

PONDICHERRY UNIVERSITY

(A Central University)



SCHOOL OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

MASTER OF COMPUTER SCIENCE

NAME : M. ARUL

REG. NO. : 19384102

SEMESTER : 10TH

SUBJECT : MAIN PROJECT REPORT

GUIDED BY : Dr. T. CHITHRALEKHA

RESPONSIBLE AI

by

ARUL.M

(Registration Number: 19384102)

**Project report submitted in partial fulfilment of the
requirements for the award of the degree of**

MASTER OF COMPUTER SCIENCE



**DEPARTMENT OF COMPUTER SCIENCE
SCHOOL OF ENGINEERING TECHNOLOGY**

PONDICHERRY UNIVERSITY-605014

MAY 2024

**DEPARTMENT OF COMPUTER SCIENCE SCHOOL
OF ENGINEERING AND TECHNOLOGY**

PONDICHERRY UNIVERSITY: PUDUCHERRY 605014

BONAFIDE CERTIFICATE

This is to certify that this project work entitled "**RESPONSIBLE AI**" is a bonafide record of work done by **Mr. M. ARUL, Reg. No. 19384102**, in partial fulfillment of the requirements for the Degree of **Integrated Master of Science** in Computer Science in the Department of Computer Science, School of Engineering and Technology of Pondicherry University.

This work has not been submitted elsewhere for the award of any other degree to the best of our knowledge.

GUIDE

Dr. T. CHITHRALEKHA

Professor

Department of Computer Science

School of Engineering & Technology

Pondicherry University

HEAD OF THE DEPARTMENT

Dr. S.K.V. JAYAKUMAR

Professor & Head

Department of Computer Science

School of Engineering & Technology

Pondicherry University

Submitted for the Viva-Voce Examination held on _____

INTERNAL EXAMINER

EXTERNAL EXAMINER

Date: 29/05/2024

Internship Certificate

This is to certify that:

Mr. Arul Murugan
MSc, Pondicherry University
Puducherry

was a Bonafide Intern at Intrust Innovation Labs from March 2024 to August 2024.

During this internship period, Arul actively participated in various projects and activities in the AI CoE department. Their role included:

- Research & Developing PoC on Responsible AI

This certificate is issued at the request of Arul for the purpose of fulfilling academic requirements related to the internship project undertaken during their tenure at Intrust Innovation Labs.

We acknowledge Arul's contribution to the company and wish them all the best in their future endeavors.

Best Regards,



Sahida Begame N
Head – Human Resource



INTRUST INNOVATION LABS PRIVATE LIMITED

01 FA, IIT Madras Research Park, Kanagam Road, Taramani, Chennai – 600 113, India
Tel: +91-9944981398 Email: info@intrustlabs.com Website: www.intrustlabs.com

TABLE OF CONTENT

TITLE	PAGE NO
ACKNOWLEDGEMENT	1
ABSTRACT	2
INTRODUCTION	3-6
ABOUT THE PROJECT	7
PROBLEM DEFINITION	8
FEASIBILITY ANALYSIS	9-10
SOFTWARE REQUIREMENT SPECIFICATION	11
FAIRNESS AND PRIVACY & SECURITY EVALUATION	12-18
SYSTEM DESIN	19-22
SCREENSHOTS	23-31
CONCLUSION	32
FUTURE WORK	33
REFERENCES	34

ACKNOWLEDGEMENT

Every project, big or small, is successful largely due to the effort of several wonderful people who have always given their valuable advice or lent a helping hand. I sincerely appreciate the inspiration, support, and guidance of all those people who have been instrumental in making this project a success.

I express my heartfelt gratitude to my Project Guide **Dr. T. CHITHRALEKHA**, Professor, Department of Computer Science, Pondicherry University, for her whole - hearted assistance and direction not only for the duration of the project but for the entire duration of the course. I will always remain grateful to her and whose constant care about me has provided a new direction to work.

I express my gratitude to **Dr. S.K.V. JAYAKUMAR**, Professor and Head, Department of Computer Science, Pondicherry University, for his support and arranging the project in a good schedule.

Finally, I would like to express my regards for all the faculty members of Department of Computer Science and others involved in this project, directly or indirectly.

- **ARUL. M**

ABSTRACT

Responsible AI (RAI) is a web application that allows users to understand the core principles which govern the use of Artificial Intelligence in a responsible manner. This application enables users to understand the dataset on which the AI model is built. There are 6 core principles to be verified in order to access whether the dataset complies to RAI. However, in this project only Fairness and Privacy & Security are verified in the dataset.

Fairness enables users to analyze their AI models for potential biases, ensuring that the outcomes are fair across diverse user groups. By evaluating fairness of AI systems, the application helps prevent discriminatory practices and promotes fairness in AI operations.

The privacy and security principle ensure protecting individual data. It provides mechanisms to secure sensitive data and ensures that AI operations are up to privacy standards and regulation. This feature is crucial for maintaining trust and safeguarding user data against unauthorized access and breaches.

INTRODUCTION

Responsible AI is about using artificial intelligence (AI) in ways that are fair, transparent, inclusive, reliable & safety, privacy & security and accountable. This is because it is important to make sure that AI systems do not unfairly favor certain groups of people or make decisions that are hard to understand. Responsible AI is essential for building trust in AI technology, by ensuring fairness, transparency, and accountability in how AI systems are developed and used. Thereby we can use the full potential of AI while minimizing its risks. Through collaboration and commitment to RAI principles, we can make the way for a future where AI benefits everyone.

Responsible AI starts with how we collect and use data. Care should be taken to avoid biases and making sure that the data we use is accurate and representative of all people. It also involves designing AI system in a way that makes their decisions understandable and predictable. Building responsible AI requires collaboration and collectiveness between different groups, including developers, policymakers, and communities affected by AI systems. By working together, we can create guidelines and standards that ensure AI is used ethically and in the best interest of society.

Responsible AI is concerned specifically with establishing ethical principles and human values to reduce biases and promote fairness, offer interpretability and explainability of outcomes, and to ensure robustness and security. The goal of building AI technologies based on Responsible AI principles helps to avoid negative consequences on human and societal well being.

Responsible AI Core Principles [2]:

- Fairness
- Inclusiveness
- Reliable & Safety
- Transparency
- Privacy & Security
- Accountable

Fairness:

AI can potentially create consequences such as biases, discrimination, errors that may lead to unexpected results. AI should not lead to discriminatory impacts on people in relation to race, native origin, religion, gender, sexual orientation, disability, or any other personal condition.

Biases can be present in the training data, from labeling or uneven distribution through under- or over-sampling, can result in models getting undesired biases.

By integrating fairness metrics into the model-building process to understand regarding whether, and how much, different groups of people are impacted when using the model to make predictions. We can apply constraints in the algorithm so that the model is bound to satisfy with certain fairness standards.

