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

nhanes

CDC

CENTERS FOR DISEASE CONTROL AND PREVENTION

Data Preparation

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

Scoring with F1

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

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

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
- Network mapping - how are features related to each other?

Thank You - Any Questions?

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

Chisum Lindauer

Rod of Asclepius
Greek Symbol For Healing

Correlation Between Risk Factors
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
Moving or speaking slowly or too fast [P_DPQ]
Have little interest in doing things [P_DPQ]
Root caries restoration [P_OHXDEN]
Self-reported weight-age 25 (pounds) [P_WHQ]
Current self-reported weight (pounds) [P_WHQ]
Main food shopper [P_DBQ]
Are you now reducing salt in diet [P_MCQ]
Doctor told you to exercise [P_MCQ]
Ever told you had any liver condition MCQ160L [P_MCQ]
Ever told you had heart attack [P_MCQ]
Ever told you had COPD, emphysema, ChB [P_MCQ]
Moderate work activity [P_PAQ]
Marital status [P_DEMO]
How do you consider your weight [P_WHQ]
Weight (kg) [P_BMX]
Alanine Aminotransferase (ALT) (U/L) [P_BIOPRO]
Ever told you have Hepatitis B? [P_HEQ]
Coded cuff size - oscillometric [P_BPXO]
Diastolic - 1st oscillometric reading [P_BPXO]
UIBC, Serum (ug/dL) [P_FETIB]
Total Protein (g/L) [P_BIOPRO]
Sodium (mmol/L) [P_BIOPRO]
GGT Comment Code [P_BIOPRO]
Exam wand type [P_LUX]
HS C-Reactive Protein Comment Code [P_HSCRP]
Tooth Count: #23 [P_OHXDEN]
Tooth Count: #10 [P_OHXDEN]
Tooth Count: #21 [P_OHXDEN]
Tooth Count: #13 [P_OHXDEN]
Coronal Caries: Tooth Count #4 [P_OHXDEN]
Coronal Caries: Tooth Count #25 [P_OHXDEN]
Coronal Caries: Tooth Count #7 [P_OHXDEN]
Coronal Caries: Tooth Count #27 [P_OHXDEN]
Tooth Count: #18 [P_OHXDEN]
Tooth Count: #16 [P_OHXDEN]
Tooth Count: #19 [P_OHXDEN]
Salt used at table yesterday? [P_DR1TOT]
Dentition Status Code [P_OHXDEN]
Blood manganese (nmol/L) [P_PBCD]
Past 30 day milk product consumption [P_DBQ]
Albumin, refrigerated serum (g/dL) [P_BIOPRO]
Total Bilirubin (umol/L) [P_BIOPRO]
Full sample interview weight [P_DEMO]
Mean platelet volume (fL) [P_CBC]
Total plain water drank yesterday (gm) [P_DR1TOT]
Total tap water drank yesterday (gm) [P_DR1TOT]
PFA 22:6 (Docosahexaenoic) (gm) [P_DR1TOT]
Beta-carotene (mcg) [P_DR1TOT]
Blood lead (ug/dL) [P_PBCD]
Hydroxycotinine, Serum (ng/mL) [P_COT]
Blood Urea Nitrogen (mmol/L) [P_BIOPRO]
Added vitamin B12 (mcg) [P_DR1TOT]
Dietary recall status [P_DR1TOT]
Gender [P_DEMO]
White blood cell count (1000 cells/uL) [P_CBC]
Chloride (mmol/L) [P_BIOPRO]
Last 7-d spent time in another home? [P_SMQSHS]
# of ready-to-eat foods in past 30 days [P_DBQ]
Masked variance pseudo-PSU [P_DEMO]
Globulin (g/dL) [P_BIOPRO]
Basophils number (1000 cells/uL) [P_CBC]
Received Hepatitis B 3 dose series [P_IMQ]
Moisture (gm) [P_DR1TOT]
Energy (kcal) [P_DR1TOT]
Number of foods/beverages reported [P_DR1TOT]
Retinol (mcg) [P_DR1TOT]
Niacin (mg) [P_DR1TOT]
Thiamin (Vitamin B1) (mg) [P_DR1TOT]
Selenium (mcg) [P_DR1TOT]
MFA 16:1 (Hexadecenoic) (gm) [P_DR1TOT]
SFA 16:0 (Hexadecanoic) (gm) [P_DR1TOT]
SFA 4:0 (Butanoic) (gm) [P_DR1TOT]
Coronal Caries: Tooth Count #14 [P_OHXDEN]
Lymphocyte percent (%) [P_CBC]

