Project Steps

- Data Preparation:
  - Variable Encoding on Train/Validation/Test sets (See slide 2 for example)
  - Feature “State” -> OneHotEncoding
- Feature Selection
  - Run exhaustive search for feature selection using Logistic Regression model
  - Check for multicollinearity using VIF method
- Modeling (LR)
  - Run Logistic Regression on selected features
  - Check for fairness and do debiasing if needed
  - Report weights and confusion matrix
- Modeling (RF)
  - Run Random Forest on the selected features
  - Check for fairness and do debiasing if needed
  - Report feature importance and confusion matrix
  - (If time permits) Run RF on all features, and find overlapping features with LR model
- Model Selection
  - Choose between RF and LR based on Accuracy/Fairness trade-off
- Investigate ”bank_xyz” treatment
  - Answer the related question accordingly.
- Describe the rejection scenario
  - We use contrastive explanation for that.
- (If Time Permits) create a simple API for reporting the credit
- Writing Report and creating slides

- All predictors' values should be encoded into numbers 1,2,3,4 and 5. This can be done via percentiles.
  - If any predictors have NaN values, number "0" should be assigned.
  - Variable "ind_acc_XYZ" should be remained untouched (0,1).
  - Variable "States" should be one hot encoded.
  - Variable "Income" should be encoded within corresponding State.

Dataset 1

P1	P2	P3	Ind_acc_XYZ	isAZ	isNC	…	Default_ind
1	2	3	0	0	1	…	1
2	4	2	1	0	0		0
5	1	1	0	1	0		0

Dataset 2

P1	P2	P3	Ind_acc_XYZ	isAZ	isNC	…	Num_Defaulted	Num_Acc
1	2	3	0	0	1	…	38	90
2	4	2	1	0	0		58	120
5	1	1	0	1	0		90	200