Predicting the Recurrence of Thyroid Cancer for Post-Treatment Patients

Alex Lai

Content

- Problem Statement
- Description of Dataset
- Data Wrangling
- Exploratory Data Analysis (EDA)
- Feature Engineering
- Modeling
- Result Analysis & Limitations
- Conclusion, Future Work, & Improving the model

Problem Statement

- To make a predictive model for thyroid cancer patients to see how likely after treatment their cancer will reoccur.
  - **New patient eligibility test for the treatment**
- Accuracy
  - At least 95%
- Stakeholders
  - company developing the treatment, doctors, patients
- Solution Space
  - Decision Tree based model using Pandas.
- Constraints
  - limited amount of data
- Time Frame
  - This will be completed within the next 3 months

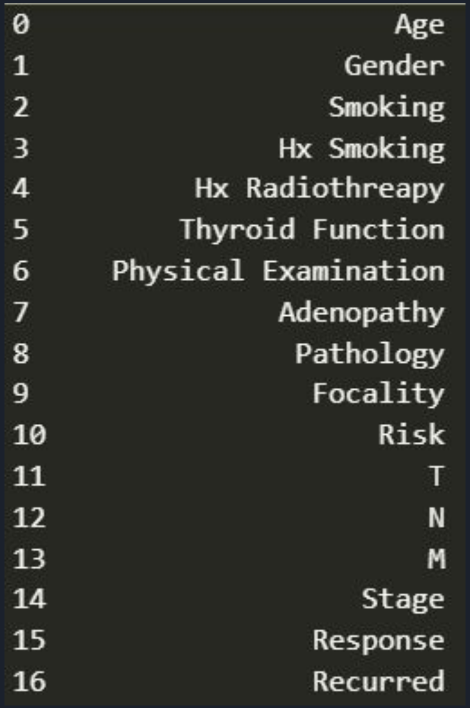
Description of Dataset

This data was provided by the UCI Machine Learning Repository:
https://archive.ics.uci.edu/dataset/915/differentiated+thyroid+cancer+recurrence

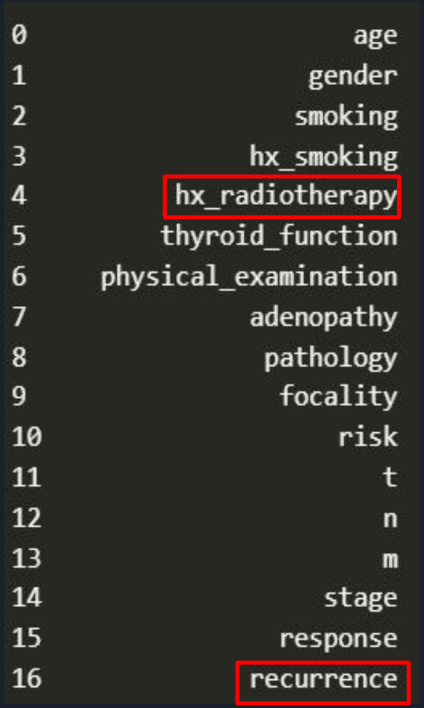
And is available on Kaggle:
https://www.kaggle.com/datasets/jainaru/thyroid-disease-data

This data under the license CC BY 4.0 ATTRIBUTION 4.0 INTERNATIONAL Deed:
https://creativecommons.org/licenses/by/4.0/

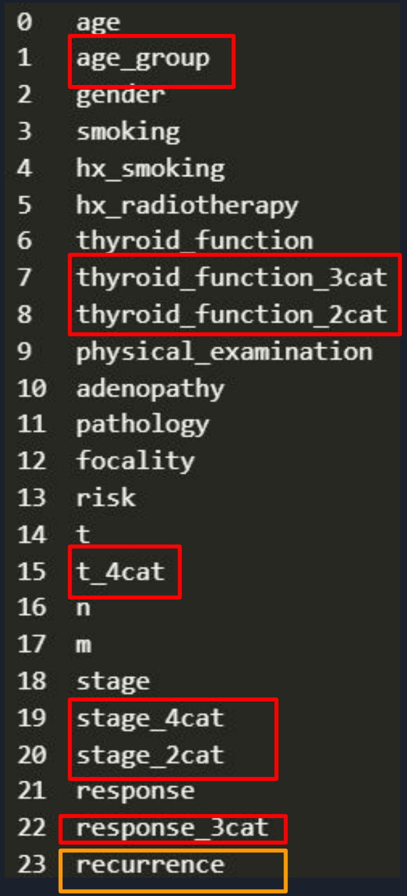
Data Wrangling



```
0                  Age
1               Gender
2              Smoking
3           Hx Smoking
4      Hx Radiothreapy
5     Thyroid Function
6 Physical Examination
7           Adenopathy
8            Pathology
9             Focality
10                Risk
11                   T
12                   N
13                   M
14               Stage
15            Response
16            Recurred
```

