**PROJECT**

**REMOVE ALGORITHM BIAS IN ANTIDEPRESSANT RESPONSE PREDICTION**

**Professor: Dr. Farrokh Alemi**

**HAP-786 – WORKSHOP IN HEALTH IFORMATICS**

**Authors:**

**K. Lakshmi Indira**

**M. Naga Gayathri**

**P. V. Siva Durga**

**INTRODUCTION**

This study examines the extent of racial bias in the knowledge base of an AI algorithm developed to guide the selection of antidepressants for managing Major Depressive Disorder, with a focus on the Hispanic population. The AI’s knowledge base was built from the experiences of 3,678,082 patients and 10,221,145 antidepressant treatments. This initial knowledge base was developed without differentiating between the needs of racial and ethnic groups, leading to what is referred to as a race-blind algorithm. The AI system comprises several components, including a conversational intake, a survey, and Blue Button services, which together collect the patient's medical history online. The patient’s history is then matched to 16,700 subgroups within the algorithm's knowledge base, and the AI system provides recommendations based on what has worked best for at least 100 patients with similar profiles. This paper assesses whether the AI system exhibits biases when advising Hispanic patients.

Depression remains a significant health issue, and the prevalence of depression among adults in the U.S. continues to grow. However, disparities in treatment persist, particularly for Hispanic individuals, who are often less likely to be diagnosed and treated for depression. Contributing factors include social determinants of health, such as poverty, language barriers, cultural stigma, low education levels, and exposure to discrimination, which are more prevalent in minority populations, including Hispanics. Furthermore, medical comorbidities, such as obesity, diabetes, hypertension, and substance use disorders, also occur more frequently among Hispanics, and can complicate treatment outcomes. These factors do not imply that race itself is a direct risk factor for depression, but rather that race is associated with other determinants that contribute to depression risk and outcomes.

The cumulative effect of these racial and ethnic differences in health and social experiences is that treatment response can vary significantly among minority populations. Current clinical guidelines often do not provide specific advice on differential treatment for minority groups, including Hispanics. One major issue is that existing empirical evidence on the effectiveness of antidepressants among Hispanic populations is limited, as many large-scale studies, including those funded by the NIH, primarily focus on White populations. As a result, Hispanics and other minorities are often not included in sufficient sample sizes to yield conclusive findings. This analysis aims to evaluate the response to antidepressants among Hispanics and determine whether the race-blind AI guidelines are biased in their recommendations for this population.

**OBJECTIVES**

1. **Identify Algorithm Bias:** Using hierarchical analysis, we aim to uncover existing biases in AI models that predict the response to antidepressants. This involves examining discrepancies in predictions among various demographic subgroups, particularly racial and ethnic minorities.
2. **Remove Algorithm Bias:** We will develop and implement conditional models tailored to different subgroups. This step involves employing advanced statistical methods to adjust the predictions of antidepressant responses, ensuring that the models provide accurate and fair outcomes across all demographics.
3. **Analyse Antidepressant Data:** The project will leverage data from the All of Us initiative, which includes a diverse population of participants. We will specifically examine how different groups (e.g., Hispanics, African Americans) respond to various antidepressants and identify the factors contributing to these responses.

**METHODS**

The project methodology is broken down into two tasks:

**Task 1: Describe the Population (Table 1)**

A detailed breakdown of Hispanic population baseline characteristics using All of Us database. It includes the following key points:

* Number of Individuals Examined: This refers to the total participants included in the study, providing a sense of the dataset's size and the diversity of the population analysed
* Number of Antidepressants Used: This metric highlights how many different antidepressant medications were considered in the analysis, helping to understand the scope of medication usage within the population.
* Baseline Medical Conditions: These are the health conditions present in participants at the start of the study, which are essential for understanding the health status of individuals before they began using antidepressants. This could include conditions like anxiety, chronic pain, or other mental health disorders.
* Other Relevant Metrics: This may cover demographic details (e.g., age, gender, race/ethnicity), lifestyle factors (e.g., smoking, alcohol use), and other health indicators that might influence antidepressant usage and overall treatment outcomes.

Overall, the table serves to provide a comprehensive view of the participants, giving researchers insights into how various factors might impact the effectiveness or patterns of antidepressant use across different subgroups within the population.

**Table-1**

**A screenshot of a computer

Description automatically generated**

This table represents the baseline characteristics of participants involved in the analysis, including antidepressant usage history and the number of medical conditions at baseline. These factors are crucial for understanding the population being studied.

