

Capstone Project 2 - Milestone Report

Predicting Health using the National Health and Nutrition Survey

Problem Statement

There are many factors that impact the health and wellness of human society. It should be possible to reduce the occurrence of particular health concerns if those factors with high impact can be identified. The goal of this project is to use statistical inference and machine learning to explore if predictions in certain illnesses can be made related to habits, nutrition, BMI, and bloodwork.

Proposed Solution

Perform exploratory analysis and predictive modeling from the National Health and Nutrition Survey (NHANES) dataset that assesses the health and nutritional status of people in the United States.

- Accurate predictive models can help people in general to benefit if looking for ways to better their life.
- It can extend to professionals in the medical, nutritional, and fitness arenas to discuss the potential effects of life choices relating to health conditions.
- This could lead to people wanting to research particular findings in more depth in order to better their lives. Potentially adding longevity and making a positive difference for our population at large.

Datasets

A collection of datasets from the National Health and Nutrition Survey (NHANES) was obtained from Kaggle. <https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey>.

This information was used in the project to study variables that could potentially help predict health conditions and improve human lives.

More information on the NHANES survey can be found on the Center for Disease Control and Prevention's website https://www.cdc.gov/Nchs/Nhanes/about_nhanes.htm.



Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™

Wrangling Steps Performed:

Data, Demographics, and Labs

1. Three csv files were downloaded and read into a normalized pandas dataframe. They were merged together to create one dataframe.

2. Columns were explored to determine which would be useful on this project. Those identified for use were then renamed from their initial codes to identifiable word strings.
3. A feature generation for BMI at age 25 was created using the height and weight found in the data.
4. The data was checked and scanned for any null information present. This was later used when looking at categories for those having asthma and those having arthritis. Null values had their entire rows removed when those columns were used for explorations.

Dealing with Outliers

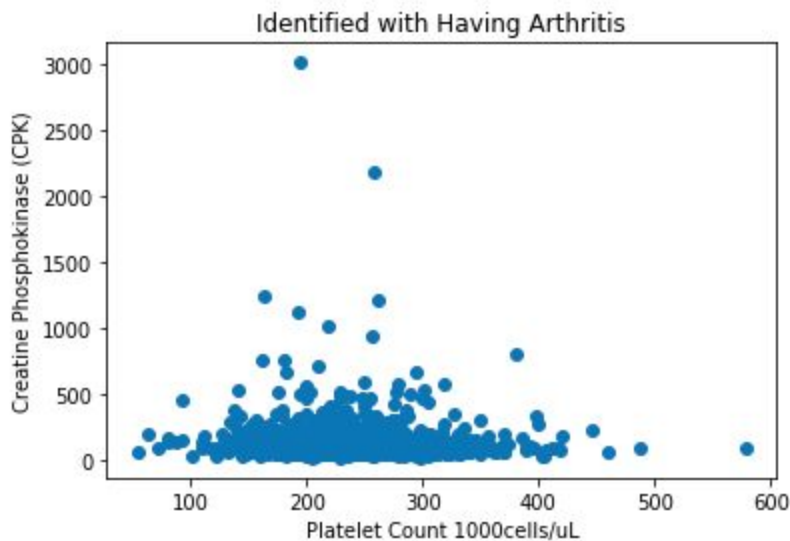
The dataset had categories that had significantly less data to use. Categories with more data were used in order to provide a substantial sampling to match the illnesses chosen to look at of asthma and arthritis.

Exploratory Data Analysis

Correlations can be found among the data.

