers. The fully connected layers range in size from 128 neurons to 68 to 32 to 1, creating a hierarchical feature extraction that allows the model to extract simpler representations from complex input. The dense layers used relu and sigmoid activations with L2 regularization. The batch normalization layers increase the training stability and the gradient flow. The dropout layers are added for model regularization, to ensure that the model does not overfit on the training data and can generalize well. The optimizer used was an Adam optimizer with a learning rate of $5e \exp -4$. The resulting model is a lightweight model with only 48645 parameters and can easily be run with limited computational power. While this is the model architecture used for this study, it may be replaced with a more complex architecture and larger number of parameters to improve model performance in terms of F1-score and accuracy. For the purpose of this study, we use a fundamental model and focus on the bias detection and mitigation strategies.

Training the Model

In an attempt to simulate the real world application of a machine learning model to automate admissions prediction, we split up the dataset in a chronological manner. The train set consisted of 7094 observations from 2014 to 2022, the validation set consisted of 735 observations from 2023 and the test set consisted of 1486 observations from 2024. The model was developed and trained using the keras library. The validation loss was used to actively monitor the model performance through the training epochs and perform early stopping. The learning rate was also reduced if the validation loss reached a plateau point. The model was trained for a total of 100 epochs with a batch size of 64 to enable more stable gradients.

Model Performance

The resulting model achieved a test accuracy of 73.82% with the training accuracy and loss graph as shown in Figure 5.3.

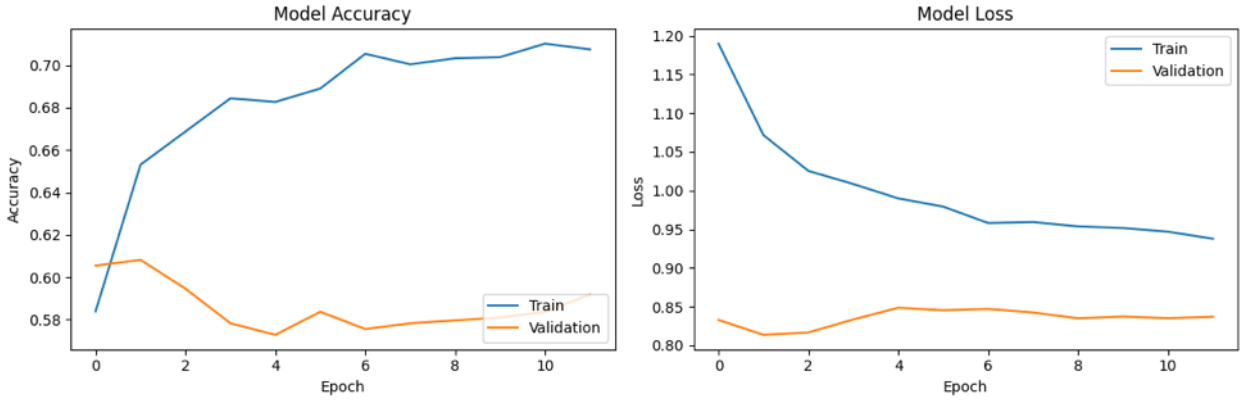


Figure 5.3: Model Accuracy and Loss (training and validation)

5.3.3 Fairness Evaluation

Fairness evaluation requires the user to define specific features as protected attributes and create a custom dataset to adhere to the format of the fairness evaluation toolkit. Fairness evaluation can be broken down into the following steps:

- Defining protected attributes
- Evaluate dataset to obtain baseline values for fairness metrics
- Evaluate model fairness and compare with the baseline
- Identify biased features
- Apply bias mitigation: preprocessing bias mitigation to tackle data bias and post-processing bias mitigation to tackle algorithmic bias

Selecting an Evaluation Tool

While there are several fairness toolkits available as detailed in 2.3.2, we found the AIF360 toolkit offered by IBM [8] to be the most suitable for our study, since it is widely used and comes with inbuilt fairness metrics and mitigation strategies. Due to the detailed documentation and examples available for this tool, it became our preferred option. Its accessibility through the Python libraries made it a convenient choice without having additional dependencies, which the some of the other tools had. Since it allowed us to create a custom model and only required an AIF360 compatible dataset to be created, it was seamlessly integrated into our machine learning pipeline.

Configuring the AIF360 dataset

The first step in fairness evaluation involves preparing the dataset for AIF360. The dataset was preprocessed such that categorical attributes were mapped to corresponding numerical representations. Gender was encoded by directly mapping 'male' to 0, 'female' to 1 and 'neutral' to 2. The age feature was binned into five quantiles with the intervals [(18.999, 25.0] < (25.0, 28.0] < (28.0, 31.0] < (31.0, 33.0] < (33.0, 66.0]], so that it may be easier to define protected attributes.

The challenge in fairness analysis was defining protected attributes. AIF360 requires the user to define protected value groups for each protected attribute, which can largely vary depending on use case and the user's definition of privileged or underprivileged. For this experiment, we concluded that the following attributes are protected, with the privileged and underprivileged groups defined in 5.2.

The privileged races were identified by those with a higher than average percentage composition in the dataset, which yielded Asian and Caucasian as privileged groups. For all other

Table 5.2: Protected Attributes defined for AIF360

Protected Attribute	Privileged Group
Person Age (binned)	0, 1 (25-31 years old)
Gender	0 (Male)
US Citizen	1
Hispanic	0
Race	Asian, Caucasian

attributes, it was similarly defined based on higher presence in the pool of applicants.

Baseline Bias Detection

The fairness evaluation pipeline begins with a comprehensive assessment of baseline bias present in the admissions dataset before any model training occurs. This critical first step utilizes AIF360's BinaryLabelDatasetMetric to quantify inherent distributional disparities across protected attributes including gender, race, Hispanic status, citizenship, and age. For each attribute, the pipeline calculates disparate impact (the ratio of favorable outcomes between unprivileged and privileged groups) and statistical parity difference (the absolute difference in selection rates). These metrics establish a statistical foundation for understanding pre-existing societal biases embedded in the historical admissions data. The implementation strategically separates this baseline analysis from model evaluation to enable subsequent bias amplification measurements, providing insight into whether algorithmic decision-making exacerbates existing inequities. By establishing these baseline disparities, the pipeline creates a reference point against which all subsequent fairness interventions can be measured.

Fairness Evaluation of Trained Model

After training the neural network on preprocessed admissions data, the pipeline conducts a thorough fairness assessment of the model's predictions using AIF360's ClassificationMetric. This evaluation extends beyond baseline metrics to include error-based fairness measures

that assess discriminatory patterns in model performance. The implementation calculates four key metrics for each protected attribute: disparate impact, statistical parity difference, equal opportunity difference, and average odds difference as explained in 4.2. These metrics provide a multi-dimensional view of algorithmic fairness, addressing both outcome distribution and error rate disparities. The pipeline then performs a novel bias amplification analysis, calculating the percentage increase or decrease in bias compared to the original dataset. This analysis explicitly quantifies how much the model magnifies or reduces pre-existing biases, with positive percentages indicating bias amplification and negative values showing bias reduction. Visualizations of these comparisons provide intuitive understanding of the model's impact on fairness across different demographic groups.

5.3.4 Bias Mitigation

Identifying Biased Features

The implementation employs a systematic threshold-based approach to identify which protected attributes exhibit significant bias in the trained model. Four distinct fairness metrics are evaluated against established thresholds: disparate impact must not deviate from the ideal value of 1.0 by more than 0.2, while statistical parity difference, equal opportunity difference, and average odds difference must not exceed an absolute value of 0.1. Any attribute violating these thresholds is classified as "biased" and flagged for mitigation. This threshold-based classification aligns with legal and ethical standards in anti-discrimination frameworks while providing practical decision boundaries for intervention. The pipeline generates a comprehensive bias assessment summary, detailing exactly which attributes exceed fairness thresholds and by how much. In our implementation, this analysis identified several biased features, with gender, race, and age demonstrating the most significant disparities. The identification process provides transparency about which demographic groups

face algorithmic discrimination and quantifies the severity of bias for each attribute, enabling prioritization of mitigation efforts.

Bias Mitigation Approach

Our pipeline consists of a two-pronged approach to bias mitigation that combines preprocessing and post-processing techniques. This strategy addresses bias at multiple points in the machine learning pipeline, providing a more robust and comprehensive solution than single-method approaches. The implementation first identifies all biased attributes, then applies a strategic combination of data transformation and prediction adjustment techniques to mitigate discrimination while preserving model performance.

Pre-Processing For preprocessing bias mitigation, the implementation employs the Reweighing algorithm from IBM's AI Fairness 360 (AIF360) toolkit, applied to the attribute with the highest statistical parity difference. The algorithm is implemented through the aif360.algorithms.preprocessing.Reweighing class, which transforms the original AIF360 dataset using the fit_transform() method. This data-centric approach assigns compensatory weights to training examples based on their protected attribute values and outcomes. These instance weights, accessed via the instance_weights property of the transformed dataset, are then integrated into the model training process using the sample_weight parameter in the fit() method. A new neural network with identical architecture is trained with these weights, addressing bias at its source. The implementation explicitly maps AIF360 instance weights to their corresponding training examples using numpy array operations to ensure correct weight application. The effectiveness of this preprocessing approach is measured using the ClassificationMetric class from AIF360, comparing fairness metrics before and after intervention.

Post-Processing The postprocessing component implements Calibrated Equalized Odds using the CalibratedEqOddsPostprocessing class from aif360.algorithms.postprocessing. This

technique operates on the model's predictions rather than the data or model itself. The implementation creates compatible AIF360 datasets containing the original test data and model predictions, then applies the fit() and predict() methods of the CalibratedEqOddsPostprocessing class to find optimal threshold adjustments. These adjusted predictions equalize both true positive and false positive rates across demographic groups. The pipeline ensures dataset compatibility between the ground truth and prediction datasets through careful construction of AIF360's BinaryLabelDataset objects, maintaining identical feature structures while only modifying label values. Fairness metrics after postprocessing are computed using AIF360's ClassificationMetric class, with specific focus on the equal_opportunity_difference() and average_odds_difference() methods to evaluate error rate disparities. The implementation handles potential dataset structure mismatches through careful reconstruction of prediction datasets that match the structure of the original test datasets, ensuring successful application of the postprocessing algorithm.

5.3.5 SHAP Anlalysis for Model Explainability

To gain deeper insights into the model's decision-making using feature importance patterns, we conducted SHAP analysis [28]. We perform a direct comparison of feature importance distributions between the original biased model with a neural network architecture, the bias-mitigated model retrained after reweighing, and an alternative model architecture, which is a Random Forest model trained on the same input data. For the original neural network and the preprocessed mitigated model, we used the DeepExplainer function of the shap library to generate explanations and visualize feature contributions towards the model decisions. For the Random Forest model, we used the TreeExplainer function, which is also available with the shap library. The attribution values were visualized using SHAP summary plots, which display both the magnitude and direction of each feature's impact on model predictions.

