[MUSIC PLAYING]

AUDACE NAKESHIMANA: Hi, my name is Audace Nakeshimana. I am an undergraduate student and researcher at MIT. In this video, we'll continue exploring fairness in machine learning by looking at techniques for mitigating bias. Throughout the course, we'll start by illustrating bias in machine learning.

Then we'll look at techniques for mitigating bias, specifically we'll explore two types of techniques-- the database techniques, where we'll look at how to calibrate and augment our data set to mitigate bias in machine learning. And then we'll look at model-based techniques, in which you explore different model types and architectures that help us to get to a less biased model. And we'll do this by applying these techniques on the UCI Adult Data Set.

In this module, we'll explore different steps and principles involved in building less biased machine learning applications. We look at two main classes of techniques, specifically data and model-based techniques, for mitigating bias in machine learning. We will be applying these techniques on the UCI Adult Data Set with the purpose of mitigating gender bias in predicting income category.

This module is comprised of seven main parts. In part 1, we're going to look at an overview of algorithmic bias. In the second part, we will explore the UCI Adult Data Set. In the third part, we look at different data preparation steps for machine learning. In part 4, we're going to look at an example of gender bias.

And in part 5, we're going to look at different data-based approaches for mitigating gender bias. And in part 6, we're going to look at different model-based approaches. And in the last part, we'll conclude by looking at the possible next steps.

Recommended prerequisites for this module are familiarity with the fields of data science, statistics, or machine learning and familiarity with the programming tools that we'll be using. These are Python, Pandas, and the Scikit-Learn Library.

In part 1 of this module, we will start by understanding algorithmic bias. We will

define it and look at its sources and implications. Throughout the module, we will use the term bias, algorithmic bias, or model bias to describe systematic errors in algorithms or models that could lead to potentially unfair outcomes. We will identify bias qualitatively and quantitatively by looking at model errors, disparities across different gender demographics. And notice that, throughout the module, we will use gender to describe biological sex at birth.

So what are some potential sources of algorithmic bias? First, bias can come during direct collection. And this could happen when the data that you collect already contains some systematic biases or stereotypes about some demographics. This could also happen if different demographics in our data set are not equally represented.

The second example of how bias could come in machine learning is in the training process, when our models are not penalized for being biased. Algorithmic bias is a problem because of different reasons. It leads to unfair outcomes toward some individuals or demographics and it leads to further bias propagation, creating a feedback cycle of bias.

In the second part of this module, we'll explore the UCI Adult Data Set by establishing familiarity with the data set and looking at different distributions in the data set. The UCI Adult Data Set is one of the most popular machine learning data sets. It is available on the internet on the UCI Machine Learning Repository. The data set is comprised of more than 48,000 data points that were extracted from the 1994 Census database in the United States.

Each data point in the data set it is comprised of 15 features. These include age, work class, education, relationship, race, sex, salary, and others. If you look at the gender distribution in the data set, you can see that about 16,000 individuals identify as female. And about 32,000 individuals identify as male. If you look at the race distribution, you can see that slightly more than 40,000 individuals identify as white. And about 4,000 to 5,000 individuals identify as black. The rest is other minorities.

If we look at the distribution of income category in the general population, we can see that about 37,000 individuals earn less or equal to $50,000. And only about

between 12,000 and 13,000 individuals earn more than $50,000. By looking at the income level distribution across different levels, we can see that the ratio of male individuals that make more than $50,000 is about a third. But for the female demographic, this ratio drops to about 20%.

An important observation from what we've seen so far is that the number of data points in the male population is significantly higher than the number of data points in the female population, exceeding it by more than three times in the higher income category. Therefore, it is very important to think about how this representation disparity might affect predictions of a model trained from this data.

In the third part of this module, we are going to explore different steps involved in transforming our data from raw representation to appropriate numerical or categorical representation in order to be able to perform machine learning tasks. An example of transformation to be made is the conversion of native country from raw representation to binary. In this example, we decided to assign a binary label to individuals who come from the United States and another binary label to individuals who come outside of the United States.

We applied the same transformation to the sex and salary attribute, since each one of these attributes has to possible values in the data set, therefore making binary representation appropriate. There are more than two possible values that the relationship attribute can take. Therefore, for this attribute, we use one-hot encoding, which is more powerful than binary encoding because it can encode an alphabet of any size.

We applied the same transformation from raw representation to binary or one-hot to all other categorical attributes. In most cases, we chose binary encoding for simplicity. But this is often a decision that has to be made on a case-by-case basis, depending on the application. It is also important to note that converting features like work class to binary can be problematic if individuals from different categories have systematically different levels of income. On the other hand, not doing this might be a problem if one category has very few people in it that we can generalize from.