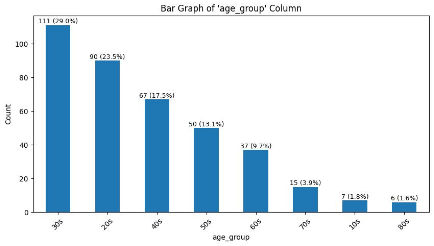


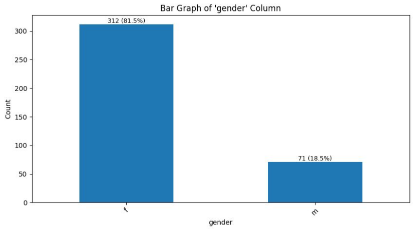
```
0                    age
1                 gender
2                smoking
3             hx_smoking
4       hx_radiotherapy
5       thyroid_function
6   physical_examination
7             adenopathy
8              pathology
9               focality
10                  risk
11                     t
12                     n
13                     m
14                 stage
15              response
16            recurrence
```

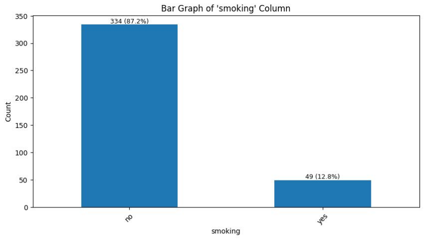


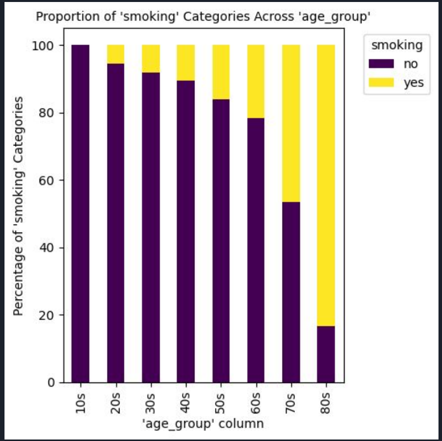
```
0   age
1   age_group
2   gender
3   smoking
4   hx_smoking
5   hx_radiotherapy
6   thyroid_function
7   thyroid_function_3cat
8   thyroid_function_2cat
9   physical_examination
10  adenopathy
11  pathology
12  focality
13  risk
14  t
15  t_4cat
16  n
17  m
18  stage
19  stage_4cat
20  stage_2cat
21  response
22  response_3cat
23  recurrence
```

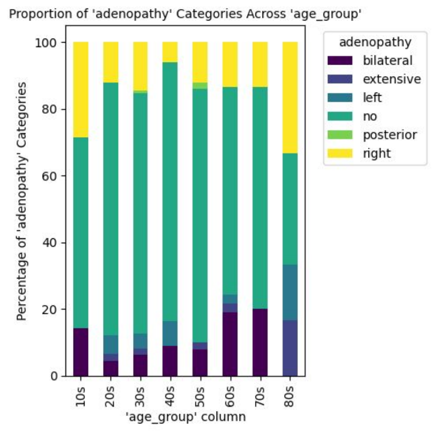
Exploratory Data Analysis (EDA)

