Noname manuscript No.
(will be inserted by the editor)

Machine Learning Techniques with Fairness for Prediction of Completion of Drug and Alcohol Rehabilitation

Karen Roberts-Licklider · Theodore Trafalis

Received: date / Accepted: date

Abstract The aim of this study is to look at predicting whether a person will complete a drug and alcohol rehabilitation program and the number of times a person attends. The study is based on demographic data obtained from Substance Abuse and Mental Health Services Administration (SAMHSA) from both admissions and discharge data from drug and alcohol rehabilitation centers in Oklahoma. Demographic data is highly categorical which led to binary encoding being used and various fairness measures being utilized to mitigate bias of nine demographic variables. Kernel methods such as linear, polynomial, sigmoid, and radial basis functions were compared using support vector machines at various parameter ranges to find the optimal values. These were then compared to methods such as decision trees, random forests, and neural networks. Synthetic Minority Oversampling Technique Nominal (SMOTEN) for categorical data was used to balance the data with imputation for missing data. The nine bias variables were then intersectionalized to mitigate bias and the dual and triple interactions were integrated to use the probabilities to look at worst case ratio fairness mitigation. Disparate Impact, Statistical Parity difference, Conditional Statistical Parity Ratio, Demographic Parity, Demographic Parity Ratio, Equalized Odds, Equalized Odds Ratio, Equal Opportunity, and Equalized Opportunity Ratio were all explored at both the binary and multiclass scenarios.

K. Roberts-Licklider
University of Oklahoma School of Industrial and Systems Engineering
E-mail: Karen.R.RobertsLicklider-1@ou.edu

T. Trafalis
University of Oklahoma School of Industrial and Systems Engineering E-mail: ttrafalis@ou.edu

Keywords Support Vector Machines · Kernel Methods · Fairness Measures · SMOTEN · Decision Trees · Random Forests · Neural Networks

1 Introduction

Substance abuse is one of the leading causes for mental illness and these issues are dealt with in the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Classification of Diseases (ICD) dsm (2021). 36.2% of adults age 18-25 had a mental illness, 29.4% for age 26-49 and 13.9% 50 or older with 11.6% being serious mental illness for age 18-25, 7.6% for 26-49 years old and 3.0% for 50 years or older National Survey on Drug Use and Health (2022). The COMPAS assessment has been used in criminal sentencing to predict whether an offender will reoffend, however it has proven to be racially biased. COMPAS is the Correctional Offender Management Profiling for Alternative Sanctions. It was developed in 1998 and uses a recidivism risk scale and has been used since 2000. It predicts a defendant's risk of committing a misdemeanor or felony within 2 years of assessment for 137 features about an individual and the individual's past criminal record Dressel and Farid (2018). The COMPAS assessment incorrectly predicted that whites would reoffend at a rate of 47.7% which was twice the rate of blacks at 28.0%. It favored white defendants over blacks. Its accuracy for white defendants was 67% and 63.8% for black defendants Dressel and Farid (2018). For this reason, considering fairness in these types of assessments is important. Also looking at whether an offender would complete rehab instead of going to prison is important as an alternative to sending them to prison due to overcrowding in prisons especially in Oklahoma. Oklahoma itself ranks 3rd if looking at the world's incarceration rate considering every state as a country

Wildra and Herring (2021). For the scope of this paper, treatment episode data from the Substance Abuse and Mental Health Services website Substance Abuse and Mental Health Services Administration (2021) is used to look at predicting whether a person completed a drug and alcohol rehabilitation program, the number of prior treatments a person has been to when entering treatment, and then the concatenation of both variables.Chawla et al. (2002) thoroughly reviewed
SMOTE: Synthetic Minority Over-Sampling Technique for both the continuous and Nominal case. The nominal case is what we will use from this work. The algorithms and use of the value distance metric were described for SMOTE-N and SMOTE-NC and how these relate to the ROC curve. Dual and three-way interactions of both the additive and multiplicative type are discussed in Veenstra (2011) and the intersecting of bias terms in shown in this work. Demographic Parity, Disparate Impact, Statistical Parity Difference, Equal Opportunity, Equalized Odds, and Min-Max Fairness were all discussed in length in Irfan et al. (2023). They all compared models such as Support Vector Classifiers, Gaussian Process Classifiers, Gaussian Naïve Bayesian, and Linear Discriminant Analysis. Worst-case fairness metrics such as Demographic Parity Ratio, Disparate Impact Ratio, Conditional Statistical Parity Ratio, Equal Opportunity Ratio, and Equal Odds Ratio as well as a multiclass example was shown in Ghosh et al. (2021). These fairness measures can be used when there are multiple classes or when the reweighting or intersectionality has been done and there are more than two ratios to evaluate between. This paper is organized as follows: in section 2 we discuss the methodology. In sections 3, 4 and 5 we discuss issues of data cleaning, encoding, balancing and data sensitivities and distributions. Section 6 discusses fairness measures applied to our data. In sections 7 and 8 we explore several machine learning models and interpretation of the results. Sections 9 and 10 discusses reweighting and new fairness calculations. Finally sections 11 and 12 discuss the conclusions and future work.

2 Methodology

Various machine learning algorithms were used on this data with a focus on kernel methods for support vector machines. With this data being highly categorical in nature, encoding techniques were used to transform the data to make it more manageable. The data was balanced to make predictions more accurate, and the missing data was imputed. Fairness measures were compared before and after weighing the variables using two- and three-way interactions with chi squared significance and intersectionalizing the nine bias variables. The nine bias variables explored were gender, veteran status, marital status, education, age, employment, pregnancy status, race, and ethnicity. Four kernels were compared; linear, polynomial, radial basis function, and sigmoid. These were tried at a range of degrees, c values, gammas, and r values. Decision trees, random forests, and neural networks were compared. The models were then compared to each other to see which models outperformed each other.

3 Data Clean Up Process

The data set was filtered down to Oklahoma and down to 30-day rehabilitation centers only. A calculated column was created called COMPLETED in which a value of 'COMPLETE' was taken if the REASON variable took on a value of 1 and 'INCOMPLETE' if REASON took on a value of anything else. The reason code translations can be seen in figure 1 below.

Value	Label
1	Treatment completed
2	Dropped out of treatment
3	Terminated by facility
4	Transferred to another treatment program or facility
5	Incarcerated
6	Death
7	Other

Fig. 1: Reason Codes

Some user defined functions were created to clean up the data. One removes unneeded columns that just have a constant value in them or had an identification number for the CASEID, admit year (ADMYR), REGION and STFIPS could be removed since the data was filtered for Oklahoma. STFIPS was the FIPS code for each state. This was filtered for 40 for Oklahoma as mentioned above.

4 Encoding and Balancing Data

Next, three data sets were created. One was for the completed predictions, one was for the reasons predictions, and one was for the no-priors (number of times in treatment) predictions. Each data set went through an encoding and data balancing processes. One hot encoding was used to encode the categorical variables into new binary variables. Each class within that variable is created into a new column which takes on the value of 1 if it appears for that record and 0 otherwise. An Example of how the services classes in figure 2 are translated

to binary variables in figure 3 can be seen below. A description of what each service class is, is included.

	COMPLETED	SERVICES
26508	INCOMPLETE	7
28648	INCOMPLETE	7
34402	COMPLETE	4
35205	INCOMPLETE	2
186619	COMPLETE	7
...	...	...
1640231	INCOMPLETE	7
1642319	INCOMPLETE	4
1644465	COMPLETE	7
1645948	COMPLETE	4
1671427	INCOMPLETE	7

	SERVICES_4	SERVICES_5	SERVICES_7
4853	0.0	0.0	1.0
2278	0.0	0.0	1.0
270	0.0	0.0	1.0
1420	0.0	0.0	1.0
10855	0.0	0.0	0.0

variable	meaning
SERVICES_4	Rehab/residential, short term (30 days or fewer)
SERVICES_5	Rehab/residential, long term (more than 30 days)
SERVICES_7	Ambulatory, non-intensive outpatient

Fig. 2: Services One Hot Encoding

Next, we look at how the data is imbalanced in all three data sets. We use the SMOTEN (Synthetic Minority Over-Sampling Technique Nominal) package in Python which uses K-Nearest neighbors Gajawada (2021) algorithm to balance the data. See the figures below. This is done for the three data frames used in prediction.

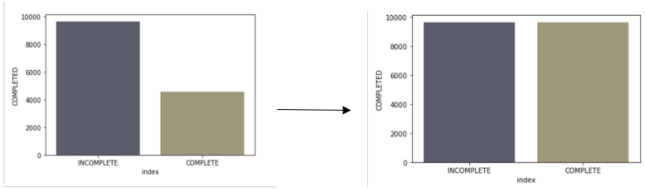


Fig. 3: Completed SMOTE Applied

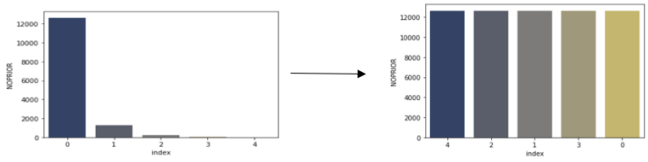


Fig. 4: NOPRIOR SMOTE Applied

Typically SMOTE is used which helps to improve the prediction accuracy. It distributes the instances of the majority class and the minority class equally. SMOTE technique increases the predictive accuracy over the minority class by creating synthetic instances of that minority class, Jishan et al. (2015). SMOTEN extends SMOTE for nominal features by getting the nearest

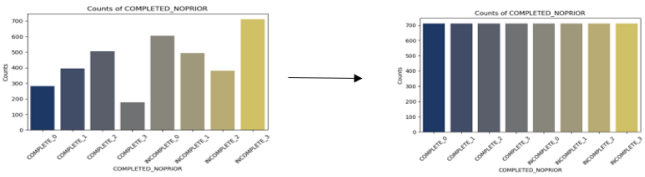


Fig. 5: Completed_NORPIOR SMOTE Applied

neighbors by using the modified version of the Value Difference Metric which looks at the overlap of feature values over all feature vectors. A matrix of features for all feature vectors is created and the distance between those features is defined in the following equation:

$$\delta(V_1, V_2) = \sum_{i=1}^{n} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right|^k \tag{1}$$

Where V_1 and V_2 are the corresponding feature values and C_1 is the total number of times V_1 occurs and C_{1i} is the total occurrences of feature V_1 for class i. The same holds for V_2, C_2, and C_{2i}. k is some constant that is typically set to 1. This gives the matrix of value differences for each nominal value in the set of feature vectorsChawla et al. (2002). An example given can be seen on the next page.

Let F1 = Red , Orange, Yellow, Green, Blue

Let F2 = Red, Purple, Yellow, Teal, Brown

Let F3 = Black, Orange, Yellow, Green, Brown

SMOTEN would create the following:

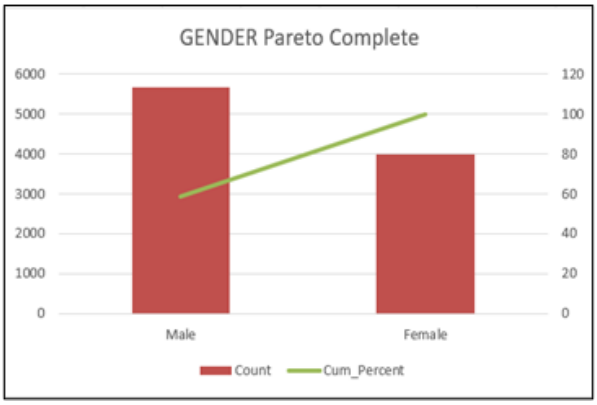
FS= Red, Orange, Yellow, Green, Brown

Fig. 6: SMOTEN Example

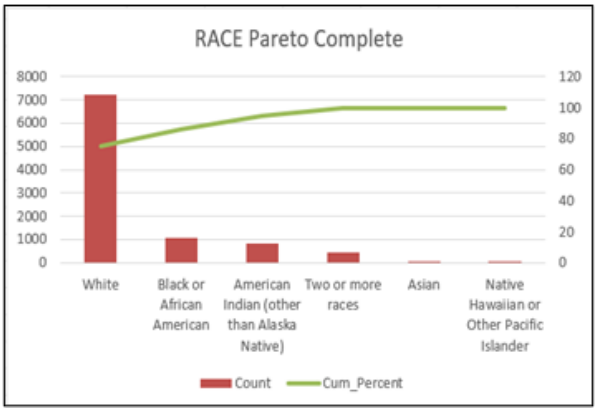
5 Data Sensitivities/Distributions

5.1 Pareto Charts

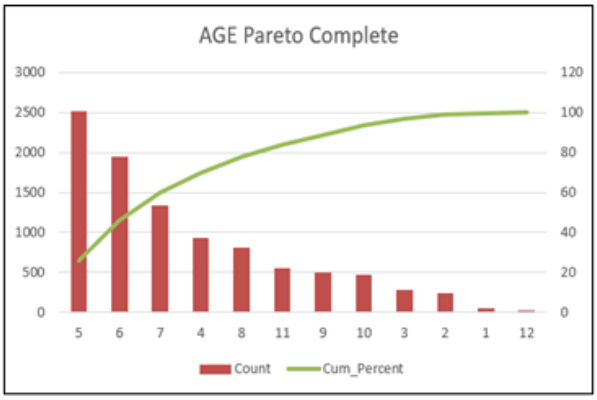
Nine variables were chosen and distributed by Pareto charts to see where bias occurred. These variables were gender, race, age, ethnic, veteran status, education status, marital status, employment status, and pregnancy status. These are shown for completion rehab outcomes in the figures below.



