Identifying Health Conditions Related to Cancer

By Aviva Mazurek

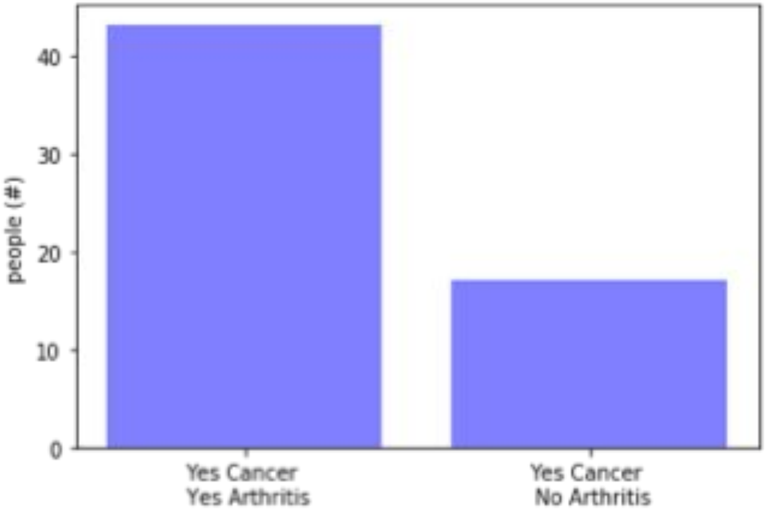
Objectives

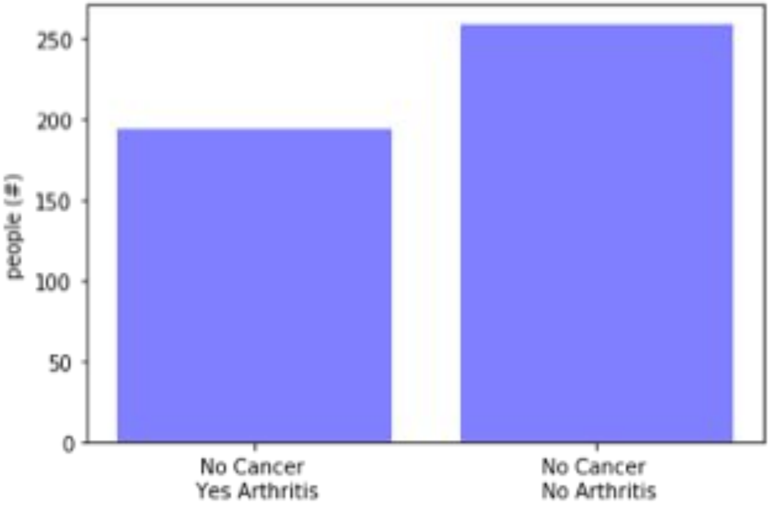
- Classify cancer as “yes” or “no” based on other health conditions such as: Asthma, Arthritis, Psoriasis, Celiac, Gouts, Heart Failure, Coronary Heart Disease, Congestive Heart Failure, Angina, Angina/Pectoris, Heart Attack, Strokes, Emphysema, COPD, Jaundice

  - ➔ Analyzed data from 2013-2014 NHANES surveys
    - ◆ 453 respondents without cancer, and 60 with cancer

EDA – Arthritis greatest indicator

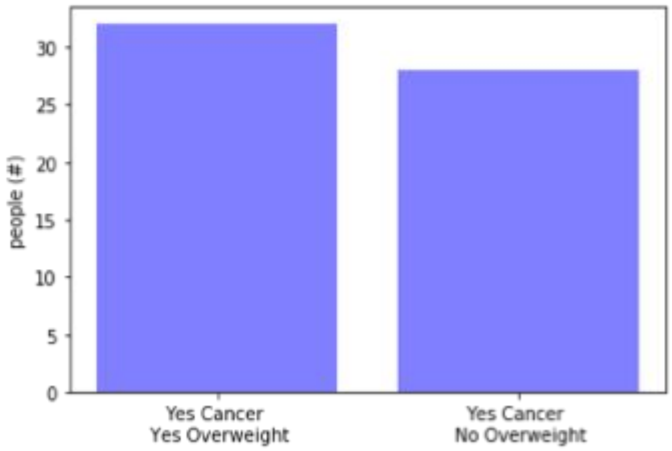
- Respondents with arthritis twice as likely to have cancer
  - consistent with current research in Sweden and France