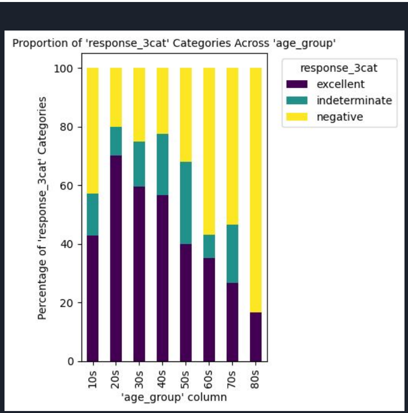




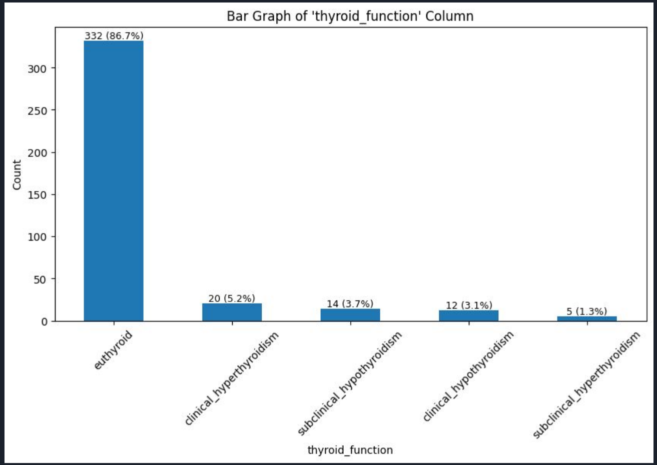


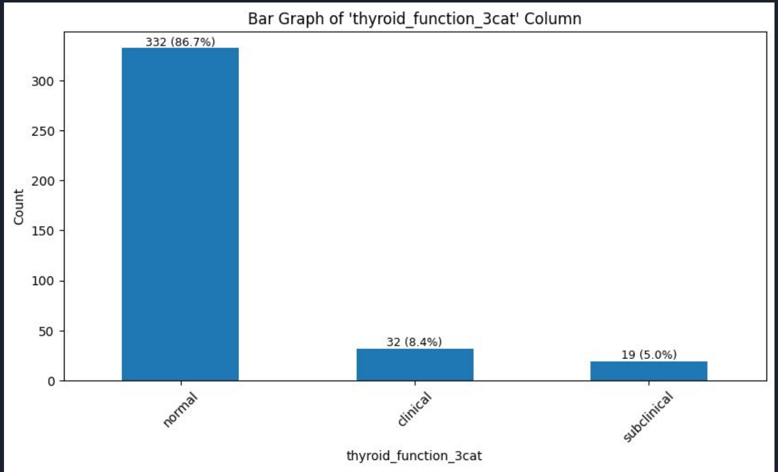


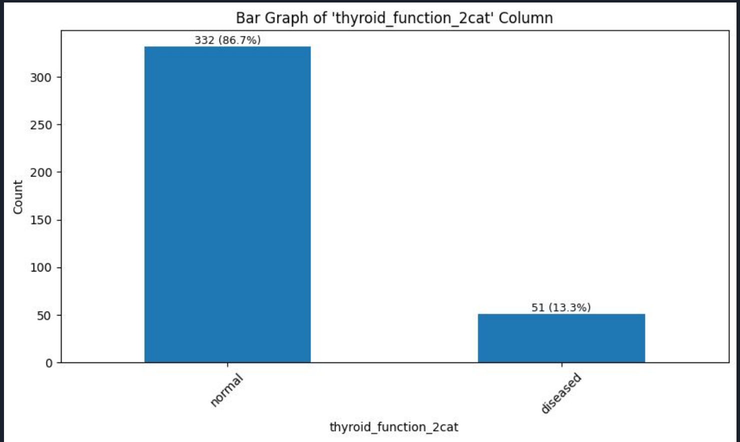




