Bias Analysis in the Diabetes Health Indicators Dataset.

William Holt, Vaishnavi Paniki, and Sebastian Segura

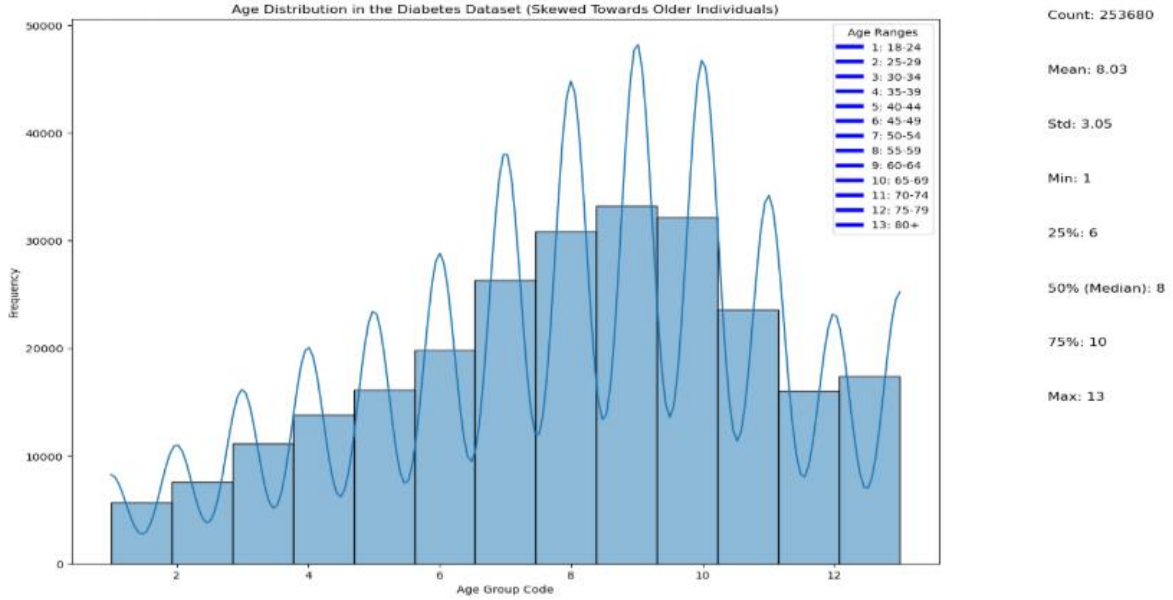
The purpose of this project is to identify, analyze, and report on biases found in the Diabetes Health Indicators Dataset from the Behavioral Risk Factor Surveillance System 2015. The dataset contains health factors for over 250,000 individuals. This dataset has been widely used for diabetes research, despite being known to have present biases. This project aims to explore these biases and better understand their implications for ethics in data science.

The dataset contains 250,000+ records and several health indicator variables. The key variable in the dataset is Diabetes_binary, which allows our model training to study biases further. Some known biases in this dataset include age bias, gender bias, and various other types of sampling bias.

Bias Analysis

Age Bias

This dataset is very heavily skewed towards older people, which poses multiple issues for a dataset intended to explore public health issues. Through initial exploratory data analysis, we visualized the distribution of age to confirm that it is in fact showing a much higher average age than the general population.



Based on this initial analysis, the average age in the dataset is from 60-64, with only 10% under 45 years old. The left-skewed distribution confirms that younger age groups are underrepresented significantly. Should data biased in this way be used to shape public health policy, the results would limit the amount of preventive care available to younger individuals. It would be very difficult to acquire sufficient information on young people from this dataset.

To further investigate the impact of this age bias, two random forest models were trained, one predicting diabetes with age, and one without age as a variable.

