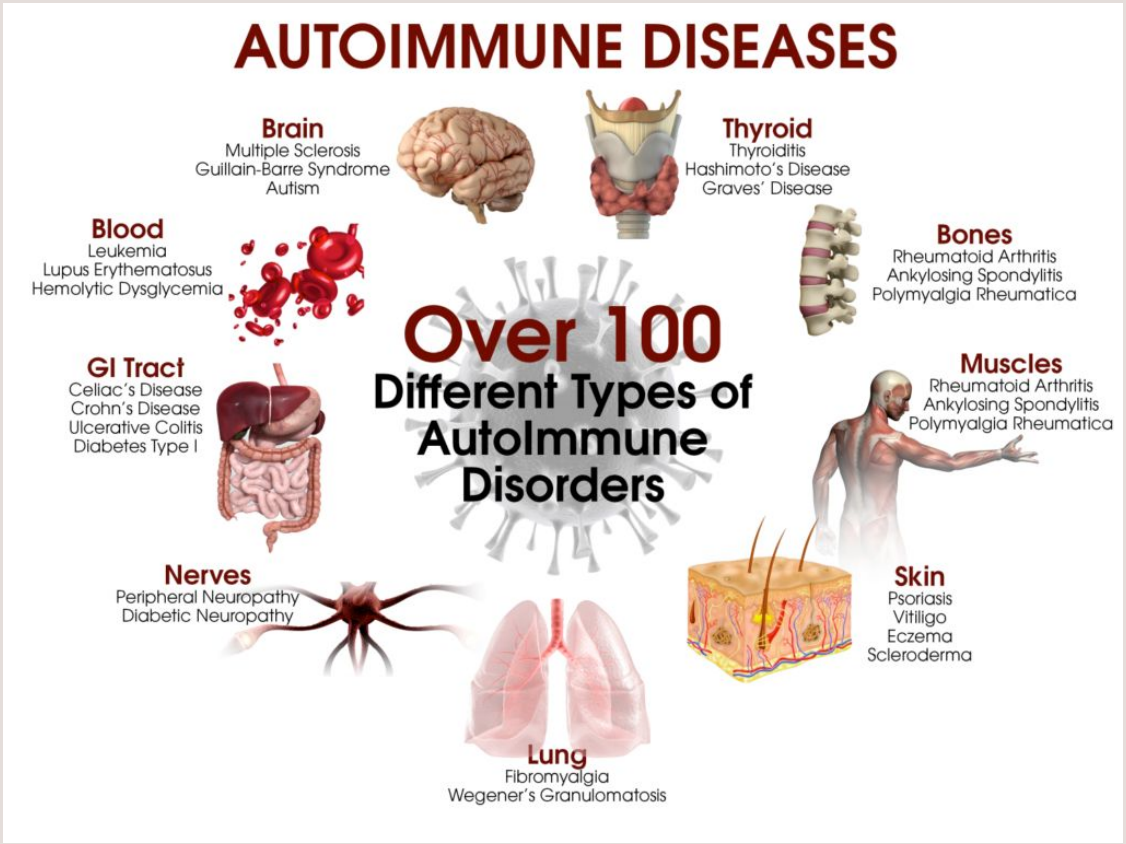
Risk Factors: Arthritis

What increases risk of an autoimmune disease?

