LONG COVID PREDICTION ALGORITHM

By Devin Gee and Juan Pablo Selame Fernandez

Background Long COVID Prediction

Long COVID

- COVID symptoms lasting more than 4-12 weeks
- No available diagnostic test or treatment options
- The risk factors for long COVID are poorly understood

Behavioral Risk Factor Surveillance System

- Annual telephone survey conducted by the CDC for noninstitutionalized adults
- 2022 Data is the first year with COVID-19 data
- Previously used in machine learning projects to predict diabetes

Previous Work

- Machine learning models have been employed to predict COVID mortality and severity of symptoms
- Other studies focused on more clinical variables such as medications taken and types of providers seen to predict long COVID

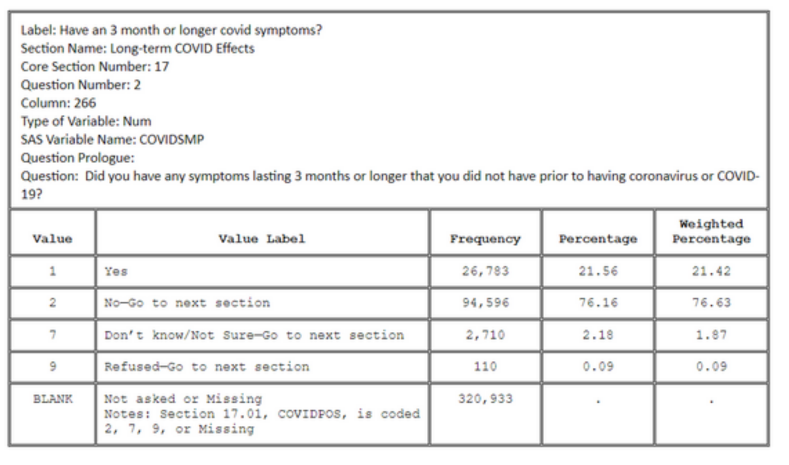
Novelty

- We are specifically using behavioral and demographic data to identify vulnerable populations, and lifestyle risk factors
- Similar studies have focused primarily on chronic diseases such as diabetes and heart disease

DATA DESCRIPTION

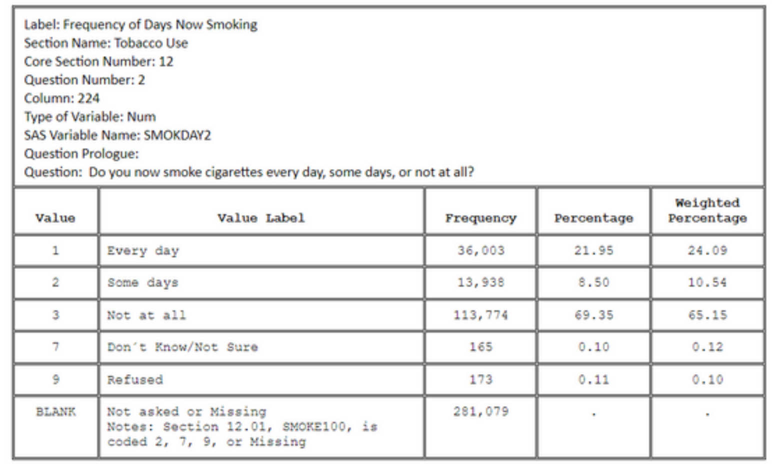
- Largely categorical and ordinal survey data
- 445,132 individuals surveyed total with 100s of questions asked. (Questions varied by participants)
- 110,877 individuals reported that they have contracted COVID
- 86,901 Individuals in the final dataset after cleaning missing values

- 29 features selected for initial dataset. Ranging from demographics (gender, income, race, education), lifestyle (exercise, sleep time), health (self reported mental and physical health), health history (diabetes, asthma, other chronic diseases), and more.