Classification Report (With Age):

	precision	recall	f1-score	support
0.0	0.88	0.97	0.92	43667
1.0	0.47	0.16	0.24	7069
accuracy			0.86	50736
macro avg	0.67	0.57	0.58	50736
weighted avg	0.82	0.86	0.83	50736

Classification Report (Without Age):

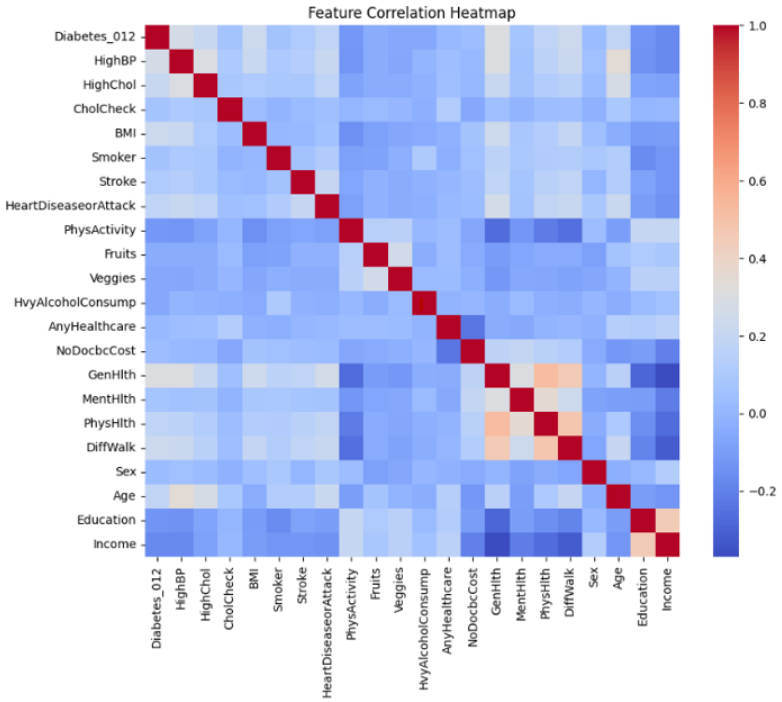
	precision	recall	f1-score	support
0.0	0.88	0.95	0.91	43667
1.0	0.37	0.18	0.25	7069
accuracy			0.84	50736
macro avg	0.63	0.57	0.58	50736
weighted avg	0.81	0.84	0.82	50736

The first model, with age, was 85% accurate in predicting “no diabetes” but only 16% accurate in “diabetes” cases. Age was the second most important factor following BMI, indicating a strong influence on the model’s predictive power.

The second model, without age, maintained a high accuracy (82%) for “no diabetes” and a slight improvement (18%) in accuracy for “diabetes” cases. BMI remained most important, followed by various other demographic variables.

These results show that age is a significant predictor of diabetes, but less so than other factors in measuring these models’ performances. The dataset’s overrepresentation of older individuals skews the model’s predictive ability, lessening its accuracy among younger people. If a model was trained on similarly biased data and used to shape health policy, it would lead to the neglect of underrepresented groups. For this dataset, there would be a lack of information on young people leading to lessened attention to their preventative healthcare needs, particularly those at risk of diabetes. We will now explore the reasons behind this bias by examining more variables.

Bias for Underrepresented Groups



This section is dedicated to analyzing multiple factors linked with diabetes like sex, age, education, and income to find a reason behind the oversaturation of these individuals in the dataset. Diabetes is a growing concern in America, so with this research we wanted to show the relation between these factors and diabetes were biased and explain the implication of using such a dataset anyway. This section's main goal is to show the patterns that the underrepresented populations may exhibit when it comes to diabetes and describe this dataset's bias toward underrepresented parties.

Sex	count	mean	std	min	25%	50%	75%	max
0.0	141974.0	0.129679	0.335951	0.0	0.0	0.0	0.0	1.0
1.0	111706.0	0.151603	0.358638	0.0	0.0	0.0	0.0	1.0

Age	count	mean	std	min	25%	50%	75%	max
1.0	5700.0	0.013684	0.116187	0.0	0.0	0.0	0.0	1.0
2.0	7598.0	0.018426	0.134494	0.0	0.0	0.0	0.0	1.0
3.0	11123.0	0.028230	0.165636	0.0	0.0	0.0	0.0	1.0
4.0	13823.0	0.045287	0.207940	0.0	0.0	0.0	0.0	1.0
5.0	16157.0	0.065049	0.246620	0.0	0.0	0.0	0.0	1.0
6.0	19819.0	0.087895	0.283150	0.0	0.0	0.0	0.0	1.0
7.0	26314.0	0.117352	0.321845	0.0	0.0	0.0	0.0	1.0
8.0	30832.0	0.138265	0.345184	0.0	0.0	0.0	0.0	1.0
9.0	33244.0	0.172452	0.377779	0.0	0.0	0.0	0.0	1.0
10.0	32194.0	0.203703	0.402757	0.0	0.0	0.0	0.0	1.0
11.0	23533.0	0.218459	0.413209	0.0	0.0	0.0	0.0	1.0
12.0	15980.0	0.212954	0.409408	0.0	0.0	0.0	0.0	1.0
13.0	17363.0	0.184818	0.388161	0.0	0.0	0.0	0.0	1.0