By Chisum Lindauer

Autoimmune Diseases/Disorders

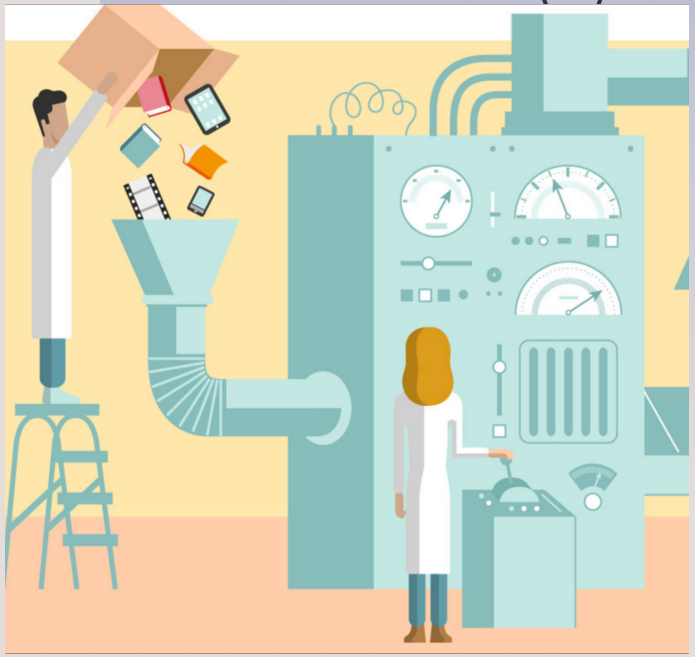


Data Set and Understanding

- National Health and Nutrition Examination Survey (NHANES), collected by the CDC
- 5,000 participants between 2017 and 2020
- Categorized and stored in XPT files, each accompanied by detailed data dictionaries that can exceed 100 pages
- Around 900 features selected

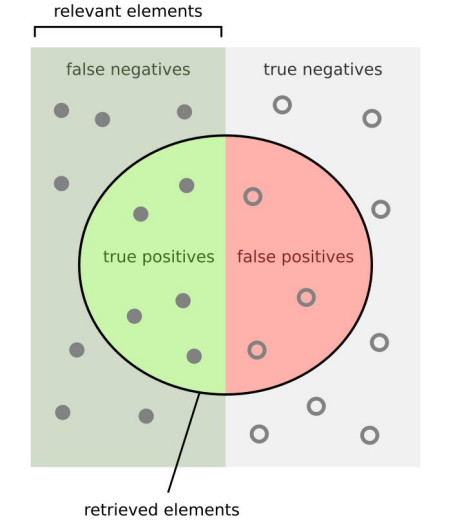
Data Preparation

- Encoded Categorical Data
- Missing Data Handled Based On Model
- Missing Sometimes Encoded
- Scikit-Learn
- Pipelines
- Custom Algorithms



Scoring with F1

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$



Random Forest Results

Classification Report:

	precision	recall	f1-score	support
-1	1.000000	1.000000	1.000000	1248.000000
0	0.810676	0.879537	0.843704	1295.000000
1	0.660131	0.532513	0.589494	569.000000
accuracy	0.864396	0.864396	0.864396	0.864396
macro avg	0.823602	0.804017	0.811066	3112.000000
weighted avg	0.859075	0.864396	0.859903	3112.000000

Feature Importances:

	feature	importance
2	Age in years at screening [P_DEMO]	0.033960
322	Ever told you had COPD, emphysema, ChB [P_MCQ]	0.030083
298	Moderate recreational activities [P_PAQ]	0.028237
324	Abdominal pain during past 12 months? [P_MCQ]	0.026185
327	Ever told you had cancer or malignancy [P_MCQ]	0.026113
..	...	...
223	UIBC, Serum Comment Code [P_FETIB]	0.000045
261	GGT Comment Code [P_BIOPRO]	0.000044
230	Blood lead comment code [P_PBCD]	0.000041
0	Interview/Examination status [P_DEMO]	0.000028
310	Questionnaire Mode Flag [P_SMQ]	0.000000

XGBoost and LightGBM

XGBoost

Classification Report:

	precision	recall	f1-score	support
0	0.82	0.87	0.84	1295
1	0.66	0.55	0.60	569
accuracy			0.78	1864
macro avg	0.74	0.71	0.72	1864
weighted avg	0.77	0.78	0.77	1864

LightGBM

Classification Report:

	precision	recall	f1-score	support
0	0.96	0.29	0.45	1295
1	0.38	0.97	0.54	569
accuracy			0.50	1864
macro avg	0.67	0.63	0.49	1864
weighted avg	0.78	0.50	0.48	1864

Deep Learning With Keras

- GPU enabled took many tries
- Ran GPU out of Memory
- Not a big model
- Low initial accuracy score of .4 terrible!
  Accuracy scores were always higher than F1 on other models so didn't spend more time on it.
- Needed a much bigger network and didn't have the horsepower for it.
- Shelved for this project.