Chapter 6

Results

Our experiments in phase 1 revealed the distribution of different demographic groups in the data, as well as independent and intersectional demographic attributes that may lead to bias in the model. In phase 2, we trained a model and evaluated the model for bias using the fairness metrics specified in 5.3.3. Using explainability tools, we further explored the underlying demographic attributes that may be influencing the model's decisions. This chapter discusses the findings of our analysis in further detail.

6.1 Phase 1

6.1.1 Exploratory Data Analysis

The analyzed dataset comprises 6,317 MS and 2,998 PhD applications in computer science, with an average acceptance rate of approximately 36%, with significantly higher acceptance in 2019, 2020, 2021, and 2023. This might be related to the lower number of applications received during the pandemic or due to the expansion of the programs in recent years. Figure 6.1 reveals the acceptance rate of the CS graduate programs of our case university.

Since this study focuses on graduate data, notably, there is a wide range of age values in the population, as many applicants may have had some industry experience or completed multiple degrees before applying for their graduate degree. By visualizing the distribution

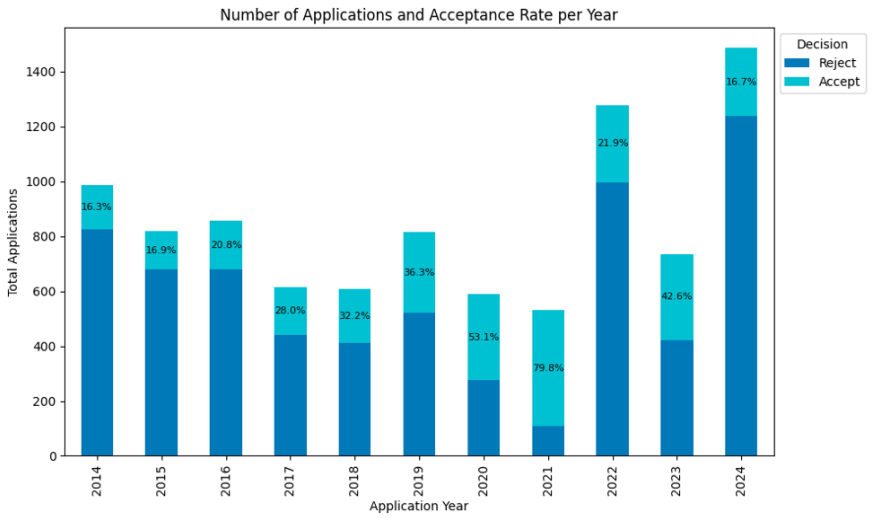


Figure 6.1: Acceptance Rate of Applications per Year

as shown in Figure 6.2, we cannot draw any apparent relationship between the age and the application decision, since the accepted and rejected observations seem to span the entire range. Through further analysis in the following sections, we can examine if there may be a deeper relationship or potential bias with the age feature.

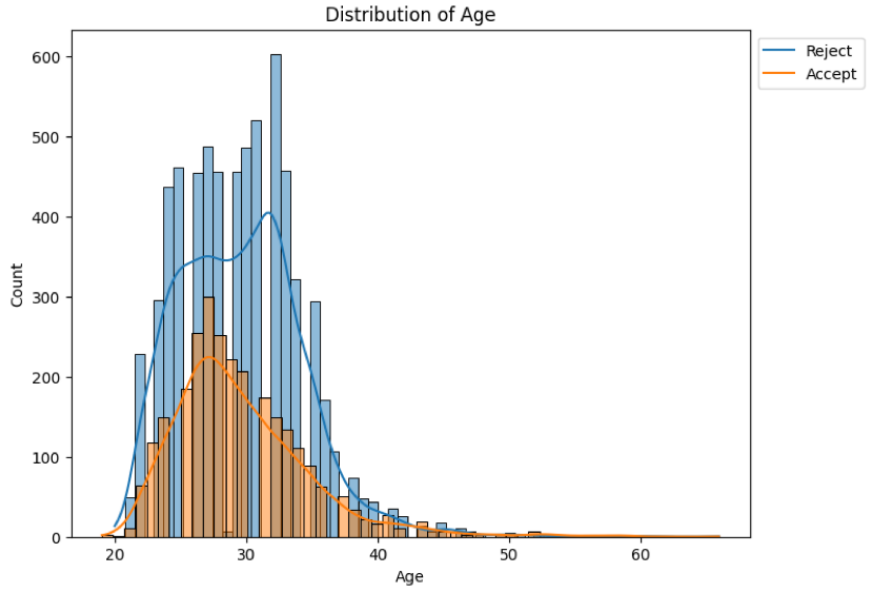


Figure 6.2: Age Distribution of Applicants

While several previous studies have focused on machine learning bias for undergraduate data [7, 11, 29, 41], few have conducted similar studies on graduate data [24, 55]. One of the peculiarities of experimenting with graduate application data is that it includes a significant number of applications from outside the United States. This can result in findings that are vastly different from undergraduate data due to the majority of undergraduate applications typically being from within the United States with similar prior education. However, in the case of graduate applications, there has been an increasing trend in the number of international applicants. This is evident in Figure 6 which shows the 10 countries with the highest number of applications, of which the United States comes third with 883 applications over 10 years. This aligns with reports by the NCES [4] that 44% of STEM master's degrees conferred in 2019-20 were by international students.

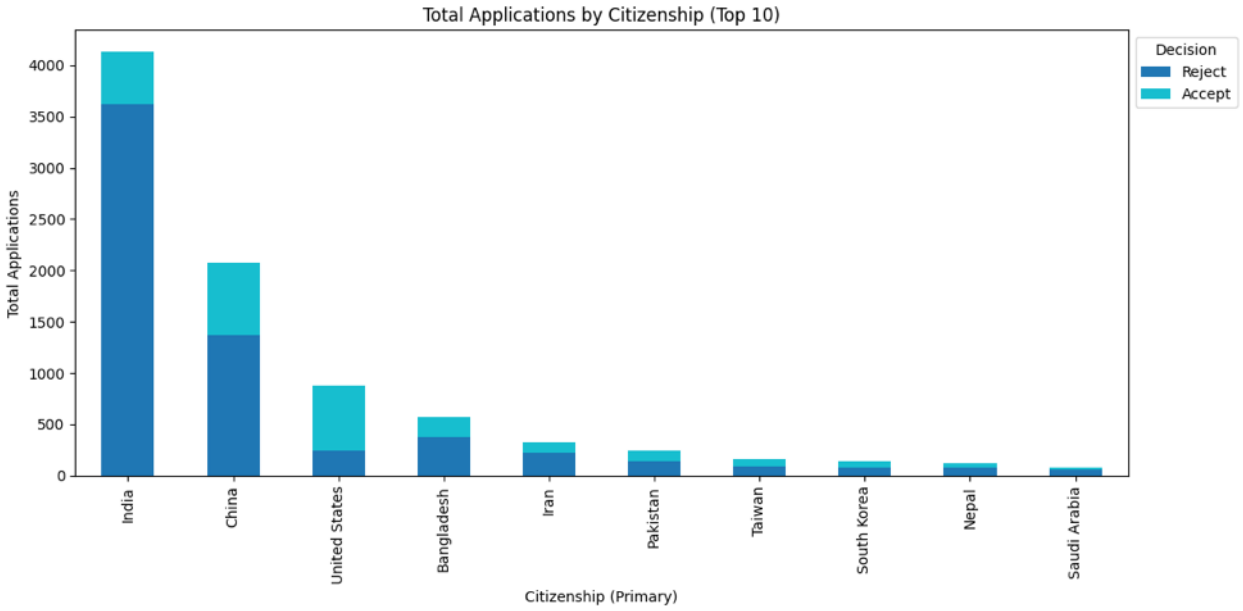
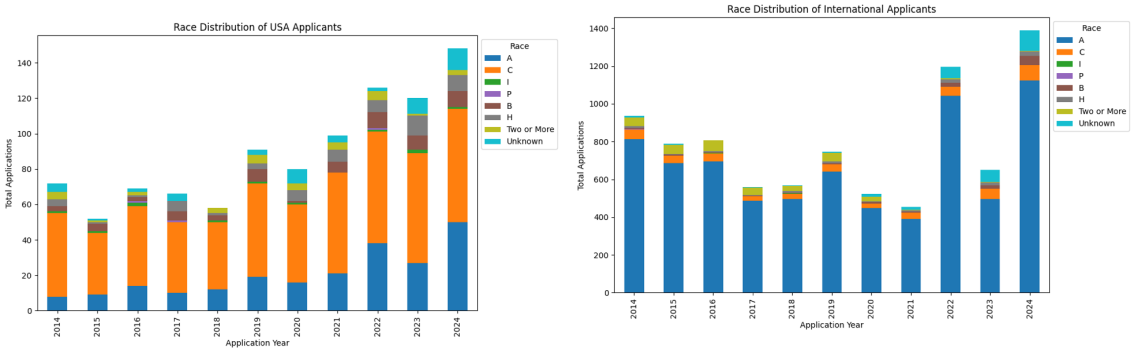


Figure 6.3: Applications by Citizenship (Top 10)

When we examined the distribution of race as reported in the applications, we found an increasing trend of applicants opting out of reporting race[1] as seen in 6.4, aligning with

[1]Race notations: A – Asian, B – Black, C – Caucasian, H – Hispanic, I – American Indian, P – Pacific

findings from [44]. The majority of U.S. applicants reported their race as Caucasian (63%), followed by those reporting as Asian (25%). Among the URMs, around 6% of applicants reported race as Black in the dataset and 5% of applicants reported race as Hispanic within the US. The American Indian and Pacific Islander groups have very low percentages of 1% and 0.3% respectively. The majority (86%) of the international applicants reported race as Asian, which aligns with findings from Figure 6.3, where the top countries include India and China.