Education	count	mean	std	min	25%	50%	75%	max
1.0	174.0	0.270115	0.445300	0.0	0.0	0.0	1.0	1.0
2.0	4043.0	0.292605	0.455015	0.0	0.0	0.0	1.0	1.0
3.0	9478.0	0.242245	0.428465	0.0	0.0	0.0	0.0	1.0
4.0	62750.0	0.176351	0.381121	0.0	0.0	0.0	0.0	1.0
5.0	69910.0	0.148105	0.355206	0.0	0.0	0.0	0.0	1.0
6.0	107325.0	0.096902	0.295826	0.0	0.0	0.0	0.0	1.0

Income	count	mean	std	min	25%	50%	75%	max
1.0	9811.0	0.242891	0.428851	0.0	0.0	0.0	0.0	1.0
2.0	11783.0	0.261903	0.439689	0.0	0.0	0.0	1.0	1.0
3.0	15994.0	0.223084	0.416327	0.0	0.0	0.0	0.0	1.0
4.0	20135.0	0.201341	0.401012	0.0	0.0	0.0	0.0	1.0
5.0	25883.0	0.174014	0.379129	0.0	0.0	0.0	0.0	1.0
6.0	36470.0	0.145078	0.352184	0.0	0.0	0.0	0.0	1.0
7.0	43219.0	0.121821	0.327083	0.0	0.0	0.0	0.0	1.0
8.0	90385.0	0.079604	0.270681	0.0	0.0	0.0	0.0	1.0

BMI Mean and Standard Deviation by Gender:

Sex	mean	std
0	28.130587	7.088173
1	28.702362	5.928355

Figure(s) 1: Shows that women seem to have a higher prevalence to diabetes than men. Men = 0 and Women = 1. This could be related to women being more susceptible to obesity-related implications which contribute to increased diabetes risk.