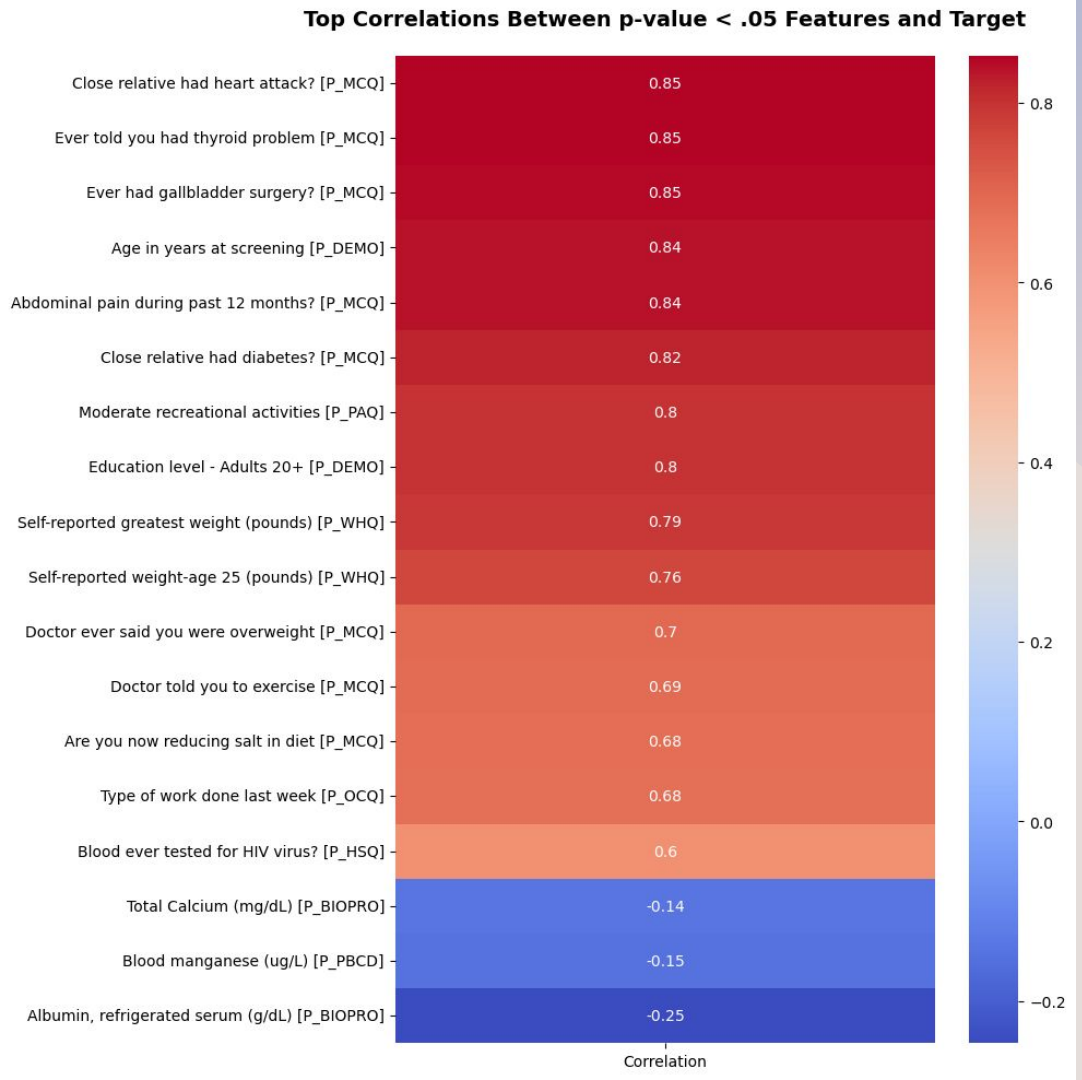
Logistic Regression Results

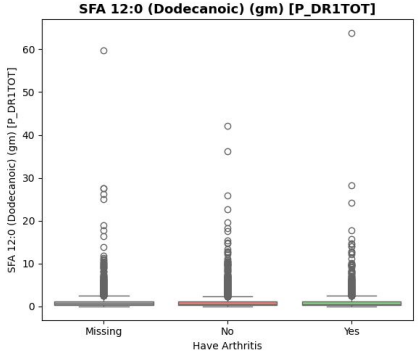
01 **Feature List with p-values < .05!**

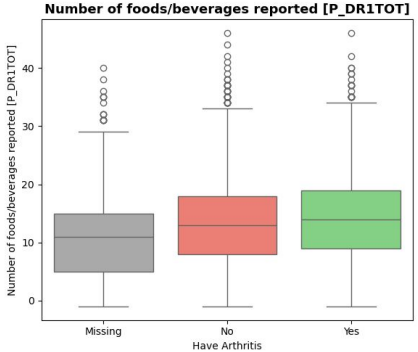
02

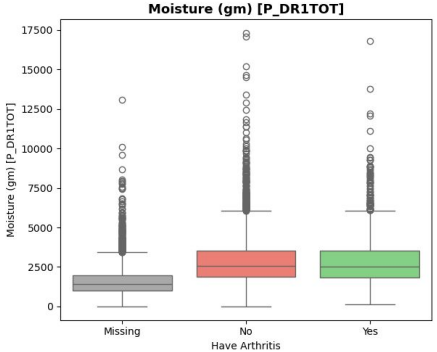
Logistic Regression Classification Report:

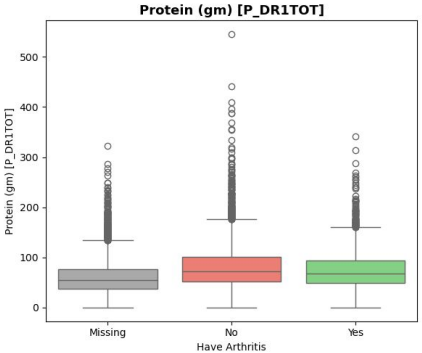
	precision	recall	f1-score	support
0	0.85	0.74	0.79	1295
1	0.55	0.71	0.62	569
accuracy			0.73	1864
macro avg	0.70	0.73	0.71	1864
weighted avg	0.76	0.73	0.74	1864

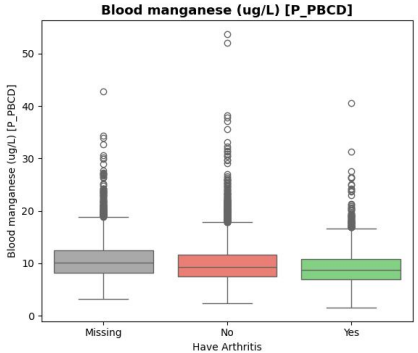

Top Correlations Between p-value < .05 Features and Target
Close relative had heart attack? [P_MCQ]
Ever told you had thyroid problem [P_MCQ]
Ever had gallbladder surgery? [P_MCQ]
Age in years at screening [P_DEMO]
Abdominal pain during past 12 months? [P_MCQ]
Close relative had diabetes? [P_MCQ]
Moderate recreational activities [P_PAQ]
Education level - Adults 20+ [P_DEMO]
Self-reported greatest weight (pounds) [P_WHQ]
Self-reported weight-age 25 (pounds) [P_WHQ]
Doctor ever said you were overweight [P_MCQ]
Doctor told you to exercise [P_MCQ]
Are you now reducing salt in diet [P_MCQ]
Type of work done last week [P_OCQ]
Blood ever tested for HIV virus? [P_HSQ]
Total Calcium (mg/dL) [P_BIOPRO]
Blood manganese (ug/L) [P_PBCD]
Albumin, refrigerated serum (g/dL) [P_BIOPRO]
0.85
0.85
0.85
0.84
0.84
0.82
0.8
0.8
0.79
0.76
0.7
0.69
0.68
0.68
0.6
-0.14
-0.15
-0.25
Correlation
0.8
0.6
0.4
0.2
0.0
−0.2