1. Environmental Factors

Behavioral and Lifestyle Factors

- **Are you now reducing salt in diet [P_MCQ]:** (Not reducing salt increases risk) High salt intake can influence inflammation and immune response, potentially contributing to arthritis.
- **Moderate recreational activities [P_PAQ]:** (Less activity increases risk) Lack of physical activity can contribute to obesity and joint stress, increasing arthritis risk.
- **SFA 12:0 (Dodecanoic) (gm) [P_DR1TOT]:** (Higher intake increases risk) Higher intake of saturated fatty acids can contribute to inflammation, though the data shows pretty similar class distributions for it on for those reporting arthritis, it is found to be predictive.
- **Number of foods/beverages reported [P_DR1TOT]:** (Lower diversity or higher quantity increases risk) Limited dietary diversity can affect inflammation and overall health. Additionally, consuming a higher quantity of certain foods might also contribute to inflammation.
- **Moisture (gm) [P_DR1TOT]:** (Lower moisture content increases risk) Lower moisture content in diet can affect hydration status, impacting joint health.
- **Protein (gm) [P_DR1TOT]:** (High intake can be beneficial) Adequate protein intake is essential for muscle and joint health, potentially reducing arthritis risk.
- **Total polyunsaturated fatty acids (gm) [P_DR1TOT]:** (Higher intake can be beneficial) Polyunsaturated fatty acids have anti-inflammatory properties.
- **Copper (mg) [P_DR1TOT]:** (Imbalance increases risk) Copper plays a role in maintaining healthy connective tissues.
- **Potassium (mg) [P_DR1TOT]:** (Adequate intake can be beneficial) Proper potassium levels are important for muscle and nerve function.

Occupational and Environmental Exposures

- **Type of work done last week [P_OCQ]:** (Physically demanding jobs increase risk) Certain types of work can stress joints and increase arthritis risk.
- **Blood manganese (ug/L) [P_PBCD]:** (Abnormal levels increase risk) Abnormal levels of manganese can affect joint health.
- **Blood lead comment code [P_PBCD]:** (High exposure increases risk) Exposure to lead and other heavy metals can be associated with inflammatory responses.
- **Blood cadmium comment code [P_PBCD]:** (High exposure increases risk) Cadmium exposure is linked to inflammation.

2. Well Known Risk Factors

- **HS C-Reactive Protein (mg/L) [P_HSCRP]:** (High levels increase risk) Elevated levels of C-reactive protein are associated with inflammation, a key factor in the development and progression of arthritis.
- **Doctor ever said you were overweight [P_MCQ]:** (Yes increases risk) Obesity is a significant risk factor for arthritis.
- **Ever told you had thyroid problem [P_MCQ]:** (Yes increases risk) Thyroid dysfunction is associated with autoimmune conditions.

3. Possible Early Warning Signs

- **Iron, refrigerated serum (ug/dL) [P_BIOPRO]:** (Low levels increase risk) Disrupted iron metabolism can be a sign of inflammation.
- **Iron, refrigerated serum (umol/L) [P_BIOPRO]:** (Low levels increase risk) Disrupted iron metabolism can be a sign of inflammation.
- **Lactate Dehydrogenase (LDH) (IU/L) [P_BIOPRO]:** (High levels increase risk) Elevated LDH levels can indicate tissue damage.
- **Red cell distribution width (%) [P_CBC]:** (High levels increase risk) High variability in red blood cell size can indicate inflammation.
- **Monocyte number (1000 cells/uL) [P_CBC]:** (High levels increase risk) Elevated monocytes are a sign of inflammation.
- **Mean cell hemoglobin concentration (g/dL) [P_CBC]:** (Low levels increase risk) Changes can indicate anemia related to chronic disease.
- **Mean cell volume (fL) [P_CBC]:** (Low levels increase risk) Altered levels can suggest anemia.
- **Albumin, refrigerated serum (g/dL) [P_BIOPRO]:** (Low levels increase risk) Lower levels suggest chronic inflammation.
- **Blood Urea Nitrogen (mg/dL) [P_BIOPRO]:** (High levels increase risk) Indicates kidney function.
- **Potassium (mmol/L) [P_BIOPRO]:** (High or low levels increase risk) Imbalance can affect muscle and nerve function.
- **Globulin (g/dL) [P_BIOPRO]:** (High levels increase risk) Higher levels can be a marker of chronic inflammation.
- **Total Calcium (mg/dL) [P_BIOPRO]:** (High or low levels increase risk) Calcium imbalance can affect bone health.
- **Creatinine, refrigerated serum (mg/dL) [P_BIOPRO]:** (High levels increase risk) Indicates kidney function.
- **Iron frozen, Serum (umol/L) [P_FETIB]:** (Low levels increase risk) Reflects iron storage and metabolism.
- **Trouble sleeping or sleeping too much [P_DPQ]:** (Yes increases risk) Sleep disturbances are common in arthritis patients.
- **Moving or speaking slowly or too fast [P_DPQ]:** (Yes increases risk) Psychomotor changes can be associated with chronic pain and depression in arthritis.
- **Poor appetite or overeating [P_DPQ]:** (Yes increases risk) Dietary habits can influence arthritis risk.
- **Abdominal pain during past 12 months? [P_MCQ]:** (Yes increases risk) Chronic pain can be an early sign of arthritis.
- **Self-reported greatest weight (pounds) [P_WHQ]:** (High weight increases risk) Weight history can indicate risk of obesity-related arthritis.
- **Self-reported weight-age 25 (pounds) [P_WHQ]:** (High weight increases risk) Historical weight can provide insights into long-term risk factors.