Inclusiveness:

Inclusiveness mandates that AI should consider all humans races and experience. Determine design practices that can help developers to understand and address potential barriers that could intentionally exclude people from accessing the AI model.

Reliable & safety:

Reliability ensures that the AI system will consistently be able to produce the right outcomes, with the same satisfactory performance under various conditions. This can be achieved through testing, verification, and quality assurance to reduce errors and uncertainties.

AI system should be volatile to adversarial attacks, errors, and unforeseen circumstances. This includes testing AI systems security across various scenarios and ensuring they do not cause harm to the users or society.

Reliability and safety in AI development is key factor, ensuring that technologies can be trusted to work properly lowering concern among users.

Transparency:

Transparency in AI involves openness in how AI systems operate, make decisions, and affect users and the society by explaining data used, how the algorithms are used. Transparency helps users understand and trust the AI systems.

Privacy & Security:

This principle ensures safeguarding individual's privacy and ensuring the AI system to handle personal data in a responsible and proper manner. Implementing measures to protect sensitive information and ensure relevant data protection regulation. Both privacy and security mechanisms together ensure the confidentiality, and reliability of AI systems.

Accountable:

Accountable ensures being answerable and responsible for the decisions, actions, and impacts of AI systems. It involves ensuring that individuals, organizations, or systems involved in use of AI technologies can be held accountable for their actions and the consequences of AI-driven outcomes. This includes establishing clear responsibility, implementing error handling mechanisms, and providing route assistance in case of errors, biases, or harm caused by AI systems.

DIAGRAM:

The below diagram explains the RAI principles in a nutshell in the form of a mind map from the above detailed explanation.

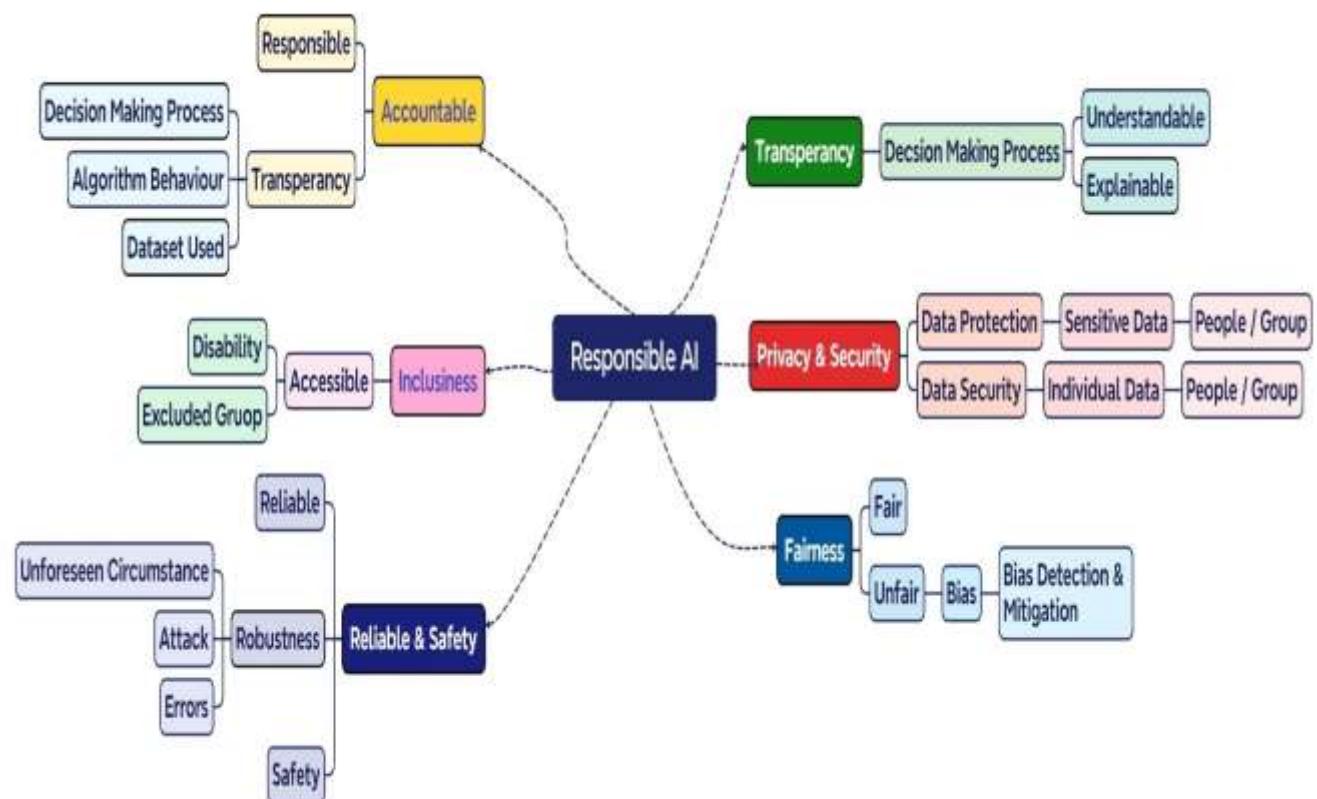


Figure: Represents breakdown of RAI principles in the form of mind map

ABOUT THE PROJECT

The project on Responsible AI is an initiative aimed to ensure the development and deployment of artificial intelligence (AI) technology in a manner that is ethical, transparent, and aligned with societal values. As AI systems become increasingly integrated into various aspects of daily life from healthcare and finance to transportation and entertainment it is important in creating frameworks and guidelines for their responsible use.

During this project phase I have developed a simple web application that allows users to understand the core principles which governs Artificial Intelligence in a responsible manner.

I have also written a research paper on Responsible AI which discusses about the core principles on responsible AI, survey on the available tools and application areas of RAI.

PROBLEM DEFINITION

PROBLEM DEFINITION:

The fast development and deployment of AI systems raise significant ethical concerns across various aspects. These concerns include biases in decision making, lack of transparency in how AI functions, and challenges in ensuring safety, accountability, and individual privacy.

EXISTING SYSTEM

Various organizations, governments, and institutions have established frameworks aimed for promoting the use of AI in a responsible manner. These systems compose of guidelines, regulations, ethical principles, and technical standards designed to ensure that AI technologies are developed and used in ways that are ethical, transparent, and beneficial to society.

PROPOSED SYSTEM

This web application demonstrates all 6 core principles of responsible AI within a single window. With the help of provided dataset in the web application, users can understand about the dataset with respect to the core RAI principles. In this project, I worked on 2 principles Fairness and Privacy & Security in which the calculations are done on the dataset and when the dataset is not Fair or Secured, we go for mitigation to make the dataset better.