**Task 2: Regression Models in Subgroups**

The code implements a regression analysis on a dataset concerning antidepressant usage and associated patient demographics. The analysis is segmented by different antidepressants, allowing for subgroup comparisons. The main goal is to understand how various factors, such as age and gender, influence the likelihood of remission from depression, as indicated by the target variable Remission.

1. **Subgroup Analysis**: Each antidepressant (e.g., Duloxetine, Venlafaxine) is analyzed independently. The code iterates through a list of antidepressants, creating a separate analysis subset for each. This allows for tailored insights into the effects of each medication on patient outcomes.
2. **Logistic Regression**: The code employs logistic regression as the primary statistical method to model the relationship between the independent variables (such as age, gender, and previous medication history) and the binary outcome variable Remission. This model is suitable for binary outcomes, which in this case is whether a patient experiences remission from depression.

**Bias Detection**

Bias detection is crucial in ensuring that the model's predictions are fair and not skewed by any demographic factors or pre-existing conditions. The code includes several strategies to identify and mitigate bias:

1. **Data Collection**: The code collects data on various patient demographics (e.g., age, gender, ethnicity) and their treatment history. This comprehensive data collection helps identify any potential biases based on demographic factors.
2. **Feature Engineering**: The code creates several features that may indicate bias, such as the number of previous antidepressants used, the number of episodes, and the number of remissions. These features are crucial in understanding how past treatment history might influence current outcomes.
3. **Distribution Check**: The code checks the distribution of the target variable remission identify any imbalances. This is done using the **value\_counts()** method, which allows for the detection of any significant disparities in the outcomes based on different demographic groups.

**Post-De-biasing Results**

Post-de-biasing results refer to the analysis conducted after implementing measures to reduce bias in the dataset. The code performs the following actions to ensure the results are unbiased:

1. **Stratification**: The analysis is stratified by antidepressant type, ensuring that each subgroup is analyzed independently. This helps in understanding the specific effects of each medication without the influence of other medications.
2. **Dummy Variables**: The code creates dummy variables for categorical features, such as age groups and disease codes, which helps in quantifying the influence of these factors on the outcome variable.
3. **Model Evaluation**: After fitting the logistic regression model, the code evaluates the model's performance using metrics like accuracy and classification reports. This evaluation helps in assessing whether the model's predictions are consistent across different demographic groups.

**De-biasing Algorithm Implementation**

The de-biasing algorithm is implemented through a combination of data preprocessing, feature engineering, and careful model training. Key steps include:

1. **Data Preprocessing**: The code removes duplicate entries and filters out records that do not meet certain criteria (e.g., conditions that began after the start of antidepressant treatment). This step ensures that the dataset is clean and relevant.
2. **Feature Creation**: The code creates new features that capture the history of antidepressant use and the number of episodes and remissions. This helps in understanding the patient's treatment journey and its impact on their current health status.
3. **Regression Model Training**: The logistic regression model is trained in the processed data, with features selected based on their relevance to the outcome variable. The model is evaluated for its ability to predict remission accurately.
4. **Bias Mitigation**: The code includes checks for multicollinearity and ensures that features with low sample sizes are removed, which helps in reducing bias in the model.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

**Data Processing and Tools**

The code utilizes various data processing techniques and tools to achieve its objectives:

1. **Pandas**: The primary library used for data manipulation and analysis. It provides powerful data structures like DataFrames, which are used to store and process the data.
2. **Numpy**: Utilized for numerical operations and handling arrays, especially when creating dummy variables and performing mathematical calculations.
3. **Scikit-learn**: This library is used for building and evaluating the logistic regression model. It provides tools for model fitting, prediction, and performance evaluation.
4. **SQL Queries**: The code includes SQL queries for data extraction from a database, specifically targeting relevant datasets for conditions and drug exposures.
5. **Data Cleaning**: The code includes several steps for cleaning the data, such as removing duplicates, filtering out irrelevant records, and handling missing values.
6. **Feature Engineering**: The process of creating new features from existing data to improve model performance and capture more information about the patient's treatment journey.

**RESULTS**

Key results from our project include:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Precision | Recall | F1-Score | Support |
| Class 0 | 0.98 | 1.00 | 0.99 | 1179 |
| Class 1 | 0.99 | 0.98 | 0.98 | 124 |
| Accuracy | - | - | 0.98 | 1303 |
| Macro Avg | 0.49 | 0.50 | 0.50 | 1303 |
| Weighted Avg | 0.96 | 0.98 | 0.97 | 1303 |

McFadden’s R-Squared: 0.7111848885666219

A screenshot of a computer

Description automatically generated

It provides the confusion matrix metrics (precision, recall, F1-score, and support) for two classes. Additionally, the overall accuracy, macro average, and weighted average are shown. The McFadden’s R-squared value is given, indicating the goodness-of-fit of the model. The output also contains warnings related to ill-defined metrics, suggesting that there are classes with no predicted samples, and recommends handling this using the `zero\_division` parameter.