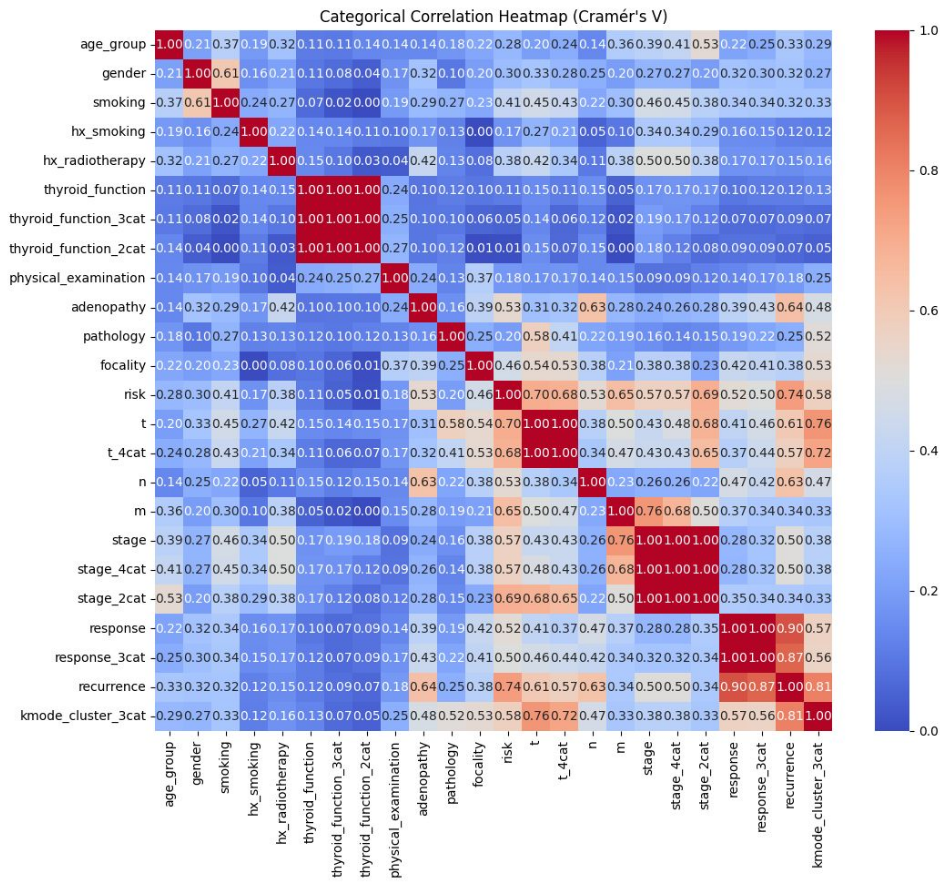
Exploratory Data Analysis (EDA)







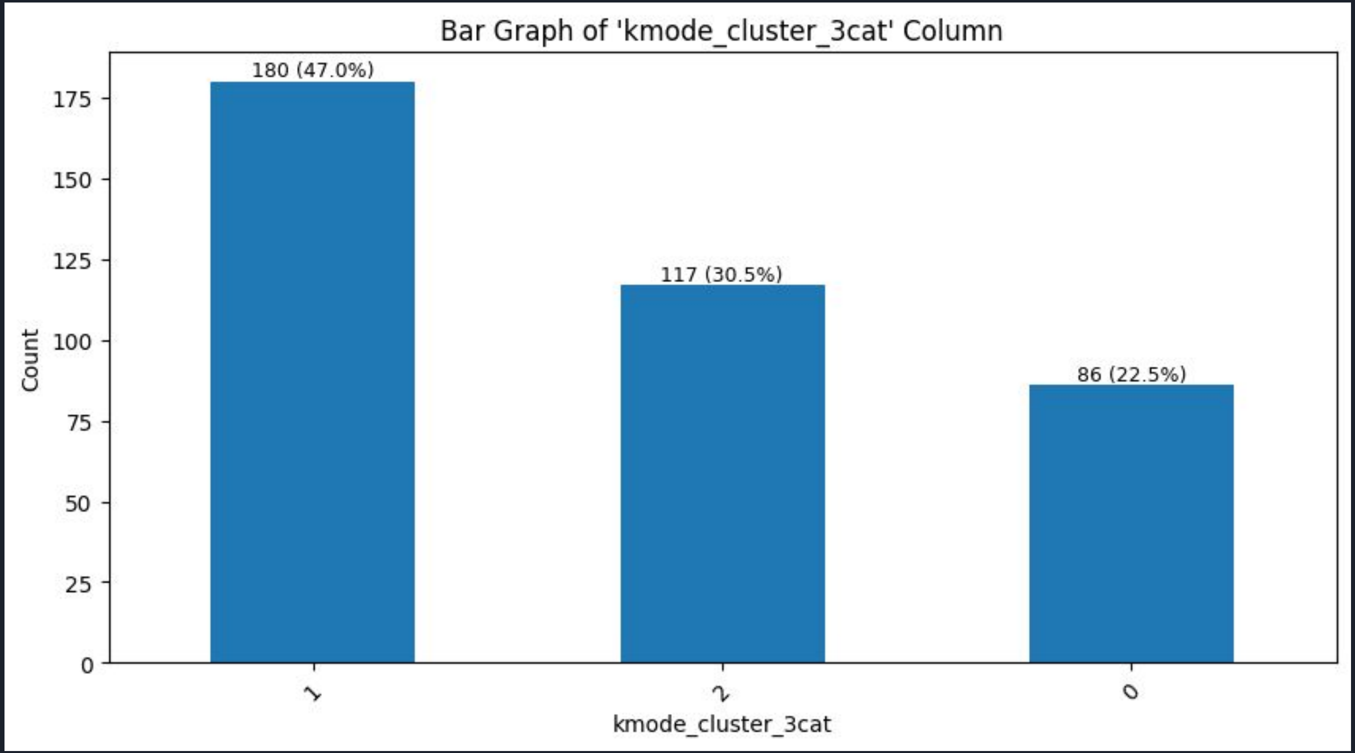
Exploratory Data Analysis (EDA)