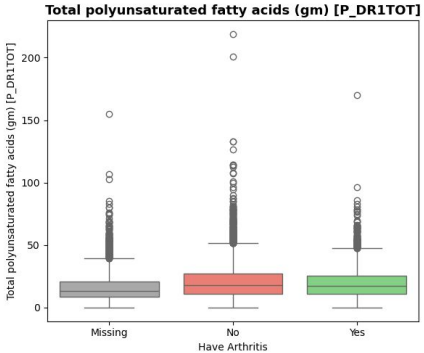

SFA 12:0 (Dodecanoic) (gm) [P_DR1TOT]
SFA 12:0 (Dodecanoic) (gm) [P_DR1TOT]
Missing
No
Yes
Have Arthritis

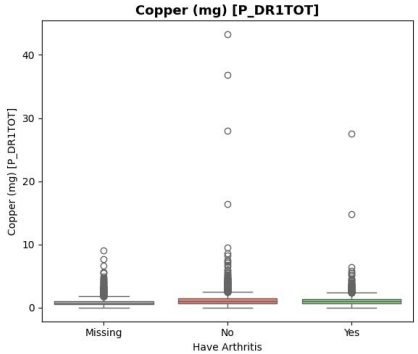

Number of foods/beverages reported [P_DR1TOT]
Number of foods/beverages reported [P_DR1TOT]
Missing
No
Yes
Have Arthritis

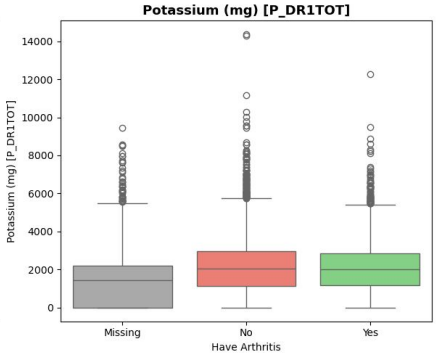

Moisture (gm) [P_DR1TOT]
Moisture (gm) [P_DR1TOT]
Missing
No
Yes
Have Arthritis

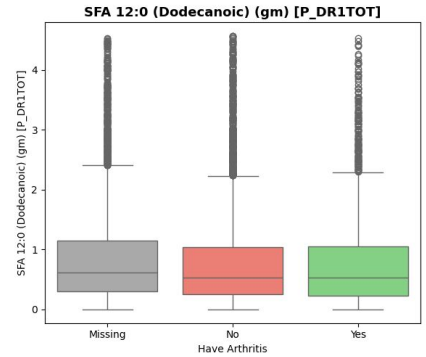

Protein (gm) [P_DR1TOT]
Protein (gm) [P_DR1TOT]
Missing
No
Yes
Have Arthritis

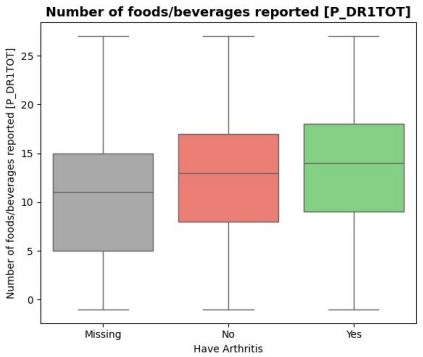

Blood manganese (ug/L) [P_PBCD]
Blood manganese (ug/L) [P_PBCD]
Missing
No
Yes
Have Arthritis