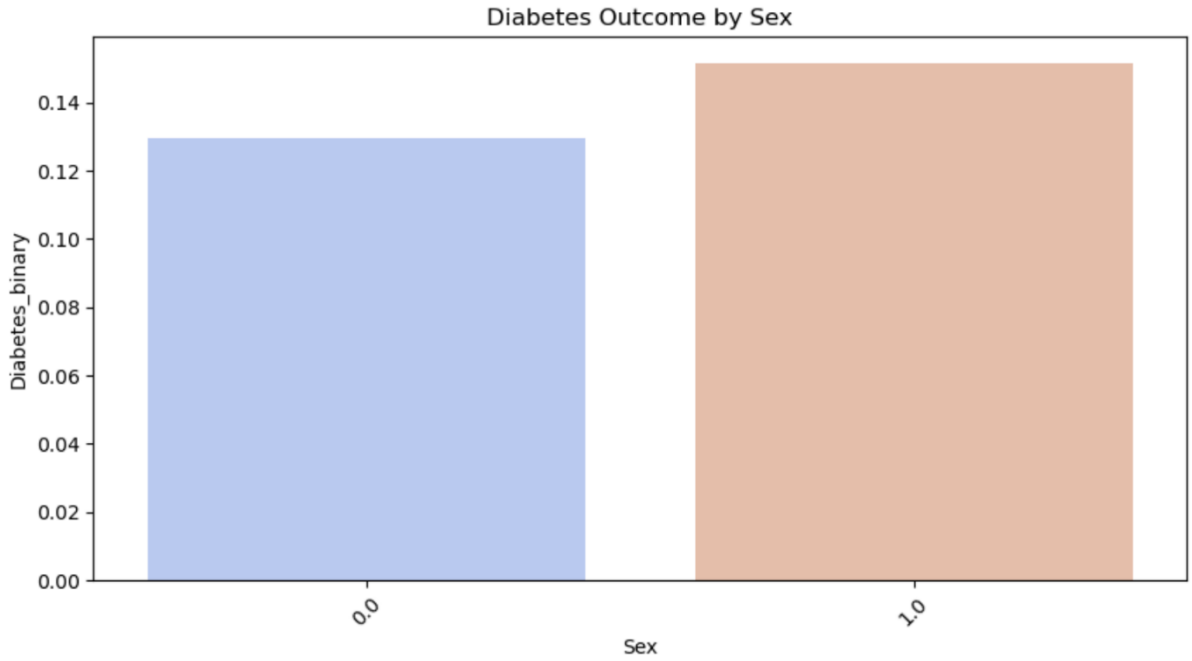


Figure 2: Shows women are more prevalent to diabetes than men.

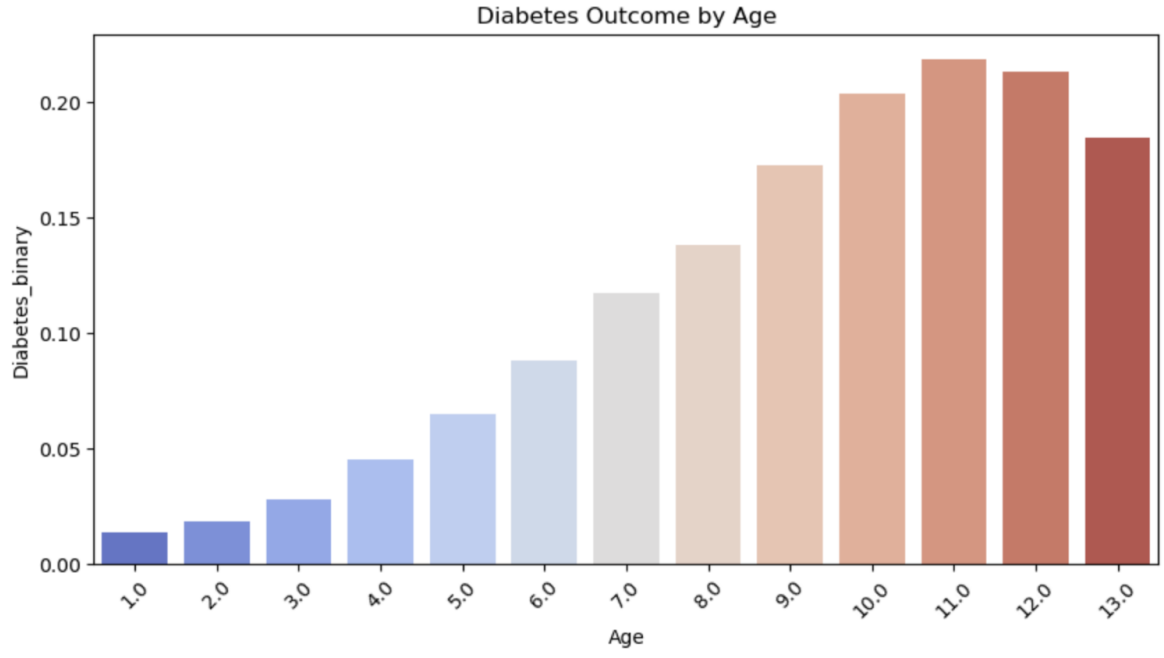


Figure 3: Young people (1-3) have the smallest predictor for diabetes showing that age plays a role in diabetes. Middle age groups (6-10) have an increased risk, and older adults have the highest risk for diabetes. Insulin decreases with age which generates an increase in metabolic issues.