(a) Gender Pareto

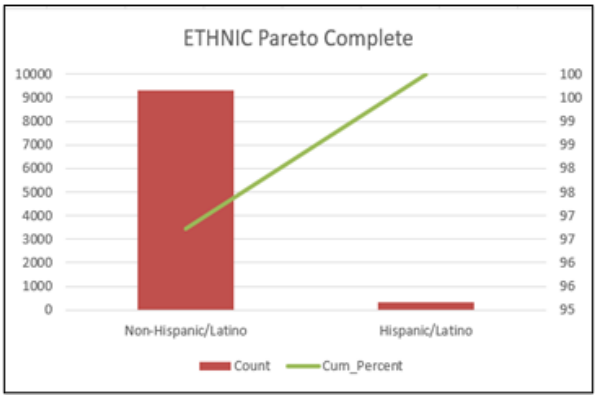


(b) Race Pareto

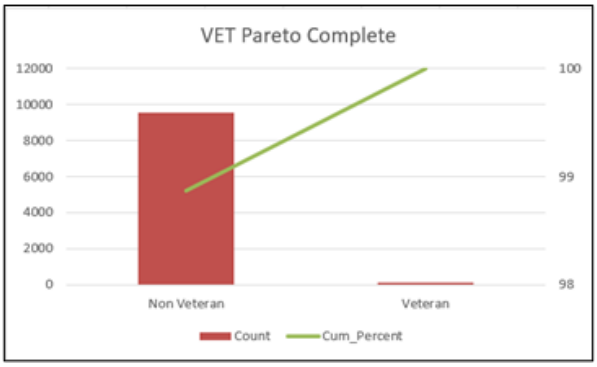


(c) Age Pareto

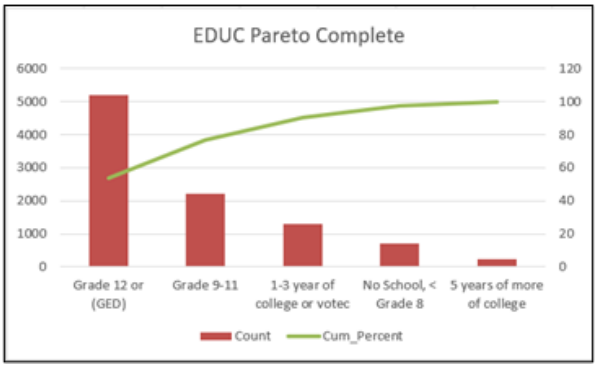
Fig. 7: Pareto Charts for Gender, Race, and Age



(a) ETHNIC Pareto

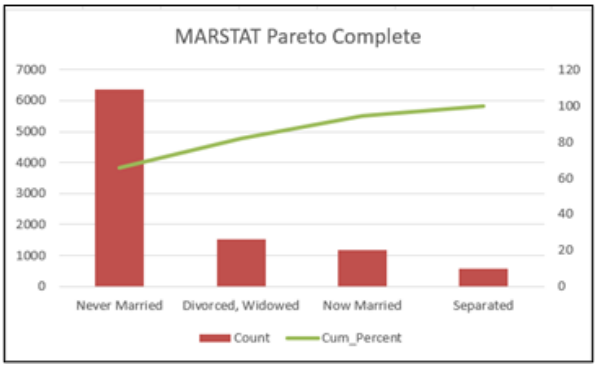


(b) VET Pareto

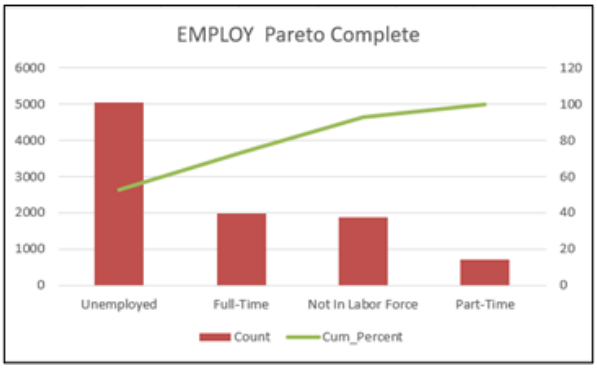


(c) EDUC Pareto

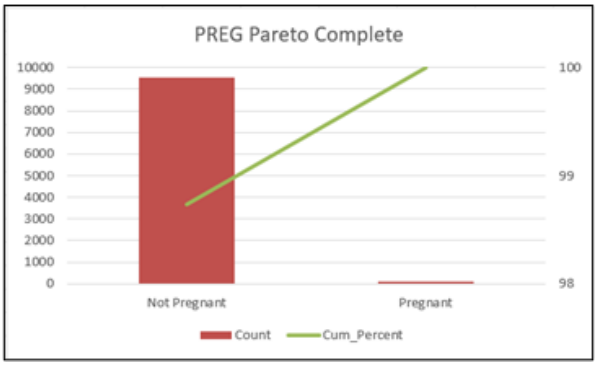
Fig. 8: Pareto Charts for Ethnicity, Veteran Status, and Education Status



(a) MARSTAT Pareto



(b) EMPLOY Pareto



(c) PREG Pareto

Fig. 9: Pareto Charts for Marital Status, Employment Status, and Pregnancy Status

5.2 Bucketized Categories

Based on the Pareto charts above, each variable was bucketized into dichotomous variables. Variables such as gender, pregnancy, ethnicity, and veteran status were already dichotomous so they did not change. Race became either white or non-white, employment status became employed or not employed, age became under 40 or 40 plus, marital status became never married or married/previously married, and education status became college or no college.

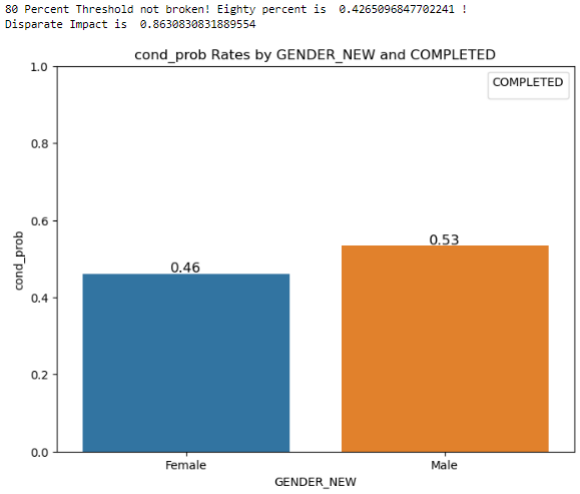
6 Fairness Measures

6.1 Disparate Impact

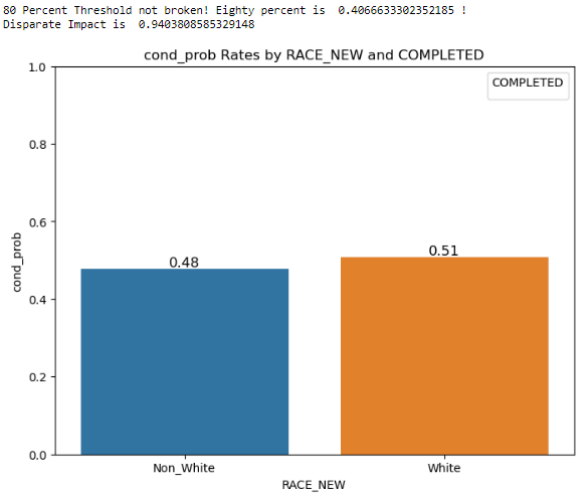
To find where the discrimination occurs, we first look at disparate impact in the dichotomous variables.

$$DI = \frac{P(\hat{Y} = 1 \mid A = 0)}{P(\hat{Y} = 1 \mid A = 1)} \tag{2}$$

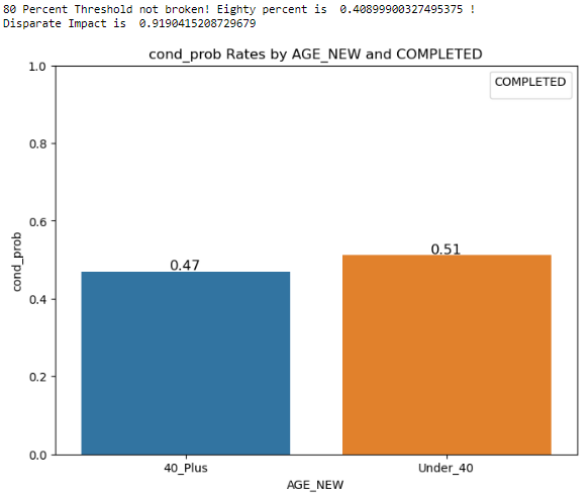
This compares the proportion of individuals receiving a favorable outcome for a privileged and underprivileged group. The closer to 1 it is, the more fair it is. $\hat{Y}$ is the model predictions and A is the protected attribute, with 0 being the underprivileged class and 1 being the privileged class Irfan et al. (2023). The resulting disparate impact charts for the completed model can be seen in the figures below. Typically the 80% rule is followed for the threshold to be considered discriminatory Ghosh et al. (2021).