(a) Race Distribution of USA Applicants

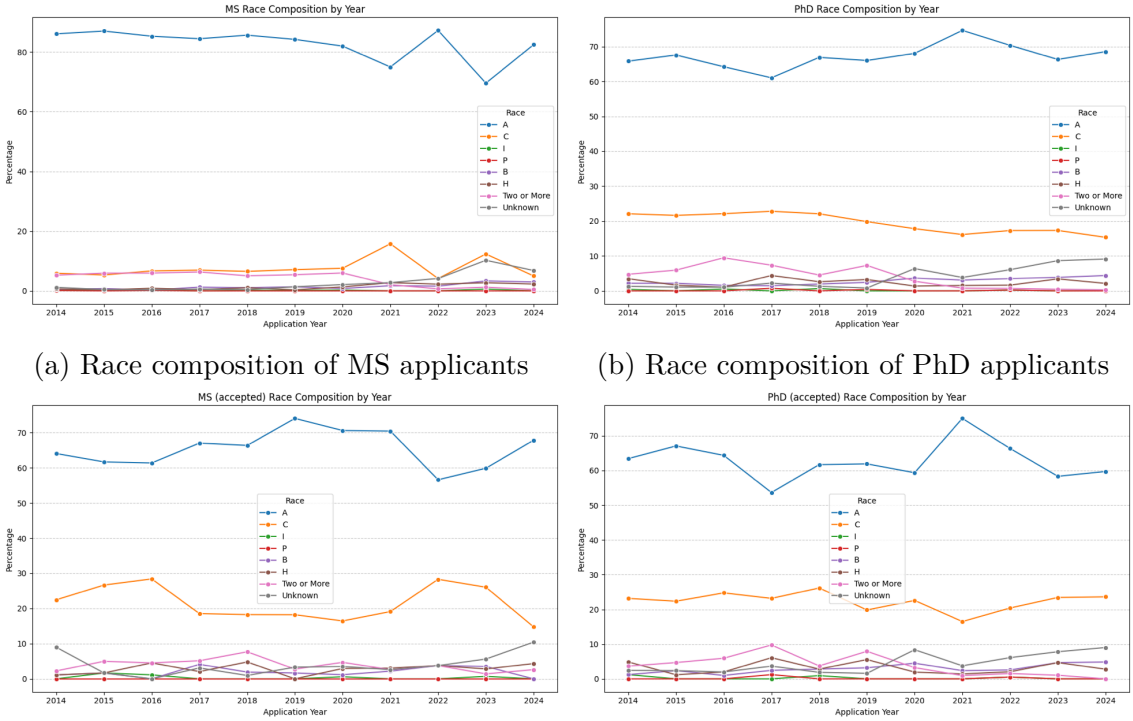
(b) Race Distribution of International Applicants

Figure 6.4: Race Distribution of USA and International Applicants

The number of applicants opting out of specifying race or marking it 'Unknown' has been increasing in recent years. While it is too early to draw conclusions about the effect of the 2023 Supreme Court ruling for affirmative action ban [54], we observe that the number of applicants opting out of reporting race has risen by 66% from 2023 to 2024. Several universities across the U.S. have reported a similar trend [12], indicating increased concern around disclosing race among applicants after the ban of affirmative action. We also examined the race composition of the two graduate computer science degrees offered by the university. Among the underrepresented minority groups (URMs), we observe a decline in the percentage of applicants who reported race as Black by 12.8% and Hispanic by 17.9% for the MS

Islander or Alaskan Native

degree. However, for the PhD degree, the increasing trend of students reporting race as Black continues, with the percentage increase changing from 11.3% to 13% in the 2022-23 and 2023-24 application cycles, whereas there is a decline in Hispanic. When comparing U.S. and international applicants with reported race Black, there is a positive change for U.S. applicants from 5.4% to 5.6% whereas international applicants declined from 3.58% to 2.94%. The percentage of PhD applicants reporting race as Hispanic sharply declined after the affirmative action ban from 3.6% to 2.2%, a decline of 38%. However, the admitted class has contrasting results with a sharp decline of Black students for the MS program but a steady increase for the PhD program. Hispanic students have contrasting results in the admitted class, with an increase in the MS program and a slight decrease in the PhD program. These findings are evident in Figure 6.5.



(a) Race composition of MS applicants (b) Race composition of PhD applicants

(c) Race composition of MS admitted class (d) Race composition of PhD admitted class

Figure 6.5: Race Distribution of MS and PhD students (applicants and admitted class)

Finally, we also examined the gender distribution of applicants for the 2014-2024 period.

We found that the gender ratio of applicants has remained nearly constant through the years, with the percentage of females (F) averaging 26.6% as shown in Figure 6.5. There are minimal observations (0.3%) that have marked their gender as neutral (N).

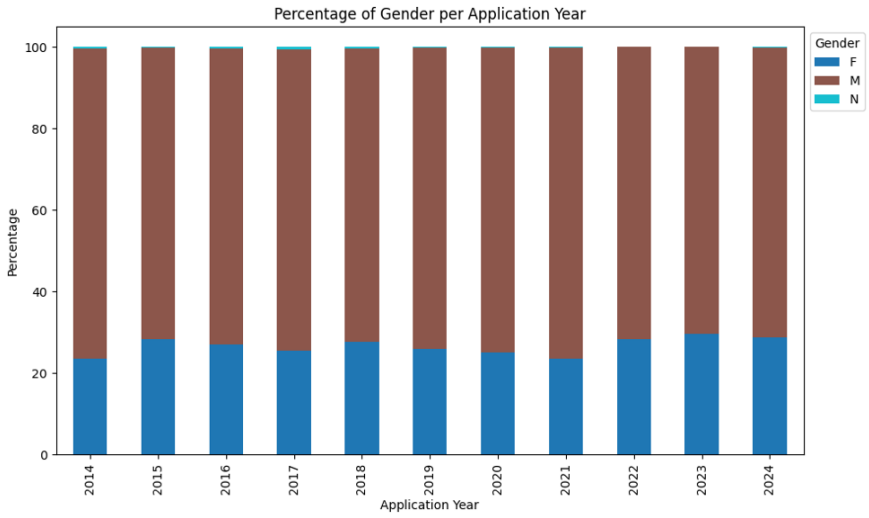


Figure 6.6: Gender Distribution of Applicants

6.1.2 Clustering

The initial clustering algorithm yielded 13 distinct clusters with 71% coverage of the dataset, leaving only 29% of the data points as noise. However, this achieved a low Density Based Clustering Validation (DBCV) score of 0.15, indicating that the clustering was not optimal. After fine-tuning the parameters of the HDBScan algorithm, we achieved an improved DBCV score of 0.33 with 7 dense clusters shown in Figure 6.7, although this resulted in a reduced coverage of 56%.

Among the seven clusters, we noticed three potential types of clusters: a. homogenous clusters with a majority of rejected applications, b. heterogeneous clusters with a mixture of accepted and rejected applications, and c. homogenous clusters with a majority of accepted applications. For the type a cluster, we noticed that 'Birth Nation' India has a negative

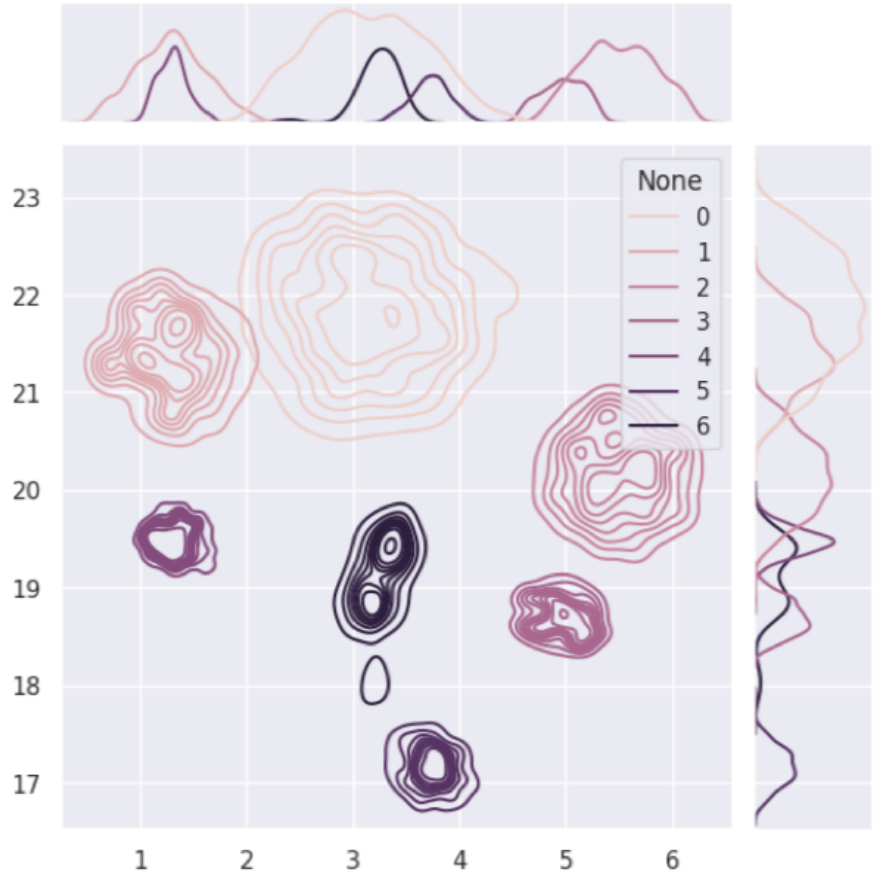


Figure 6.7: HDBScan Clustering Visualization

correlation with the decision, while 'Birth Nation' South Korea and Pakistan have a positive correlation with it. 'GPA' also has a positive correlation with the decision. This could imply that being from India might lead to a negative decision but may also reflect the applicants in this cluster since the cluster contains 96% applications with 'Birth Nation' India, 4 applications with 75% acceptance from Pakistan, and 2 applications with 100% acceptance from South Korea. Type b cluster had an overall acceptance rate of 36%, with the decision having a negative correlation to 'Age' and a positive correlation to 'GPA'. Type c clusters had an acceptance rate of 71%, with positive correlations for 'GPA' and 'Birth Nation' United States. It had negative correlations for 'TOEFL' with 'Birth Nation' being Bangladesh or Jordan and 'Race' being 'Two or More'. Since all the correlations mentioned

are in the absolute range of [0.08, 0.2], they are only mildly correlated with the decision. It is also important to note that the correlations depend on the demographic features of the subset of data within each specific dense cluster. Nonetheless, when a machine learning model trains on such data, it may consider these apparent correlations as rubrics for decision-making, which can have severe negative impacts on class diversity.

6.1.3 Subgroup Discovery

Our analysis revealed weak bias towards the target variable, as evidenced by low quality values, though this finding may be limited by the dataset's distribution and size. We extracted the top ten subsets with the highest quality and listed those with quality above the threshold of 0.01 in Table 6.1 and Table 6.2.