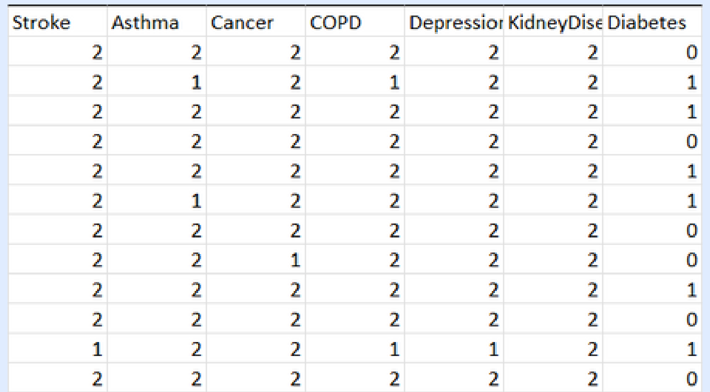
Label: Have an 3 month or longer covid symptoms?
Section Name: Long-term COVID Effects
Core Section Number: 17
Question Number: 2
Column: 266
Type of Variable: Num
SAS Variable Name: COVIDSMP
Question Prologue:
Question: Did you have any symptoms lasting 3 months or longer that you did not have prior to having coronavirus or COVID-19?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	26,783	21.56	21.42
2	No—Go to next section	94,596	76.16	76.63
7	Don’t know/Not Sure—Go to next section	2,710	2.18	1.87
9	Refused—Go to next section	110	0.09	0.09
BLANK	Not asked or Missing Notes: Section 17.01, COVIDPOS, is coded 2, 7, 9, or Missing	320,933	.	.



Label: Frequency of Days Now Smoking
Section Name: Tobacco Use
Core Section Number: 12
Question Number: 2
Column: 224
Type of Variable: Num
SAS Variable Name: SMOKDAY2
Question Prologue:
Question: Do you now smoke cigarettes every day, some days, or not at all?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Every day	36,003	21.95	24.09
2	Some days	13,938	8.50	10.54
3	Not at all	113,774	69.35	65.15
7	Don´t Know/Not Sure	165	0.10	0.12
9	Refused	173	0.11	0.10
BLANK	Not asked or Missing Notes: Section 12.01, SMOKE100, is coded 2, 7, 9, or Missing	281,079	.	.



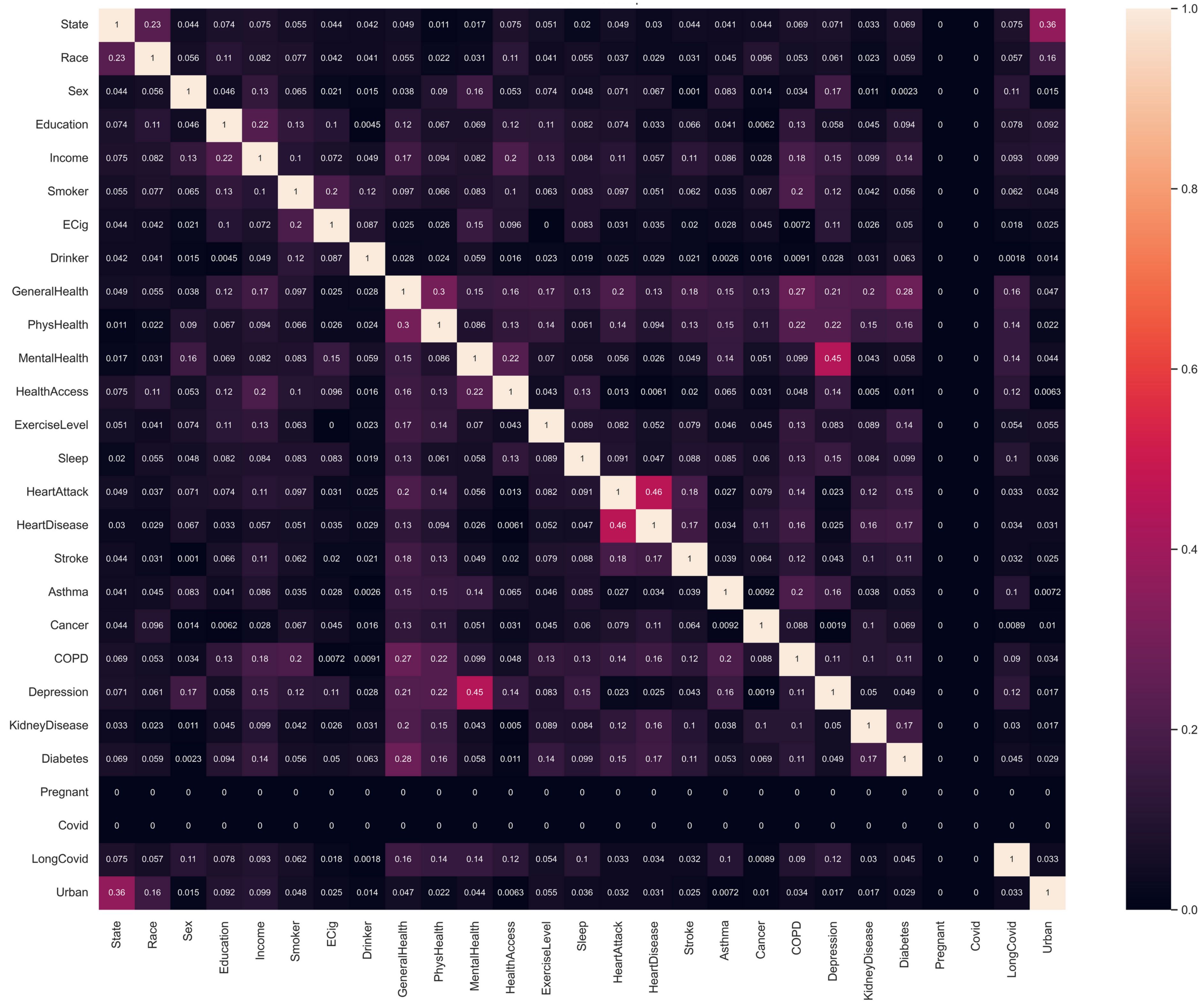
Stroke	Asthma	Cancer	COPD	Depressior	KidneyDise	Diabetes
2	2	2	2	2	2	0
2	1	2	1	2	2	1
2	2	2	2	2	2	1
2	2	2	2	2	2	0
2	2	2	2	2	2	1
2	1	2	2	2	2	1
2	2	2	2	2	2	0
2	2	1	2	2	2	0
2	2	2	2	2	2	1
2	2	2	2	2	2	0
1	2	2	1	1	2	1
2	2	2	2	2	2	0

19163 Positive for Long COVID (~22%)
67738 Negative for Long COVID
Imbalanced Dataset

FEATURE SELECTION

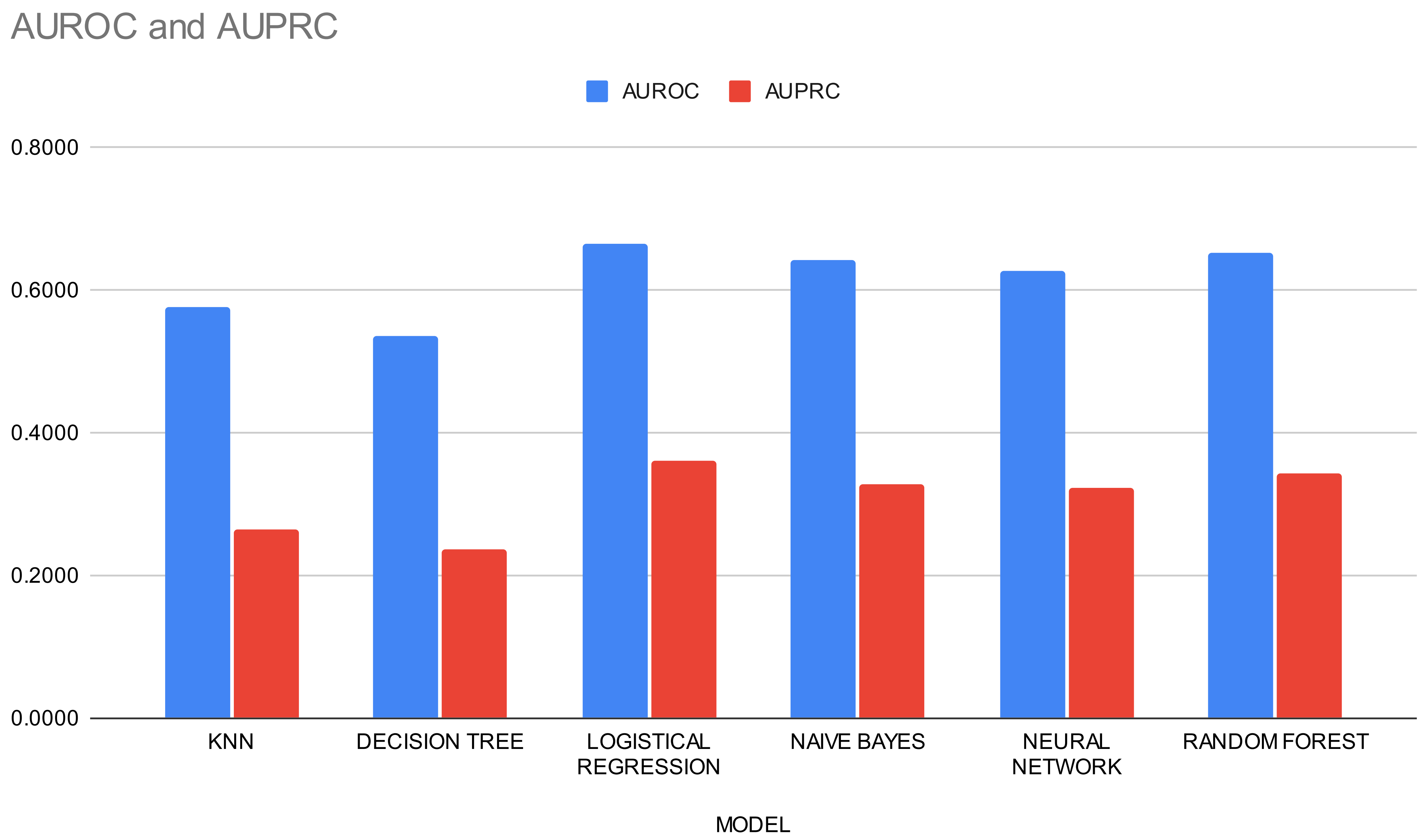
Cramer’s V Correlation Test

- Extremely weak correlations filtered from data set (<0.1)
- No features were very highly correlated
- Pregnancy and COVID features were removed due to only 1 value across the dataset