FEASIBILITY ANALYSIS

After the problem is clearly understood and the solutions are proposed, the next step is to conduct the feasibility study, which is a part of system analysis well as system design process. The main objective of the study is to determine whether the proposed system is feasible or not. There are three main types of feasibility studies that a proposed system typically undergoes:

- Operational Feasibility
- Technical Feasibility
- Economic Feasibility

OPERATIONAL FEASIBILITY:

Target Users: The RAI web application is designed to help a broad range of users, including students, data analyst, data scientist and AI developers.

User Adoption: The web application is user-friendly whereby each step is guided so that it is easy for the user to use the application.

Implementation Process: The web application will be readily available when the all principles are incorporated.

Maintenance and Updates: The web application's maintenance, updates, and content management will fall under my supervision in collaboration with my fellow friends and my guide. This collaborative approach ensures timely updates, bug fixes, and the addition of new features as needed.

TECHNICAL FEASIBILITY:

Technology Stack: The web application is developed in visual Studio with HTML which provides the skeleton design of the web page, CSS used for the design of the web page and JavaScript is used for the logic and Google charts API is used for graphical visualization.

ECONOMICAL FEASIBILITY:

Development Costs: There was no cost involved in development as it was developed using open-source software.

Ongoing Costs: The ongoing costs for maintenance, updates, and hosting are also calculated to be minimal, as these tasks will be done after the full potential of the web application is met.

Overall Assessment:

The RAI web application is highly feasible in terms of operation, technology, and economy. It effectively meets the needs of its diverse users with an easy-to-use interface with guided steps. The technology stack, including HTML, CSS, JavaScript, and Google Charts API, is reliable and appropriate for the application. Development and ongoing costs are minimal, making it a cost-effective solution. Maintenance and updates will be managed collaboratively, ensuring the application remains valuable and up-to-date.

SOFTWARE REQUIREMENT SPECIFICATION

SYSTEM REQUIREMENTS

Hardware Specification

- Processor - Intel(R) Core (TM) i3
- Memory - 4 GB

Software Specification

- Operating System - Windows 11 Home
- Visual Studio
- HTML, CSS, JS

FAIRNESS AND PRIVACY & SECURITY EVALUATION

Fairness Evaluation:

Naïve Bayes Classification: (Training Model)

Naive Bayes classification is a probabilistic machine learning algorithm that applies Bayes' Theorem to classify data into categories. It involves calculating the probability of a given data belonging to a specific category, assuming that all features of the data point are independent of each other.

During the training phase, the algorithm learns the prior probabilities of each category and the likelihood of each feature given each category from a set of labels. When predicting the category of a new data point, the algorithm calculates the likelihood of each category. It then combines these likelihoods to determine the overall probability (posterior probability) for each category. Finally, it assigns the data point to the category with the highest overall probability. This method is efficient and works well for many practical applications, especially when the independence assumption is reasonably accurate.

FAIRNESS METRICS: [1]

STATICAL PARITY DIFFERENCE:

Statistical parity difference measures the fairness in machine learning models. It checks if the model's predictions are evenly distributed across different groups, like gender or race.

Example:

Imagine you have a machine learning model that decides if people get approved for a loan. Statistical parity difference helps you see if one group (men) gets approved more often than another group (women).

How It Works:

1. Calculate Approval Rates: Find the approval rate for each group. For example:
 - Group A (men): 40 out of 100 approved (40%).
 - Group B (women): 30 out of 100 approved (30%).
2. Compute the Difference: Subtract the approval rate of one group from the other.

Using the example:

$$SPD = 40\% - 30\% = 10\%$$

Interpretation

- Zero Difference: If the difference is 0%, it means both groups are treated equally by the model.
- Non-zero Difference: If the difference is not 0%, it indicates potential bias. For example, a 10% difference means one group is more likely to be approved than the other.

Why It Matters

Statistical parity difference helps ensure that a model treats all groups fairly. It is an important metric for checking and improving the fairness of machine learning decisions.

THEIL INDEX:

Theil index serves as a metric to measure inequality.

Concept:

- It helps understand how evenly resources are distributed among people.
- A lower Theil index means more equal distribution, while a higher index indicates greater inequality.

Interpretation:

- A Theil index of 0 means perfect equality.
- As the index increases, inequality gets worse.

Why It is Important

- Helps policymakers and economists understand and address inequality issues.
- Allows comparison of inequality levels across different populations or over time.

Example

- Imagine a small group of people with different incomes: \$10, \$20, and \$30.
- The Theil index would be calculated based on how these incomes deviate from the average income.

AVERAGE ODDS DIFFERENCE:

The Average Odds Difference (AOD) is used to check if a model treats different groups fairly. It looks at how often the model makes mistakes for each group, such as different genders or races.

Interpret the Result:

- If the average difference is 0, it means the model treats all groups the same.
- If it is positive, one group is more likely to be affected by false alarms or missed opportunities.
- If it is negative, the other group is more affected.

Why It is Important:

- Helps ensure fairness in decisions made by the model.
- Helps identify and fix biases that could unfairly affect certain groups.

Average Odds Difference is a useful tool for making sure machine learning models treat everyone fairly, regardless of their background.

DISPARATE IMPACT:

Computed as the ratio of rate of favorable outcome for the unprivileged group to that of the privileged group.

Example

Imagine a company has a promotion test. Let us look at how different groups perform.

Scenario:

- Privileged group (men): 100 take the test, and 80 pass (80% pass rate).
- Unprivileged group (women): 100 take the test, and 50 pass (50% pass rate).

Calculation:

$$\text{Impact Ratio} = \text{Pass Rate of Unprivileged Group} / \text{Pass Rate of privileged Group} = 50 / 80 = 0.625$$

Interpretation

- The impact ratio is 0.625 (or 62.5%).
- According to the 80% Rule, if the impact ratio is less than 80%, there is evidence of disparate impact.

Since 62.5% is less than 80%, this means the promotion test has a disparate impact against the unprivileged group (women).

Disparate impact helps identify if a policy (like a promotion test) unintentionally disadvantages the unprivileged group more than the privileged group.

Why It is Important:

Disparate impact matters because it highlights unintentional biases in policies that can disadvantage certain groups. Addressing these biases promotes fairness, legal compliance, diversity, trust, performance, and social responsibility.

EQUAL OPPORTUNITY:

Equal opportunity ensures that all individuals have the same chance to succeed, regardless of their background or characteristics such as race, gender, age, or socioeconomic status. It focuses on providing everyone with the same starting point and access to opportunities.

Example

Imagine a company is hiring for a new position:

Fair Hiring Practices:

The company evaluates all applicants based on their skills, experience, and qualifications, not on their race, gender, or other irrelevant characteristics.

Why It Matters:

Equal opportunity ensures everyone has the same chances to succeed, focusing on fairness.