Table 6.1: Subgroup discovery results for Decision 'Accepted' (quality >=0.01)

Index	Subset	Quality	Subgroup Coverage	Lift
1	Birth Nation=='United States'	0.04	0.22	2.46
2	Birth Nation=='United States' AND Gender=='M'	0.03	0.18	2.47
3	Race=='C'	0.03	0.19	2.08
4	Birth Nation=='United States' AND Race=='C'	0.02	0.13	2.67
5	Gender=='M' AND Race=='C'	0.02	0.15	2.14
6	Birth Nation=='United States' AND Gender=='M' AND Race=='C'	0.02	0.11	2.62
7	Birth Nation=='China'	0.01	0.26	1.17
8	Birth Nation=='China' AND Race=='A'	0.01	0.24	1.17

The results indicate a weak bias towards the target variable since the quality values are low. However, this may be limited by the distribution of data in the dataset we used, as well as its small size. The quality metric represents the deviation of observed outcomes from expected outcomes for each subset, hence indicating bias. For the positive outcome subsets, i.e., with observations that had the decision 'Accepted', we observe that the common intersections of features are with 'Birth Nation' being United States or China, 'Gender' being male, and 'Race' being 'C' or 'A'. The subgroups with the highest coverage of more than 20% are Birth Nation=='United States', Birth Nation=='China', and (Birth Nation=='China' AND Race=='A'). Subset coverage indicates how much of the overall dataset constitutes

Table 6.2: Subgroup discovery results for Decision 'Rejected' (quality >=0.01)

Index	**Subset**	**Quality**	**Subgroup Coverage**	**Lift**
1	Birth Nation==‘India’	0.07	0.55	1.23
2	Birth Nation==‘India’ AND Race==‘A’	0.06	0.46	1.23
3	Birth Nation==‘India’ AND Gender==‘M’	0.06	0.39	1.25
4	Birth Nation==‘India’ AND Gender==‘M’ AND Race==‘A’	0.05	0.32	1.25
5	Race==‘A’	0.03	0.77	1.06
6	Gender==‘M’ AND Race==‘A’	0.02	0.56	1.07
7	Birth Nation==‘India’ AND Gender==‘F’	0.02	0.15	1.20
8	Birth Nation==‘India’ AND Gender==‘F’ AND Race==‘A’	0.02	0.13	1.20

that particular subset. The lift metric represents the ratio of the target class in the subset as against its prevalence in the entire dataset. The results show that an applicant in subgroups

1 and 2 is nearly 2.5 times more likely to receive an admit, an applicant in subgroup 6 is 2.6 times more likely to receive an admit and an applicant in subgroup 4 is nearly 2.7 times more likely to receive an admit.

We extrapolated similar findings for the negative outcome subsets, i.e., when the application decision is 'Rejected'. The features identified through subset scanning include 'Birth Nation' being India, 'Race' being Asian, and 'Gender' being male or female. Table 3 is sorted by quality and displays the top subgroups discovered with quality greater than or equal to 0.01. The lift value is highest for subgroups 3 and 4 with an increased likelihood of rejection of 1.25 times for features including birth nation, gender, and race.

Overall, we observe low bias for both the target outcomes but identify features that could lead to a biased outcome. We also note that the increased likelihood is higher for the positive target class (accept) than the negative class (reject).

6.1.4 Random Forest Feature Importance

The resulting feature importances with a significance greater than 0.01 are displayed in Figure 6.8. While we observe the GRE, TOEFL, and GPA scores among the top features with high importance, demographic features like age, citizenship, birth nation, and race have also been identified by the model as highly important. First-generation and military experience features, though present in the feature importance graph, may also be there due to the majority of their values being imputed missing values with –1. The model may have presumed that these features having value –1 correlate to a decision class. The major cause for concern from the feature importance graph is that features including age, the applicant's citizenship and birth nation being India or the United States, and race being Asian are all inferred as important features by the model. These may simply be the demographic

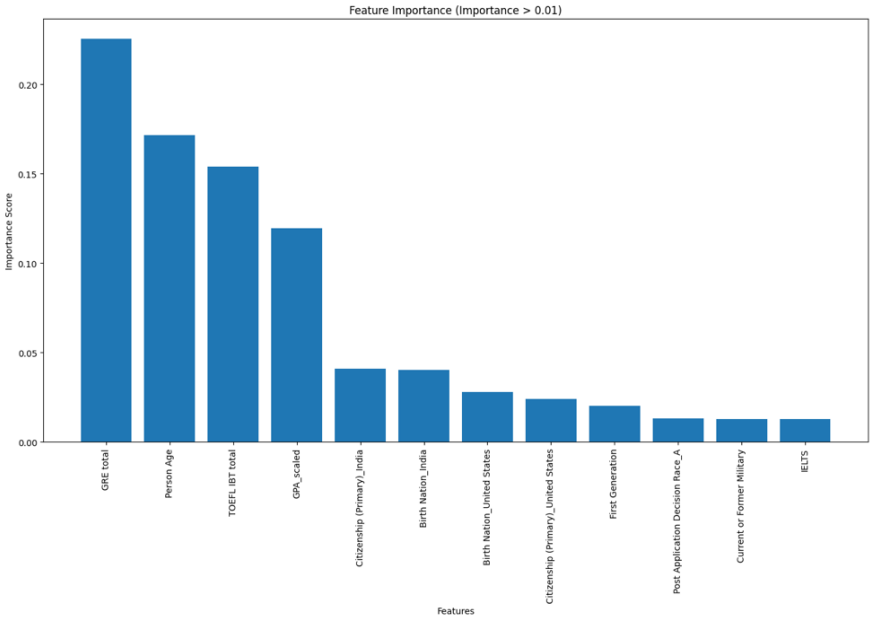


Figure 6.8: Feature Importance of Random Forest Model (all features)

features of the 'accept' and 'reject' subgroups after the admission decision was made and are unlikely to be decision-making factors that were considered by the admission review committee. However, they are inferred as the most important independent features by the random forest model, which illustrates that the model may be learning some biases.

6.2 Phase 2

6.2.1 Fairness Evaluation

In this section, we present the results of our fairness evaluation across multiple protected attributes: Gender, Race, Hispanic status, US Citizenship status, and Age. We analyze the presence of bias in the original dataset, examine how the neural network model affects this bias, and evaluate the effectiveness of various bias mitigation techniques.

Baseline Data Bias

Our initial analysis of the dataset revealed varying degrees of bias across different protected attributes:

- **Gender**: The original dataset exhibited a slight bias against female and gender-neutral applicants with a Disparate Impact of 0.9553 and a Statistical Parity Difference of -0.0135. These metrics indicate that female and gender-neutral applicants had approximately 4.5% lower probability of favorable outcomes compared to male applicants (the privileged group). According to the "80% rule" commonly used in bias evaluation [2], this level of disparity (95.53%) is well above the 0.8 threshold and would not be considered highly biased from a legal or regulatory perspective. While perfect fairness would yield a DI of 1.0, this relatively small deviation suggests a slight bias in favor of male applicants in the original dataset.

- **Race**: The data showed bias favoring unprivileged racial groups with a DI of 1.2387 and SPD of 0.0680. The privileged groups in this analysis were identified as categories 'A' (Asian), 'C' (Caucasian), and 'Asian - Other', while all other racial categories were considered unprivileged. These metrics indicate that applicants from presumed unprivileged racial groups had approximately 23.9% higher probability of favorable outcomes compared to applicants from the privileged racial groups. While this value exceeds the ideal DI of 1.0, it represents a reverse disparity that benefits the unprivileged groups. However, the remaining groups also consist of varying combinations of Asian and Caucasian applicants and could therefore mislead our analysis. This suggests a moderate bias in the admissions dataset that favors students that did not declare race as only Asian or Caucasian.

- **Hispanic**: We detected bias against Hispanic applicants with a DI of 0.8587 and SPD

of -0.0477, showing that Hispanic applicants had a 14.1% lower probability of favorable outcomes.

- **US Citizen**: The most significant bias was observed for non-US citizens with a DI of 0.3406 and SPD of -0.4781, indicating that non-US citizens had a 65.9% lower probability of favorable outcomes compared to US citizens.

- **Age**: The data showed bias against older applicants (age bins 2 and 3) with a DI of 0.8253 and SPD of -0.0556, indicating approximately 17.5% lower probability of favorable outcomes compared to younger applicants.

These baseline metrics highlight significant disparities in acceptance rates across different demographic groups, with citizenship status showing the most pronounced bias.

Neural Network Model Bias

After training our neural network model on the original dataset, we evaluated how the model affected these biases:

Table 6.3: Bias Metrics After Neural Network Training

Protected Attribute	**EOD**	**AOD**	**DI**	**SPD**
Gender	0.2129	0.1323	1.2910	0.0712
Post App. Decision Race	-0.0073	0.0519	1.3502	0.0989
Hispanic	0.0193	-0.0265	0.8418	-0.0545
US Citizen	-0.5363	-0.5977	0.2518	-0.6924
Age Bin	-0.1087	-0.0470	0.9559	-0.0131

where EOD = Equal Opportunity Difference, AOD = Average Odds Difference, DI = Disparate Impact, and SPD = Statistical Parity Difference.

The neural network model demonstrated significant bias amplification for several attributes:

- **Gender**: The model amplified the existing bias by 551.19% in terms of Disparate Impact and 428.40% for Statistical Parity Difference, now strongly favoring male applicants over female and non-binary applicants.

- **Race**: The model increased the bias by 46.70% for Disparate Impact and 45.53% for Statistical Parity Difference, further broadening the disparity between unprivileged and privileged groups (which include 'A', 'C', and 'Asian - Other').

- **US Citizen**: The model further amplified the already substantial bias against non-US citizens by 13.47% for Disparate Impact and 44.82% for Statistical Parity Difference, widening the disparity between citizens and non-citizens.

- **Age**: Interestingly, the model significantly reduced the bias present in the original data by 74.76% for Disparate Impact and 76.48% for Statistical Parity Difference, diminishing age-based disparities between younger applicants and older applicants.