4. Possible Diagnosis Markers

- **HS C-Reactive Protein (mg/L) [P_HSCRP]:** (High levels) Inflammation marker used in diagnosis.
- **Iron, refrigerated serum (umol/L) [P_BIOPRO]:** (Low levels) Inflammation and iron metabolism marker.
- **Lactate Dehydrogenase (LDH) (IU/L) [P_BIOPRO]:** (High levels) Tissue damage marker.
- **Albumin, refrigerated serum (g/dL) [P_BIOPRO]:** (Low levels) Chronic inflammation marker.
- **Blood Urea Nitrogen (mg/dL) [P_BIOPRO]:** (High levels) Kidney function marker.
- **Potassium (mmol/L) [P_BIOPRO]:** (High or low levels) Muscle and nerve function marker.
- **Globulin (g/dL) [P_BIOPRO]:** (High levels) Chronic inflammation marker.
- **Total Calcium (mg/dL) [P_BIOPRO]:** (High or low levels) Bone health marker.
- **Creatinine, refrigerated serum (mg/dL) [P_BIOPRO]:** (High levels) Kidney function marker.
- **Iron frozen, Serum (umol/L) [P_FETIB]:** (Low levels) Iron storage and metabolism marker.
- **Red cell distribution width (%) [P_CBC]:** (High levels) Inflammation marker.
- **Monocyte number (1000 cells/uL) [P_CBC]:** (High levels) Inflammation marker.
- **Mean cell hemoglobin concentration (g/dL) [P_CBC]:** (Low levels) Anemia marker.
- **Mean cell volume (fL) [P_CBC]:** (Low levels) Anemia marker.

5. Other Relevant Factors

Demographic Factors

- **Age in years at screening [P_DEMO]:** (Older age increases risk) Age is a well-known risk factor for arthritis.
- **Gender [P_DEMO]:** (Women are at higher risk) Gender can influence arthritis risk, with women being more prone to certain types.
- **Country of birth [P_DEMO]:** (Varies) Geographical and genetic factors may play a role in arthritis prevalence.
- **Education level - Adults 20+ [P_DEMO]:** (Lower education increases risk) Lower education levels are often associated with higher arthritis risk due to socioeconomic factors.
- **Family monthly poverty level category [P_INQ]:** (Lower income increases risk) Lower socioeconomic status is associated with higher arthritis risk due to limited access to healthcare and healthy lifestyle options.

Medical History and Screening

- **Blood ever tested for HIV virus? [P_HSQ]:** (Positive result increases risk) Indicates the history of medical screening and potential immunocompromised status.
- **Total # of Antacids Taken [P_DSQTOT]:** (Higher usage increases risk) High usage of antacids can reflect gastrointestinal issues linked to arthritis.
- **Total # of Dietary Supplements Taken [P_DSQTOT]:** (Higher usage) Supplement use can indicate attempts to manage chronic conditions.
- **Received Hepatitis A vaccine [P_IMQ]:** (Vaccination status) While not a direct risk factor, vaccination status can reflect overall health management.

Oral Health Indicators

- **Tooth Count: #10, #12, #24, #28 [P_OHXDEN]:** (Fewer teeth increases risk) Reduced tooth count can indicate poor oral health, which is linked to systemic inflammation.
- **Coronal Caries: Tooth Count #24 [P_OHXDEN]:** (More caries increases risk) Dental caries are indicators of oral health, affecting overall health.
- **Dental Implant: yes / no? [P_OHXDEN]:** (Yes increases risk) Presence of dental implants indicates prior oral health issues.

Family Medical History

- **Close relative had heart attack? [P_MCQ]:** (Yes increases risk) Family history of cardiovascular disease can be linked to inflammatory conditions.
- **Close relative had asthma? [P_MCQ]:** (Yes increases risk) Family history of asthma suggests a predisposition to autoimmune conditions.
- **Close relative had diabetes? [P_MCQ]:** (Yes increases risk) Family history of diabetes is linked to metabolic syndrome and inflammation.

Self-Reported Health Conditions

- **Ever been told you have asthma [P_MCQ]:** (Yes increases risk) Asthma and other autoimmune conditions are associated with higher arthritis risk.
- **Ever had gallbladder surgery? [P_MCQ]:** (Yes increases risk) History of gallbladder issues can indicate metabolic problems.