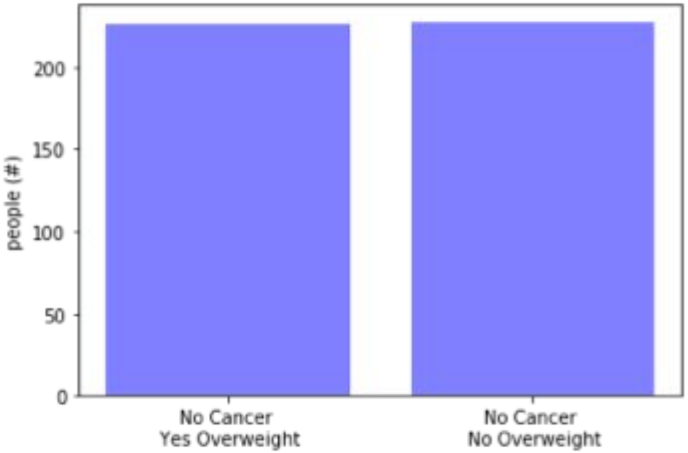




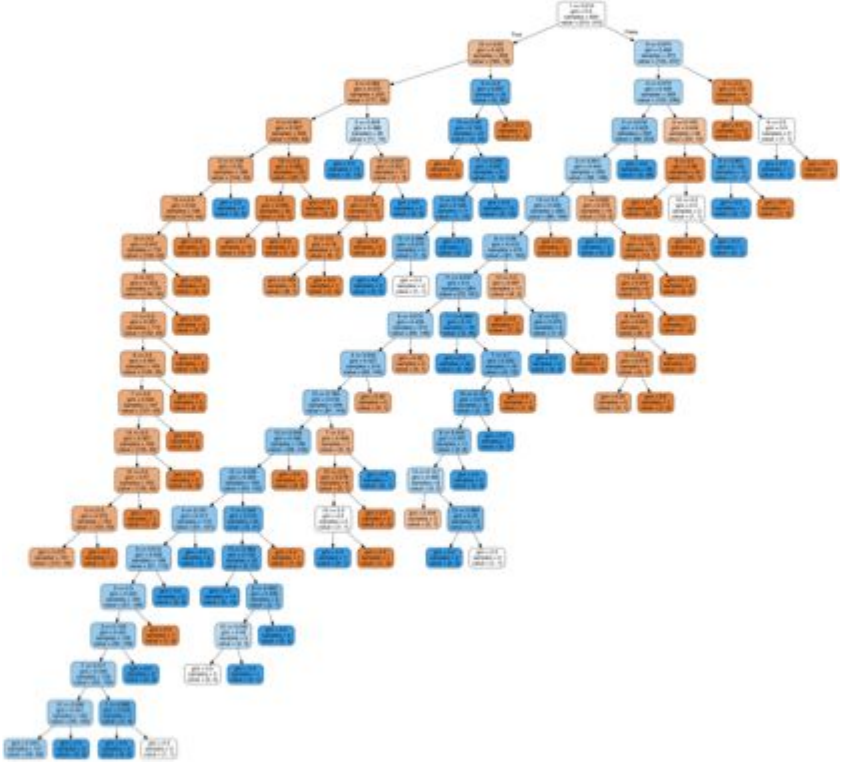
EDA - removed “overweight” from dataset

- Evenly distributed among respondents with and without cancer





Initial Tree Model



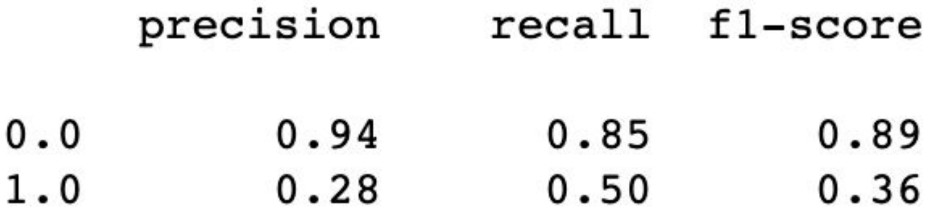
Accuracy = 77%

```
        precision    recall  f1-score

0.0          0.91      0.83      0.87
1.0          0.15      0.25      0.19
```