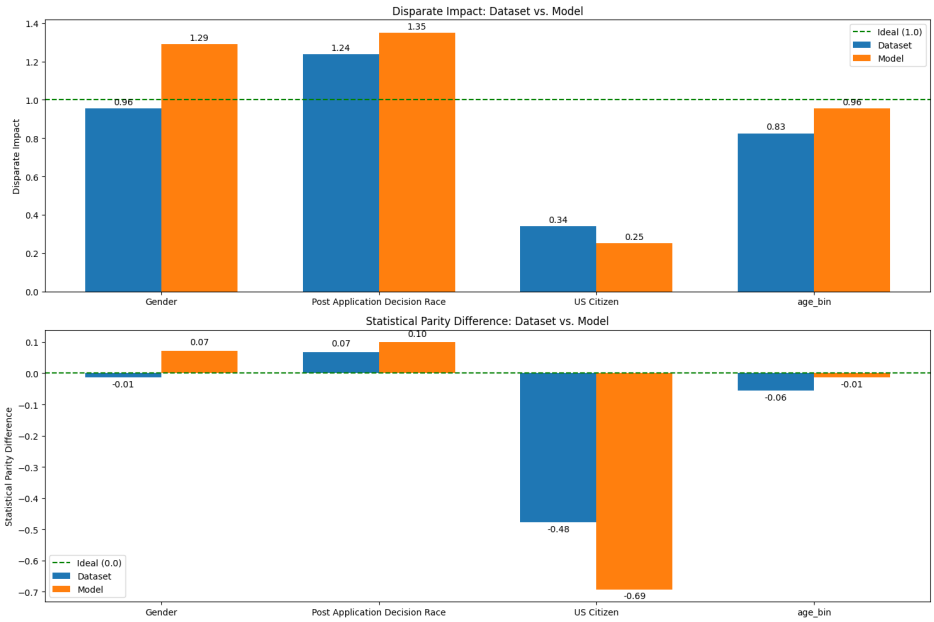


Figure 6.9: Dataset and Model Bias: Disparate Impact and Statisitcal Parity Difference

The Equal Opportunity Difference and Average Odds Difference metrics further confirmed the presence of significant bias in the model's predictions, with particularly concerning values for Gender (0.2129, 0.1323) and US Citizen (-0.5363, -0.5977).

Bias Mitigation

Based on our findings, we implemented a pre and post processing bias mitigation approach targeting the most problematic attributes. Our strategy involved:

1. **Pre-processing mitigation**: We applied Reweighing to the training data, focusing on the most biased attribute (US Citizen).

2. **Post-processing mitigation**: We implemented Calibrated Equalized Odds for each biased attribute separately.

This approach allowed us to address bias at multiple stages of the machine learning pipeline.

Pre-processing Results

Upon reweighing the data for pre-processing mitigation, we re-trained the model on the adjusted dataset and achieved a test accuracy of 81%, a slight improvement from the original model. Applying the Reweighing pre-processing technique to address bias for the US Citizen attribute resulted in the following changes, where EOD = Equal Opportunity Difference,

Table 6.4: Bias Metrics After Pre-processing Mitigation

Protected Attribute	**EOD**	**AOD**	**DI**	**SPD**
Gender	-0.0382	-0.0424	0.7335	-0.0516
Post App. Decision Race	0.1581	0.1570	2.2072	0.1635
US Citizen	-0.8652	-0.8677	0.0785	-0.8940
Age Bin	0.1304	0.1696	2.3531	0.1902

AOD = Average Odds Difference, DI = Disparate Impact, and SPD = Statistical Parity

Difference.

The pre-processing mitigation yielded mixed results, since we preprocessed for the attribute with highest bias detected:

- **Gender**: Pre-processing improved the Equal Opportunity Difference by 0.2511, bringing it closer to the ideal value of 0. However, it worsened the Disparate Impact.
- **Race**: Pre-processing worsened the bias metrics, increasing the Equal Opportunity Difference from -0.0073 to 0.1581 and substantially amplifying the Disparate Impact from 1.3502 to 2.2072.
- **US Citizen**: While attempting to address the significant bias for this attribute, pre-processing actually worsened the Equal Opportunity Difference from -0.5363 to -0.8652 and dramatically reduced the Disparate Impact from 0.2518 to 0.0785.
- **Age**: The pre-processing approach reversed and amplified the bias for age, changing the Equal Opportunity Difference from -0.1087 to 0.1304 and significantly increasing the Disparate Impact from 0.9559 to 2.3531.

These results indicate that pre-processing alone, while effective for certain attributes like Gender, can sometimes exacerbate bias for other attributes, highlighting the challenging nature of multi-attribute bias mitigation. This needs to be further improved by experimenting with various pre-processing bias mitigation strategies and optimizing for scenarios where bias emerges from multiple attributes.

Post Processing Results

We applied the Calibrated Equalized Odds post-processing technique to each biased attribute separately, where EOD = Equal Opportunity Difference, AOD = Average Odds Difference,

Table 6.5: Bias Metrics After Post-processing Mitigation

Protected Attribute	**EOD**	**AOD**	**DI**	**SPD**
Gender	0.2256	0.1415	1.3290	0.0782
Post App. Decision Race	0.0000	0.0000	1.1678	0.0274
US Citizen	-0.5363	-0.5977	0.2518	-0.6924
Age Bin	0.5000	0.3739	∞	0.2835

DI = Disparate Impact, and SPD = Statistical Parity Difference.

The post-processing results showed varying effectiveness when compared to both the original neural network model and the preprocessing mitigation approach:

- **Gender**: After neural network training showed significant bias, preprocessing mitigation made modest improvements. However, post-processing slightly worsened the Equal Opportunity Difference to 0.2256 and Average Odds Difference to 0.1145, though it did further improve Disparate Impact to 0.8299, bringing it closer to the ideal value of 1.0.

- **Race**: The neural network amplified the original bias, and pre-processing mitigation made minimal improvements. Post-processing was remarkably effective, achieving perfect Equal Opportunity Difference (0.0000) and Average Odds Difference (0.0000), while also improving Disparate Impact from 1.3502 to a more equitable 1.1678.

- **US Citizen**: The neural network showed substantial bias, which pre-processing mitigation slightly improved. Post-processing maintained a similar level of bias as the pre-processing approach, offering no significant additional improvement.

- **Age**: While the neural network showed moderate bias, pre-processing mitigation made significant improvements. However, post-processing dramatically worsened the bias metrics, with Equal Opportunity Difference increasing to 0.5000 and Disparate Impact

reaching infinity, indicating extreme bias that completely favors one group.

Overall, our post-processing approach showed inconsistent results across different attributes. It was successful for Race, bringing its fairness metrics to nearly ideal values, but proved ineffective or even detrimental for other attributes. This highlights the challenges of finding a universal bias mitigation approach and suggests that attribute-specific strategies may be necessary for optimal fairness outcomes.

In conclusion, our fairness evaluation demonstrates that there is an amplification of bias from the data to the algorithm. We find that bias mitigation requires careful consideration of multiple approaches tailored to specific protected attributes. Our approach showed the most promising results for racial attributes, while pre-processing alone was more effective for gender and age-related bias, highlighting the need for adaptive mitigation strategies in fair machine learning systems.

6.2.2 SHAP Analysis Results

Neural Network Model

The SHAP analysis reveals that academic metrics exert the strongest influence on the neural network model's predictions. GPA stands out as the most influential factor, with higher values substantially increasing acceptance probabilities, followed by English proficiency scores (IELTS and TOEFL) and GRE results. Age demonstrates mixed effects depending on specific values, while US citizenship status clearly impacts predictions—being a US citizen generally increases acceptance probability, confirming the citizenship bias identified in our fairness evaluation.

Demographic features show varying degrees of influence, with gender and race categories demonstrating smaller but notable impacts. Male gender appears to slightly decrease accep-

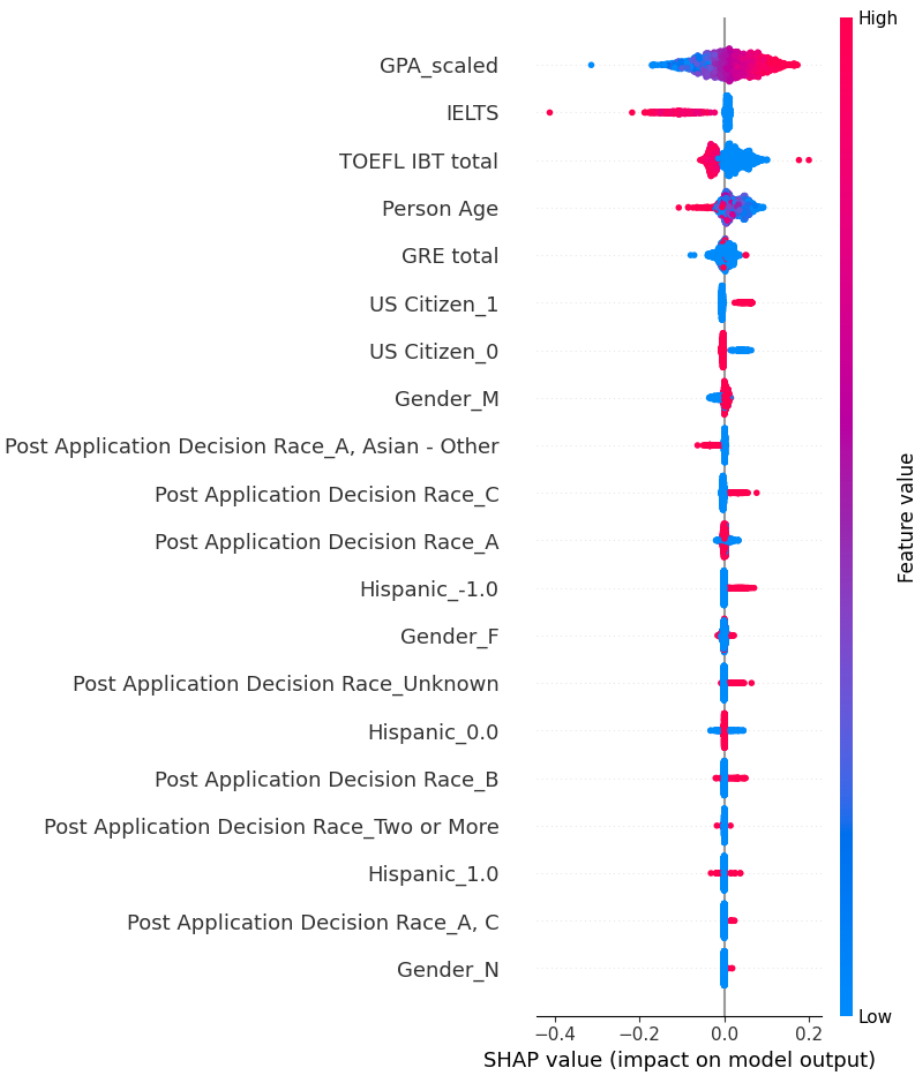


Figure 6.10: SHAP Analysis of NN Model

tance probability compared to other gender categories. Race classifications show diverse effects, with Race_C (Caucasian) having a moderately positive influence and Race_A (Asian) showing a slightly negative impact. Hispanic status indicators demonstrate varying influences, with most having minimal effect. These findings align with our fairness metrics, which identified significant bias related to citizenship status and more moderate bias across other demographic attributes.

Bias Mitigated Neural Network Model

The model from 6.2.2 was retrained after the preprocessing bias mitigation method. We performed further SHAP analysis on this retrained model to observe changes in feature importance and data spread of the values of different attributes. The new model maintains

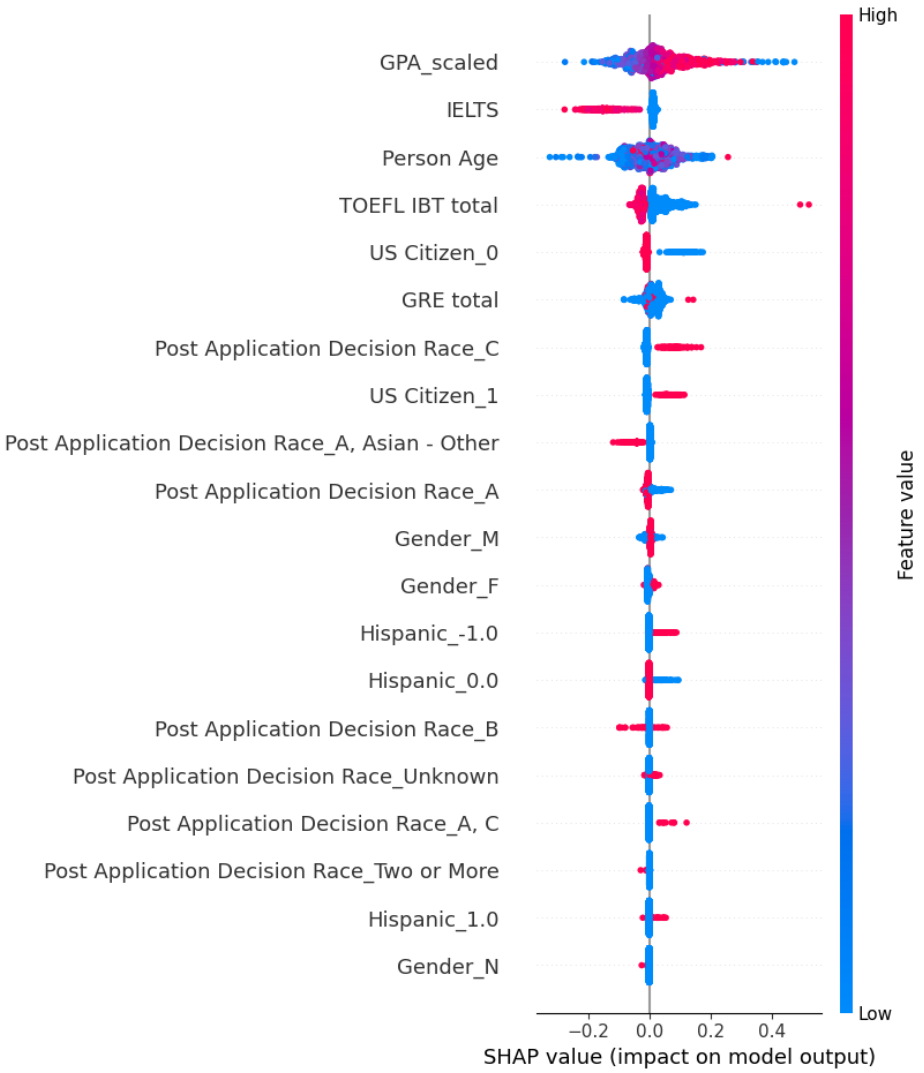


Figure 6.11: SHAP Analysis of Preprocessed NN Model

GPA as the most influential predictor, though with a more balanced distribution of positive and negative impacts compared to the original model. IELTS scores have moved up in impor-

tance, while Person Age has gained significance, now ranking third instead of fourth. Most notably, US Citizen status shows significant changes - US Citizen_0 (non-US citizens) has shifted from having a predominantly negative impact to showing a mix of positive influences, indicating the bias mitigation has partially addressed citizenship-based discrimination.

The race-related features show interesting shifts, with Race_C (Caucasian) gaining importance and now having a stronger positive influence. Gender variables show reduced impact overall, suggesting the preprocessing has diminished gender-based bias. The distribution of SHAP values appears more balanced across demographic features, with narrower spreads and fewer extreme impacts. This aligns with our fairness metrics that showed improvements in bias metrics for several protected attributes, though the visualization confirms that demographic factors still influence predictions even after mitigation techniques were applied.

6.3 Summary

Our comprehensive analysis reveals significant biases within graduate admissions data across multiple demographic attributes. The EDA in phase 1 showed variations in acceptance rates across time periods and demographic groups, with international applications dominating the dataset and racial distributions shifting following policy changes. Clustering analysis identified distinct application groups with varying acceptance rates, while subgroup discovery revealed that US birth nation, Caucasian race, and male gender correlated with higher acceptance probabilities by 2.5 to 2.7 times, whereas Indian birth nation and Asian race correlated with higher rejection probabilities of 1.25 times. Our neural network model amplified existing data biases, particularly regarding gender (551.19% increase in Disparate Impact) and citizenship status (13.47% increase). Bias mitigation techniques yielded mixed results: preprocessing was effective in reducing equal opportunity difference for gender bias

and disparate impact for the U.S citizen attribute, while post-processing achieved near-perfect fairness metrics for race but worsened age-related bias. SHAP analysis confirmed that academic metrics (GPA, IELTS, TOEFL scores) exerted the strongest influence on predictions, with citizenship status also significantly impacting outcomes. After preprocessing bias mitigation, the model maintained similar feature importance rankings but showed more balanced distributions across demographic attributes, suggesting partial success in addressing demographic-based discrimination. Overall, it is evident that our bias mitigation approaches are moderately effective on the admissions prediciton model, but require more experimentation to achieve fairness.

Chapter 7

Discussion

7.1 Potential for Data Bias

Through various experiments, we explored RQ 1 and demonstrated that several features could contribute to the data bias of a machine learning model. The feature importance ranking with the Random Forest classifier in 5.2.3 rendered demographic features like age, birth nation, and race as highly important features. This indicates that the model's generalized algorithm for deciding if a particular application should be granted admission factors in these features with a high weightage. Though a deep learning model or a neural network may improve performance, the Random Forest classifier offers interpretability, which is paramount when modeling human-centric processes. The density-based clustering 5.2.1 reveals results that coincide with our findings from the feature importance ranking, such as the birth nation India having a higher tendency to be rejected and the birth nation United States having a higher tendency to be accepted. Another finding from the clustering was that 'Age' had a negative correlation to the decision for some clusters, which aligns with our finding from the feature importance ranking that age had an importance of around 0.17. This implies that age is another feature that may be used for decision-making by the machine learning model.

Using subgroup discovery 5.2.2 for intersectional bias detection (RQ 2), we found various

permutations of birth nation being India, gender being male or female, and race being Asian displaying bias towards an application being rejected. This augments our findings from 5.2.1 and 5.2.3, where experiments revealed the same demographic features as having a higher likelihood of being rejected. Likewise, the subgroup discovery yielded permutations of the birth nation being United States or China, gender being male and race being Caucasian or Asian to have a higher likelihood of being accepted. In a discussion detailed in [17], the author mentions that members of the admission review committee are highly skilled at triangulation - a process in which they can easily determine if the characteristics of an applicant remain consistent throughout different application materials, and are therefore able to tactfully make decisions on the student's potential success in university. This means that while reviewers might be looking for certain specific qualities, they do not process applications by strictly adhering to an algorithm, which allows for flexibility of human judgment based on context. However, as seen in the case of GRADE [52], machine learning models do not have a similar ability, and might harm diversity by measuring the success of students based directly on historical admissions trends of accepted students. Therefore, it is critical to be aware of the potential biases that a model may infer from the dataset and rectify these before the model is applied in practice.

7.2 Bias or a Reflection of the Applicant Population

The experiments in 5.2.1 demonstrated different approaches to examine the dataset for biases that may be inferred by a machine learning model. However, we must also note that many of these biases stem from distributions in the dataset of historical admits and rejects of applications. This raises the question of whether these are truly biased outcomes or simply a reflection of the data itself. For example, the negative class, i.e., the decision being 'rejected', was often biased towards applications having birth nation India. Though

the data indicates a bias, this does not necessarily imply that the admission reviewers will reject applications on the basis of the birth nation being India. Figure 6.3 illustrates that the dataset contains around 4000 observations falling in the category of birth nation India, which accounts for nearly 43% of the dataset. Since the dataset itself is not evenly distributed and certain demographic qualities exist in larger proportions in the dataset, there is a higher tendency for the model to learn these as decision rules, thereby leading to biased outcomes. Nonetheless, Figure 6.1 also shows us that the acceptance rate is higher for applications from the United States and China, despite the total number of applications being lower. This concurs with findings from our analysis that there exists a bias in the model (RQ 3) for applications based on citizenship towards the 'accepted' class.

We must also note that the dataset is from a university in the United States, which may be a reason that there is a higher acceptance rate for applicants from the United States, since there may be students who were previously enrolled in the same university or other well-recognized universities in relevant programs. If the results of our bias detection experiments are a reflection of the data, then one might ask why a machine learning model learning the same patterns might be problematic. This is because human reviewers make multiple considerations based on the unique attributes of each application and do not adhere to a specific formula for making decisions despite using a rubric of minimum requirements that indicate the applicant's potential to succeed in the program. On the contrary, the machine learning algorithm learns patterns in the data as rules, which determine how it will classify observations in the future. This claim is demonstrated through our experiments in phase 2, where we find that the machine learning model further exacerbates the bias in the data as algorithmic bias (RQ 3). When such models are deployed in society, they can produce unfair outcomes and potentially harm the diversity of the graduate student population. Therefore, it is imperative to examine the data for potential biases that a model might infer

and minimize the risk of reduced diversity caused by a potentially biased model. Once a model is developed, universities could also apply explainability methods as done in 5.3.5 to uncover the biases learned by the machine learning model during training and seek methods to rectify them through bias correction measures. Our work demonstrated this process in section 5 and found that demographic attributes were indeed influencing model decisions at a high level of importance, particularly features like age and citizenship (RQ 4).

7.3 The Cycle of Bias

This study focused on identifying the various sources of bias in automating admission reviews and explored some fairness evaluation and mitigation approaches. Throughout the study, we demonstrate how bias is a complex problem that cannot be solved unilaterally and requires interventions from multiple stages and dimensions.

In our fairness evaluation, we demonstrated how bias can exist in the dataset itself due to historical data containing human bias in selecting candidates, which then seeps into the dataset upon which a model is trained. The model infers this bias from the dataset and may additionally develop its own algorithmic biases which could come from the model's learning ability and architecture. When these models are deployed in society without much attention to bias mitigation, they may further propagate biased outcomes, which feeds into a vicious cycle of bias in the data and the algorithm. The decisions made by such models may once again be used as historical data for retraining the model in the future, creating an endless loop of bias.

Therefore, it is critical to consider the various stages in applied machine learning in the context of admissions where bias can emerge or be amplified and perform appropriate interceptive methods to mitigate this bias. It is also important to ensure that it is a continuous

development process, where machine learning models employed for socio-technical processes are re-evaluated for bias periodically and calibrated to ensure fairness.

7.4 Impact of the Affirmative Action Ban

Affirmative action was introduced in the early 2000s in the United States as a national initiative to address historical injustices in the lives of women and ethnic minorities by guaranteeing them some advantage in college admissions and employment possibilities. Though its historical significance was to provide more opportunities for underrepresented communities to excel in academia and industry, it did not imply that universities would explicitly use race as a criterion for scoring applications but rather use it as an additional characteristic after consolidating already highly qualified applicants, within constitutional limits [31]. In 2023, the Supreme Court ruled against affirmative action practices and declared that universities may not conduct race-conscious admission reviews [54]. Preliminary effects of this decision have already been analyzed for various universities based on their published 2024 admissions data [12, 33, 48]. A study published in 2024 analyzed the enrollment changes for the undergraduate class of 2028 across different U.S. universities that previously practiced race-conscious admissions [32]. They provide fair warning that these trends may be the typical changes that occur year-on-year with the admissions cycles and therefore it is not conclusive if they are a result of the affirmative action ban. Nonetheless, inferences can be made about the change in the class population of URMs. A vast majority of the universities in the study do reflect a decline in the number of Black and Hispanic students, with a few universities having contrasting outcomes.

Our findings from 5.2.1 indicate similar changes in our graduate admissions dataset, where Black and Hispanic students' applications have reduced in the 2024 admissions cycle for

the MS program compared to 2023. Since the applications from the American Indian and Alaskan Native populations are already low, we could not draw significant conclusions for these groups in the 2024 cycle. A positive observation is that the rate of change year on year has increased from 2023 to 2024 by 15% in the case of Black PhD applicants. We observed a significant increase in the number of applicants opting out of reporting race in 2024, categorized as 'unknown'. Though this had a pre-existing increasing trend for applications from within the US, we observed a sudden increase among international applicants. This aligns with reports [44] that an increasing number of students are choosing to not disclose their race after the affirmative action ban, especially in highly competitive universities. In conclusion, with respect to RQ 5, we observed trends in our dataset that concur with observations from across competitive universities in the US that there is a common hesitation to report race and a decline in URM applicants. However, it remains premature to make strong claims with a single year of data post the affirmative action ban and will be more significant after a few years of observation.

7.5 Opportunities to Increase Efficiency and Improve Diversity

With the affirmative action ban in place and based on the findings in this study regarding the use of a machine learning model to efficiently conduct admission reviews, we identified two main issues concerning diversity for universities to consider: a. demographic composition of the applicant pool and b. demographic composition of the admitted population. We propose that universities tackle both these issues in order to ensure that diversity is facilitated in their programs. The first issue may be tackled by conducting an increased number of inclusivity initiatives that encourage applicants from URMs to apply, along with bridge courses and

workshops that may provide support for students with a poor academic history or lack of access to resources for standardized tests to boost their profile. Additional financial support initiatives may also encourage students of these groups to apply. When applying machine learning models for admissions processing, the second issue can be tackled by thoroughly investigating data bias from data used as input to the machine learning models by using a framework similar to ours shown in 3.1. They may also conduct post-modeling analysis by following our framework in 3.2 and implement bias correction methods to ensure that machine learning models do not just repeat history and instead give fair consideration to all applicants, especially those from URMs. It is also critical to develop these models with incremental improvements with human evaluation or develop a human-in-the-loop system such that the model decisions are validated by a reviewer.

Chapter 8

Limitations and Future Work

Bias in the admissions process is a challenging problem to tackle, as it can emerge in various stages as shown in 2.1. Studying bias in the context of graduate admissions data, for graduate computer science programs in our case, brings added complexity. Previous studies have often focused on bias in machine learning for admission in undergraduate data, which tends to be less complex than graduate data due to the uniformity in application details since they are often derived through the Common App [1]. Since most applicants apply from within the United States, they have a similar educational background, without different GPA scales or standardized testing requirements based on country of origin. This leads to more consistent and less sparse data when considering undergraduate applications. Graduate applications, however, may require additional essays and standardized tests that are required for a specific subset of applicants or are optional. In this work, we focused solely on non-essay data and may not have considered valuable information such as hidden demographic attributes from essays in our fairness evaluation.

Feature distributions and admission review criteria also differ vastly between undergraduate and graduate data. For example, age is usually consistent across undergraduate applications since most prospective students apply immediately after the completion of high school, but the same feature has a much higher variance in graduate data since applicants may pursue graduate education directly after their undergraduate education or at any stage of their ca-

reer. Graduate admissions data also has a significantly lower volume of data per admissions cycle, owing to its significantly lower intake compared to undergraduate programs. In addition to this, the process of admission review varies not only between different universities but also between the undergraduate and graduate programs in the same university. Undergraduate applications are typically reviewed centrally by the university whereas graduate admission review may be conducted by a specific department's professors and staff since essays can be specific to the field. Therefore, it is difficult to generalize decision-making criteria as application reviews may be unique to each department. Our experiments focused on the dataset we acquired from one university and one department, and our findings are applicable to this specific dataset and not generalizable across all universities.

One of the challenges in this work was fairness evaluation. Defining protected attributes is subjective to each context and the type of fairness being considered. For example, we may consider different metrics when analyzing group vs. individual fairness and when analyzing independent vs. intersectional fairness. Therefore, it is a complex and highly context-dependent process and can require domain-specific knowledge to determine the protected attributes, along with the privileged and underprivileged groups. De-biasing approaches are dependent on the fairness metrics we are evaluating, and can also vary drastically based on the context. In this study we had limited the work to pre-processing and post-processing approaches to tackle data and algorithmic bias, but there several other approaches such as in-processing mitigation and proxy muting, as described in 4.3, whose effectiveness we have not evaluated.

Our machine learning model achieved an accuracy of approximately 74%, which is similar to some fo the previous works, but can be improved with more focus and experimentation in the model development phase. Changing the training parameters or model architecture may yield improved accuracy and more robust models.

This study discusses the potential bias that may be inferred from the data by a machine learning model, and evaluates for bias and fairness in the model but was limited in that it only analyzed non-essay data. Future work could include the various essay texts in the analyses as input data for the machine learning model, much like the human review process would. This might reveal further potential biases hidden in text data [7]. Since admissions data is typically confidential and not available online, we limited our study to the one dataset we were able to access. Future work could apply our framework to different datasets and assess how effectively bias can be detected and mitigated using the same framework. It would also be interesting to explore if the results on different datasets align with our findings in this work. The study could also be extended by training multiple machine learning models and analyzing their outcomes to validate if the determined biases exist in the outcomes. Another direction of research could explore fairness assessment suites apart from the IBM AIF360 used in this study and attempt to mitigate bias through different pre, in and post processing approaches, which may provide an additional range of metrics that can help assess fairness and bias in the model outcomes.

Chapter 9

Conclusion

The increasing number of graduate CS applications necessitates AI-based admissions processing to increase efficiency and allow universities to allocate limited resources towards other critical tasks. While such solutions benefit universities immensely, they also present risks to diversity due to models training on historical data and perpetuating bias. This concern is particularly acute following recent policy changes surrounding affirmative action in university admissions.

With respect to RQ1, our feature importance analysis identified that academic metrics, particularly GPA, had the strongest influence on prediction outcomes, while demographic features such as citizenship status also significantly impacted model decisions. The SHAP analysis confirmed that higher GPA values strongly contributed to positive prediction outcomes, with language proficiency scores (IELTS and TOEFL) following in importance. Addressing RQ2, our subgroup discovery technique uncovered significant intersectional bias effects, particularly for applicants from India, and combinations of gender and race categories, with lift values up to 2.67 for certain intersections.

Our fairness evaluation framework (RQ3) revealed substantial bias amplification in the neural network model from the baseline bias found in the dataset, particularly for gender and citizenship status. We demonstrated that a using preprocessing (Reweighing) and post-processing (Calibrated Equalized Odds) techniques could significantly improve fairness met-

rics for some protected attributes, though effects varied across different protected attributes. For RQ4, our explainability analysis showed that citizenship status had particularly strong influence on the model's decisions before mitigation, with non-US citizens facing substantial disadvantage, while after mitigation, the impact of demographic features became more balanced, though still present.

Our analysis revealed that trends post-affirmative action ban (RQ 5) showed a decline in applications from Black and Hispanic students for MS programs, aligning with broader national trends. For PhD programs, we observed an increase in Black applicants, though this was not reflected in acceptance rates. Notably, we found a significant 66% increase in applicants opting out of reporting race from 2023 to 2024, indicating widespread diversity concerns following the ban. However, with only one year of post-ban data, these conclusions are preliminary and will need data collected through extended periods to confirm the effects of the ban.

The key contribution of this work is our comprehensive two-phase methodology for bias management in admissions systems. Phase 1 enables preemptive bias detection through exploratory analysis, clustering, and subgroup discovery before model deployment, while Phase 2 provides a framework for fairness evaluation and bias correction using both preprocessing and postprocessing approaches coupled with explainability tools. We also provide a novel analysis of bias amplification, to detect the increase in bias from the dataset to the algorithm.This systematic approach allows institutions to understand potential sources of algorithmic bias and implement targeted mitigation strategies. Future work should incorporate textual application data into the bias detection framework, experiment with multiple datasets and explore more sophisticated mitigation techniques for multi-attribute fairness. Our research highlights the critical importance of proactive bias detection and mitigation in AI-based admissions systems to ensure diverse and equitable outcomes in higher education.

Bibliography

[1] The common application. URL https://www.commonapp.org/.

[2] Common fairness metrics. URL https://fairlearn.org/main/user_guide/assessment/common_fairness_metrics.html#assessment-four-fifths.

[3] Comparing python clustering algorithms — hdbscan 0.8.1 documentation. URL https://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html.

[4] Coe - graduate degree fields. URL https://nces.ed.gov/programs/coe/indicator/ctb.

[5] 8 in 10 colleges will use ai in admissions by 2024, September 2023. URL https://cset.georgetown.edu/article/levers-for-improving-diversity-in-computer-science/.

[6] AI Fairness 360. average_odds_difference - AI Fairness 360 Toolkit. https://aif360.readthedocs.io/en/stable/modules/generated/aif360.sklearn.metrics.average_odds_difference.html, 2024.

[7] A.J. Alvero, Noah Arthurs, Anthony Lising Antonio, Benjamin W. Domingue, Ben Gebre-Medhin, Sonia Giebel, and Mitchell L. Stevens. Ai and holistic review: Informing human reading in college admissions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 200–206, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375871. URL https://doi.org/10.1145/3375627.3375871.