ENHANCEMENT ALGORITHM:

Based on the evaluation of the above metrics it would be possible to conclude whether the dataset meets the fairness principle if it does not meet the required criteria then appropriate enhancement method needs to be initiated.

Below described algorithm is used for fairness enhancement.

Reweighting Algorithm:

Reweighting algorithm for fairness metrics is techniques used to mitigate bias in machine learning models by adjusting the weights of the training instances to achieve fairness. The algorithm aims to ensure that the model's predictions are not biased against certain demographic groups or protected attributes (such as race, gender, or age).

These reweighting algorithms help promote fairness and reduce bias in machine learning models, leading to more fair outcomes for all individuals or groups involved

Steps involved in applying Reweighting Algorithm:

Identifying Bias:

Before applying reweighting algorithms, it is essential to identify the sources of bias in the dataset or model. This involves analyzing the data and assessing whether certain groups are disproportionately affected by the model's predictions.

Defining Fairness Metrics:

Fairness metrics quantify the degree of bias in a model's predictions. The fairness metrics provide a basis for evaluating the fairness of the model and guiding the reweighting process.

Calculating Weights:

Reweighting algorithms compute weights for each training instance based on its attributes in different groups. The aim is to assign higher weights to instances from unprivileged groups and lower weights to instances from privileged groups. This rebalancing helps the model learn from all groups equally and reduces bias in its predictions.

Training with Reweighted Data:

Once the weights are calculated, the training data is reweighted accordingly, and the machine learning model is trained using this adjusted dataset. By incorporating the reweighted data, the model learns to make predictions that are fairer across different groups.

Evaluation and Iteration: After training the model with the reweighted data, we can evaluate its performance using fairness metrics. If the model still exhibits bias, additional iterations of reweighting and training may be necessary until satisfactory fairness levels are achieved.

Privacy & Security Evaluation:

ATTACK:

Membership Attack:

Membership attack refers to a scenario where an adversary attempts to manipulate or influence the data used to train a machine learning model. This manipulation aims to either compromise the stability of the model or bias it towards specific outcomes.

A membership attack is when someone tries to figure out if a specific piece of data was used to train a machine learning model. It is a privacy concern on individual data. To defend against membership attacks in training datasets, robust security measures should be implemented.

Mitigation Algorithm:

Based on the evaluation of the attack accuracy we will go for mitigation if it exceeds the predefined threshold value of the attack accuracy.

Below described algorithm is used for mitigating Privacy & Security concerns.

Differential Privacy:

Differential privacy is a concept in data privacy that aims to protect sensitive information about individuals while still allowing useful analysis of the data.

Adding noise to the output helps achieve differential privacy by hiding the exact values of individual data points. This noise is carefully added to ensure that individual data remain hidden.

Laplace noise, based on the Laplace distribution, is a common method for implementing differential privacy. The concept of adding random noise to the data in such a way that it maintains statistical usefulness while ensuring the privacy of individuals in the dataset.

SYSTEM DESIGN

Fairness:

Step by Step Procedure for fairness Evaluation:

1. The dataset (compas) shown in developed web application should be chosen to proceed for the further step.
2. The dataset is split into 2 parts,
 - a. Training Data
 - b. Testing Data
3. The training dataset is used to train Naïve Bayes Classifier and the testing data is used to predict if an individual is likely to reoffend or not.
4. The Naïve Bayes classifier also calculates the accuracy of the testing data.
5. The dataset is used to evaluate the fairness metrics which include:
 - a. Statistical Parity Difference
 - b. Theil Index
 - c. Equal Opportunity Difference
 - d. Disparate Impact
 - e. Average Odds Difference
6. There is a predefined threshold value for each metrics;
 - a. Statistical Parity Difference (Fairness between: -0.1 and 0.1)
 - b. Theil Index (Fairness lower scores)
 - c. Equal Opportunity Difference (Fairness between: -0.1 and 0.1)
 - d. Disparate Impact (Fairness between: 0.8 and 1.25)
 - e. Average Odds Difference (Fairness between: -0.1 and 0.1)

If the threshold values exceed mitigation is required.

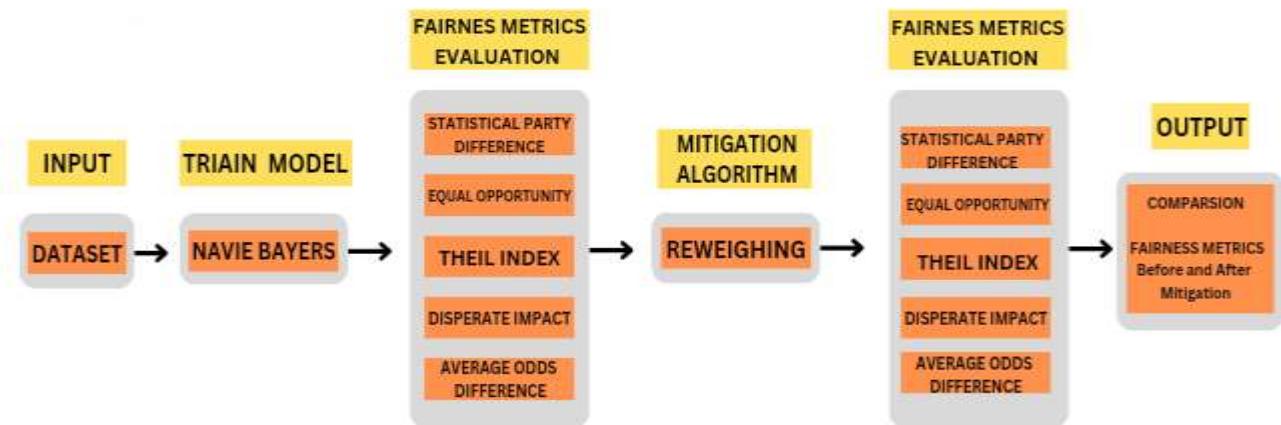
7. For mitigation we use Reweighting Algorithm:

The dataset is taken as input and perform reweighting on the sensitive attributes (sex).The weights are adjusted depending upon the privileged or unprivileged group.

A new dataset is returned with the adjusted weight.

8. The New dataset is taken as an input and steps 4,5 and 6 performed.
9. Finally, a comparison between the fairness metrics value before and after mitigation is shown using charts and tables.

Flow Diagram:



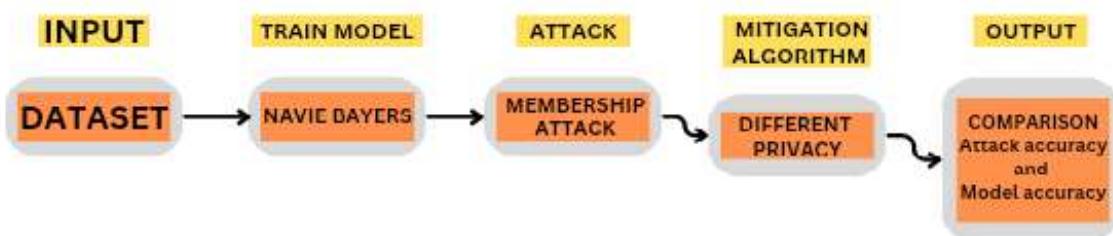
Privacy & Security:

Step by Step Procedure for Privacy & Security:

1. The dataset (German Credit Scoring) shown in developed web application should be chosen to proceed for the further step.
2. The dataset is split into 2 parts,
 - a. Training Data
 - b. Testing Data
3. The training dataset is used to train Naïve Bayes Classifier and the testing data is used to predict if an individual has good credit risk or a bad one.
4. A membership attack is performed on the dataset. The attack accuracy is calculated comparing the predicted labels with true labels which is used to determine if the training dataset is a member of testing dataset.
5. The summary of both Naïve bayes classification accuracy and membership attack accuracy is displayed.
6. There is predefined threshold for attack accuracy if it exceeds 40%, we should go for mitigation.
7. For mitigation we use Differential Privacy Algorithm:

DP is achieved by adding noise to the dataset.
It takes epsilon and sensitive values as parameters and returns a noised dataset.
8. A wide range of epsilon values (0.1, 0.5, 1, 5, 10) is used for adding noise and evaluating the attack accuracy.
9. Finally, a visualization is shown on the dataset the attack accuracy and model accuracy various for various epsilon values.

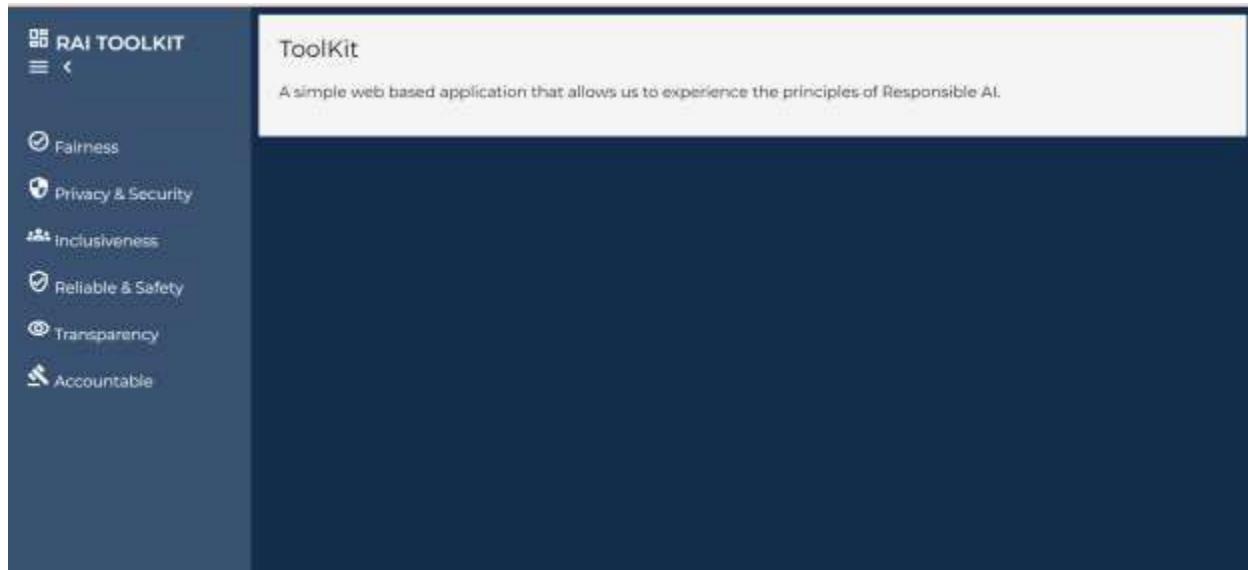
Flow Diagram:



SCREENSHOTS

Fairness:

Dashboard: Displays all the principles, allowing users to experience the respective RAI principle when clicked.



Fairness: When the fairness principle is clicked, it will lead us to the AI Fairness Evaluation page, where we need to choose the given dataset for further evaluation.

A screenshot of the AI Fairness Evaluation page. The title "AI Fairness Evaluation" is at the top. Below it, there is a radio button next to the text "Compas (ProPublica recidivism)" and a description "Predict a criminal defendant's likelihood of reoffending.". Underneath, there is a section titled "Protected Attributes:" with a note "- Sex, privileged: Female, unprivileged: Male". At the bottom right is a blue "Next" button.

Classification: When "Next" is clicked, Naïve Bayes classification is performed on the dataset, and the accuracy is displayed.

AI Fairness Evaluation

● Compas (ProPublica recidivism)
Predict a criminal defendant's likelihood of reoffending.

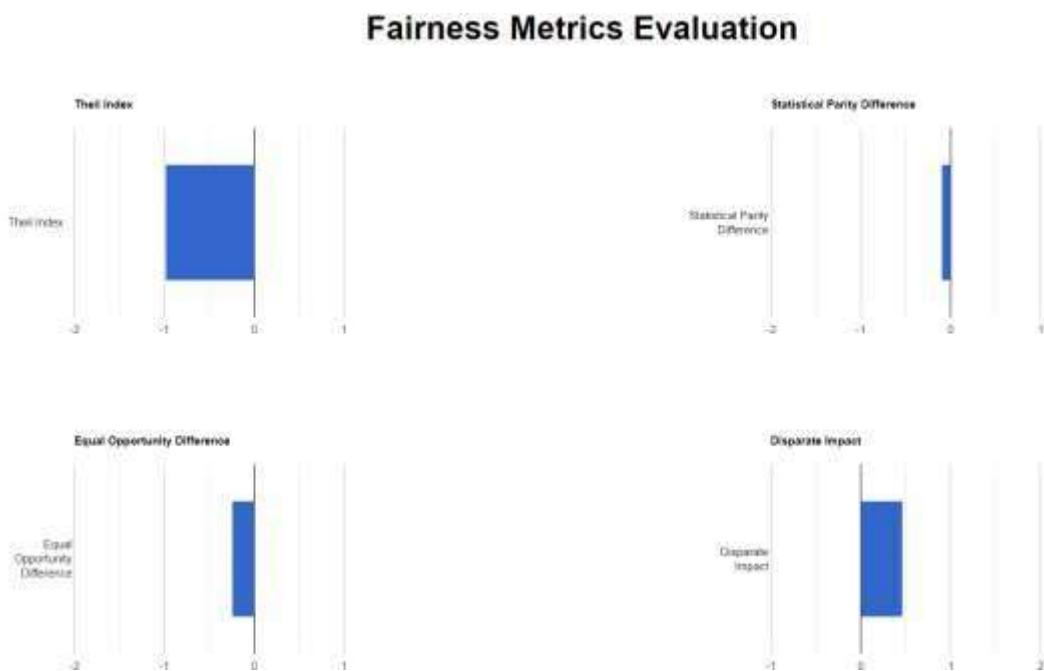
Protected Attributes:
- Sex, privileged: Female, unprivileged: Male