(a) DI Dichotomous Gender

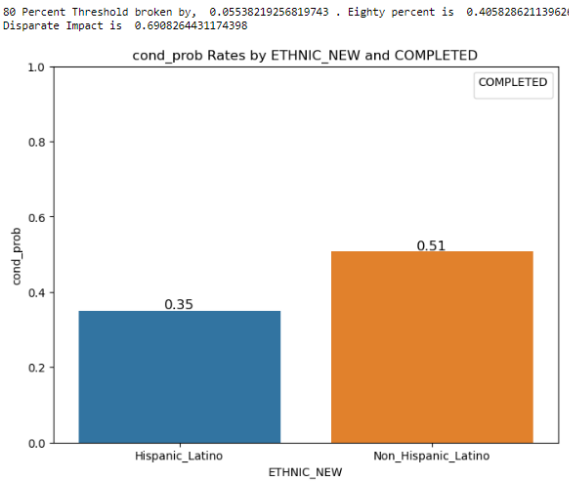


(b) DI Dichotomous Race

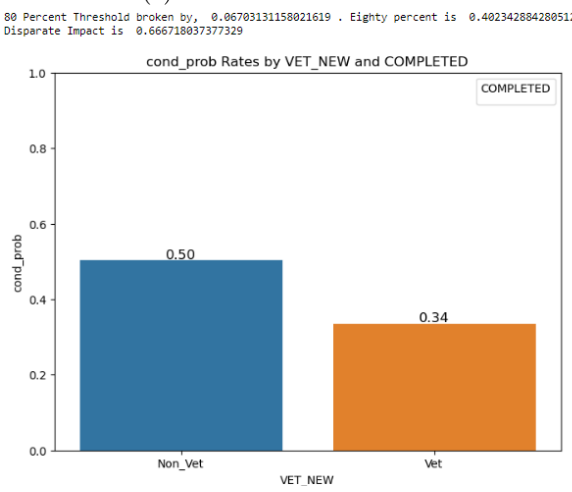


(c) DI Dichotomous Age

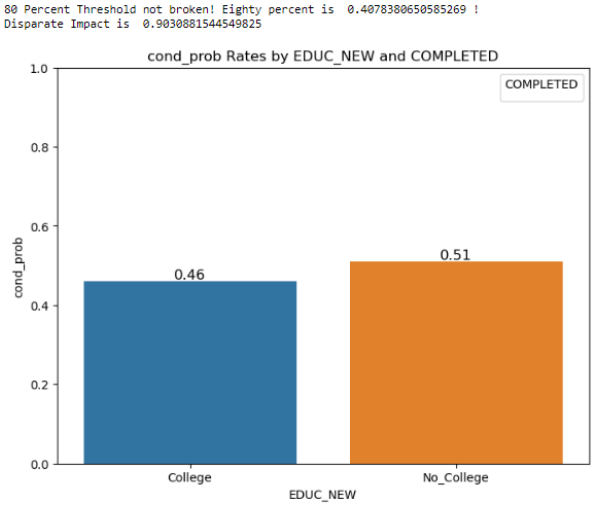
Fig. 10: DI Dichotomous Charts for Gender, Race, and Age



(a) DI Dichotomous ETHNIC

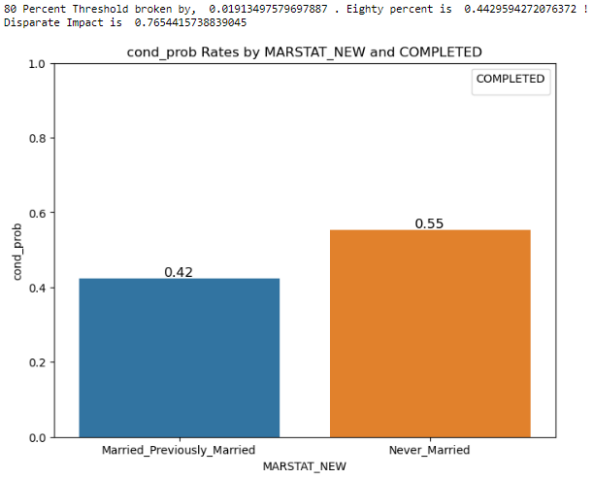


(b) DI Dichotomous VET

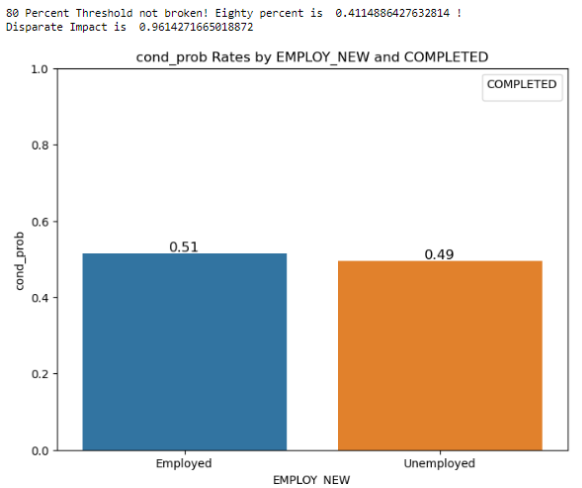


(c) DI Dichotomous EDUC

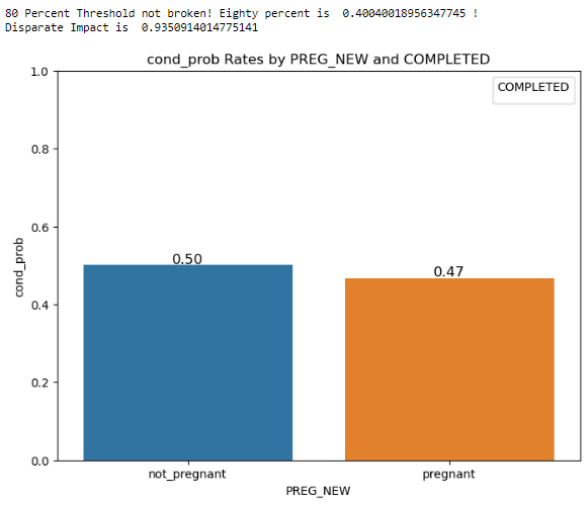
Fig. 11: DI Dichotomous Charts for Ethnicity, Veteran Status, and Education Status



(a) DI Dichotomous MARSTAT Pareto



(b) DI Dichotomous EMPLOY



(c) DI Dichotomous PREG

Fig. 12: DI Dichotomous Charts for Marital Status, Employment Status, and Pregnancy Status

6.2 Disparate Impact – Multiclass

Next we look at the multiclass case in which we find the disparate impact for each class and divide the minimum conditional probability when the target matches the outcome by the maximum conditional probability when the target matches the outcome.

$$\text{DI}_{\text{Multiclass}} = \frac{\min\Big(P(\hat{Y} = 1 \mid A = 1)\Big)}{\max\Big(P(\hat{Y} = 1 \mid A = 1)\Big)} \tag{3}$$

The outputs for noprior model for each of the variables can be seen in the table below.

Table 1: Disparate Impact Analysis

Variable	Class	DI	80% Threshold Broken
GENDER	0	0.43	UNFAIR
	3	0.78	UNFAIR
	2	0.02	UNFAIR
	1	0.81	FAIR
AGE	0	0.56	UNFAIR
	3	0.01	UNFAIR
	2	0.65	UNFAIR
	1	0.97	FAIR
VET	0	0.28	UNFAIR
	3	1.00	FAIR
	2	0.17	UNFAIR
	1	0.33	UNFAIR
EDUC	0	0.80	UNFAIR
	3	1.00	FAIR
	2	0.38	UNFAIR
	1	0.73	UNFAIR
MARSTAT	0	0.58	UNFAIR
	3	0.42	UNFAIR
	2	0.67	UNFAIR
	1	0.77	UNFAIR
EMPLOY	0	0.36	UNFAIR
	3	1.00	FAIR
	2	1.00	FAIR
	1	0.75	UNFAIR
RACE	0	0.92	FAIR
	3	0.00	UNFAIR
	2	0.34	UNFAIR
	1	0.81	FAIR
ETHNIC	0	0.64	UNFAIR
	3	1.00	FAIR
	2	0.43	UNFAIR
	1	0.20	UNFAIR
PREG	0	0.30	UNFAIR
	3	0.17	UNFAIR
	2	1.00	FAIR
	1	0.52	UNFAIR

6.3 Statistical Parity Difference

For the completed model we next look at Statistical Parity Difference (SPD). The closer the result is to 0, the more fair it is.

$$SPD = P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1), \tag{4}$$

where $\hat{Y}$ is the models predictions, $A = 0$ is the protected attribute for the unprivileged class and $A = 1$ is the protected attribute for the privileged class. Irfan et al. (2023). The results can be seen for each variable in the table below.

Table 2: Statistical Parity Difference (SPD) by Variable

Variable	Class	SPD	SPD Difference
GENDER	Male	0.53	0.07
	Female	0.46	
AGE	Under 40	0.51	0.04
	40 Plus	0.47	
VET	Veteran	0.34	0.16
	Non-Veteran	0.50	
EDUC	No College	0.51	0.05
	College	0.46	
MARSTAT	Married/Previously Married	0.55	0.13
	Never Married	0.42	
EMPLOY	Unemployed	0.49	0.02
	Employed	0.51	
RACE	White	0.51	0.03
	Non-White	0.48	
ETHNIC	Hispanic/Latino	0.51	0.16
	NonHispanic/Latino	0.35	
PREG	Pregnant	0.47	0.03
	Not Pregnant	0.50	

6.4 Statistical Parity Difference – Multiclass

This was done again for each class of noprior and the results can be seen in the table below. The max threshold is the maximum threshold for which a fairness is achieved for the spd value for each class. If at least one of the classes has an SPD of zero, then SPD is satisfied for that variable, otherwise it is not.

Table 3: Statistical Parity Difference (SPD) Analysis

Variable	Class	SPD	Max Threshold	At least one SPD = 0
GENDER	0	0.27	0.26	Not Satisfied for All Classes
	3	0.07	0.06	
	2	0.29	0.26	
	1	0.05	0.01	
AGE	0	0.17	0.16	Not Satisfied for All Classes
	3	0.30	0.26	
	2	0.12	0.11	
	1	0.01	0.01	
VET	0	0.63	0.61	Satisfied for At Least One Class
	3	0.00	0.00	
	2	0.21	0.21	
	1	0.17	0.16	
EDUC	0	0.05	0.01	Satisfied for At Least One Class
	3	0.00	0.00	
	2	0.29	0.26	
	1	0.08	0.06	
MARSTAT	0	0.14	0.11	Not Satisfied for All Classes
	3	0.18	0.16	
	2	0.10	0.06	
	1	0.06	0.06	
EMPLOY	0	0.42	0.41	Satisfied for At Least One Class
	3	0.00	0.00	
	2	0.00	0.00	
	1	0.08	0.06	
RACE	0	0.02	0.01	Not Satisfied for All Classes
	3	0.35	0.31	
	2	0.32	0.31	
	1	0.05	0.01	
ETHNIC	0	0.14	0.11	Satisfied for At Least One Class
	3	0.00	0.00	
	2	0.32	0.31	
	1	0.20	0.16	
PREG	0	0.58	0.56	Satisfied for At Least One Class
	3	0.21	0.21	
	2	0.00	0.00	
	1	0.12	0.11	

6.5 Equal Opportunity

When the true positive rate (TPR) is the same for both the privileged and unprivileged groups, it is considered fair Irfan et al. (2023).

$$P\left(\hat{Y} = 1 \mid Y = 1, A = 0\right) = P\left(\hat{Y} = 1 \mid Y = 1, A = 1\right) \tag{5}$$

$$\mathrm{TPR}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} \tag{6}$$

$$\mathrm{EqOpp}_{\mathrm{diff}} = \max_i(\mathrm{TPR}_i) - \min_i(\mathrm{TPR}_i) \tag{7}$$

This fairness measure was ran at four different C values with the default values for gamma and r, where $gamma = scale = 1/n$ and $r = 0$. C was tried at .1, 1, 10, and 100. The optimal results are shown in the table below. Notice again that this is tried at different thresholds. The difference is found and then the maximum threshold value for which the $EqOpp_{diff} < threshhold$ is recorded in the table for each model.