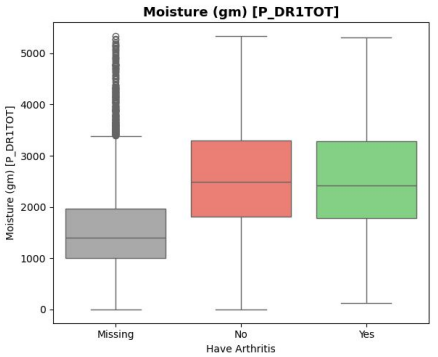

Total polyunsaturated fatty acids (gm) [P_DR1TOT]
Total polyunsaturated fatty acids (gm) [P_DR1TOT]
Missing
No
Yes
Have Arthritis

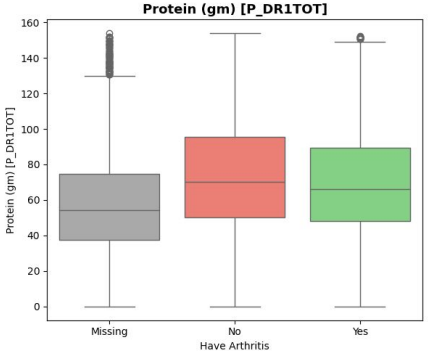

Copper (mg) [P_DR1TOT]
Copper (mg) [P_DR1TOT]
Missing
No
Yes
Have Arthritis

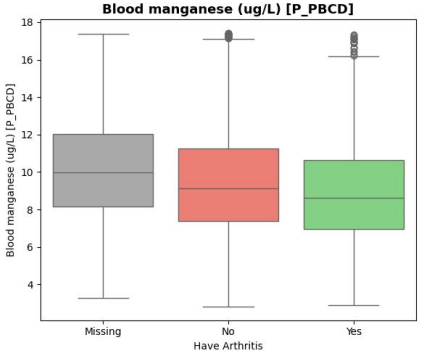

Potassium (mg) [P_DR1TOT]
Potassium (mg) [P_DR1TOT]
Missing
No
Yes
Have Arthritis

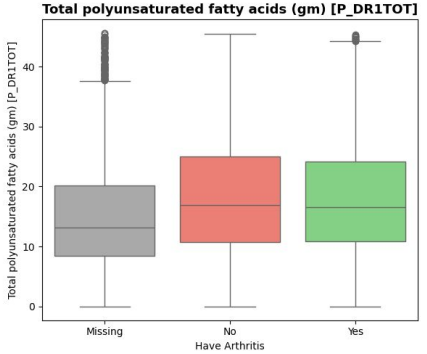

SFA 12:0 (Dodecanoic) (gm) [P_DR1TOT]
SFA 12:0 (Dodecanoic) (gm) [P_DR1TOT]
Missing
No
Yes
Have Arthritis

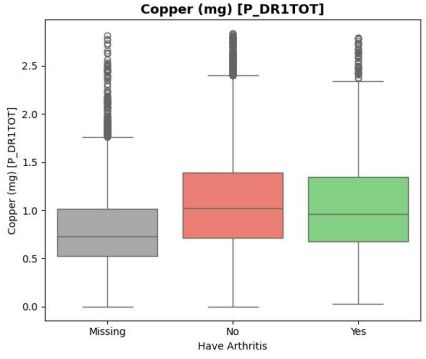

Number of foods/beverages reported [P_DR1TOT]
Number of foods/beverages reported [P_DR1TOT]
Missing
No
Yes
Have Arthritis