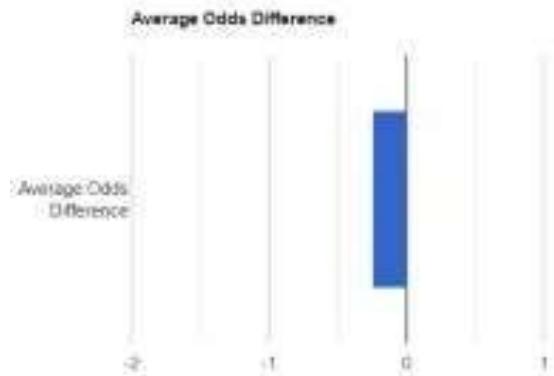
[Next](#)

Accuracy: 0.63

Classification details: Naive Bayes Classification

Fairness Metrics Evaluation: The predefined fairness metrics are evaluated on the dataset, and the values are visualized in graphs.





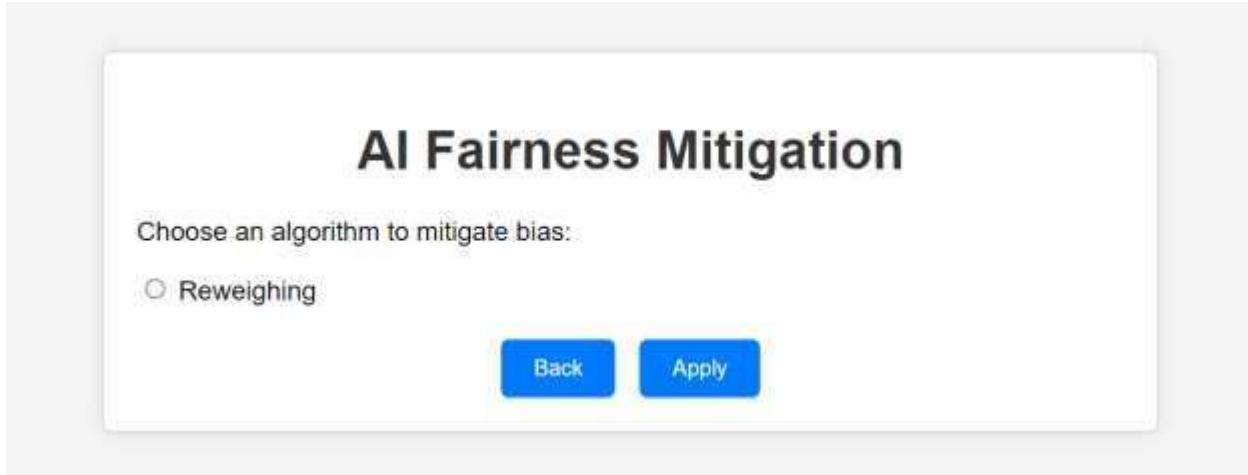
Fairness Metrics Evaluation: We evaluate predefined fairness metrics on the dataset and present the values in a table format, indicating whether mitigation is necessary based on predefined threshold values.

Metric	Value	Mitigation
Total Population	51	
Privileged Population	23	
Unprivileged Population	28	
Privileged Proportion	0.4510	
Unprivileged Proportion	0.5490	
Theil Index	-0.9931	
Statistical Parity Difference	-0.0980	
Equal Opportunity Difference	-0.2469	Mitigation required
Disparate Impact	0.4682	Mitigation required
Average Odds Difference	-0.2469	Mitigation required
Mitigation Required	3	

Back

Next

Mitigation: If bias is found, we proceed with mitigation by performing the Reweighting algorithm to enhance the dataset.

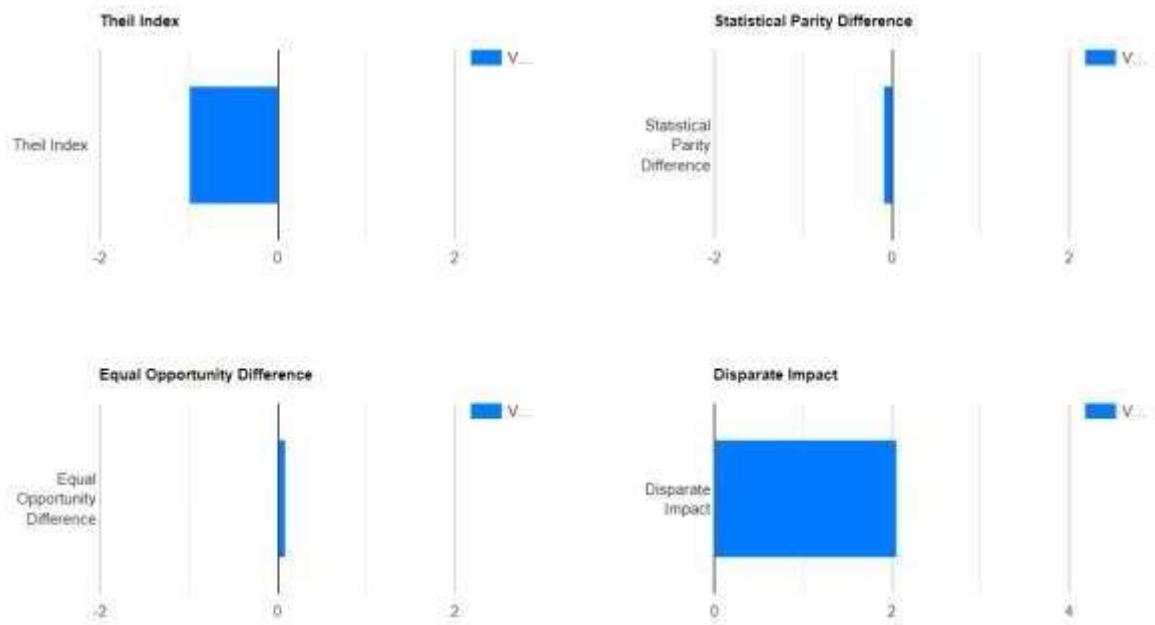


Comparison: Following the mitigation technique to enhance the dataset, we will display a comparison of the fairness metrics before and after mitigation in both table and graph formats.