Next we do the same thing but for each class and take the max difference and the minimum difference to calculate the fairness in each class where c is in each class. The results can be seen in the table below. The

Table 4: Equal Opportunity - COMPLETED

Variable	Optimal C Value	Model	Max TPR	Min TPR	EqOpp TPR Diff	Fairness
GENDER	0.1	Linear	1	0	1	UNFAIR
	10	Poly	1	0	1	UNFAIR
	10	RBF	1	0	1	UNFAIR
	1	Sigmoid	1	0	1	UNFAIR
AGE	0.1	Linear	1	0	1	UNFAIR
	10	Poly	1	0	1	UNFAIR
	10	RBF	1	0	1	UNFAIR
	1	Sigmoid	1	0.23	0.77	UNFAIR
VET	0.1	Linear	0.99	0	0.99	UNFAIR
	10	Poly	0.99	0	0.99	UNFAIR
	10	RBF	0.99	0	0.99	UNFAIR
	1	Sigmoid	1	0	1	UNFAIR
EDUC	0.1	Linear	0.99	0	0.99	UNFAIR
	10	Poly	1	0	1	UNFAIR
	10	RBF	1	0	1	UNFAIR
	1	Sigmoid	1	0	1	UNFAIR
MARSTAT	0.1	Linear	1	0.7	0.3	UNFAIR
	10	Poly	1	0.7	0.3	UNFAIR
	10	RBF	1	0.7	0.3	UNFAIR
	1	Sigmoid	1	0.36	0.64	UNFAIR
EMPLOY	0.1	Linear	0.99	0	0.99	UNFAIR
	10	Poly	1	0	1	UNFAIR
	10	RBF	1	0	1	UNFAIR
	1	Sigmoid	1	0	1	UNFAIR
RACE	0.1	Linear	1	0	1	UNFAIR
	10	Poly	1	0	1	UNFAIR
	10	RBF	1	0	1	UNFAIR
	1	Sigmoid	1	0.47	0.53	FAIR
ETHNIC	0.1	Linear	1	0	1	UNFAIR
	10	Poly	1	0	1	UNFAIR
	10	RBF	1	0	1	UNFAIR
	1	Sigmoid	1	0	1	UNFAIR
PREG	0.1	Linear	1	0	1	UNFAIR
	10	Poly	1	0	1	UNFAIR
	10	RBF	1	0	1	UNFAIR
	1	Sigmoid	1	0	1	UNFAIR

TPR difference is shown in the table and this is the maximum value that the threshold can be taken at. Again this is shown for the optimal C value with the default values of gamma, r, and degree.

$$TPR_{sc} = \frac{TP_{sc}}{TP_{sc} + FN_{sc}} \tag{8}$$

$$TPR_{\mathrm{diff}} = \max(TPR_{sc}) - \min(TPR_{sc}) \tag{9}$$

$$TPR_{sc} = \frac{TP_{sc}}{TP_{sc} + FN_{sc}} \tag{10}$$

$$FPR_{sc} = \frac{FP_{sc}}{FP_{sc} + TN_{sc}} \tag{11}$$

$$\mathrm{Max\_TPR}_{\mathrm{diff}} = \max_c\left(\max_s(TPR_{sc})\right) \tag{12}$$

$$\mathrm{Min\_TPR}_{\mathrm{diff}} = \min_c\left(\min_s(TPR_{sc})\right) \tag{13}$$

$$\mathrm{EqOpp}_{\mathrm{diff}} = |\mathrm{Max}_{\mathrm{TPR}_{\mathrm{diff}}} - \mathrm{Min}_{\mathrm{TPR}_{\mathrm{diff}}}| \tag{14}$$

Table 5: Equalized Opportunity - NOPRIOR

Variable	**Optimal C Value**	**Model**	**TPR Difference**	**FPR Difference**	**Equalized Odds Diff**	**Fair/Unfair**
Variable	C=Optimal	Model	Max TPR	Min TPR	Equalized Opp Diff	Fair/Unfair
GENDER	10	Linear	0.69	0.37	0.32	UNFAIR
	1	Poly	0.70	0.56	0.14	FAIR
	10	RBF	0.69	0.52	0.17	FAIR
	1	Sigmoid	0.65	0.40	0.25	UNFAIR
AGE	10	Linear	0.69	0.43	0.26	UNFAIR
	1	Poly	0.73	0.58	0.15	FAIR
	10	RBF	0.71	0.54	0.17	FAIR
	1	Sigmoid	0.67	0.44	0.23	UNFAIR
	10	Linear	0.50	0.00	0.50	UNFAIR
	1	Poly	0.63	0.00	0.63	UNFAIR
	10	RBF	0.50	0.00	0.50	UNFAIR
	1	Sigmoid	0.51	0.00	0.51	UNFAIR
EDUC	10	Linear	0.55	0.23	0.32	UNFAIR
	1	Poly	0.66	0.41	0.25	UNFAIR
	10	RBF	0.63	0.38	0.25	UNFAIR
	1	Sigmoid	0.54	0.32	0.22	UNFAIR
MARSTAT	10	Linear	0.50	0.49	0.01	FAIR
	1	Poly	0.62	0.62	0.00	FAIR
	10	RBF	0.60	0.56	0.04	FAIR
	1	Sigmoid	0.54	0.45	0.09	FAIR
EMPLOY	10	Linear	0.51	0.12	0.39	UNFAIR
	1	Poly	0.62	0.50	0.12	FAIR
	10	RBF	0.60	0.38	0.22	UNFAIR
	1	Sigmoid	0.51	0.25	0.26	UNFAIR
RACE	10	Linear	0.50	0.48	0.02	FAIR
	1	Poly	0.60	0.62	0.02	FAIR
	10	RBF	0.60	0.54	0.06	FAIR
	1	Sigmoid	0.54	0.49	0.05	FAIR
ETHNIC	10	Linear	0.50	0.00	0.50	UNFAIR
	1	Poly	0.63	0.00	0.63	UNFAIR
	10	RBF	59.00	0.00	59.00	UNFAIR
	1	Sigmoid	0.50	0.50	0.00	FAIR
PREG	10	Linear	0.50	0.00	0.50	UNFAIR
	1	Poly	0.62	0.00	0.62	UNFAIR
	10	RBF	0.59	0.00	0.59	UNFAIR
	1	Sigmoid	0.51	0.00	0.51	UNFAIR

6.6 Equalized Odds

Let $A = 1$ and $A = 0$ be the privileged and unprivileged groups respectively. When $Y = 1$, the equations shows the TPR and when $Y = 0$ it shows the FPR. This shows the ROC where TPR and FPR are equal for the privileged and unprivileged demographics.
Irfan et al. (2023).

$$P\left[\hat{Y} = 1 \middle| Y = 1, A = 0\right] = P[\hat{Y} = 1 | Y = 1, A = 1] \tag{15}$$

The table below shows the TPR and FPR difference as well as the Equalized odds difference and whether it is fair or unfair. The differences are calculated by subtracting the two values from each class from each other for the TPF and FPR respectively for each variable. Fairness is determined based on a .2 threshold. Any threshold could be chosen. Since equality determines fairness, a threshold closer to zero is wanted.

Table 6: Equal Odds - COMPLETED

Variable	Optimal C Value	Model	TPR Difference	FPR Difference	Equalized Odds Diff	Fair/Unfair
GENDER	0.1	Linear	0.32	0.03	0.29	UNFAIR
	10	Poly	0.14	0.09	0.05	FAIR
	10	RBF	0.17	0.06	0.11	FAIR
	1	Sigmoid	0.25	0.06	0.19	FAIR
AGE	0.1	Linear	0.26	0.03	0.23	UNFAIR
	10	Poly	0.15	0.05	0.10	FAIR
	10	RBF	0.16	0.08	0.08	FAIR
	1	Sigmoid	0.22	0.01	0.21	UNFAIR
VET	0.1	Linear	0.51	0.22	0.29	UNFAIR
	10	Poly	0.63	0.18	0.45	UNFAIR
	10	RBF	0.60	0.16	0.44	UNFAIR
	1	Sigmoid	0.51	0.22	0.29	UNFAIR
EDUC	0.1	Linear	0.32	0.29	0.03	FAIR
	10	Poly	0.25	0.48	0.23	UNFAIR
	10	RBF	0.25	0.36	0.11	FAIR
	1	Sigmoid	0.22	0.00	0.22	UNFAIR
MARSTAT	0.1	Linear	0.02	0.03	0.01	FAIR
	10	Poly	0.01	0.02	0.01	FAIR
	10	RBF	0.04	0.01	0.03	FAIR
	1	Sigmoid	0.09	0.02	0.07	FAIR
EMPLOY	0.1	Linear	0.39	0.03	0.36	UNFAIR
	10	Poly	0.12	0.04	0.08	FAIR
	10	RBF	0.22	0.04	0.18	FAIR
	1	Sigmoid	0.26	0.01	0.25	UNFAIR
RACE	0.1	Linear	0.02	0.00	0.02	FAIR
	10	Poly	0.00	0.02	0.02	FAIR
	10	RBF	0.06	0.03	0.03	FAIR
	1	Sigmoid	0.05	0.04	0.01	FAIR
ETHNIC	0.1	Linear	0.50	0.17	0.33	UNFAIR
	10	Poly	0.63	0.13	0.50	UNFAIR
	10	RBF	0.59	0.16	0.43	UNFAIR
	1	Sigmoid	0.00	0.25	0.25	UNFAIR
PREG	0.1	Linear	0.50	0.25	0.25	UNFAIR
	10	Poly	0.62	0.19	0.43	UNFAIR
	10	RBF	0.59	0.21	0.38	UNFAIR
	1	Sigmoid	0.51	0.25	0.26	UNFAIR

6.7 Equalized Odds - Multiclass

We do the same thing for the multiclass except now we take the maximum and minumum of each class to get the TPR difference and FPR difference for each class where $s \in$ *sense variable*, $c \in$ *class label*.

$$TPR_{sc} = \frac{TP_{sc}}{TP_{sc} + FN_{sc}} \tag{16}$$

$$FPR_{sc} = \frac{FP_{sc}}{FP_{sc} + TN_{sc}} \tag{17}$$

$$TPR_{diff} = \max(TPR_{sc}) - \min(TPR_{sc}) \tag{18}$$

$$FPR_{diff} = \max(FPR_{sc}) - \min(FPR_{sc}) \tag{19}$$

$$\begin{aligned} EqOdds_{diff} = &|\max(TPR_{sc}) - \min(TPR_{sc})| \\ &- |\max(FPR_{sc}) - \min(FPR_{sc})| \end{aligned} \tag{20}$$

The table with the TPR difference, FPR difference, Equalized Odds difference and whether it is fair or unfair for each variable can be seen below. Again this was done at a .2 threshold and the default values for gamma, degree, and r with optimal C values for the multiclass noprior model. Note that closer to zero for an $EqOdds_{diff}$ is more fair.

Table 7: Equalized Odds NOPRIOR

ETHNIC							
C-Value	**Model**	**Max TPR**	**Min TPR**	**Max FPR**	**Min FPR**	**EOD Diff**	**Fairness**
GENDER							
10	Linear	1.00	0.00	0.26	0.00	0.74	UNFAIR
1	Poly	1.00	0.00	0.14	0.00	0.86	UNFAIR
10	RBF	1.00	0.00	0.12	0.00	0.88	UNFAIR
1	Sigmoid	1.00	0.00	0.29	0.00	0.71	UNFAIR
AGE							
10	Linear	0.99	0.00	0.20	0.00	0.79	UNFAIR
1	Poly	1.00	0.00	0.07	0.00	0.93	UNFAIR
10	RBF	0.99	0.00	0.07	0.00	0.92	UNFAIR
1	Sigmoid	1.00	0.27	0.28	0.00	0.45	UNFAIR
VET							
10	Linear	0.99	0.00	0.17	0.00	0.82	UNFAIR
1	Poly	0.99	0.00	0.17	0.00	0.82	UNFAIR
10	RBF	0.99	0.00	0.17	0.00	0.82	UNFAIR
1	Sigmoid	1.00	0.00	0.22	0.00	0.78	UNFAIR
EDUC							
10	Linear	0.99	0.00	0.09	0.00	0.90	UNFAIR
1	Poly	1.00	0.00	0.06	0.00	0.94	UNFAIR
10	RBF	1.00	0.00	0.07	0.00	0.93	UNFAIR
1	Sigmoid	1.00	0.00	0.31	0.00	0.69	UNFAIR
MARSTAT							
10	Linear	1.00	0.69	0.11	0.00	0.20	UNFAIR
1	Poly	1.00	0.70	0.11	0.00	0.19	FAIR
10	RBF	1.00	0.71	0.12	0.00	0.17	FAIR
1	Sigmoid	1.00	0.29	0.34	0.00	0.37	UNFAIR
EMPLOY							
10	Linear	0.99	0.00	0.25	0.00	0.74	UNFAIR
1	Poly	1.00	0.00	0.06	0.00	0.94	UNFAIR
10	RBF	1.00	0.00	0.06	0.00	0.94	UNFAIR
1	Sigmoid	1.00	0.00	0.89	0.00	0.11	FAIR
ETHNIC							
10	Linear	1.00	0.00	0.09	0.00	0.91	UNFAIR
1	Poly	1.00	0.00	0.06	0.00	0.94	UNFAIR
10	RBF	1.00	0.00	0.06	0.00	0.94	UNFAIR
1	Sigmoid	1.00	0.34	0.28	0.00	0.38	UNFAIR
ETHNIC							
10	Linear	1.00	0.00	0.08	0.00	0.92	UNFAIR
1	Poly	1.00	0.00	0.10	0.00	0.90	UNFAIR
10	RBF	1.00	0.00	0.07	0.00	0.93	UNFAIR
1	Sigmoid	1.00	0.00	1.00	0.00	0.00	FAIR
PREG							
10	Linear	1.00	0.00	1.00	0.00	0.00	FAIR
1	Poly	1.00	0.00	1.00	0.00	0.00	FAIR
10	RBF	1.00	0.00	1.00	0.00	0.00	FAIR
1	Sigmoid	1.00	0.00	1.00	0.00	0.00	FAIR

7 The Model

7.1 Support Vector Machines

The data sets are split 70/30 training and test sets with shuffle set at true which splits the data randomly. Support Vector Machines were run with four kernels. Linear, Polynomial, Radial Basis Function, and Sigmoid. The formulas for these kernels can be seen below.