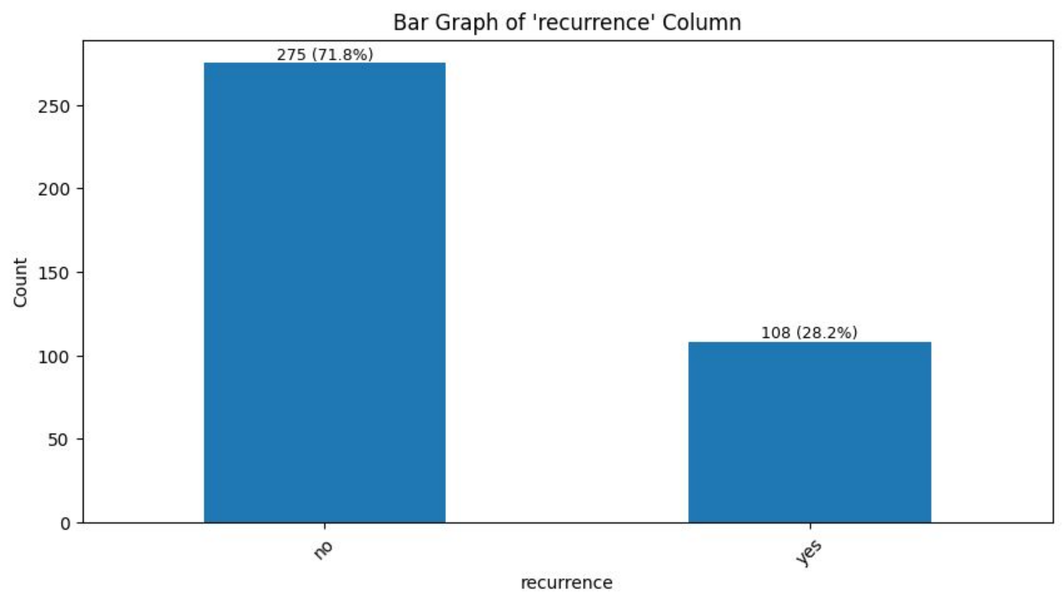
Categorical Correlation Heatmap (Cramér's V)