	State	Race	Sex	Education	Income	Smoker	ECig	Drinker	GeneralHealth	PhysHealth	MentalHealth	HealthAccess	ExerciseLevel	Sleep	HeartAttack	HeartDisease	Stroke	Asthma	Cancer	COPD	Depression	KidneyDisease	Diabetes	Pregnant	Covid	LongCovid	Urban
State	1	0.23	0.044	0.074	0.075	0.055	0.044	0.042	0.049	0.011	0.017	0.075	0.051	0.02	0.049	0.03	0.044	0.041	0.044	0.069	0.071	0.033	0.069	0	0	0.075	0.36
Race	0.23	1	0.056	0.11	0.082	0.077	0.042	0.041	0.055	0.022	0.031	0.11	0.041	0.055	0.037	0.029	0.031	0.045	0.096	0.053	0.061	0.023	0.059	0	0	0.057	0.16
Sex	0.044	0.056	1	0.046	0.13	0.065	0.021	0.015	0.038	0.09	0.16	0.053	0.074	0.048	0.071	0.067	0.001	0.083	0.014	0.034	0.17	0.011	0.0023	0	0	0.11	0.015
Education	0.074	0.11	0.046	1	0.22	0.13	0.1	0.0045	0.12	0.067	0.069	0.12	0.11	0.082	0.074	0.033	0.066	0.041	0.0062	0.13	0.058	0.045	0.094	0	0	0.078	0.092
Income	0.075	0.082	0.13	0.22	1	0.1	0.072	0.049	0.17	0.094	0.082	0.2	0.13	0.084	0.11	0.057	0.11	0.086	0.028	0.18	0.15	0.099	0.14	0	0	0.093	0.099
Smoker	0.055	0.077	0.065	0.13	0.1	1	0.2	0.12	0.097	0.066	0.083	0.1	0.063	0.083	0.097	0.051	0.062	0.035	0.067	0.2	0.12	0.042	0.056	0	0	0.062	0.048
ECig	0.044	0.042	0.021	0.1	0.072	0.2	1	0.087	0.025	0.026	0.15	0.096	0	0.083	0.031	0.035	0.02	0.028	0.045	0.0072	0.11	0.026	0.05	0	0	0.018	0.025
Drinker	0.042	0.041	0.015	0.0045	0.049	0.12	0.087	1	0.028	0.024	0.059	0.016	0.023	0.019	0.025	0.029	0.021	0.0026	0.016	0.0091	0.028	0.031	0.063	0	0	0.0018	0.014
GeneralHealth	0.049	0.055	0.038	0.12	0.17	0.097	0.025	0.028	1	0.3	0.15	0.16	0.17	0.13	0.2	0.13	0.18	0.15	0.13	0.27	0.21	0.2	0.28	0	0	0.16	0.047
PhysHealth	0.011	0.022	0.09	0.067	0.094	0.066	0.026	0.024	0.3	1	0.086	0.13	0.14	0.061	0.14	0.094	0.13	0.15	0.11	0.22	0.22	0.15	0.16	0	0	0.14	0.022
MentalHealth	0.017	0.031	0.16	0.069	0.082	0.083	0.15	0.059	0.15	0.086	1	0.22	0.07	0.058	0.056	0.026	0.049	0.14	0.051	0.099	0.45	0.043	0.058	0	0	0.14	0.044
HealthAccess	0.075	0.11	0.053	0.12	0.2	0.1	0.096	0.016	0.16	0.13	0.22	1	0.043	0.13	0.013	0.0061	0.02	0.065	0.031	0.048	0.14	0.005	0.011	0	0	0.12	0.0063
