Outlook

Re: DATA 6550 - Module 2 Group 4 Discussion

From Moses Kipkemboi <mk7v@mtmail.mtsu.edu>
Date Fri 2/7/2025 8:26 PM
To Pallavi Suram <pgs2q@mtmail.mtsu.edu>; Savi Alwis <sma7b@mtmail.mtsu.edu>; Muqtasid Qureshi <mnq2a@mtmail.mtsu.edu>

Savi and Team,

Did a preview of the information. It is a good one. It will give us good talking point and well aligned with the objectives. @Savi Alwis 👌.

Moses

Get Outlook for Android

From: Pallavi Suram <pgs2q@mtmail.mtsu.edu>
Sent: Friday, February 7, 2025 3:37:41 PM
To: Savi Alwis <sma7b@mtmail.mtsu.edu>; Moses Kipkemboi <mk7v@mtmail.mtsu.edu>; Muqtasid Qureshi <mnq2a@mtmail.mtsu.edu>
Subject: RE: DATA 6550 - Module 2 Group 4 Discussion

Hi everyone,
I agree with Savi's suggestion. What are your thoughts, everyone?

Do you have any recommendations for the dataset we should choose?

Thanks,
Pallavi.

From: Savi Alwis <sma7b@mtmail.mtsu.edu>
Sent: Thursday, February 6, 2025 7:00 PM
To: Moses Kipkemboi <mk7v@mtmail.mtsu.edu>; Muqtasid Qureshi <mnq2a@mtmail.mtsu.edu>; Pallavi Suram <pgs2q@mtmail.mtsu.edu>
Subject: Re: DATA 6550 - Module 2 Group 4 Discussion

Hey everyone,

The Heart Disease dataset that I discussed in class is AI generated and 70% male and 30% female like the other brain stroke dataset. So, I think this has good biases we can discuss.

Everyone let me know which dataset you'd like to work with and we can continue from there.

https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset
https://www.kaggle.com/code/touatousmane/demographic-group-bias-in-brain-stroke-dataset

Best,
Savi

From: Moses Kipkemboi <mk7v@mtmail.mtsu.edu>
Sent: Thursday, February 6, 2025 5:22 PM
To: Muqtasid Qureshi <mnq2a@mtmail.mtsu.edu>; Savi Alwis <sma7b@mtmail.mtsu.edu>; Pallavi Suram <pgs2q@mtmail.mtsu.edu>
Subject: Re: DATA 6550 - Module 2 Group 4 Discussion

Qureshi,

This is great. Thank you

Moses

Get Outlook for Android

From: Muqtasid Qureshi <mnq2a@mtmail.mtsu.edu>
Sent: Thursday, February 6, 2025 1:46:27 PM
To: Moses Kipkemboi <mk7v@mtmail.mtsu.edu>; Savi Alwis <sma7b@mtmail.mtsu.edu>; Pallavi Suram <pgs2q@mtmail.mtsu.edu>
Subject: Re: DATA 6550 - Module 2 Group 4 Discussion

Hello everyone.

I found a good dataset for bias: it involves obesity, gender,and race.

I also made a GitHub account for us to use for the class.

The repository is prepared for this project due next week.

Please note that the weekly group discussions are divided into weekA and weekB.

As such, this email chain is weekA.
And we will make a new email chain for weekB

I will send the login information for our group GitHub in a separate email.

Get Outlook for Android

From: Muqtasid Qureshi <mnq2a@mtmail.mtsu.edu>
Sent: Wednesday, February 5, 2025 1:11:06 PM
To: Moses Kipkemboi <mk7v@mtmail.mtsu.edu>; Savi Alwis <sma7b@mtmail.mtsu.edu>; Pallavi Suram <pgs2q@mtmail.mtsu.edu>
Subject: Re: DATA 6550 - Module 2 Group 4 Discussion

Hello everyone.

My apologies for not responding sooner.