Linear Kernel :

$$K(X, Y) = X^T Y \tag{21}$$

Polynomial Kernel :

$$K(X, Y) = (X^T Y + C)^d \tag{22}$$

Radial Basis Function Kernel :

$$K(X, Y) = e^{-\gamma \|X-Y\|^2}, \quad \gamma > 0 \tag{23}$$

Sigmoid Kernel :

$$K(X, Y) = \tanh(\gamma X^T Y + C) \tag{24}$$

$$\text{where } \gamma = \text{auto} \quad = \frac{1}{n} \tag{25}$$

$$\text{where } \gamma = \text{scale} \quad = \frac{1}{n \times \text{Var}(X)} \tag{26}$$

$$\text{where Var}(x) \quad = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{27}$$

After further research, it was determined that with algorithms such as SVM, feature importance is not that useful. It is more useful to select an appropriate C value instead. The C value is a parameter that sets how much you want to penalize your model for each misclassified point. This means that the larger the C value, the smaller the margin, the smaller the C value, the larger the margin. The goal is to maximize the margin while minimizing the classification error. Next, we looked at different C values. In the models above the C value was set to four different C-values, .1, 1, 10, and 100 and the optimal values were chosen for each model. Note in the figure below Felipe (2019) the difference between large and small C values.

Large values of C	Small Values of C
Large effect of noisy points.	Low effect of noisy points.
A plane with very few misclassifications will be given precedence.	Planes that separate the points well will be found, even if there are some misclassifications

Fig. 13: The Influence of C Parameter ?

The results of using the different C-Values with the default values of *gamma*, *r*, and *degree* can be seen on the COMPLETED and NOPRIOR datasets for all four kernels below. The best performing model was the polynomial model with a C value of 10 having a prediction accuracy of 63.70% for the COMPLETED data set. The confusion matrix can be seen in the figure below. The radial basis function kernel model was a close second with an accuracy of 61.73% with a $C = 10$. $C = 10$ performed better in all models except in the sigmoid and linear kernel models.

COMPLETED		
Model	**C-value**	**Accuracy**
Linear	0.1	59.13
Poly	10	63.7
RBF	10	61.78
Sigmoid	1	58.41

Fig. 14: Accuracy COMPLETED

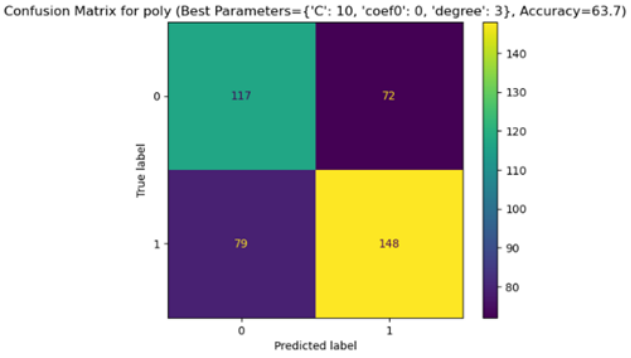


Fig. 15: Optimal Confusion COMPLETED

For the multiclass model, the accuracy was much higher with RBF coming in first with an accuracy of 91.83% with a C-value of 10 and polynomial coming in second with an accuracy of 90.89% and a C-value of 1. The linear model was also high at 88.09% with a C-value of 10. The confusion matrix can be seen below.

NOPRIOR		
Model	**C-value**	**Accuracy**
Linear	10	88.08
Poly	1	90.89
RBF	10	91.83
Sigmoid	1	73.15

Fig. 16: Accuracy NOPRIOR

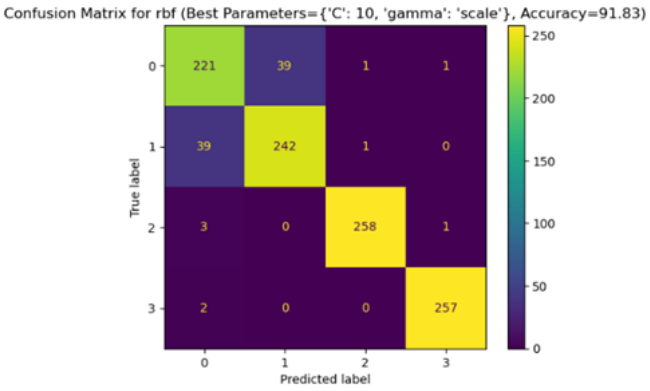


Fig. 17: Optimal Confusion NOPRIOR

7.2 Decision Trees

A grid search approach was used for decision trees with the following parameters:

- **Max Depth**: [None, 2, 4, 6, 8, 10]
- **Min Samples Split**: [2, 5, 10]
- **Min Samples Leaf**: [1, 2, 4]

The optimal results for each data set can be seen below:

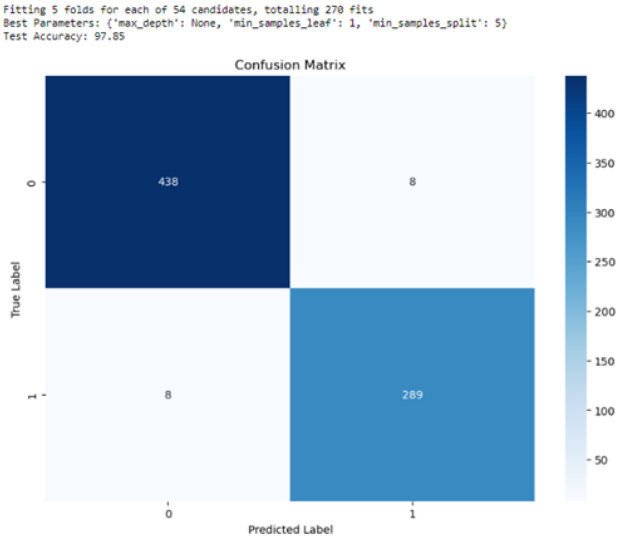


Fig. 18: Optimal Confusion COMPLETED DT

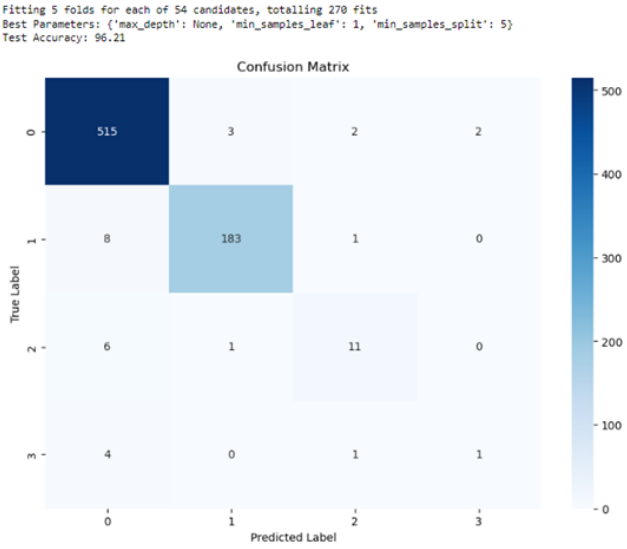


Fig. 19: Optimal Confusion NOPRIOR DT

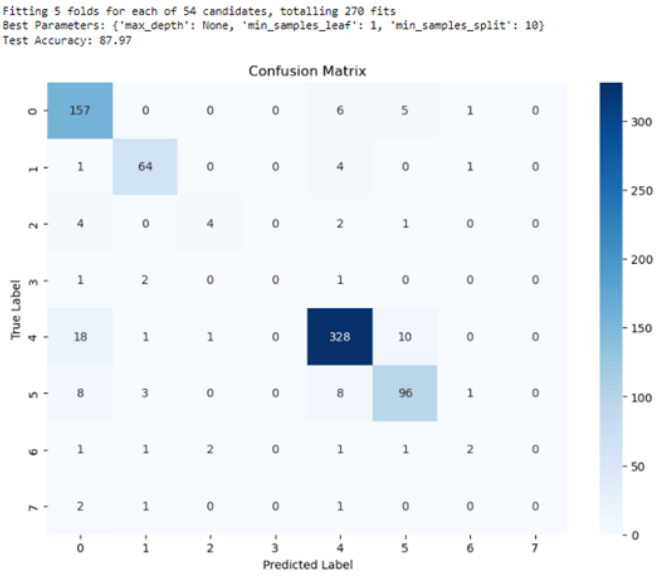


Fig. 20: Optimal Confusion COMPLETED_NOPRIOR DT

7.3 Random Forests

A parameter grid approach was used for random forests with the following parameters:

- **Number of Estimators**: [10, 50, 100, 200]
- **Max Features**: ['auto', 'sqrt', 'log2']
- **Max Depth**: [None, 5, 10, 20]
- **Min Samples Split**: [2, 5, 10]
- **Min Samples Leaf**: [1, 2, 4]

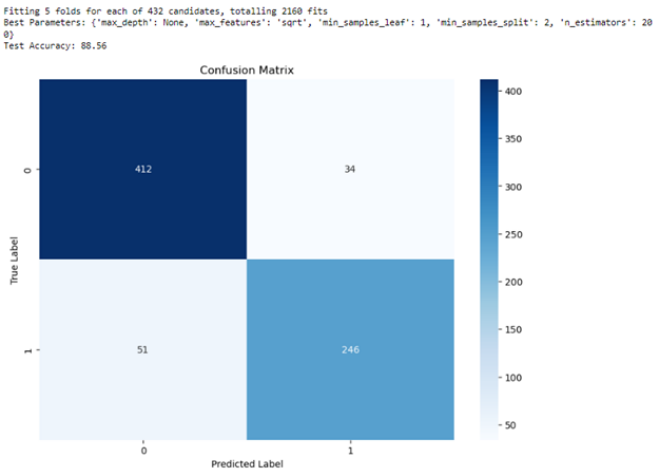


Fig. 21: Optimal Confusion COMPLETED RF

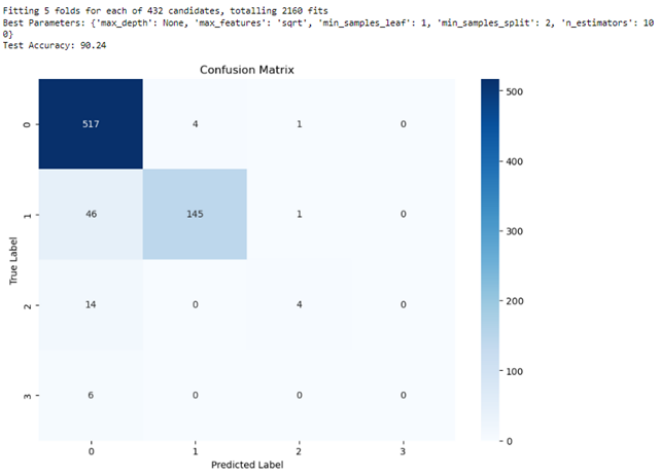


Fig. 22: Optimal Confusion NOPRIOR RF

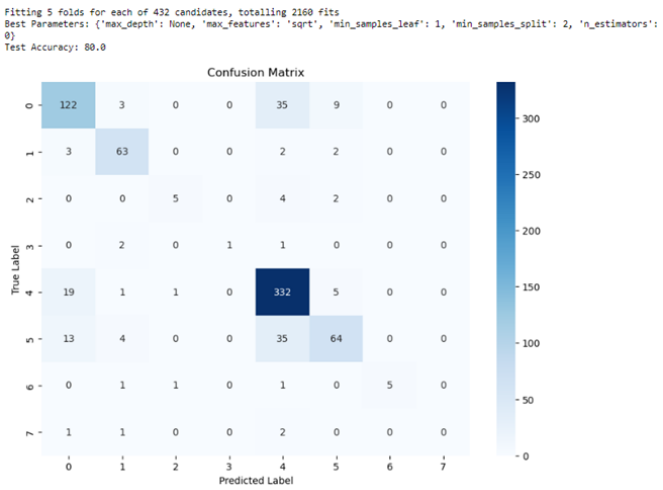


Fig. 23: Optimal Confusion COMPLETED_NOPRIOR RF

7.4 Neural Networks

The parameter grid for the neural network model is as follows:

- **Units in Layer 1**: [8, 10, 20, 30]
- **Units in Layer 2**: [8, 10, 20, 30]
- **Activation Function in Input Layers**: ['relu', 'tanh', 'sigmoid']
- **Optimizer**: ['adam', 'sgd']
- **Loss Function**: ['categorical_crossentropy', 'mean_squared_error']

The resulting optimal parameters along with their confusion matrices for each data set can be seen below:

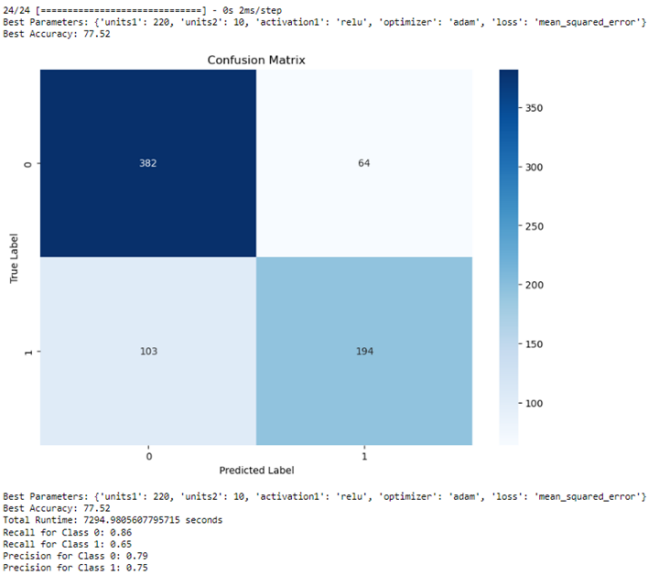


Fig. 24: Optimal Confusion COMPLETED NN

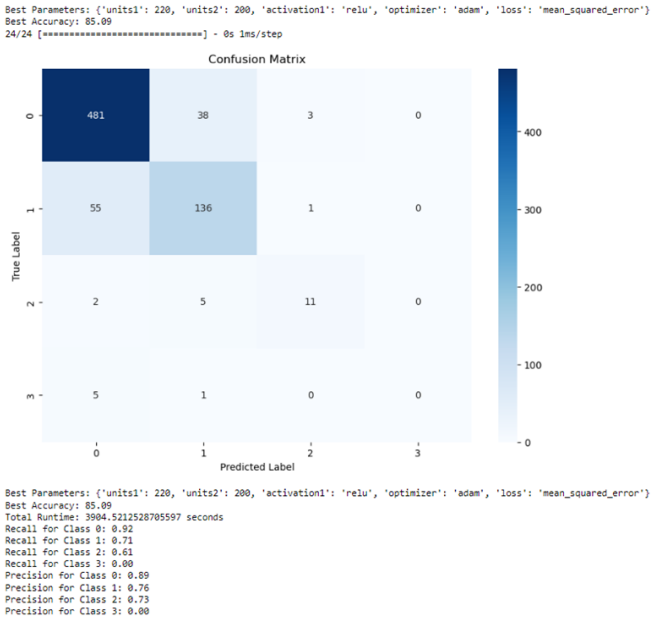


Fig. 25: Optimal Confusion NOPRIOR NN

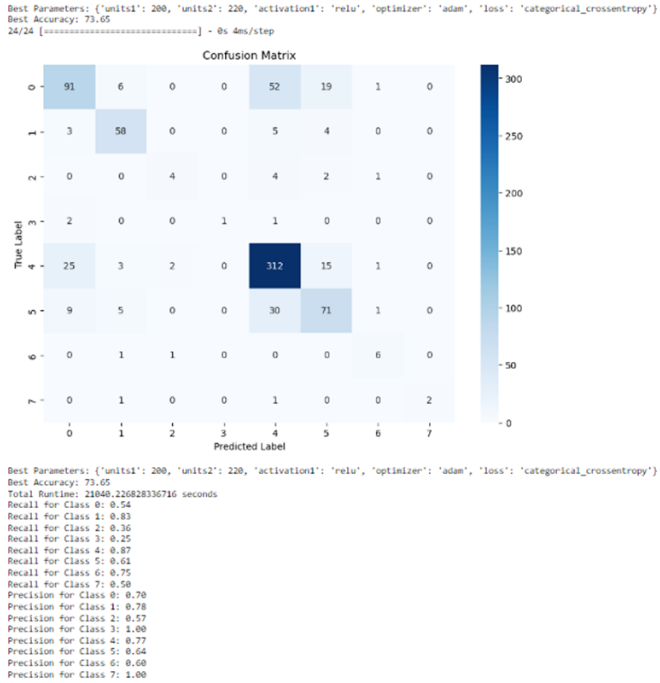


Fig. 26: Optimal Confusion COMPLETED_NOPRIOR NN

8 Interpretation of Results

The NOPRIOR variable translation can be seen below. This is the translation to what is shown in the x-axis of the confusion matrix above.

Value	Label
0	No prior treatment episodes
1	One prior treatment episode
2	Two prior treatment episodes
3	Three prior treatment episodes
4	Four prior treatment episodes
5	Five or more prior treatment episodes

Fig. 27: NOPRIOR Translation

The encoded variables used in the data sets and their translated meanings can be seen in the following tables.

variable	meaning
SERVICES_4	Rehab/residential, short term (30 days or fewer)
SERVICES_5	Rehab/residential, long term (more than 30 days)
SERVICES_7	Ambulatory, non-intensive outpatient
PSOURCE_2	Alcohol/drug use care provider
PSOURCE_3	Other health care provider
PSOURCE_4	School (educational)
PSOURCE_5	Employer/EAP
PSOURCE_6	Other community referral
PSOURCE_7	Court/criminal justice referral/DUI/DWI
METHUSE_2	No medication-assisted opioid therapy
LIVARAG_2	Dependent living
LIVARAG_3	Independent living
ALCFLG_1	alcohol reported at admissions
COKEFLG_1	cocaine/crack reported at admissions
MARFLG_1	marijuana/hashish reported at admissions
HERFLG_1	heroin reported at admissions
METHFLG_1	non-rx methadone reported at admissions
OPSYNFLG_1	Other opiates/synthetics reported at admission
PCPFLG_1	PCP reported at admission
HALLFLG_1	Hallucinogens reported at admission
MTHAMFLG_1	Methamphetamine/speed reported at admission
AMPHFLG_1	Other amphetamines reported at admission
STIMFLG_1	Other stimulants reported at admission
BENZFLG_1	Benzodiazepines reported at admission
TRNQFLG_1	Other tranquilizers reported at admission
BARBFLG_1	Barbiturates reported at admission
SEDHPFLG_1	Other sedatives/hypnotics reported at admission
INHFLG_1	Inhalants reported at admission
OTCFLG_1	Over-the-counter medication reported at admission
OTHERFLG_1	Other drug reported at admission
ALCDRUG_1	alcohol only
ALCDRUG_2	other drugs only
ALCDRUG_3	alcohol and other drugs
DETNLF_3	retired, disables not in labor force
DETNLF_4	resident of institution not in labor force
DETNLF_5	other not in labor force
DETCRIM_2	Formal adjudication process
DETCRIM_3	probation/parole
DETCRIM_6	prison
NOPRIOR_1	one prior treatment
NOPRIOR_2	two prior treatment
NOPRIOR_3	3 prior treatment
PSOURCE_2	Alcohol/drug use care provider
PSOURCE_3	Other health care provider
PSOURCE_6	Other community referral
PSOURCE_7	Court/criminal justice referral/DUI/DWI
ARRESTS_1	once
ARRESTS_2	two or more times
SUB3_10	methamphetamine tertiary
SUB3_11	other amphetamines tertiary
SUB3_12	other stimulants tertiary
SUB3_13	benzodiazepines tertiary
SUB3_16	other sedatives or hypnotics tertiary
SUB3_19	other drugs tertiary
SUB3_2	alcohol tertiary
SUB3_3	cocaine/crack tertiary
SUB3_4	marijuana/hashish tertiary
SUB3_5	heroin tertiary
SUB3_6	non prescription methadone tertiary
SUB3_7	other opiates and synthetics tertiary
SUB3_8	PCP tertiary
SUB3_9	hallucinogens tertiary

Table 8: Variable Descriptions

variable	meaning
METHUSE_2	no medication-assisted opioid therapy
PSYPROB_2	no coocurring mental and substance abuse disorders
DSMCRIT_10	cannabis use
DSMCRIT_11	other substance use
DSMCRIT_12	opioid abuse
DSMCRIT_13	cocaine abuse
DSMCRIT_2	substance-induced disorder
DSMCRIT_3	alcohol intoxication
DSMCRIT_4	alcohol dependency
DSMCRIT_5	opioid dependency
DSMCRIT_6	cocaine dependency
DSMCRIT_7	cannabis dependency
DSMCRIT_8	other substance dependency
DSMCRIT_9	alcohol abuse
SUB1_10	Methamphetamine/speed primary
SUB1_11	other amphetamines primary
SUB1_12	other stimulants primary
SUB1_13	benzodiazepines primary
SUB1_19	other drugs primary
SUB1_2	alcohol primary
SUB1_3	cocaine/crack primary
SUB1_4	marijuana/hashish primary
SUB1_5	heroin primary

Table 9: Variable Descriptions Cont'd

9 Reweighting

9.1 Interactions

The Chi Squared test was used to test significant interactions between bias variables.

$$\chi^2 = \frac{(E-O)^2}{E}, \tag{28}$$

where E is the expected value and O is referring to the observed frequencies in the contingency table respectively. We looked at both dual and three-way interactions and then combined the datasets and calculated the probability that at least one of these interactions occurred as our new reweighted weight. If there is a significant interaction, then the interaction counts for each x1, x2 for the dual interaction or x1, x2, and x3 for the three-way interaction are then divided by the total value counts for x1, x2, with the target or x1, x2, and x3 with the target. This gives the probability. These would be multiplicative models Veenstra (2011).

9.1.1 Dual Interaction

$$S_1 = \{x_1, x_2, \text{target}\}$$
$$S_2 = \{x_1, x_2\}$$
$$\text{DualInteraction}_{\text{Probability}} = \frac{|S_1|}{|S_2|} \tag{29}$$

Example:

$$\frac{P(\text{Complete} \cap \text{White} \cap \text{Male})}{P(\text{White} \cap \text{Male})} \tag{30}$$

9.1.2 Three Way Interaction

$$S_3 = \{x_1, x_2, x_3, \text{target}\} \tag{31}$$
$$S_4 = \{x_1, x_2, x_3\} \tag{32}$$
$$\text{3wayInteraction}_{\text{Probability}} = \frac{|S_3|}{|S_4|} \tag{33}$$

Example:

$$\frac{P(\text{Complete} \cap \text{White} \cap \text{Female} \cap \text{Pregnant})}{P(\text{White} \cap \text{Female} \cap \text{Pregnant})} \tag{34}$$

9.2 Intersectionalization

By taking the case for which at least one interaction occurs, we are reweighting the fairness. Then we combined the two-way and three-way interaction datasets together to get our new df dataset. This allows for the case when a person can be in two groups at once. An example would be veteran and male or veteran and male under 40.

$$\text{FinalProbability} = 1 - \left(1 - \prod \text{probability}_{\text{interactions}}\right) \tag{35}$$

$$\text{probability}_{\text{interactions}} = (\text{DualInteraction}_{\text{Probability}}) \cup (\text{3wayInteraction}_{\text{Probability}}) \tag{36}$$

10 New Fairness Calculations

10.1 Worst Case Demographic Parity

Demographic parity compares the pass rate of (positive outcome) of two groups. It is satisfied if:

$$P\left(\hat{Y} \middle| A \in \text{sg}_i\right) = P\left(\hat{Y} \middle| A \in \text{sg}_i\right) \forall i, j \in \mathbb{N},\ i \neq j \tag{37}$$