In the fourth part of this module, we are going to illustrate gender bias. We will

apply the standard machine learning approach to our data and then evaluate the bias in the task of predicting income category. We start by splitting the data set into the training and test data. We then feed MLPClassifier on training data, then use the model to make prediction on the test data.

In case you're not familiar with the multi-layer perceptron or MLPClassifier, this model belongs to the class of feedforward neural networks. Each node uses a non-linear activation function, giving the model ability to separate non-linear data. The model is trained using backpropagation technique. However, a few downsides is that the model suffers overfitting, and it is not easily interpretable.

Before we evaluate our model, let's start by establishing some terminology. Throughout the rest of the module, we will refer to the positive category as the group of individuals that earn more than $50,000 a year, or the high-income category. And we will refer to the negative category as the group of individuals that earn $50,000 a year or less. We will also refer to it as the low-income category.

Now that we've established important terminology, let's look at different error rate metrics for the model that we trained previously across different gender demographics. If you look at accuracy, you can see that the accuracy for the male demographic is about 0.8, while the accuracy for the female demographic is about 90%, or 0.9. If you look at the positive rate and the true positive rate, you can see that both of these metrics are higher for the male demographic than for the female demographic. However, if you look at the negative rate and the true negative rate, you can see that these metrics are higher for the female demographic than for the male demographic instead.

The metrics that we just saw indicate consistent disparity in error rate between the male and the female demographic. This is what will define as gender bias. Mitigating gender bias, in this case, is equivalent to using different techniques to minimize this disparity. And this will be the focus of the rest of the module.

In part 5 of our module, we are going to explore different database debiasing techniques. More specifically, we will look at different ways we can recalibrate or augment our data set in a way that makes predictions less biased. The motivation behind this is from our hypothesis that the gender bias could come from unequal

representation of male and female demographics in our data set. We therefore make an attempt to recalibrate or augment the data set with the aim of equalizing gender representation in our training data.

The first database technique that you're going to explore is called debiasing by unawareness. And in this technique, we mitigate gender bias by removing gender from the attributes that we train on. The code snippet shows our implementation.

By looking at the results for our debiasing by unawareness technique, we can see that, although there was not significant improvement in reducing the gap for accuracy, we were able to reduce the gap for other metrics, like the positive rate, the negative rate, the true positive rate, and the true negative rate. And this is an example of how a debasing technique might not be able to achieve an improvement in all the metrics, although it might see a significant reduction in the gap for other metrics of interest.

Debiasing by unawareness can be one approach to mitigate gender bias to some extent, as we saw in our results. However, studies have shown that this method can be ineffective, especially if there are other features in the data set that have some correlation with the protected attributes that we are dropping. These type of attributes are referred to as proxy variables.

The second database technique that you're going to explore is equalizing the number of data points. And in this approach, we will attempt to equalize representation by using equal number or equal ratio of male and female individuals in our data set or within each income category.

We will start by attempting to equalize the number of data points per gender category. And in this approach, we're going to draw a sample in which there is equal number of data points from the male demographic and from the female demographic. Feel free to pause the video to understand the implementation.

These are the results that we got from training an MLPClassifier on a data set in which there is an equal number of data points per gender category. Feel free to pause the video to understand the results in more detail.

The next attempt will be to equalize the number of data points per income level in

each gender category. And what this means is that the number of high-income and low-income earners is the same in the male and the female demographic for the sample that you're going to use to train. Here are the key metrics from a model that was trained on a sample of the data set in which there is an equal number of data points per income level in each gender category.

One downside to this methodology of equalizing the number of data points per demographic is that the size of the resulting data set depends on the size of the smallest demographic. Therefore, if the smallest demographic has a very small number of data points, you're going to end up with a very small training set. Therefore, in some cases, you might find that equalizing the ratio instead of the number of data points by demographic can lead to a higher resulting sample size. And that's what we're going to look at in the next approach.

In this methodology of equalizing the ratio of the number of data points per income level in each category, we equalize the ratio of male individuals with high income to male individuals with low income. And we do this for the female demographic, as well. This results into a higher sample size. And to see how this is the case, I encourage you to look at the notebook for this work.

This is the plot for the results of this methodology. And you can see that, although there is some gap for the accuracy and true positive rate, the gap is way smaller for positive rate and negative rate or true negative rate.

The next technique that you're going to look at is counterfactual augmentation. And in this approach, for each data point Xi with a given gender, we generate a new data point Yi that differs with Xi only at the gender attribute. And we add Yi to our training data set. I encourage you to pause a little bit and convince yourself that the resulting data set from counterfactual augmentation will satisfy all the following constraints shown here.

The code snippet shown here shows our implementation of counterfactual augmentation on the data set. By looking at the results of counterfactual augmentation on our data set, you can see that the gap for all the metrics that you're looking at for the male and female demographic is pretty much gone. And this is what we want and expect from a fair machine learning model.