I found an example of analyzing for bias on a dataset.

https://www.kaggle.com/code/touatousmane/demographic-group-bias-in-brain-stroke-dataset

Get Outlook for Android

From: Moses Kipkemboi <mk7v@mtmail.mtsu.edu>
Sent: Monday, February 3, 2025 11:25:53 PM
To: Savi Alwis <sma7b@mtmail.mtsu.edu>; Pallavi Suram <pgs2q@mtmail.mtsu.edu>; Muqtasid Qureshi <mnq2a@mtmail.mtsu.edu>
Subject: Re: DATA 6550 - Module 2 Group 4 Discussion

Thank you @Savi Alwis. This is really insightful.

Regards
Moses

From: Savi Alwis <sma7b@mtmail.mtsu.edu>
Sent: Tuesday, 4 February 2025 06:13
To: Pallavi Suram <pgs2q@mtmail.mtsu.edu>; Moses Kipkemboi <mk7v@mtmail.mtsu.edu>; Muqtasid Qureshi <mnq2a@mtmail.mtsu.edu>
Subject: Re: DATA 6550 - Module 2 Group 4 Discussion

Hey everyone,

Here are a couple datasets I found that may have some biases:

https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset: The idea of this Heart Attack Risk Prediction dataset sounds like it could generate useful and proactive results. However, this is a synthetic AI-generated dataset that may have a selection bias. This isn't exactly representative of real-world data. Therefore, biased data leads to biased and inaccurate results.

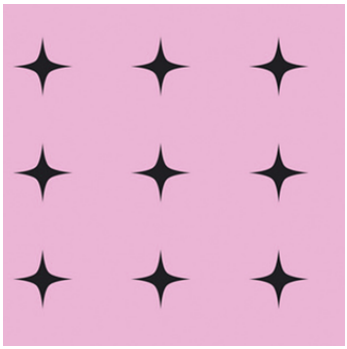
Heart Attack Risk Prediction

Dataset - Kaggle

Unlocking Predictive Insights with Multifaceted Synthetic Heart Attack Dataset

www.kaggle.com

https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who: Health related data was collected from the World Health Organization for the purpose of health analysis. Although this seems like a reliable, comprehensive dataset that spans 15 years, and has data on 193 countries, it leaves out data from 'lesser known' countries since finding all the data for those countries was difficult. This may demonstrate a selection-participate bias. From our reading, this "occurs when the data is unrepresentative due to participations gaps in the data collection process."



Life Expectancy (WHO) - Kaggle

Statistical Analysis on factors influencing Life Expectancy. Kaggle uses cookies from Google to deliver and enhance the quality of its services and to analyze traffic.

www.kaggle.com

Let me know your thoughts and other datasets you find that may be biased.

Best,
Savi

From: Savi Alwis
Sent: Monday, February 3, 2025 5:17 PM
To: Pallavi Suram <pgs2q@mtmail.mtsu.edu>; Moses Kipkemboi <mk7v@mtmail.mtsu.edu>; Muqtasid Qureshi <mnq2a@mtmail.mtsu.edu>
Subject: DATA 6550 - Module 2 Group 4 Discussion

Hello all,

I hope everyone had a great weekend!

Module 2 is about bias, representation, and fairness in datasets. Since we have no deliverables outside of the reading quiz, we should prepare for our project due on February 12th.

Here is the link to the Group Project description:
https://elearn.mtsu.edu/d2l/le/content/9996230/viewContent/104467544/View

We will essentially be analyzing a biased dataset. To start, we should each find a few biased datasets online, share them, and discuss the bias. With our options, we will be prepared to discuss our selection during this week's class, as well as have a solid dataset we can use for our project.

1. **Engage** with Your Group on the *Exploring Data Bias* Assignment:
    - Look up your assigned group in the course platform and start your group discussion.
    - Discuss potential datasets with known biases.
    - Be prepared to share your group’s selection during the Zoom session.

I will get to looking for datasets. Talk to y'all soon!

Best,
Savi