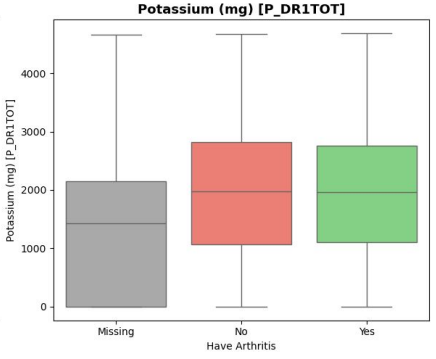

Moisture (gm) [P_DR1TOT]
Moisture (gm) [P_DR1TOT]
Missing
No
Yes
Have Arthritis

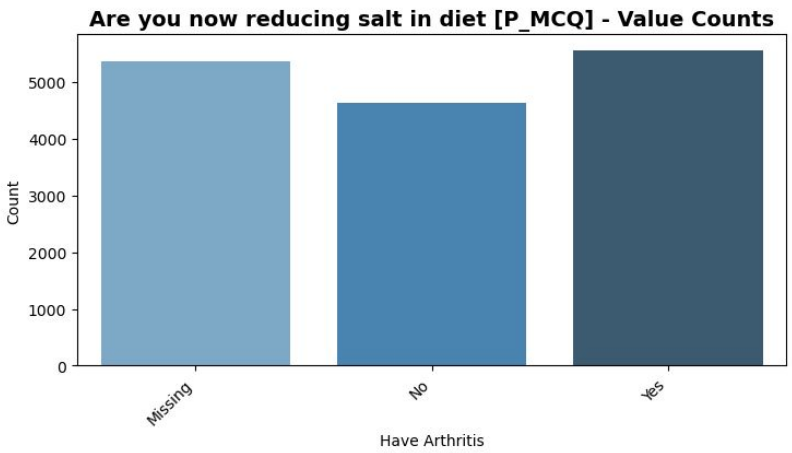

Protein (gm) [P_DR1TOT]
Protein (gm) [P_DR1TOT]
Missing
No
Yes
Have Arthritis

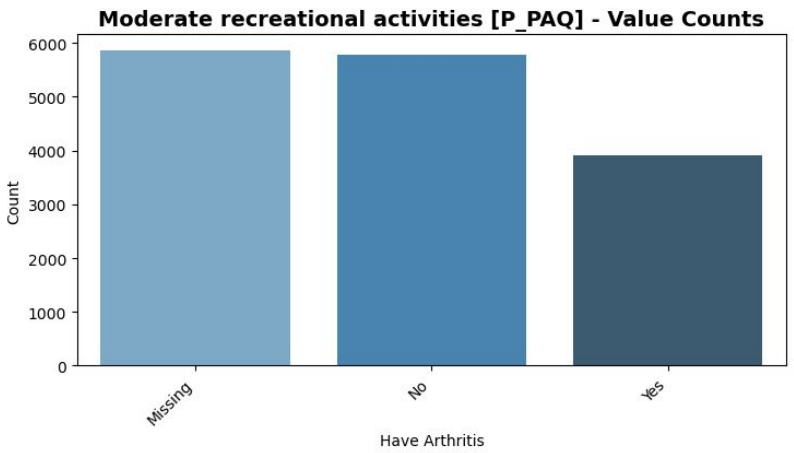

Blood manganese (ug/L) [P_PBCD]
Blood manganese (ug/L) [P_PBCD]
Missing
No
Yes
Have Arthritis

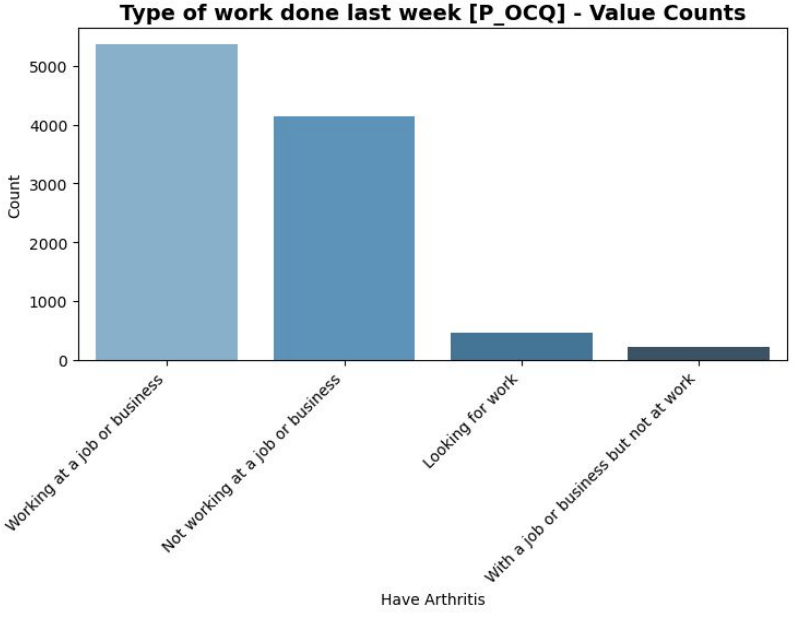

Total polyunsaturated fatty acids (gm) [P_DR1TOT]
Total polyunsaturated fatty acids (gm) [P_DR1TOT]
Missing
No
Yes
Have Arthritis