Where N is the total number of subgroups. When using the worst-case ratio formula, it is also known as demographic parity ratio, Ghosh et al. (2021):

$$\text{DPR} = \frac{\min\left\{P\left(\hat{Y} \middle| A \in \text{sg}_i\right) \forall i \in \mathbb{N}\right\}}{\max\left\{P\left(\hat{Y} \middle| A \in \text{sg}_i\right) \forall i \in \mathbb{N}\right\}} \tag{38}$$

```
Minimum Demographic Parity Ratio: 0.9524058315188466
```

	COMPLETED	Number_of_Terms	min	max	Demographic_Parity_Ratio
0	0	2	0.999484	0.999981	0.999503
1	0	3	0.951284	0.998822	0.952406
2	1	2	0.999836	1.000000	0.999837
3	1	3	0.974790	0.999861	0.974925

Fig. 28: Merged - COMPLETED

```
Minimum Demographic Parity Ratio: 0.7218107423676939
```

	NOPRIOR	Number_of_Terms	min	max	Demographic_Parity_Ratio
0	0	2.0	0.981420	1.000000	0.981420
1	0	3.0	0.903449	0.999998	0.903451
2	1	2.0	0.985489	0.994492	0.990947
3	1	3.0	0.920917	0.985644	0.934330
4	2	2.0	0.964407	0.999969	0.964437
5	2	3.0	0.721809	0.999998	0.721811
6	3	2.0	0.841075	0.999551	0.841453
7	3	3.0	0.831436	0.999546	0.831813

Fig. 29: Merged - NOPRIOR

```
Minimum Demographic Parity Ratio: 0.02458828773206945
```

	COMPLETED_NOPRIOR	Number_of_Terms	min	max	Demographic_Parity_Ratio
0	0	2.0	0.645950	0.962295	0.671259
1	0	3.0	0.487470	0.901409	0.540787
2	1	2.0	0.524531	0.895904	0.585477
3	1	3.0	0.581854	0.823778	0.706324
4	2	2.0	0.859057	0.967239	0.888154
5	2	3.0	0.571817	0.937434	0.609981
6	3	2.0	0.630716	0.998639	0.631576
7	3	3.0	0.819125	0.998596	0.820277
8	4	2.0	0.762554	0.941340	0.810073
9	4	3.0	0.417672	0.981201	0.425674
10	5	2.0	0.024203	0.984348	0.024588
11	5	3.0	0.636053	0.959108	0.663171
12	6	2.0	0.562646	0.899390	0.625586
13	6	3.0	0.355316	0.499445	0.711422
14	7	2.0	0.918633	0.981632	0.935823
15	7	3.0	0.838083	0.906130	0.924903

Fig. 30: Merged - COMPLETED_NOPRIOR

10.2 Worst Case Demographic Parity – Multiclass

$$\text{DPR}_{\text{multiclass}} = \min \left\{ \frac{\min \left\{ P\left(\hat{Y} \middle| A \in \text{sg}_i\right) \forall i \in N \right\}}{\max \left\{ P\left(\hat{Y} \middle| A \in \text{sg}_i\right) \forall i \in N \right\}} \right\} \tag{39}$$

The same is done for the multiclass DPR except the overall minimum is taken to get the minimum over all classes. [Gosh et al., 2022]Ghosh et al. (2021). The results can be seen in the figures below for COMPLETED, NORPIOR, and COMPLETED_NORPIOR merged on two-way and three-way interactions.

Minimum Demographic Parity Ratio: 0.02458828773206945

	COMPLETED_NOPRIOR	Number_of_Terms	min	max	Demographic_Parity_Ratio
0	0	2.0	0.645950	0.962295	0.671259
1	1	2.0	0.524531	0.895904	0.585477
2	2	2.0	0.859057	0.967239	0.888154
3	3	2.0	0.630716	0.998639	0.631576
4	4	2.0	0.762554	0.941340	0.810073
5	5	2.0	0.024203	0.984348	0.024588
6	6	2.0	0.562646	0.899390	0.625586
7	7	2.0	0.918633	0.981632	0.935823

Fig. 31: 2-Way Interaction NOPRIOR

Minimum Demographic Parity Ratio: 0.4256741722674935

	COMPLETED_NOPRIOR	Number_of_Terms	min	max	Demographic_Parity_Ratio
0	0	3.0	0.487470	0.901409	0.540787
1	1	3.0	0.581854	0.823778	0.706324
2	2	3.0	0.571817	0.937434	0.609981
3	3	3.0	0.819125	0.998596	0.820277
4	4	3.0	0.417672	0.981201	0.425674
5	5	3.0	0.636053	0.959108	0.663171
6	6	3.0	0.355316	0.499445	0.711422
7	7	3.0	0.838083	0.906130	0.924903

Fig. 32: 3-Way Interaction NOPRIOR

10.3 Worst Case Disparate Impact

$$\text{DI} = \min \left\{ \frac{P(\hat{Y}|A \in \text{sg}_i)}{P(\hat{Y}|A \in \text{sg}_i)}; \forall i, j \in N, i \neq j \right\}, \tag{40}$$

where in this case i and j are unique pairs in the series. The data is grouped by number of terms (2- or 3-way interactions), The probabilities are calculated between the pairs of ratios for each group of target class and pair of probabilities. The minimum ratios are all aggregated across all pairs and then minimum amongst these aggregated ratios is the worst-case scenario for disparate impact. The 80% rule has to be passed for fairness to be achieved. This is also known as the four fifths rule. The pass rate of group 1 to group 2 has to be greater than 80% with group 1 and 2 being interchangeable Ghosh et al. (2021). Again, the results are shown below.

Minimum Min_Ratio_Pairs: 0.9524058315188466

	COMPLETED	Number_of_Terms	Min_Ratio_Pairs
0	0	2	0.999503
1	0	3	0.952406
2	1	2	0.999837
3	1	3	0.974925

Fig. 33: Merged COMPLETED DI

Minimum Min_Ratio_Pairs: 0.7218107423676939

	NOPRIOR	Number_of_Terms	Min_Ratio_Pairs
0	0	2.0	0.981420
1	0	3.0	0.903451
2	1	2.0	0.990947
3	1	3.0	0.934330
4	2	2.0	0.964437
5	2	3.0	0.721811
6	3	2.0	0.841453
7	3	3.0	0.831813

Fig. 34: Merged NOPRIOR DI

```
Minimum Min_Ratio_Pairs: 0.02458828773206945
```

	COMPLETED_NOPRIOR	Number_of_Terms	Min_Ratio_Pairs
0	0	2.0	0.671259
1	0	3.0	0.540787
2	1	2.0	0.585477
3	1	3.0	0.706324
4	2	2.0	0.888154
5	2	3.0	0.609981
6	3	2.0	0.631576
7	3	3.0	0.820277
8	4	2.0	0.810073
9	4	3.0	0.425674
10	5	2.0	0.024588
11	5	3.0	0.663171
12	6	2.0	0.625586
13	6	3.0	0.711422
14	7	2.0	0.935823
15	7	3.0	0.924903

Fig. 35: Merged COMPLETED_NOPRIOR DI

Note that, COMPLETED is the fairest, NOPRIOR is second most fair, and COMPLETED_NOPRIOR is the least fair. It is fair individually, however in a worst-case scenario it has the lowest score.

10.4 Conditional Statistical Parity Ratio

This uses the worst-case min-max ratio, with predictor $\hat{Y}$, member A, and legitimate attribute L Corbett-Davies et al. (2017).

$$P\left(\hat{Y} \mid L=1, A \in \text{sg}_i\right) = P\left(\hat{Y} \mid L=1, A \in \text{sg}_j\right) \quad \forall i, j \in N, i \neq j \tag{41}$$

$$CSPR = \frac{\min\left\{P(\hat{Y} \mid L=1, A \in \text{sg}_i)\ \forall i \in N\right\}}{\max\left\{P(\hat{Y} \mid L=1, A \in \text{sg}_i)\ \forall i \in N\right\}} \tag{42}$$

$$CSPR_{\text{Multiclass}} = \min\left\{\frac{\min\left\{P(\hat{Y} \mid L=1, A \in \text{sg}_i)\ \forall i \in N\right\}}{\max\left\{P(\hat{Y} \mid L=1, A \in \text{sg}_i)\ \forall i \in N\right\}}\right\} \tag{43}$$

This uses the worst-case min-max ratio, with predictor $\hat{Y}$, member A, and legitimate attribute L Corbett-Davies et al. (2017).

```
Minimum CSPR value: 0.9759400182323508
```

	Number_of_Terms	CSPR
0	2	0.999856
1	3	0.975940

Fig. 36: CSPR COMPLETED

```
Minimum CSPR value: 0.7218110561977474
```

Fig. 37: CSPR NOPRIOR

```
Minimum CSPR value: 0.4968096639431072
```

Fig. 38: CSPR NOPRIOR_COMPLETED

10.5 Equal Opportunity Ratio - Multiclass

The TPR is compared for the protected and unprotected group. The minimum of these overall is taken for the multiclass case Ghosh et al. (2021).

$$P(\hat{Y}=1|A \in \text{sg}_i, Y=1) = P(\hat{Y}=1|A \in \text{sg}_j, Y=1) \quad \forall i, j \in N, i \neq j \tag{44}$$

$$\text{EOppR} = \frac{\min\left\{P(\hat{Y}=1|A \in \text{sg}_i, Y=1) \forall i \in N\right\}}{\max\left\{P(\hat{Y}=1|A \in \text{sg}_i, Y=1) \forall i \in N\right\}} \tag{45}$$

(46)

$$\text{EOppR}_{\text{Multi}} = \min\left\{\frac{\min\left\{P(\hat{Y}=y_k|A \in \text{sg}_i, Y=y_k), \forall i \in N, \forall k \in K\right\}}{\max\left\{P(\hat{Y}=y_k|A \in \text{sg}_i, Y=y_k), \forall i \in N, \forall k \in K\right\}}\right\} \tag{47}$$

```
Minimum Equal Opportunity Ratio: 0.95

{(2, 0): 1.0, (2, 1): 1.0, (3, 0): 0.95, (3, 1): 0.97}
```

Fig. 39: Merged COMPLETED Equal Opporunity

```
Minimum Equal Opportunity Ratio: 0.72

{(2.0, 0): 0.98,
 (2.0, 1): 0.99,
 (2.0, 2): 0.96,
 (2.0, 3): 0.84,
 (3.0, 0): 0.9,
 (3.0, 1): 0.93,
 (3.0, 2): 0.72,
 (3.0, 3): 0.83}
```

Fig. 40: Merged NOPRIOR Equal Opportunity

```
Minimum Equal Opportunity Ratio: 0.02

{(2.0, 0): 0.67,
 (2.0, 1): 0.59,
 (2.0, 2): 0.89,
 (2.0, 4): 0.81,
 (2.0, 3): 0.63,
 (2.0, 5): 0.02,
 (2.0, 6): 0.63,
 (2.0, 7): 0.94,
 (3.0, 0): 0.54,
 (3.0, 1): 0.71,
 (3.0, 2): 0.61,
 (3.0, 3): 0.82,
 (3.0, 4): 0.43,
 (3.0, 5): 0.66,
 (3.0, 7): 0.92,
 (3.0, 6): 0.71}
```

Fig. 41: Merged COMPLETED_NOPRIOR Equal Opportunity

10.6 Equal Odds Ratio

K is the set of all possible output classes. The closer the value is to 1, the lower the disparity is. For this case we calculate the worst-case odds for each output class y and then take the minimum of all those values Ghosh et al. (2021). This is the multiclass case for equalized odds ratio.

$$P\left(\hat{Y} = y_k \middle| A \in \text{sg}_i\right) = P\left(\hat{Y} = y_k \middle| A \in \text{sg}_j\right) \quad \forall i, j \in N, i \neq j, \forall k \in K \tag{48}$$

$$\text{EOddR}_{\text{Multiclass}} = \min\left\{\frac{\min\left\{P\left(\hat{Y} = y_k \middle| A \in \text{sg}_i\right)\right\}}{\max\left\{P\left(\hat{Y} = y_k \middle| A \in \text{sg}_i\right)\right\}}\right\} \forall i \in N, \forall k \in K \tag{49}$$

```
Equalized Odds Ratio for class '0' in term group '2': 1.0
Equalized Odds Ratio for class '1' in term group '2': 1.0
Equalized Odds Ratio for class '0' in term group '3': 0.98
Equalized Odds Ratio for class '1' in term group '3': 1.02
Minimum Equalized Odds Ratio: 0.98
```

