Module 2: Bias, Representation, and Fairness

2/12/2025

Exploring Data Bias

By: Group 4

Bias itself is a double-edged sword. On the one hand, bias leads to motivation (finding information that supports a narrative on a topic). On the other, it clouds judgement and limits one's perspective, which can lead to errors in judgement and even injustice. Though it is obvious that bias has affected humans throughout history, in the modern age bias also affects tools made by humans. Historically, it is known that photo cameras were originally made for white skin and couldn't properly make images for dark-skinned individuals. Recently, however, it has become evident that AI tools themselves are not immune to bias, either.

As data scientists, we aim to reduce bias in datasets to avoid adverse impacts in data analysis. Bias in datasets can lead to misleading conclusions, inaccurate predictions, and unfair outcomes in machine learning models. This can contribute to existing inequalities and outcomes that disproportionately affect certain groups and demographics, instead of addressing the needs of the true population. This analysis explores the various biases present in the Heart Attack Risk Prediction Dataset from Kaggle. This is a synthetic dataset that provides comprehensive information related to heart health, including demographic, lifestyle, socioeconomic, and lifestyle data from patients worldwide. Despite its comprehensive data and diverse features, the dataset displays a variety of biases like sampling bias, selection bias, historical bias, and geographic bias, affecting the data's reliability and its outcomes.

This report will discuss the identified biases, the methods we used to identify them, our results, and real-world implications these biases can have on predicting heart disease on different groups of people. Our analysis contains an Exploratory Data Analysis (EDA) to identify trends and patterns, data visualizations, and statistical distribution comparisons to effectively identify bias in the dataset. Addressing these biases in this dataset will bring attention to the disparities that exist in the collection of healthcare data when predicting heart disease on certain groups and demographics, leading to the actionable insights into more fair and accurate predictive modeling.

I. Initial Explanation of the Data

As mentioned earlier, the data is called "Heart Attack Risk Prediction Dataset", provided by S. Banerjee on Kaggle.com. This biomedical dataset is a synthetic dataset, a type of dataset that is entirely AI-generated. Synthetic datasets are made from analysis of organic datasets and mimic the true data. One major advantage they have is anonymity; since they are based on true human metrics but are not actually the true human metrics, there is little-to-no risk of privacy violations from using AI-generated data. This is especially important for fields that require handling sensitive data, such as healthcare and finance {Mostly.ai}.

This dataset, which contains 8763 non-null variables, exists to help researchers in biostatistics to analyze and predict factors that are associated with heart attacks. The data includes demographics of patients (such as age, gender, & nationality), preexisting medical conditions (such as diabetes, obesity, & family history), as well as lifestyle (including stress levels, dietary habits, and exercise hours). The data also includes socioeconomic and geographic conditions of the patients as well. Although synthetic data is made to help reduce bias, it may in fact increase bias depending on the conditions it was made under, which includes the original data used to make the synthetic data.

II. Imbalance in Variables

One issue affecting datasets, both organic and synthetic, is the issue of data balance. A balanced dataset would have similar proportions of a classification for variables (such as 'yes' and 'no' for an illness). If one or more classifications are overrepresented and the others are underrepresented, then the dataset is imbalanced. Imbalanced datasets often don't have enough datapoints for training models. If a disease dataset, for example, has over 90% 'yes' and less than 10% 'no', then predictive models (such as classification models) may overfocus on the majority label and disregard the minority label altogether {Saikat}. Depending on the context, this would vastly increase either the false positive or false negative rate of the model, which undermines the model's predictive power.

One example here is the gender imbalance in the 'sex' variable. There are only 2652 females (about 30%) compared to 6111 males. Though the imbalance is not extreme, there is a risk that statistical models made from this dataset will not be able to accurately predict heart risk factors for female patients. Currently, medical institutions have acknowledged that female patients are underrepresented in medical research {Balch}. In organic data, low female representation is a form of selection bias called sampling bias, when a group of the population is less likely to be included in a study {Balch} Therefore, for the sake of women's health, it is important to sample enough females for medical analysis.

Another example of imbalance is the continent variable. This dataset is meant to describe factors affecting cardiology patients internationally. However, the continent variable shows that there are only 873 Africans (around 10%) compared to 2543 Asians and 2242 Europeans. Due to this, there is a risk that, though the dataset is meant for global analysis, it may not accurately examine risk factors for African patients. One can also see from the data that the diabetes variable is

somewhat unbalanced. There are 3047 non-diabetics (around 35%), compared to 5716 diabetic patients. Imbalance here could lead to the models focusing too much on factors for diabetic patients, and neglect factors that are more important for non-diabetics. Smoking is even more imbalanced than diabetes. There are 904 non-smokers (around 10%) compared to 7859 smokers. Having a dataset this imbalanced has a strong risk of focusing only on conditions that affect smokers and not on those that affect non-smokers.

III. Revealed Imbalance after Preprocessing

In some cases, imbalance in a variable is not obvious without doing preprocessing beforehand. In the case of BMI, it appears that the dataset is balanced because one has many people whose BMI ranges from less than 20 to over 30. However, the standard international BMI definition, used by the NIH, states that underweight is less than 18.5, normal weight is from 18.5 to under 25, overweight is from 25 to under 30, and obese is 30 or more {Joslin Diabetes}. When looking at the BMI through this lens, it appears that there are only 204 underweight individuals (2.3%). With so few individuals in the category, there is a huge risk that the data analysis results won't accurately identify which heart disease factors are more significant for underweight individuals.

Another variable that is imbalanced, after preprocessing, is 'Exercise Hours Per Week'. This variable gauges how much one exercises each week. After preprocessing, it is shown that 240 people (2.7%) hardly exercise at all, and 858 people (9.8%) exercise less than 2.5 hours every week. This shows that the model made with this dataset won't be as effective at discovering the most important factors that influence heart disease risk in those who do little-to-no exercise.

IV. Non-Imbalanced Biases in Reported Variables

In addition to imbalance in the dataset, one must also be mindful of bias inherent within the variables themselves. Though data scientists take pride in their analytical abilities, the quality of the data often determines the usability of the analysis. If the data quality is poor or doesn't sufficiently explain the observation of interest, the data scientists may unintentionally mislead stakeholders and lead to erroneous decision-making.

Looking back at BMI, this dataset is meant to cover individuals from across the globe. However, the international definition of BMI does not equally apply to all ethnic groups. Those of Asian descent experience the effects of obesity at lower BMIs than non-Asians. As such, Asian countries, and Asian American health professionals often use adjusted BMI standards for Asian patients. The Asian American Diabetes Initiative states that for Asian Americans, a BMI of 23 is overweight, and a BMI of 27 is obese {Joslin Diabetes}.

From our preprocessing, there are 3881 obese individuals by the international BMI score. However, the "Obesity" variable states that there are 4394 obese individuals. After comparing the BMI to the official obese dummy variable in the dataset, the flaws became even more evident. It appears that many individuals were incorrectly classified as obese. Even underweight individuals (BMI of 18) were erroneously classified as obese. This is a huge problem for data analysis, if one predicts erroneously that obesity is not a factor in heart attack risks when they are, this could lead to physicians making incorrect judgements, which would put the lives of patients at risk.

Alcohol consumption is another variable that has inherent bias. Here, there are 3522 non-drinkers (about 40%), compared to 5241 drinkers. Though the imbalance appears to be mild, this variable does not show the extent of alcohol consumption. As such, a model made from this data won't distinguish between light drinkers and heavy drinkers and would assume that the effects of any alcohol consumption are the same regardless of the amount of alcohol consumed.

Finally, the income variable contains inherent bias. Though a histogram suggests balance, preprocessing shows that only 967 (11%) have an income of less than 50 thousand a year. Note that the global per capita income was over $23,000 a year in 2022 {World Inequality Report}, which suggests that the dataset vastly overrepresents wealthy individuals and doesn't represent the global population well.

V: Conclusion:

Artificial Intelligence is a powerful tool, when used correctly. However, as an analytical researcher, one must be aware of the inherent biases of the datasets that one uses. In this paper, the bias of the displayed dataset has been demonstrated. If one only needs to adjust for imbalance within variables, using statistical weights to represent the minority class more is ideal. However, due to the inherent flaws within some of the variables, it is not recommended to use this dataset for biostatistical analysis. In AI research, there is a concept called "AI-decay". When an AI model is trained on AI-generated synthetic data, it runs the risk of losing predictive ability {Stock}. When that AI model is continuously trained on synthetic data, then the model becomes less effective over time. The solution is to use AI models only sparingly and to make sure that the model also has access to organic data. In the end, data science, including AI-generated content, is intended to produce results or the betterment of humanity, and data scientists must examine the datasets carefully to uphold their duties to stakeholders and the public.

Sources:

Balch, B., (2024, March 26). *Why we know so little about women's health*. AAMC. https://www.aamc.org/news/why-we-know-so-little-about-women-s-health

Banerjee, S. (2024, May 11). *Heart attack risk prediction dataset*. Kaggle. https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset

Joslin Diabetes. (n.d.). *Asian BMI Calculator* . Asian American Diabetes Initiative. https://aadi.joslin.org/en/am-i-at-risk/asian-bmi-calculator

Nikolopoulou, K. (2023, May 1). *What is selection bias?: Definition & examples*. Scribbr. https://www.scribbr.com/research-bias/selection-bias/

Saikat. (2024, December 6). *5 techniques to handle imbalanced data for a classification problem*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/

Stock, P. (2023, June 17). *Degenerative AI: Researchers say training artificial intelligence models on machine-generated data leads to model collapse*. Cosmos. https://cosmosmagazine.com/technology/ai/training-ai-models-on-machine-generated-data-leads-to-model-collapse/

Mostly AI *What is synthetic data?*. MOSTLY AI. (2024, March 19). https://mostly.ai/what-is-synthetic-data

World Inequality Report. (2022, April 14). *The world inequality report 2022 presents the most up-to-date & complete data on Inequality Worldwide:*. World Inequality Report 2022. https://wir2022.wid.world/chapter-1/