The report details the process of extracting and saving coefficients from a logistic regression model using Jupyter.

1. Setting Response Variable: The code sets the target variable, "Remission," which the logistic model predicts.
2. Extracting Coefficients: It retrieves coefficients associated with each feature from the trained model and also includes the intercept term.
3. Organizing Data: The features, coefficients, and response are organized into a Pandas DataFrame, allowing easy inspection and manipulation.
4. Saving Results: The DataFrame is saved to a CSV file named `hap786\_logistic\_regression\_coefficients\_with\_features.csv` for documentation or further analysis.
5. Output Verification: The code prints the DataFrame, displaying a preview of the features and their coefficients, and confirms the file's successful creation.

This report provides a structured approach to examining and exporting logistic regression outputs, useful for understanding feature impact on predictions and sharing results.

A screenshot of a computer

Description automatically generated

**CONCLUSION**

This study highlights the importance of addressing racial bias in AI algorithms, particularly in the context of antidepressant recommendations for Hispanic patients with Major Depressive Disorder. By examining the knowledge base of a race-blind AI system, we identified potential biases and implemented conditional models to ensure fair and accurate treatment recommendations across diverse demographic groups. Our analysis utilized comprehensive data from the All of Us initiative, revealing significant insights into the response of Hispanic individuals to various antidepressants. The results underscore the necessity for tailored approaches in mental health treatment and the critical need for future research to include underrepresented populations in clinical studies. Ultimately, this work aims to enhance the efficacy of AI-driven healthcare solutions while promoting equity in mental health treatment.

**RECOMMENDATIONS**

**Recommendations for Addressing Racial Bias in AI Algorithms for Antidepressant Selection**

To effectively mitigate racial bias in AI algorithms used for antidepressant selection, several strategies can be implemented. These recommendations focus on improving the fairness and accuracy of AI-driven healthcare solutions, particularly for underrepresented populations such as Hispanic individuals.

**1. Enhance Data Diversity**

* **Inclusive Data Collection**: Ensure that datasets used to train AI algorithms include diverse populations, particularly racial and ethnic minorities. This can be achieved by actively recruiting participants from various backgrounds in clinical trials and studies.
* **Stratified Sampling**: Implement stratified sampling techniques to ensure that minority groups are adequately represented in the training datasets, allowing the AI to learn from a broader range of experiences and treatment responses.

**2. Bias Detection and Monitoring**

* **Regular Audits**: Conduct regular audits of AI algorithms to identify and assess biases in predictions. This includes analyzing the performance of the model across different demographic groups to detect disparities in treatment recommendations.
* **Bias Metrics**: Develop and utilize specific metrics to measure bias in AI outputs, such as disparate impact ratios and fairness indicators, to ensure that the model's recommendations are equitable.

**3. Algorithmic Transparency and Explainability**

* **Model Interpretability**: Enhance the interpretability of AI models by using techniques that allow clinicians to understand how decisions are made. This can help identify potential biases in the decision-making process.
* **User -Friendly Interfaces**: Design user interfaces that clearly communicate the rationale behind AI recommendations, enabling healthcare providers to critically assess and adjust treatment plans based on their clinical judgment.

**4. Tailored Treatment Models**

* **Conditional Models**: Develop conditional models that account for the unique characteristics and needs of different demographic groups. This involves creating separate algorithms or adjusting existing ones to provide personalized treatment recommendations.
* **Feedback Loops**: Implement feedback mechanisms where clinicians can report outcomes and experiences with AI recommendations, allowing for continuous improvement of the algorithms based on real-world data.

**5. Training and Education**

* **Clinician Training**: Provide training for healthcare providers on the potential biases in AI systems and the importance of considering social determinants of health when interpreting AI recommendations.
* **Cultural Competency**: Incorporate cultural competency training into medical education to help clinicians understand the specific needs and challenges faced by minority populations, improving their ability to make informed decisions.

**6. Collaboration with Community Organizations**

* **Engagement with Communities**: Collaborate with community organizations and advocacy groups to better understand the needs and preferences of minority populations. This can inform the development of more effective and culturally sensitive AI solutions.
* **Public Awareness Campaigns**: Launch campaigns to raise awareness about the importance of equitable treatment in mental health care, encouraging participation from diverse communities in research and clinical trials.