AI Fairness Evaluation

After Mitigation

Metric	Value	Mitigation
Total Population	51	
Privileged Population	5.844999999999999	
Unprivileged Population	45.155	
Privileged Proportion	0.1146	
Unprivileged Proportion	0.8854	
Theil Index	-0.9935	
Statistical Parity Difference	-0.0945	
Equal Opportunity Difference	0.0880	
Disparate Impact	2.0601	Mitigation required
Average Odds Difference	-0.3597	Mitigation required

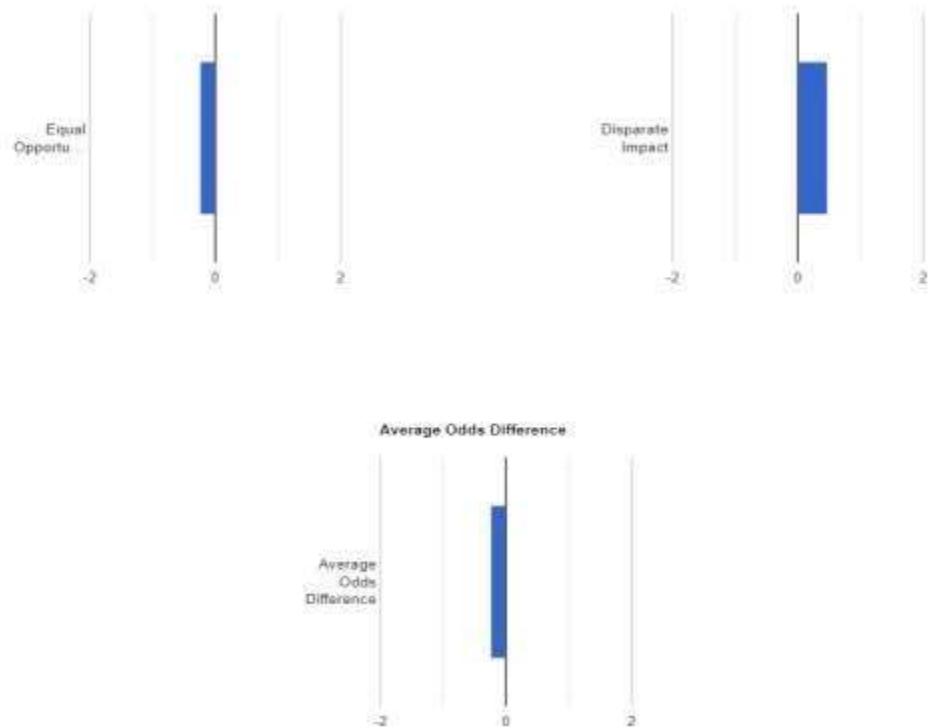


Before Mitigation

Fairness Metrics Evaluation



Before Mitigation



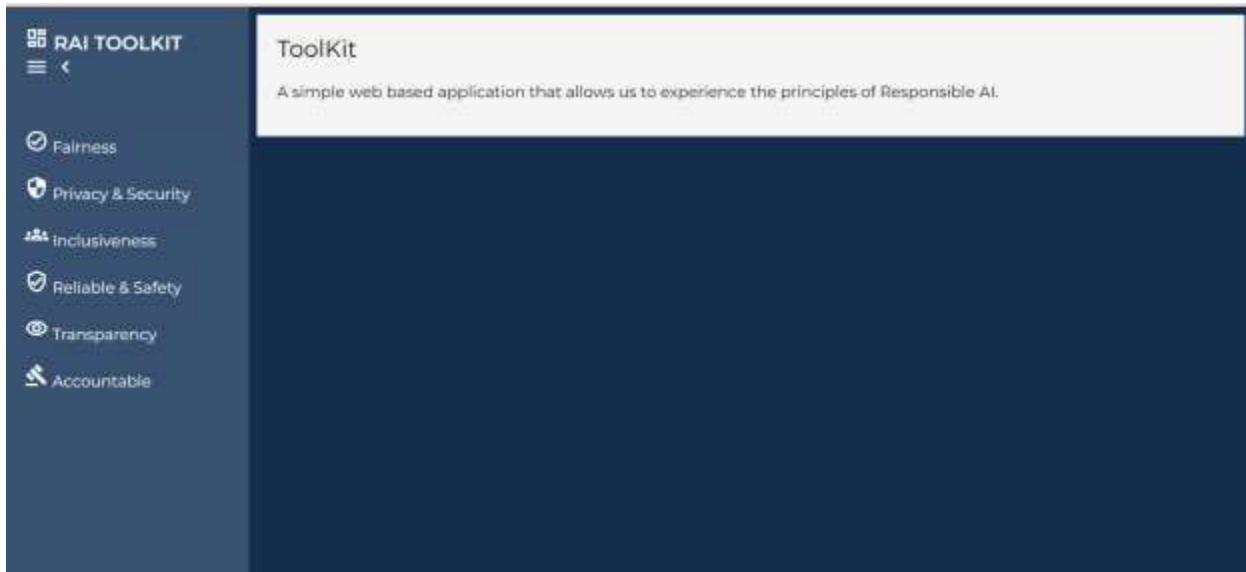
Metric	Value	Mitigation
Total Population	51	
Privileged Population	23	
Unprivileged Population	28	
Privileged Proportion	0.4510	
Unprivileged Proportion	0.5490	
Theil Index	-0.9931	
Statistical Parity Difference	-0.0980	
Equal Opportunity Difference	-0.2469	Mitigation required
Disparate Impact	0.4682	Mitigation required
Average Odds Difference	-0.2469	Mitigation required
Mitigation Required	3	

[Back](#)

[Back to Dashboard](#)

Privacy & Security:

Dashboard: Displays all the principles, allowing users to experience the respective RAI principle when clicked.



Privacy & Security: Upon selecting the Privacy & Security principle, users will be directed to the AI Privacy Evaluation page, where they can choose the designated dataset for further assessment.

A screenshot of the AI Privacy Evaluation page. The title is "AI Privacy Evaluation". Section 1 is titled "Choose dataset". It says "Choose the dataset click Next to proceed. To explore a simple understanding of Privacy & Security principle in Responsible AI." There is a radio button next to the text "German credit scoring" which is described as "Predicts an individual's credit risk.". A blue "Next" button is at the bottom right.

Classification: Upon clicking "Next," Naïve Bayes classification and a membership attack are executed on the dataset, and the accuracies are displayed.

Check attack results

Model: Naive Bayes

Dataset: German credit score

Base model Accuracy: 60.00%

Attack Type: Membership

Attack Accuracy: 70.00%

****Explanation:**** The results show that some membership information was leaked, indicating that the dataset has been affected by a membership attack. Mitigation measures are required to address this privacy risk.

Back

Next

Mitigation: If the attack accuracy surpasses the predefined threshold value, mitigation will be initiated using the Differential Privacy algorithm.