ExerciseLevel	0.051	0.041	0.074	0.11	0.13	0.063	0	0.023	0.17	0.14	0.07	0.043	1	0.089	0.082	0.052	0.079	0.046	0.045	0.13	0.083	0.089	0.14	0	0	0.054	0.055
Sleep	0.02	0.055	0.048	0.082	0.084	0.083	0.083	0.019	0.13	0.061	0.058	0.13	0.089	1	0.091	0.047	0.088	0.085	0.06	0.13	0.15	0.084	0.099	0	0	0.1	0.036
HeartAttack	0.049	0.037	0.071	0.074	0.11	0.097	0.031	0.025	0.2	0.14	0.056	0.013	0.082	0.091	1	0.46	0.18	0.027	0.079	0.14	0.023	0.12	0.15	0	0	0.033	0.032
HeartDisease	0.03	0.029	0.067	0.033	0.057	0.051	0.035	0.029	0.13	0.094	0.026	0.0061	0.052	0.047	0.46	1	0.17	0.034	0.11	0.16	0.025	0.16	0.17	0	0	0.034	0.031
Stroke	0.044	0.031	0.001	0.066	0.11	0.062	0.02	0.021	0.18	0.13	0.049	0.02	0.079	0.088	0.18	0.17	1	0.039	0.064	0.12	0.043	0.1	0.11	0	0	0.032	0.025
Asthma	0.041	0.045	0.083	0.041	0.086	0.035	0.028	0.0026	0.15	0.15	0.14	0.065	0.046	0.085	0.027	0.034	0.039	1	0.0092	0.2	0.16	0.038	0.053	0	0	0.1	0.0072
Cancer	0.044	0.096	0.014	0.0062	0.028	0.067	0.045	0.016	0.13	0.11	0.051	0.031	0.045	0.06	0.079	0.11	0.064	0.0092	1	0.088	0.0019	0.1	0.069	0	0	0.0089	0.01
COPD	0.069	0.053	0.034	0.13	0.18	0.2	0.0072	0.0091	0.27	0.22	0.099	0.048	0.13	0.13	0.14	0.16	0.12	0.2	0.088	1	0.11	0.1	0.11	0	0	0.09	0.034
Depression	0.071	0.061	0.17	0.058	0.15	0.12	0.11	0.028	0.21	0.22	0.45	0.14	0.083	0.15	0.023	0.025	0.043	0.16	0.0019	0.11	1	0.05	0.049	0	0	0.12	0.017
KidneyDisease	0.033	0.023	0.011	0.045	0.099	0.042	0.026	0.031	0.2	0.15	0.043	0.005	0.089	0.084	0.12	0.16	0.1	0.038	0.1	0.1	0.05	1	0.17	0	0	0.03	0.017
Diabetes	0.069	0.059	0.0023	0.094	0.14	0.056	0.05	0.063	0.28	0.16	0.058	0.011	0.14	0.099	0.15	0.17	0.11	0.053	0.069	0.11	0.049	0.17	1	0	0	0.045	0.029
Pregnant	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Covid	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LongCovid	0.075	0.057	0.11	0.078	0.093	0.062	0.018	0.0018	0.16	0.14	0.14	0.12	0.054	0.1	0.033	0.034	0.032	0.1	0.0089	0.09	0.12	0.03	0.045	0	0	1	0.033
Urban	0.36	0.16	0.015	0.092	0.099	0.048	0.025	0.014	0.047	0.022	0.044	0.0063	0.055	0.036	0.032	0.031	0.025	0.0072	0.01	0.034	0.017	0.017	0.029	0	0	0.033	1