	age_group	gender	smoking	hx_smoking	hx_radiotherapy	thyroid_function	thyroid_function_3cat	thyroid_function_2cat	physical_examination	adenopathy	pathology	focality	risk	t	t_4cat	n	m	stage	stage_4cat	stage_2cat	response	response_3cat	recurrence	kmode_cluster_3cat
age_group	1.00	0.21	0.37	0.19	0.32	0.11	0.11	0.14	0.14	0.14	0.18	0.22	0.28	0.20	0.24	0.14	0.36	0.39	0.41	0.53	0.22	0.25	0.33	0.29
gender	0.21	1.00	0.61	0.16	0.21	0.11	0.08	0.04	0.17	0.32	0.10	0.20	0.30	0.33	0.28	0.25	0.20	0.27	0.27	0.20	0.32	0.30	0.32	0.27
smoking	0.37	0.61	1.00	0.24	0.27	0.07	0.02	0.00	0.19	0.29	0.27	0.23	0.41	0.45	0.43	0.22	0.30	0.46	0.45	0.38	0.34	0.34	0.32	0.33
hx_smoking	0.19	0.16	0.24	1.00	0.22	0.14	0.14	0.11	0.10	0.17	0.13	0.00	0.17	0.27	0.21	0.05	0.10	0.34	0.34	0.29	0.16	0.15	0.12	0.12
hx_radiotherapy	0.32	0.21	0.27	0.22	1.00	0.15	0.10	0.03	0.04	0.42	0.13	0.08	0.38	0.42	0.34	0.11	0.38	0.50	0.50	0.38	0.17	0.17	0.15	0.16
thyroid_function	0.11	0.11	0.07	0.14	0.15	1.00	1.00	1.00	0.24	0.10	0.12	0.10	0.11	0.15	0.11	0.15	0.05	0.17	0.17	0.17	0.10	0.12	0.12	0.13
thyroid_function_3cat	0.11	0.08	0.02	0.14	0.10	1.00	1.00	1.00	0.25	0.10	0.10	0.06	0.05	0.14	0.06	0.12	0.02	0.19	0.17	0.12	0.07	0.07	0.09	0.07
thyroid_function_2cat	0.14	0.04	0.00	0.11	0.03	1.00	1.00	1.00	0.27	0.10	0.12	0.01	0.01	0.15	0.07	0.15	0.00	0.18	0.12	0.08	0.09	0.09	0.07	0.05
physical_examination	0.14	0.17	0.19	0.10	0.04	0.24	0.25	0.27	1.00	0.24	0.13	0.37	0.18	0.17	0.17	0.14	0.15	0.09	0.09	0.12	0.14	0.17	0.18	0.25
adenopathy	0.14	0.32	0.29	0.17	0.42	0.10	0.10	0.10	0.24	1.00	0.16	0.39	0.53	0.31	0.32	0.63	0.28	0.24	0.26	0.28	0.39	0.43	0.64	0.48
pathology	0.18	0.10	0.27	0.13	0.13	0.12	0.10	0.12	0.13	0.16	1.00	0.25	0.20	0.58	0.41	0.22	0.19	0.16	0.14	0.15	0.19	0.22	0.25	0.52
focality	0.22	0.20	0.23	0.00	0.08	0.10	0.06	0.01	0.37	0.39	0.25	1.00	0.46	0.54	0.53	0.38	0.21	0.38	0.38	0.23	0.42	0.41	0.38	0.53
risk	0.28	0.30	0.41	0.17	0.38	0.11	0.05	0.01	0.18	0.53	0.20	0.46	1.00	0.70	0.68	0.53	0.65	0.57	0.57	0.69	0.52	0.50	0.74	0.58
t	0.20	0.33	0.45	0.27	0.42	0.15	0.14	0.15	0.17	0.31	0.58	0.54	0.70	1.00	1.00	0.38	0.50	0.43	0.48	0.68	0.41	0.46	0.61	0.76
t_4cat	0.24	0.28	0.43	0.21	0.34	0.11	0.06	0.07	0.17	0.32	0.41	0.53	0.68	1.00	1.00	0.34	0.47	0.43	0.43	0.65	0.37	0.44	0.57	0.72
n	0.14	0.25	0.22	0.05	0.11	0.15	0.12	0.15	0.14	0.63	0.22	0.38	0.53	0.38	0.34	1.00	0.23	0.26	0.26	0.22	0.47	0.42	0.63	0.47
m	0.36	0.20	0.30	0.10	0.38	0.05	0.02	0.00	0.15	0.28	0.19	0.21	0.65	0.50	0.47	0.23	1.00	0.76	0.68	0.50	0.37	0.34	0.34	0.33
stage	0.39	0.27	0.46	0.34	0.50	0.17	0.19	0.18	0.09	0.24	0.16	0.38	0.57	0.43	0.43	0.26	0.76	1.00	1.00	1.00	0.28	0.32	0.50	0.38
stage_4cat	0.41	0.27	0.45	0.34	0.50	0.17	0.17	0.12	0.09	0.26	0.14	0.38	0.57	0.48	0.43	0.26	0.68	1.00	1.00	1.00	0.28	0.32	0.50	0.38
stage_2cat	0.53	0.20	0.38	0.29	0.38	0.17	0.12	0.08	0.12	0.28	0.15	0.23	0.69	0.68	0.65	0.22	0.50	1.00	1.00	1.00	0.35	0.34	0.34	0.33
response	0.22	0.32	0.34	0.16	0.17	0.10	0.07	0.09	0.14	0.39	0.19	0.42	0.52	0.41	0.37	0.47	0.37	0.28	0.28	0.35	1.00	1.00	0.90	0.57
response_3cat	0.25	0.30	0.34	0.15	0.17	0.12	0.07	0.09	0.17	0.43	0.22	0.41	0.50	0.46	0.44	0.42	0.34	0.32	0.32	0.34	1.00	1.00	0.87	0.56
recurrence	0.33	0.32	0.32	0.12	0.15	0.12	0.09	0.07	0.18	0.64	0.25	0.38	0.74	0.61	0.57	0.63	0.34	0.50	0.50	0.34	0.90	0.87	1.00	0.81
kmode_cluster_3cat	0.29	0.27	0.33	0.12	0.16	0.13	0.07	0.05	0.25	0.48	0.52	0.53	0.58	0.76	0.72	0.47	0.33	0.38	0.38	0.33	0.57	0.56	0.81	1.00