[8] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018. URL https://arxiv.org/abs/1810.01943.

[9] Sarah Bird, Miroslav Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, Kathleen Walker, and Allovus Design. Fairlearn: A toolkit for assessing and improving fairness in ai. 2020.

[10] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

[11] Kelly Van Busum and Shiaofen Fang. Analysis of ai models for student admissions: A case study. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pages 17–22. ACM, March 2023. doi: 10.1145/3555776.3577743. URL https://dl.acm.org/doi/10.1145/3555776.3577743.

[12] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, page 160–172, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-37456-2.

[13] Cole Claybourn. Is ai affecting college admissions?, 2023. URL https://www.usnews.com/education/best-colleges/articles/is-ai-affecting-college-admissions.

[14] Janice Cuny and William Aspray. Recruitment and retention of women graduate students in computer science and engineering: results of a workshop organized by the computing research association. *ACM SIGCSE Bulletin*, 34(2):168–174, June 2002. ISSN 0097-8418. doi: 10.1145/543812.543852. URL https://dl.acm.org/doi/10.1145/543812.543852.

[15] Janice Daniel. Diversifying graduate student enrollment: What we know and what we're learning. URL https://www.aaaspolicyfellowships.org/blog/diversifying-graduate-student-enrollment-what-we-know-and-what-were-learning.

[16] Sergio Escalera, Xavier Baró, Hugo Jair Escalante, and Isabelle Guyon. Chalearn looking at people: Events and resources. *CoRR*, abs/1701.02664, 2017. URL http://arxiv.org/abs/1701.02664.

[17] Ellen Evaristo. Balancing the potentials and pitfalls of ai in college admissions. URL https://rossier.usc.edu/news-insights/news/balancing-potentials-and-pitfalls-ai-college-admissions.

[18] Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, December 2023. ISSN 2413-4155. doi: 10.3390/sci6010003. URL http://dx.doi.org/10.3390/sci6010003.

[19] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel. "how biased are your features?": Computing fairness influence functions with global sensitivity analysis. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 138–148, New York, NY, USA, 2023. Association for Com-

puting Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3593983. URL https://doi.org/10.1145/3593013.3593983.

[20] Kate Gibson. Unpacking the earning potential of a graduate degree, 2024. URL https://graduate.northeastern.edu/knowledge-hub/earning-potential/.

[21] Xavier Gitiaux and Huzefa Rangwala. Multi-differential fairness auditor for black box classifiers. *CoRR*, abs/1903.07609, 2019. URL http://arxiv.org/abs/1903.07609.

[22] E. Grieco. Diversity and stem: Women, minorities, and persons with disabilities 2023. URL https://www.nsf.gov/reports/statistics/diversity-stem-women-minorities-persons-disabilities-2023.

[23] Sumyea Helal. Subgroup discovery algorithms: A survey and empirical evaluation. *Journal of Computer Science and Technology*, 31(3):561–576, May 2016. ISSN 1860-4749. doi: 10.1007/s11390-016-1647-1.

[24] Ghazal Kalhor, Tanin Zeraati, and Behnam Bahrak. Diversity dilemmas: uncovering gender and nationality biases in graduate admissions across top north american computer science programs. *EPJ Data Science*, 12(1):44, September 2023. ISSN 2193-1127. doi: 10.1140/epjds/s13688-023-00422-5. URL http://arxiv.org/abs/2302.00589.

[25] Faisal Kamiran and Toon Calders. Data pre-processing techniques for classification without discrimination. *Knowledge and Information Systems*, 33, 10 2011. doi: 10.1007/s10115-011-0463-8.

[26] Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek, and Sunil Aryal. The pursuit of fairness in artificial intelligence models: A survey, 2024. URL https://arxiv.org/abs/2403.17333.

[27] Florian Lemmerich and Martin Becker. pysubgroup: Easy-to-use subgroup discovery in python. In Ulf Brefeld, Edward Curry, Elizabeth Daly, Brian MacNamee, Alice Marascu, Fabio Pinelli, Michele Berlingerio, and Neil Hurley, editors, *Machine Learning and Knowledge Discovery in Databases*, page 658–662, Cham, 2019. Springer International Publishing. ISBN 978-3-030-10997-4.

[28] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[29] Thomas Lux, Randall Pittman, Maya Shende, and Anil Shende. Applications of supervised learning techniques on undergraduate admissions data. In *Proceedings of the ACM International Conference on Computing Frontiers*, CF '16, page 412–417, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341288. doi: 10.1145/2903150.2911717. URL https://doi-org.ezproxy.lib.vt.edu/10.1145/2903150.2911717.

[30] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL https://doi.org/10.1145/3457607.

[31] Donald Mitchell, Jr. and Elizabeth A. Daniele. Diversity in american graduate education admissions: Twenty-first-century challenges and opportunities. *Executives, Administrators, & Staff Publications*, (31), 2015. URL https://digitalcommons.molloy.edu/eas_pub/31.

[32] James Murphy. Tracking the impact of the sffa decision on col-

lege admissions, . URL https://edreformnow.org/2024/09/09/tracking-the-impact-of-the-sffa-decision-on-college-admissions/.

[33] James Murphy. What happened to campus diversity post-sffa? five findings, . URL https://edreformnow.org/2024/10/17/what-happened-to-campus-diversity-post-sffa-five-findings/.

[34] Barbara Martinez Neda and Sergio Gago-Masague. Feasibility of machine learning support for holistic review of undergraduate applications. In *2022 International Conference on Applied Artificial Intelligence (ICAPAI)*, pages 1–6. IEEE, May 2022. doi: 10.1109/ICAPAI55158.2022.9801571. URL https://ieeexplore.ieee.org/document/9801571.

[35] Barbara Martinez Neda, Yue Zeng, and Sergio Gago-Masague. Using machine learning in admissions: Reducing human and algorithmic bias in the selection process. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, SIGCSE '21, page 1323, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380621. doi: 10.1145/3408877.3439664. URL https://doi.org/10.1145/3408877.3439664.

[36] Daniel B. Neill, Edward McFowland III, and Huanian Zheng. Fast subset scan for multivariate event detection. *Statistics in Medicine*, 32(13):2185–2208, 2013. doi: https://doi.org/10.1002/sim.5675. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5675.

[37] National Centre of Educaiton Statistics. Degrees in computer and information sciences conferred by postsecondary institutions, by level of degree and sex of student: Academic years 1964-65 through 2021-22, 2021. URL https://nces.ed.gov/programs/digest/d23/tables/dt23_325.35.asp.

[38] Alejandro Peña, Ignacio Serna, Aythami Morales, and Julian Fiérrez. Bias in multimodal AI: testbed for fair automatic recruitment. *CoRR*, abs/2004.07173, 2020. URL https://arxiv.org/abs/2004.07173.

[39] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3), February 2022. ISSN 0360-0300. doi: 10.1145/3494672. URL https://doi-org.ezproxy.lib.vt.edu/10.1145/3494672.

[40] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. *CoRR*, abs/1709.02012, 2017. URL http://arxiv.org/abs/1709.02012.

[41] Amisha Priyadarshini, Barbara Martinez-Neda, and Sergio Gago-Masague. Admission prediction in undergraduate applications: an interpretable deep learning approach. In *2023 Fifth International Conference on Transdisciplinary AI (TransAI)*, pages 135–140, September 2023. doi: 10.1109/TransAI60598.2023.00040. URL http://arxiv.org/abs/2401.11698.

[42] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi-org.ezproxy.lib.vt.edu/10.1145/2939672.2939778.

[43] Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *CoRR*, abs/1811.05577, 2018. URL http://arxiv.org/abs/1811.05577.

[44] Zachary Schermele. At selective colleges, fewer students are disclosing race in their applications. *USA TODAY*. URL https://www.usatoday.com/story/news/education/2024/10/21/affirmative-action-ban-admissions-effect-2024/75699203007/.

[45] Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. *Towards a standard for identifying and managing bias in artificial intelligence*. Number NIST SP 1270. National Institute of Standards and Technology (U.S.), Gaithersburg, MD, March 2022. doi: 10.6028/NIST.SP.1270. URL https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf.

[46] Gizem Sogancioglu, Heysem Kaya, and Albert Ali Salah. Using explainability for bias mitigation: A case study for fair recruitment assessment. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, page 631–639, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700552. doi: 10.1145/3577190.3614170. URL https://doi.org/10.1145/3577190.3614170.

[47] Sriram Vasudevan and Krishnaram Kenthapadi. Lift: A scalable framework for measuring fairness in ml applications. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2773–2780, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3412705. URL https://doi-org.ezproxy.lib.vt.edu/10.1145/3340531.3412705.

[48] Patricia Waldron. Race-blind college admissions harm diversity without improving quality. URL https://news.cornell.edu/stories/2024/11/race-blind-college-admissions-harm-diversity-without-improving-quality.

[49] John Wamburu, Girmaw Abebe Tadesse, Celia Cintas, Adebayo Oshingbesan, Tanya Akumu, and Skyler Speakman. Systematic discovery of bias in data. In *2022 IEEE*

International Conference on Big Data (Big Data), pages 4719–4725, 2022. doi: 10.1109/BigData55660.2022.10020781.

[50] Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 336–349, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533101. URL https://doi-org.ezproxy.lib.vt.edu/10.1145/3531146.3533101.

[51] Yanchen Wang and Lisa Singh. Mitigating demographic bias of machine learning models on social media. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400703812. doi: 10.1145/3617694.3623244. URL https://doi.org/10.1145/3617694.3623244.

[52] Austin Waters and Risto Miikkulainen. Grade: Machine-learning support for graduate admissions. *AI Magazine*, 35(1):64–75, 2014. ISSN 2371-9621. doi: 10.1609/aimag.v35i1.2504. URL https://onlinelibrary.wiley.com/doi/abs/10.1609/aimag.v35i1.2504.

[53] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viegas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, page 1–1, 2019. ISSN 2160-9306. doi: 10.1109/tvcg.2019.2934619. URL http://dx.doi.org/10.1109/TVCG.2019.2934619.

[54] Sarah Wood. How does affirmative action affect college admissions? *U.S.*

News & World Report. URL https://www.usnews.com/education/best-colleges/applying/articles/how-does-affirmative-action-affect-college-admissions.

[55] Yijun Zhao, Zhengxin Qi, John Grossi, and Gary M. Weiss. Gender and culture bias in letters of recommendation for computer science and data science masters programs. *Scientific Reports*, 13(1):14367, September 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-41564-w. URL https://www.nature.com/articles/s41598-023-41564-w.