Fig. 42: Merged COMPLETED Equalized Odds

```
Equalized Odds Ratio for class '0' in term group '2.0': 0.99
Equalized Odds Ratio for class '1' in term group '2.0': 1.01
Equalized Odds Ratio for class '2' in term group '2.0': 0.98
Equalized Odds Ratio for class '3' in term group '2.0': 0.86
Equalized Odds Ratio for class '0' in term group '3.0': 0.97
Equalized Odds Ratio for class '1' in term group '3.0': 1.03
Equalized Odds Ratio for class '2' in term group '3.0': 0.8
Equalized Odds Ratio for class '3' in term group '3.0': 0.92
Minimum Equalized Odds Ratio: 0.8
```

Fig. 43: Merged NOPRIOR Equalized Odds

```
Equalized Odds Ratio for class '0' in term group '2.0': 1.15
Equalized Odds Ratio for class '1' in term group '2.0': 0.87
Equalized Odds Ratio for class '2' in term group '2.0': 1.32
Equalized Odds Ratio for class '4' in term group '2.0': 1.21
Equalized Odds Ratio for class '3' in term group '2.0': 0.94
Equalized Odds Ratio for class '5' in term group '2.0': 0.04
Equalized Odds Ratio for class '6' in term group '2.0': 0.93
Equalized Odds Ratio for class '7' in term group '2.0': 1.39
Equalized Odds Ratio for class '0' in term group '3.0': 0.77
Equalized Odds Ratio for class '1' in term group '3.0': 1.31
Equalized Odds Ratio for class '2' in term group '3.0': 1.13
Equalized Odds Ratio for class '3' in term group '3.0': 1.52
Equalized Odds Ratio for class '4' in term group '3.0': 0.79
Equalized Odds Ratio for class '5' in term group '3.0': 1.23
Equalized Odds Ratio for class '7' in term group '3.0': 1.71
Equalized Odds Ratio for class '6' in term group '3.0': 1.32
Minimum Equalized Odds Ratio: 0.04
```

Fig. 44: Merged COMPLED_NOPRIOR Equalized Odds

COMPLETED							Reweighted Fairness				
Model	C-Value	Gamma	r	degree	Accuracy	Precision	DIR	DPR	CSPR	EqOppR	EqOddsR
Linear	0.1				59.13	58.85	1	0.95	0.98	0.95	0.98
Linear-wgt	1000				67.56	67.18					
Poly	10		0	3	63.7	63.83					
Poly-wgt	10		-0.1	5	86.14	86.08					
RBF	10	scale			61.78	61.83					
RBF-wgt	10	0.1			85.46	85.4					
Sigmoid	1	scale	0		58.41	58.18					
Sigmoid-wgt	1000	0.01	-1		79.68	79.63					

Fig. 45: Completed Results

NOPRIOR							Reweighted Fairness				
Model	C-Value	Gamma	r	degree	Accuracy	Precision	DIR	DPR	CSPR	EqOppR	EqOddsR
Linear	10				88.08	88.04	0.72	0.72	0.72	0.72	0.8
Linear-wgt	1				79.4	79.66					
Poly	1		0	3	90.89	90.91					
Poly-wgt	10		-0.1	5	90.38	90.34					
RBF	10	scale			91.83	91.86					
RBF-wgt	10	0.1			90.38	90.31					
Sigmoid	1	scale	0		73.15	72.49					
Sigmoid-wgt	1000	0.01	-1		85.64	85,54					

Fig. 46: NOPRIOR Results

COMPLETED-NOPRIOR							Reweighted Fairness				
Model	C-Value	Gamma	r	degree	Accuracy	Precision	DIR	DPR	CSPR	EqOppR	EqOddsR
Linear	10				88.08	88.04	0.02	0.02	0.5	0.02	0.04
Linear-wgt	10				80.81	80.8					
Poly	1		0	3	90.89	90.91					
Poly-wgt	1		1	4	82.84	82.6					
RBF	10	scale			91.83	91.86					
RBF-wgt	1000	scale			86.22	86.18					
Sigmoid	1	scale	0		73.15	72.49					
Sigmoid-wgt	1000	0.01	-1		84.19	84.4					

Fig. 47: Completed_NOPRIOR Results

Class	Precision	Recall	F1-Score	Support
Incomplete	0.98	0.99	0.98	446
Complete	0.98	0.97	0.97	297
Accuracy: 97.98				

Table 10: Classification Report Completed DT

Class	Precision	Recall	F1-Score	Support
0	0.97	0.99	0.98	522
1	0.97	0.95	0.96	192
2	0.69	0.61	0.65	18
3	0.50	0.17	0.25	6
Accuracy: 96.07				

Table 11: Classification Report NOPRIOR DT

Class	Precision	Recall	F1-Score	Support
COMPLETE_0	0.81	0.93	0.87	169
COMPLETE_1	0.91	0.91	0.91	70
COMPLETE_2	0.57	0.36	0.44	11
COMPLETE_3	0.00	0.00	0.00	4
INCOMPLETE_0	0.94	0.91	0.93	358
INCOMPLETE_1	0.82	0.84	0.83	116
INCOMPLETE_2	1.00	0.25	0.40	8
INCOMPLETE_3	0.00	0.00	0.00	4
Accuracy: 87.97				

Table 12: Classification Report Completed_NOPRIOR DT

Class	Precision	Recall	F1-Score	Support
Incomplete	0.89	0.92	0.91	446
Complete	0.88	0.83	0.85	297
Accuracy: 88.56				

Table 13: Classification Report Completed RF

Class	Precision	Recall	F1-Score	Support
0	0.89	0.99	0.94	522
1	0.97	0.76	0.85	192
2	0.67	0.22	0.33	18
3	0.00	0.00	0.00	6
Accuracy: 90.24				

Table 14: Classification Report NOPRIOR RF

Class	Precision	Recall	F1-Score	Support
COMPLETE_0	0.77	0.72	0.75	169
COMPLETE_1	0.84	0.90	0.87	70
COMPLETE_2	0.71	0.45	0.56	11
COMPLETE_3	1.00	0.25	0.40	4
INCOMPLETE_0	0.81	0.93	0.86	358
INCOMPLETE_1	0.78	0.55	0.65	116
INCOMPLETE_2	1.00	0.62	0.77	8
INCOMPLETE_3	0.00	0.00	0.00	4
Accuracy: 80.0				

Table 15: Classification Report Completed_NOPRIOR RF

Model	Completed	NOPRIOR	Completed_NOPRIOR
Number of Layers	2	2	2
Number of Neurons Layer 1	220	220	200
Number of Neurons Layer 2	10	200	220
Activation Function	ReLU (Rectified Linear Unit)	ReLU (Rectified Linear Unit)	ReLU (Rectified Linear Unit)
Network Optimizer	adam	adam	adam
Loss Function	mean squared error	mean squared error	categorical crossentropy
Epoch	20	20	20
Accuracy	77.52	85.09	73.65
Recall Incomplete	0.86		
Recall Complete	0.65		
Precision Incomplete	0.79		
Precision Complete	0.75		
Recall 0		0.92	
Recall 1		0.71	
Recall 2		0.61	
Recall 3		0.00	
Precision 0		0.89	
Precision 1		0.76	
Precision 2		0.73	
Precision 3		0.00	
Recall 0			0.54
Recall 1			0.83
Recall 2			0.36
Recall 3			0.25
Recall 4			0.87
Recall 5			0.61
Recall 6			0.75
Recall 7			0.50
Precision 0			0.70
Precision 1			0.78
Precision 2			0.57
Precision 3			1.00
Precision 4			0.77
Precision 5			0.64
Precision 6			0.60
Precision 7			1.00

Table 16: Model performance metrics

The Rectified Linear Unit activation function is 0 when x is negative and x when x is positive $f(x) = max(x, 0)$, Ramachandran et al. (2018). This was the optimal activation function in all three models.

11 Conclusions

Initially this research set out to predict whether a person completed treatment or not and it proved difficult due to the imbalance of the data with roughly 32% of people completing treatment. When this proved to have poor accuracy with SVMs, other algorithms such as decision trees and random forest approaches were utilized, A different approach was taken to look at predicting how many prior times a person had been in treatment as well as how the combination of completed and NOPRIOR. Fairness was explored with 9 variables and after reweighting the variables, a higher fairness was achieved in all the models with only a small decrease in accuracy. The highest model for the completed data set was the poly-wgt at 86.14%. Note that the reweighted COMPLETED model is fair when it comes to all fairness measures with a fairness of 95% for demographic parity, 98% for CSPR, equal opportunity of 95% and equal odds of 98%. For the NOPRIOR data set, RBF-wgt had the best accuracy with the best fairness. It had an accuracy of 91.83% and a fairness of 72% for all fairness except equal odds which was 80%. COMPLETED_NOPRIOR was fair looking at individual fairness but not at worse case fairness, due to some classes having poor fairness. The accuracy of this model was also 91.83% for RBF-wgt. Based on fairness and accuracy, the NOPRIOR model predicts the best and COMPLETED is the next best. An RBF or poly kernel would be the best to use when it comes to SVM, c value of 10, r=-0.1, gamma=scale. This provides the best results. In comparison, decision trees and random forests performed much better overall with decision trees having an accuracy of 97.98% for COMPLETED, 96.07% for NOPRIOR and 87.97% for COMPLETED_NOPRIOR. Random forests had an accuracy of 88.56% for COMPLETED, 90.24% for NOPRIOR, and 80% for COMPLETED_NOPRIOR. Neural Networks performed the worst with an accuracy of 77.52% for COMPLETED, 85.09% for NORPIOR, and 73.65% for
COMPLETED_NOPRIOR. Overall decision trees performed the best with the COMPLETED data set being the best predicting data set, NOPRIOR being next and COMPLETED_NOPRIOR coming in last in accuracy. For this reason, we would choose to use the decision trees model to predict all three data sets.

12 Future Work

For future work, it would be beneficial to look at linear discriminant analysis for dimensionality reduction. Also, MCA or Multiple Correspondence Analysis would be good to look at for reducing dimensionality of categorical data. Researching kernelized MCA would be interesting to see how that affects the accuracy of the model. Other methods of balancing the data that are more sophisticated than SMOTE using K-nearest neighbors or deep learning methods would be interesting to investigate. Looking deeper into the data to see what other information could be extracted. Perhaps a clustering analysis of what drugs/substances occur in treatment together in a patient and how those ties to age and demographics.

References

(2021). *Diagnostic and Statistical Manual of Mental Disorders*. 5th edition. Accessed 21 Dec 2023.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, pages 321–357.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1).

Felipe (2019). The c parameter.

Gajawada, K. S. (2021). Chi-square test for feature selection in machine learning.

Ghosh, A. et al. (2021). Characterizing intersectional group fairness with worst-case comparisons. In *AIDBEI*.

Irfan, Z., McCaffery, F., and Loughran, R. (2023). *Evaluating Fairness Metrics*, volume 1840 of *Communications in Computer and Information Science*. Springer, Cham.

Jishan, S., Rashu, R., Haque, N., et al. (2015). Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decis. Anal.*, 2(1).

National Survey on Drug Use and Health (2022). Key substance use and mental health indicators in the united states: Results from the 2022 national survey on drug use and health. Substance Abuse and Mental Health Services Administration. Accessed 21 Dec 2023.

Ramachandran, P., Zoph, B., and Le, Q. (2018). Searching for activation functions.

Substance Abuse and Mental Health Services Administration (2021). Treatment episode data set (teds): 2019.

Veenstra, G. (2011). Race, gender, class, and sexual orientation: intersecting axes of inequality and self-rated health in canada. *Int J Equity Health*, 10(3).

Wildra, E. and Herring, T. (2021). States of incarceration: The global context 2021. Prison Policy Initiative. Accessed April 21, 2022.