Exploratory Data Analysis (EDA)



Exploratory Data Analysis (EDA)



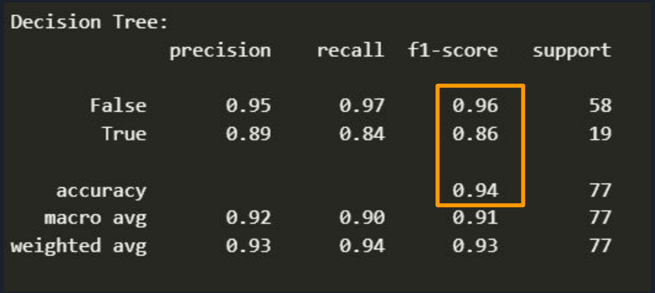
Feature Engineering

Encoding

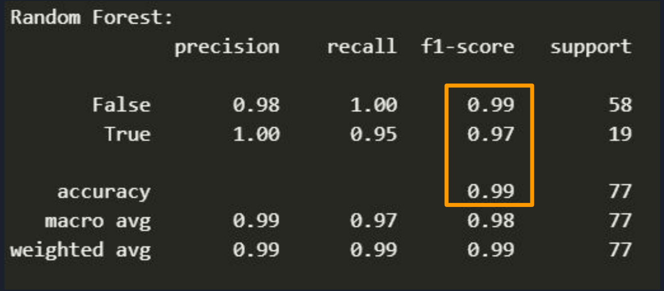
- Dummy encode binary
- One hot encode categories
- Label encode categories

Modeling

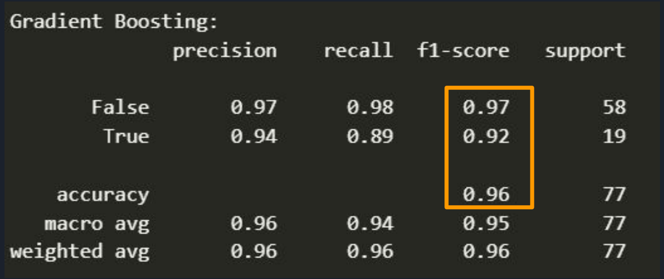
Native with imbalance data:



```
Decision Tree:
              precision    recall  f1-score   support

       False       0.95      0.97      0.96        58
        True       0.89      0.84      0.86        19

    accuracy                           0.94        77
   macro avg       0.92      0.90      0.91        77
weighted avg       0.93      0.94      0.93        77
```