3. Choose Mitigation Technique

We have provided one mitigation that can be used to reduce privacy risks for this model. Choose and then click Next to see how it would affect the model's performance.

Differential Privacy

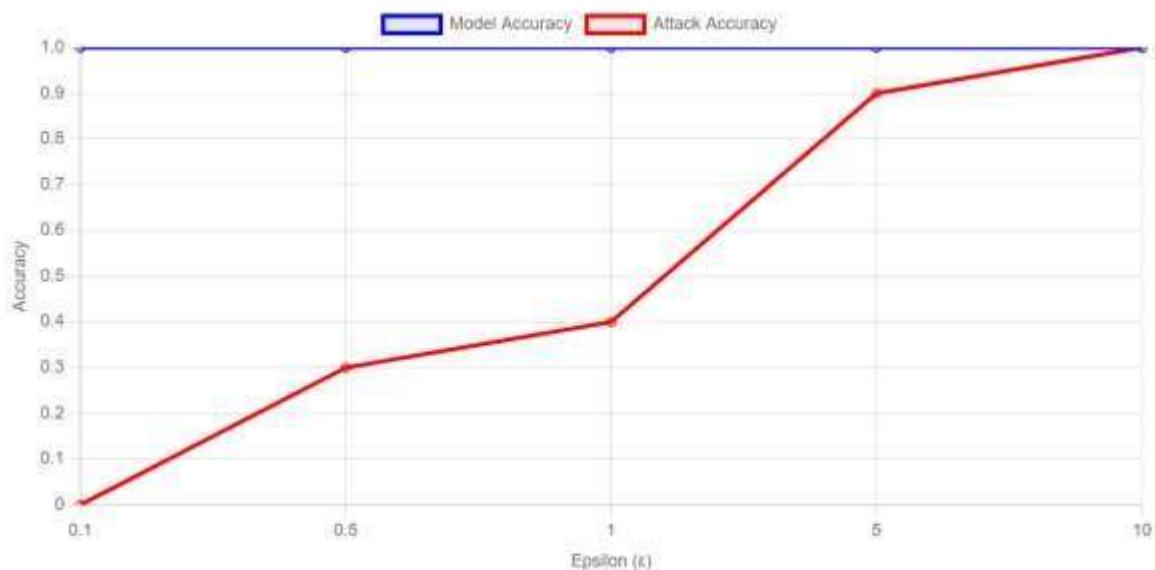
Applying differential privacy (DP) to model training entails adding carefully crafted noise during training to reduce the effect of any single individual on the model's outcome, thus protecting training sets from privacy leakage and countering membership attacks. Should be selected when a formal privacy guarantee is required and when the training procedure can be replaced. Works for tabular or non-tabular data. The privacy loss parameter is called epsilon (lower epsilon means better privacy).

Back

Next

Comparison: Following the Differential Privacy algorithm, the attack accuracy and model accuracy will be presented for various epsilon (privacy parameter) values.

Attack Accuracy (Differential Privacy)



The chart below shows how model accuracy and attack accuracy vary for different values of epsilon (ϵ) used by differential privacy mitigation.

[Back](#)[Back to Dashboard](#)

CONCLUSION

The Responsible AI web application project successfully delivers a web application that provides users with the knowledge about the dataset used develop and deploy AI systems in a responsible and ethical manner. By focusing on our 2 principles for the time being fairness and privacy & security, the application ensures that AI models are built and evaluated for ethical foundations.

Outcomes:

Enhanced Understanding: Users gain a deep understanding of the datasets of their AI models, enabling them to make informed decisions that promote fairness and mitigate biases.

Bias Detection and Mitigation: The Fairness Evaluation feature helps users identify and address potential biases in their AI models, ensuring mitigation treatment to eliminate unfairness.

Robust Data Protection: The Privacy & Security feature safeguards sensitive information, ensuring data protection standards and protecting user data from unauthorized access.

Overall, the Responsible AI web application project provides a crucial resource for the development of AItchnologies that are not only powerful and innovative but also ethical and trustworthy.

FUTURE WORK

As Responsible AI consist of 6 core principles which includes;

1. Fairness
2. Privacy & Security
3. Inclusiveness
4. Reliable & Safety
5. Accountable
6. Transparency

This project only focused on Fairness and Privacy & Security principle and web application is developed to experience those principles. Considering rest of the other principles and incorporating them in the web application are left for future work where research and development needs to be done. For the time being we have given only one dataset to experience the RAI principle but, in the future, we have planned to give multiple datasets. We also planned to enhance the product by making it up to market standards and achieve a smooth user experience with detailed guidance.

REFERENCES:

- [1] Socially Responsible AI Algorithms: Issues, Purposes, and Challenges Lu Cheng Computer Science and Engineering, Arizona State University Kush R. Varshney IBM Research – Thomas J. Watson Research Center Huan Liu Computer Science and Engineering, Arizona State University (**Research Paper**)
- [2] Responsible AI by Design in Practice Richard Benjamins, Alberto Barbado, Daniel Sierra Telefónica, Ronda de la Comunicación, 28050 Madrid, Spain {richard.benjamins, alberto.barbadogonzalez,
- Responsible AI by Design in Practice Richard Benjamins, Alberto Barbado, Daniel Sierra Telefónica, Ronda de la Comunicación, 28050 Madrid, Spain richard.benjamins, alberto.barbadogonzalez, (**Research Paper**)
- What is Google Charts and How To Create Charts – Codersarts
<https://youtu.be/1lI0Ba9vmL0?si=mQUtRAVNFPQUjHYi> (**video**)
- MDN Web Docs - HTML. (n.d.). MDN Web Docs. <https://developer.mozilla.org/en-US/docs/Web/HTML>
- MDN Web Docs - CSS. (n.d.). MDN Web Docs. <https://developer.mozilla.org/en-US/docs/Web/CSS>
- CSS-Tricks. (n.d.). CSS-Tricks. <https://css-tricks.com/>
- MDN Web Docs - JavaScript. (n.d.). MDN Web Docs. <https://developer.mozilla.org/en-US/docs/Web/JavaScript>
- **Fairness Explained: Definitions and Metrics**
<https://medium.com/ibm-data-ai/fairness-explained-definitions-and-metrics-9690f8e0a4ea>
- An end-to-end framework for privacy risk assessment of AI models, Abigail Goldstein, Shlomit Shachor, Natalia Raznikov, 15th ACM International Conference on Systems and Storage (SYSTOR), 2022 (**Research Paper**)
- Differential Privacy-enabled Federated Learning for Sensitive Health Data, O. Choudhury, A. Gkoulalas-Divanis, T. Salonidis, I. Sylla, Y. Park, G. Hsu, A. Das, NeurIPS ML4H (Machine Learning for Health), 2019, Dec, 2019 (**Research Paper**)