MODEL CREATION



70/30 Test Split

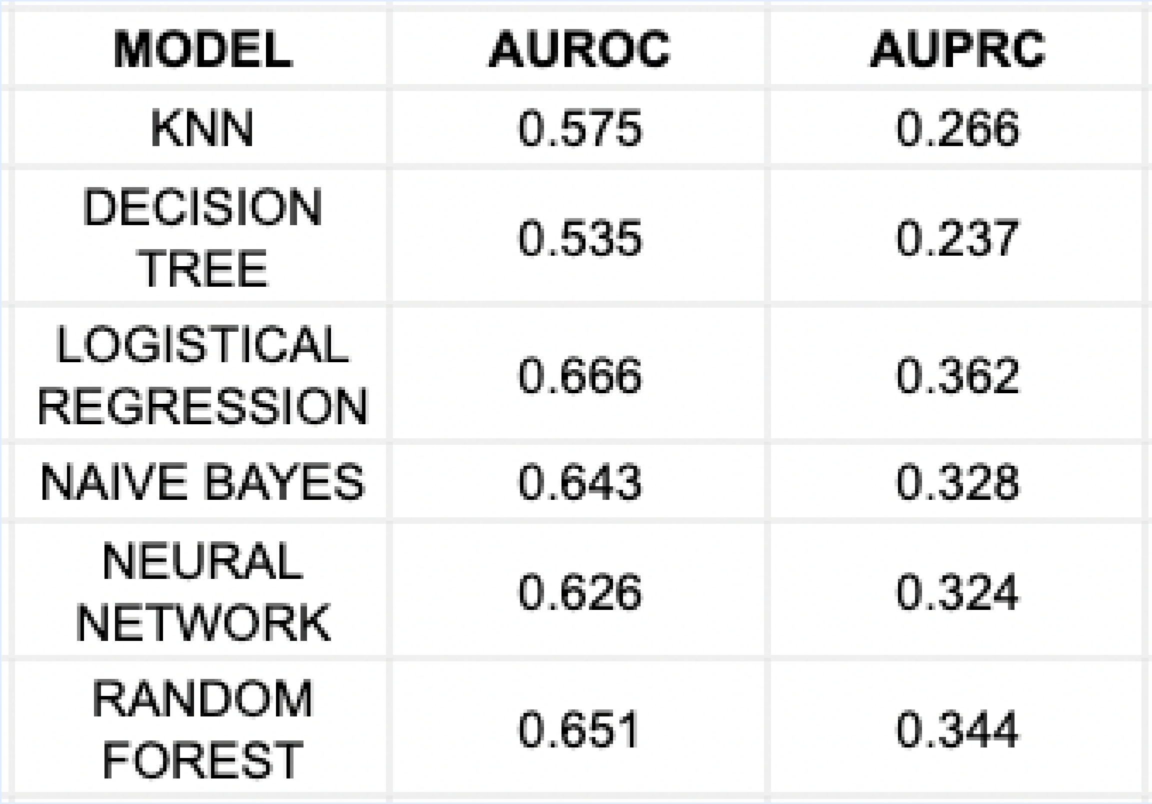
AUPRC as metric because of dataset is imbalanced