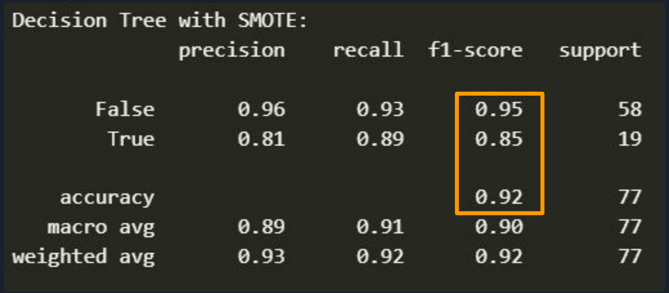


```
Random Forest:
              precision    recall  f1-score   support

       False       0.98      1.00      0.99        58
        True       1.00      0.95      0.97        19

    accuracy                           0.99        77
   macro avg       0.99      0.97      0.98        77
weighted avg       0.99      0.99      0.99        77
```

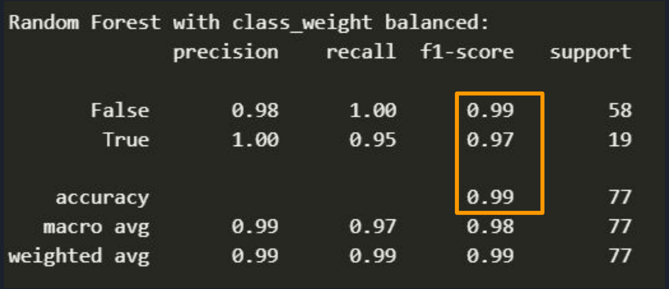


```
Gradient Boosting:
              precision    recall  f1-score   support

       False       0.97      0.98      0.97        58
        True       0.94      0.89      0.92        19

    accuracy                           0.96        77
   macro avg       0.96      0.94      0.95        77
weighted avg       0.96      0.96      0.96        77
```

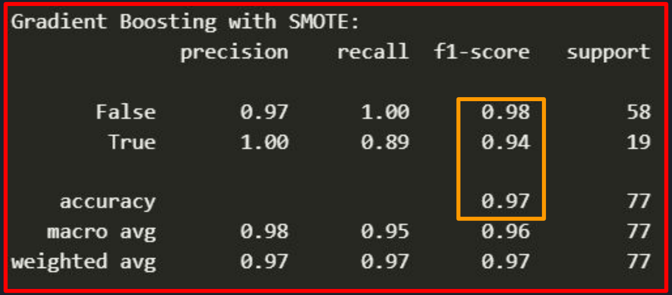
Balanced Data Modifications:



```
Decision Tree with SMOTE:
              precision    recall  f1-score   support

       False       0.96      0.93      0.95        58
        True       0.81      0.89      0.85        19

    accuracy                           0.92        77
   macro avg       0.89      0.91      0.90        77
weighted avg       0.93      0.92      0.92        77
```



```
Random Forest with class_weight balanced:
              precision    recall  f1-score   support

       False       0.98      1.00      0.99        58
        True       1.00      0.95      0.97        19

    accuracy                           0.99        77
   macro avg       0.99      0.97      0.98        77
weighted avg       0.99      0.99      0.99        77
```



```
Gradient Boosting with SMOTE:
              precision    recall  f1-score   support

       False       0.97      1.00      0.98        58
        True       1.00      0.89      0.94        19

    accuracy                           0.97        77
   macro avg       0.98      0.95      0.96        77
weighted avg       0.97      0.97      0.97        77
```

For those that need a quick read, focus on:

F1 Scores and Accuracy

The higher the better

Conclusion, Future Work, & Improving the model

Conclusion:

- Chose the Gradient Booster with SMOTE model

Future Works:

- Partner with hospitals to get more consented user data.
- Explore more if overfitting.
- Explore the Random Forest model more.
- Expand the model to include cross reactions.

Possible issues:

- CC 4.0 Attribution license

Thank You

Questions?