Grid Search for Tree Model

- Recommended no max depth

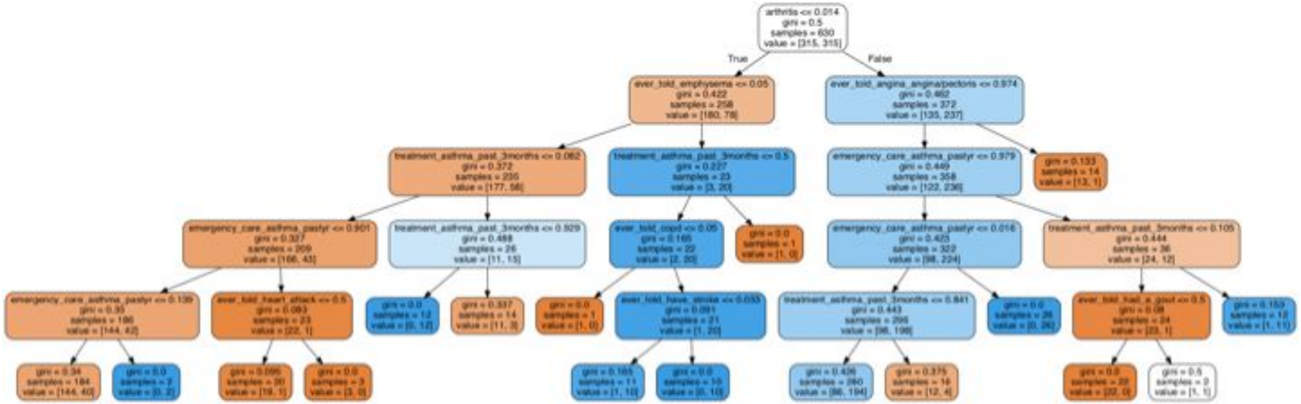
Accuracy = 81%



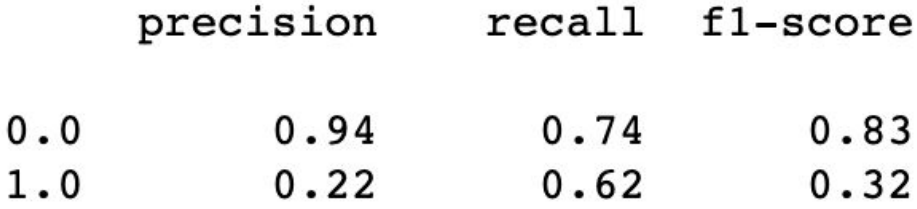
```
          precision    recall  f1-score

0.0            0.94      0.85      0.89
1.0            0.28      0.50      0.36
```

Manual Pruning

- Recall for cancer the highest
- Prefer a higher recall for cancer diagnosis over higher accuracy
  - Prefer more false positives than false negatives



Accuracy = 73%
Mean Cross Validation Score: 84.76%



```
                precision    recall  f1-score

         0.0         0.94      0.74      0.83
         1.0         0.22      0.62      0.32
```

Tree method outputs highest recall for cancer diagnosis

	Cancer/No Cancer Predictions	Precision	Recall	F1 Score
XGBoost	No Cancer (0)	0.90	0.77	0.83
XGBoost	Cancer(1)	0.23	0.43	0.30
SVM	No Cancer(0)	0.88	0.91	0.89
SVM	Cancer(1)	0.25	0.19	0.22
KNN	Total	0.27	0.19	0.22
Final tree	No Cancer (0)	0.94	0.74	0.83
Final tree	Cancer (1)	0.22	0.62	0.32

Conclusions

- Data suggests arthritis possible indicator for cancer
    - Unidentified mechanism - could be the medications people take for arthritis, or could be inflammation allows for more mutations...
- Data indicates tree model produces highest recall

Future Work

- Classify specific cancers
- Identify causality, or mechanism of causality
- Include more health conditions
- Include ages respondents diagnosed with health conditions and cancer