Previous work on predicting diabetes ~ .90 AUROC

HYPER PARAMETER TUNNING

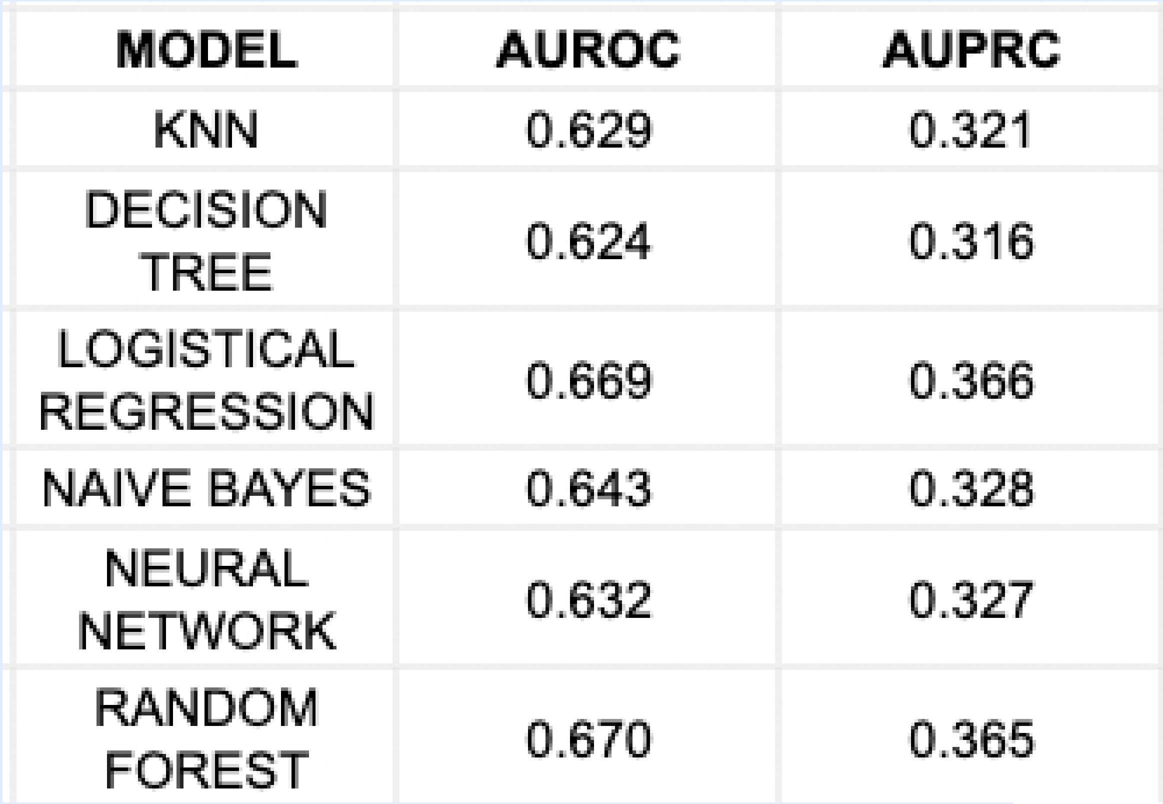
- KNN - k =30
- DECISION TREE - max_depth 10, min_samples_leaf = 1, min_samples_split = 10
- LOGISTICAL REGRESSION - C=0.001, max_iter = 100, penalty=l2, solver= saga
- NAIVE BAYES - alpha = 0.01, nb_type = gaussian
- NEURAL NETWORK - activation = relu, alpha = 0.01, batch_size = auto, learning_rate=adaptive, verbose = True, validation_fraction=0.1
- RANDOM FOREST - max_depth = 10, min_samples_leaf = 1, min_samples_split = 5, n_estimators = 200

BEFORE TUNING



MODEL	AUROC	AUPRC
KNN	0.575	0.266
DECISION TREE	0.535	0.237
LOGISTICAL REGRESSION	0.666	0.362
NAIVE BAYES	0.643	0.328
NEURAL NETWORK	0.626	0.324
RANDOM FOREST	0.651	0.344

AFTER TUNING



MODEL	AUROC	AUPRC
KNN	0.629	0.321
DECISION TREE	0.624	0.316
LOGISTICAL REGRESSION	0.669	0.366
NAIVE BAYES	0.643	0.328
NEURAL NETWORK	0.632	0.327
RANDOM FOREST	0.670	0.365

DISCUSSION & FUTURE DIRECTION

Difficulty increasing AUPRC is likely due to the nature of the problem. Long COVID is most likely not strongly associated with the included sociobehavioral, or demographic features. Other biological features such as viral strain and viral load are probably better indicators.

Future Directions:

- Add more features (COVID vaccination, former smoking status)
- Test new model (gradient boosting binary classifier)
- Shift question focus (COVID vs Long COVID)