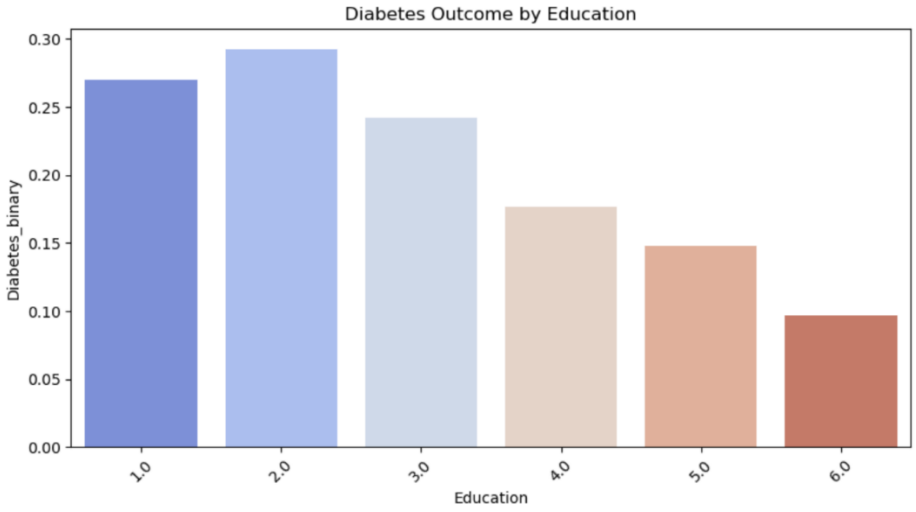


Figure 4: Shows the underrepresented people are the ones more prone to diabetes. The graph shows that the least educated (1-2) with a percentage of 26% to 29% have a way higher prevalence than most educated people (5-6) with a very low percentage of 9%. This is related to income which the next graph will explore.

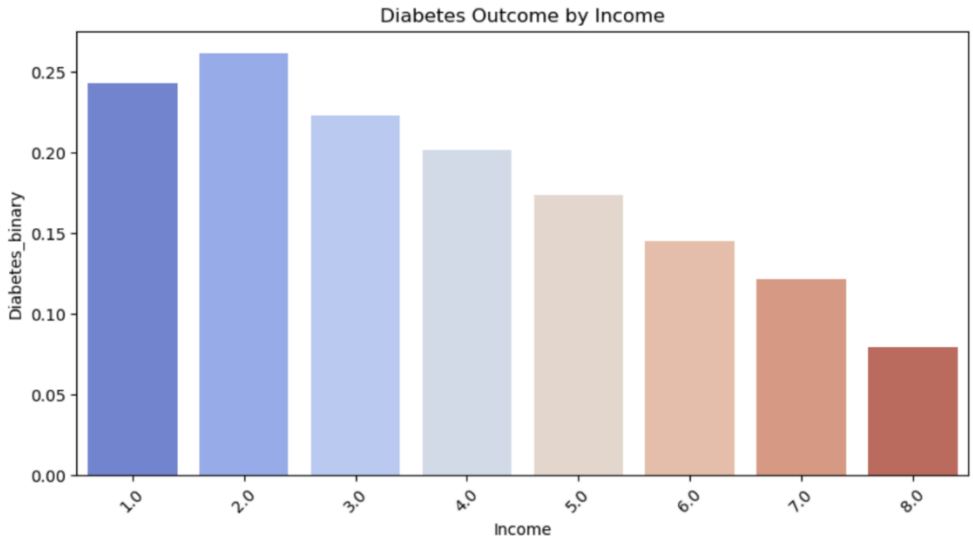


Figure 5: Economic status has a correlation with diabetes. This economic disparity provides the dataset with most of its data, effectively skewing diabetes research towards low-income individuals more so than wealthier ones.

Mitigation of Bias in Health Data

Potential mitigation techniques related to bias in healthcare data include collection of more representative data, using weighted sampling to adjust for overrepresented populations, or digging deeper to address different present biases. This project promotes the importance of deeply evaluating datasets for bias and ensuring that data-based decisions come from a place of equity and fairness for all. It is important to raise awareness in underrepresented communities about health factors to mitigate the effects of biased healthcare policy.

Closing

This project explored the presence and impact of different types of bias from the Diabetes Health Indicators Dataset. By comparing different levels of analysis and modeling we have been able to determine that age bias significantly skews the dataset, resulting in an overrepresentation of older populations. Our random forest models demonstrated that age is an important factor in diabetes prediction, but not as important as BMI, income, or other factors. Then, we explained the datasets link between underrepresented communities and diabetes, which explains why there is so much more data in certain groups than others. Together, these findings explain the ethical implications behind using biased data to inform public health decisions and their potential negative effects on the underrepresented populations.