Gender	-1	0.033	0.013	0.36	0.007	0.004	0.001	0.093	0.008	0.063	0.005	0.044	0.024	0.006	0.02	0.039	0.037	0.027	0.033	0.045	0.075	0.079	0.016	0.14	0.00	
AgeAtSurvey	-0.033	1	0.035	0.12	0.011	0.006	0.013	-0.3	0.013	0.012	0.089	0.003	0.049	0.11	-0.047	0.076	0.079	0.005	0.046	0.028	0.19	0.028	0.036	0.37	-0.03	
WeightAt25Yrs	-0.013	0.035	1	-0.67	0.48	0.011	0.003	0.058	0.026	0.012		0.077	0.029	-0.044	0.034	0.029	0.065	0.013	0.007	0.028	0.065	0.063	0.088	0.066	0.01	
BMI	0.36	0.12	-0.67	1	-0.32	0.021	0.009	0.057	0.005	0.017		0.066	0.055	0.035	-0.014	0.011	-0.038	0.011	0.025	-0.05	0.039	0.061	0.073	0.017	0.02	
HeightAt25Yrs	0.007	0.011	0.48	-0.32	1	0.012	0.011	0.1	0.006	-0.024		0.032	0.008	0.001	0.002	0.043	0.034	0.022	0.001	0.056	0.018	0.051	0.15	0.049	-0.1	
PlateletCount	0.004	0.006	0.011	0.021	0.012	1	-0.057	0.013	0.004	0.012	0.027	0.026	0.003	0.019	0.003	0.008	0.005	0.028	0.012	0.027	0.009	-0.03	-0.019	0.026	0.01	
CreatinePhosphokinase	-0.001	0.013	0.003	0.009	0.011	0.057	1	0.018	0.003	0.029	0.023	0.002	0.005	0.013	0.012	0.015	0.007	0.012	0.01	0.006	0.025	0.004	0.013	0.022	0.01	
HasArthritis	-0.093	-0.3	0.058	0.057	0.1	0.013	0.018	1	0.079	0.093		0.025	0.036	0.033	0.073	0.019	0.049	0.028	0.022	0.047	0.059	0.022	-0.05	-0.1	-0.00	
HasAsthma	-0.008	0.013	0.026	0.005	0.006	-0.004	0.003	0.079	1	0.14	0.013	0.003	0.026	0.037	0.018	-0.026	0.018	0.011	-0.01	0.01	-0.006	0.032	-0.014	0.026	0.00	
AllergyCongestionPast12Months	-0.063	0.012	0.012	0.017	0.024	-0.012	0.029	0.093	0.14	1		0.004	0.045	0.022	0.032	0.002	0.013	0.028	0.017	-0.054	0.032	0.055	-0.033	0.009	0.00	
Breast_fed	-0.005	0.089			0.027	-0.023		0.013		1					0.034	0.041										
UseMilk>=5x/wk	-0.044	0.003	0.077	0.066	-0.032	0.026	0.002	0.025	0.000	0.004		1	-0.24	-0.39	-0.017	0.039	-0.11	-0.048	0.085	-0.034	0.026	-0.032	-0.035	0.007	0.02	
DrankMilkAge13to17yrs	-0.024	0.049	0.029	0.055	-0.009	0.003	0.005	-0.036	0.026	0.045		-0.24	1	0.48	0.048	0.004	0.039	0.017	0.006	0.003	0.048	0.013	-0.023	0.04		
DrankMilkAge18to35yrs	0.006	0.11	-0.044	0.035	0.001	0.019	0.013	-0.033	0.037	0.022		-0.39	0.48	1	0.014	-0.04	-0.049	0.035	0.029	0.023	0.032	-0.035	0.024	0.018	0.01	
ChemicalPesticideUsedInHomePast7days	-0.02	0.047	0.034	-0.014	0.002	0.003	0.012	0.073	0.018	0.032	0.034	-0.017	0.048	0.014	1	0.19	-0.025	0.001	0.008	0.079	0.012	0.015	0.004	-0.04	-0.00	
ChemicalWeedKillUsedExteriorHomePast7days	-0.039	0.007	0.002	0.010	0.042	0.088	0.015	-0.019	0.020	0.020	0.041	0.039	0.004	-0.04	0.19	1	0.018	0.006	0.002	0.004	0.040	0.010	0.007	0.002	0.00	
NumberOfDays/WeekVigorousWork	-0.037	0.075	0.069	0.38	0.034	0.005	0.007	0.049	0.018	0.013		-0.11	-0.039	0.049	0.025	0.019	1	0.12	0.33	0.054	0.043	0.17	0.002	-0.01	0.04	
Minutes/DayVigorousIntensityWork	-0.027	0.005	0.013	0.011	-0.022	0.028	0.012	0.028	0.011	0.028		0.048	0.017	0.035	0.001	0.006	0.12	1	0.59	0.7	0.006	0.02	0.96	0.015	0.00	
NumberOfDays/WeekModerateWork	-0.039	0.046	0.078	0.025	0.001	0.012	0.01	0.022	-0.01	0.017		0.085	0.006	0.029	0.008	0.002	0.33	0.59	1	0.082	0.016	0.17	0.034	0.032	0.01	
Minutes/DayModerateIntensityWork	-0.045	0.028	0.028	-0.05	0.056	0.027	0.006	0.047	0.01	0.054		0.034	0.003	0.023	0.007	0.004	0.054	0.7	0.082	1	0.014	0.018	0.94	0.031	-0.03	
WalkOrBicycle10+Min/weekYN	-0.075	0.19	-0.065	0.039	-0.018	0.094	0.025	-0.059	0.006	0.032		0.026	0.003	0.032	-0.012	0.046	0.043	0.006	0.016	0.014	1			0.13	-0.00	
NumberDays/WeekWalkBicycle	-0.079	0.028	0.063	0.061	0.051	-0.030	0.048	0.022	0.032	0.055		0.032	0.048	0.035	0.015	0.011	0.17	0.02	0.17	0.018		1	0.04	0.007	0.11	
Minutes/DayWalkBicycleTransportation	-0.016	0.036	0.088	-0.073	0.15	-0.019	0.013	-0.05	-0.014	0.031		0.035	0.013	0.024	0.004	0.007	0.002	0.96	0.034	0.94		0.04	1	-0.006	0.01	
VigorousRecreationalActivities10+min/WeekYN	0.14	0.37	0.066	0.017	0.049	0.026	0.022	-0.1	0.026	0.009		0.007	0.023	0.018	-0.04	0.002	-0.01	0.015	0.032	0.031	0.13	-0.007	0.006	1		
Days/WeekVigorousRecreationalActivities	-0.003	0.035	0.017	0.023	-0.1	0.017	0.018	0.007	0.005	0.009		0.022	0.042	0.011	-0.000	0.003	0.06	-0.002	0.017	-0.036	0.046	0.11	0.013		1	