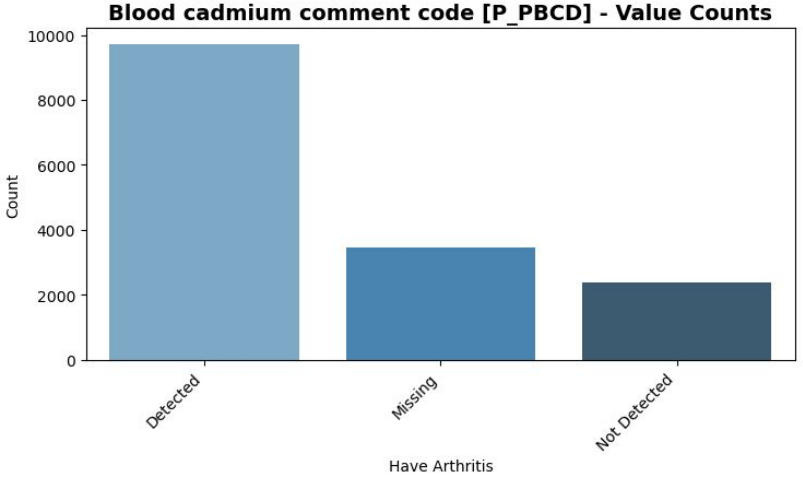

Copper (mg) [P_DR1TOT]
Copper (mg) [P_DR1TOT]
Missing
No
Yes
Have Arthritis


Potassium (mg) [P_DR1TOT]
Potassium (mg) [P_DR1TOT]
Missing
No
Yes
Have Arthritis


Are you now reducing salt in diet [P_MCQ] - Value Counts
Count
0
1000
2000
3000
4000
5000
Missing
No
Yes
Have Arthritis


Moderate recreational activities [P_PAQ] - Value Counts
Count
0
1000
2000
3000
4000
5000
6000
Missing
No
Yes
Have Arthritis


Type of work done last week [P_OCQ] - Value Counts
Count
0
1000
2000
3000
4000
5000
Working at a job or business
Not working at a job or business
Looking for work
With a job or business but not at work
Have Arthritis


Blood cadmium comment code [P_PBCD] - Value Counts
Count
0
2000
4000
6000
8000
10000
Detected
Missing
Not Detected
Have Arthritis

Risk Factor Features (p-value < .05)

Diet

Get Enough Calcium, Avoid High Salt Intake, Healthy Liver (Albumin), Don’t Overeat, Moisture Content in Diet, Enough Protein, Have Polyunsaturated Fatty Acids, Have Balanced Copper, Manganese and Potassium Levels

Exercise

Maintain Healthy Weight, Moderate Recreational Activities

Toxins

Avoid Cadmium Exposure (cigarette smoke, industrial processes), Avoid Lead and other heavy metals

Future Improvements

- Make additional models with split features
- Employ Variance Inflation Factor (VIF) to address multicollinearity
- Further test methods to address skew and imbalance classes
- Use Optuna for hyperparameter optimization

These steps will enhance the logistic model's F1 score, providing more accurate results.

Next Steps

- Analyse more datasets
- Refine models
- Investigate unknown environmental factors causing autoimmune disorders
- Investigate rising prevalence of autoimmune disorders

Thank You - Any Questions?

https://www.linkedin.com/in/chisum-lindauer/

Chisum Lindauer

Rod of Asclepius
Greek Symbol For Healing