So let's compare all the metrics of interest on all the approaches that we've carried out so far. By looking at the metric of overall accuracy, you can see that some techniques, like equal number of data points per gender or counterfactual augmentation, leads to higher degrees of accuracy. But you can see that some techniques, like gender unawareness, do not always guarantee higher accuracy.

By looking at accuracy across gender, you can see that the counterfactual augmentation technique still has the smallest gap between male and female. But you can see that some techniques like gender unawareness still have a significantly higher gap between the accuracy in male versus female demographics.

The plots shown showed the comparison between the positive rates across gender and the comparison between the negative rates across gender. The plots shown here show the comparison between true positive rate across gender and the comparison between true negative rates across gender for all the techniques that we covered so far.

In this part, we are going to explore model-based debiasing techniques. And specifically, we will look at different model types and architectures and examine how each one of them performs for the male versus female demographic. The motivation for this is that we should expect different models to have different degrees of bias. Therefore, by changing the model type or architecture, we can observe which ones tend to be inherently less biased. And these are the ones that we are going to choose in our application.

We start by examining single-model architectures. And for each of the model families shown here, we picked one model and trained it on the data that we have. This code snippet shows different models and parameters that we used. For simplicity, we used different parameters for different models. But in a practical setting, we would have to use a technique like cross-validation or hyperparameter search to find the best parameter to use for each model.

We also examined multi-model architectures. In this approach, we trained a group of different models on the same data and then make a final prediction based on consensus. We compared two types of consensus.

The hard voting consensus is the one in which the final prediction is the majority

prediction among all models. And in the soft voting consensus, the final prediction is the average prediction across all models in consideration. We use Scikit-Learn VotingClassifier to combine single models and train them all at once. The code snippet shown here shows the models that we used and how we trained the VotingClassifiers on our data.

Let us now evaluate and compare the metrics of interest on all model types and architectures that we've trained on so far. This plot shows the results for overall accuracy. You can see that the random forest classifier has the highest degree of accuracy, which is about 94%. You can also see that the Gaussian Naive Bayes model has about 72% of accuracy. And you can also see that all the other models fall in between.

This plot shows accuracy across gender. And you can see that there are different levels of the gap between the male and female demographic, depending on the model type. And this is an example that indicates that different models inherently have different levels of bias.

These plots show the positive and negative rates across gender. If you look at the plot for the positive rate, you observe that the positive rate is always higher for the male demographic than the female demographic for all the models that we've looked that. And this can be problematic if we deploy any of these models in the real world, because you end up in a scenario in which the model just systematically predicts more favorable outcome for the male demographic than the female demographic.

These plots show the true positive and true negative rates across gender. If you look at the plot for the true positive rate, you observe that the true positive rate is always higher for the male individuals than female individuals. And if you look at the plot for the true negative rate, it's the other way around.

The true negative rate is always higher for the female demographic than for the male demographic. And this can especially be problematic, because it shows that our models have learned how to better classify high-income male earners than high-income female earners and to classify low-income female earners than low-income male earners, which means they could be widening the gap between the

male earners and the female earners.

To account for randomness, we ran the previous experiment five more times in order to get a better idea of the average model behavior. To compare the metrics across multiple training sessions, we created five instances of each model type. And we trained each one of them on the data. Then, for each one of these instances, we evaluated the absolute value of the difference in each metric of interest between the male and the female demographics from the test data.

If you look at the plot for the accuracy disparity comparison, you can see that models like logistic regression or hard voting or SVC have a significantly lower accuracy disparity than Gaussian Naive Bayes or random forest. We see a similar trend by looking at the positive and negative rate disparity. If you look at models like logistic regression or SVC or had voting, you can see that they have significantly lower disparity than GNB.

Surprisingly, we see a significantly different result by looking at the true positive and negative rate disparity. If you look at the true positive rate disparity, you can see that models like logistic regression or SVC now have higher disparity and higher variability than models like GNB. But if you look at the plot for the true negative rate, you can see that it tends to follow the previous trend, where logistic regression and SVC have lower variability and lower disparity than GNB. It is therefore very important to see that these models have different inherent behaviors when it comes to bias.

In the last part, we conclude by looking at the possible next steps that will allow us to strengthen our understanding and application of ethics in machine learning from the technical perspective. Now that you've gone through the entire module, we invite you to check out our GitHub repository. This will help you deepen your understanding of the work that's being done.

In addition, we also encourage you to explore more advanced debiasing techniques. And we also recommend sharing and discussing these across your team, organization, or community. And of course, we all need to take action by applying what we learned in what we do every day. Finally, here are the references to the materials that we consulted while making the module.

Thank you for following this course on mitigating bias in machine learning. And I hope that this helps you build less biased machine learning applications in the future.

[MUSIC PLAYING]