Minutes/DayVigorousRecreationalActivities	-0.057	0.048	0.011	0.098	0.013	0.031	0.004	0.040	0.078	0.043		0.057	0.044	0.013	0.004	0.002	0.12	0.051	0.015	0.064	0.012	0.043	0.003		0.00	
ModerateRecreationalActivities10+min/WeekYN	-0.003	0.079	0.093	0.068	0.062	0.013	0.014	0.044	0.008	0.008		0.014	8e-06	0.003	0.015	0.022	0.057	0.022	0.026	0.034	0.083	0.037	0.023	0.26	0.2	
Days/WeekModerateRecreationalActivities	-0.029	0.035	0.036	0.061	0.016	0.004	0.019	0.012	0.011	0.003		0.018	0.075	0.026	0.001	0.021	0.071	0.064	0.021	0.006	0.048	0.14	5e-08	0.005	0.11	
Minutes/DayModerateRecreationalActivities	-0.13	0.033	0.073	-0.12	0.021	0.014	0.002	0.015	0.001	0.006		0.006	0.009	0.022	0.033	0.009	0.054	0.063	-0.017	0.13	0.044	0.085	0.095	0.003	0.02	
Minutes/DaySedentaryActivity	0.007	0.012	0.049	0.039	0.031	0.013	0.022	-0.02	0.039	0.065		0.046	0.024	0.028	0.006	0.009	0.060	0.980	0.42	0.45	0.063	0.003	-0.04	0.18	-0.008	0.2
Gender		Gender																								
AgeAtSurvey		AgeAtSurvey																								
WeightAt25Yrs		WeightAt25Yrs																								
BMI		BMI																								
HeightAt25Yrs		HeightAt25Yrs																								
PlateletCount		PlateletCount																								
CreatinePhosphokinase		CreatinePhosphokinase																								
HasArthritis		HasArthritis																								
HasAsthma		HasAsthma																								
AllergyCongestionPast12Months		AllergyCongestionPast12Months																								
Breast_fed		Breast_fed																								
UseMilk>=5x/wk		UseMilk>=5x/wk																								
DrankMilkAge13to17yrs		DrankMilkAge13to17yrs																								
DrankMilkAge18to35yrs		DrankMilkAge18to35yrs																								
ChemicalPesticideUsedInHomePast7days		ChemicalPesticideUsedInHomePast7days																								
ChemicalWeedKillUsedExteriorHomePast7days		ChemicalWeedKillUsedExteriorHomePast7days																								
NumberOfDays/WeekVigorousWork		NumberOfDays/WeekVigorousWork																								
Minutes/DayVigorousIntensityWork		Minutes/DayVigorousIntensityWork																								
NumberOfDays/WeekModerateWork		NumberOfDays/WeekModerateWork																								
Minutes/DayModerateIntensityWork		Minutes/DayModerateIntensityWork																								
WalkOrBicycle10+Min/weekYN		WalkOrBicycle10+Min/weekYN																								
NumberDays/WeekWalkBicycle		NumberDays/WeekWalkBicycle																								
Minutes/DayWalkBicycleTransportation		Minutes/DayWalkBicycleTransportation																								
VigorousRecreationalActivities10+min/WeekYN		VigorousRecreationalActivities10+min/WeekYN																								
Days/WeekVigorousRecreationalActivities		Days/WeekVigorousRecreationalActivities																								

One area this study further explored the correlation with those who reported having been told they had arthritis and their bloodcounts.



The scatterplot shows most participants identified with arthritis are in the lower left quadrant, which indicates lower platelet and CPK counts. Based on this visual alone, a hypothesis statement such that a participant with low levels of CPK and low platelet counts have a good chance of being told by a health professional that they have arthritis. Further, there are no participants in the upper right quadrant, which suggests that higher counts for both CPK and platelets relates to those not being diagnosed with arthritis. There are definitely a few outliers in the upper left quadrant that are not significant and would not be included when making generalizations.

There are more outliers in the lower right quadrant, perhaps significant enough to look further into whether the higher platelet counts play a role in diagnosing possible arthritis.

In general, high CPK levels in the muscle suggest the presence of inflammatory muscle disease, but they can also be caused by trauma, injection into the muscle, or muscle disease due to hypothyroidism. Conversely, low levels of CPK can be indicative of rheumatoid arthritis.

<https://www.arthritis-health.com/glossary/creatin-phosphokinase>

Hypothesis Testing

Machine Learning Algorithms K-Means Clustering

Conclusion: