Predicting Recurrence of 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

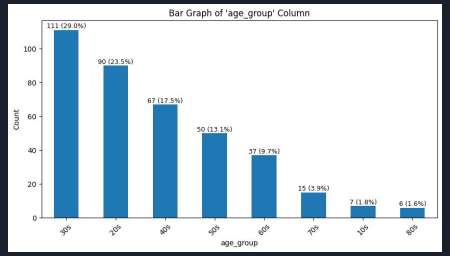
Data Wrangling

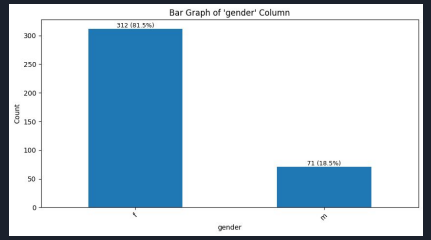
0	Age
1	Gender
2	Smoking
3	Hx Smoking
4	Hx Radiothreapy
5	Thyroid Function
6	Physical Examination
7	Adenopathy
8	Pathology
9	Focality
10	Risk
11	T
12	N
13	M
14	Stage
15	Response
16	Recurred

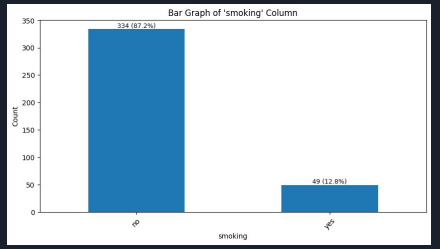
0	age
1	gender
2	smoking
3	hx_smoking
4	hx_radiotherapy
5	thyroid_function
6	physical_examination
7	adenopathy
8	pathology
9	focality
10	risk
11	t
12	n
13	m
14	stage
15	response
16	recurrence

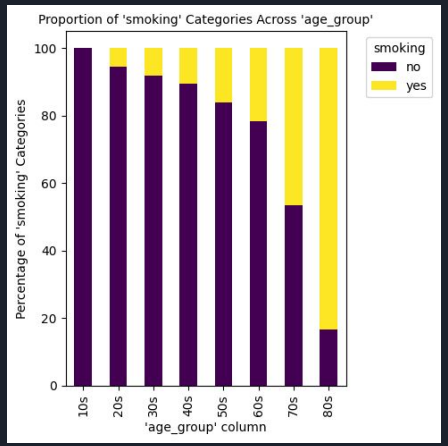
0	age
1	age_group
2	gender
3	smoking
4	hx_smoking
5	hx_radiotherapy
6	thyroid_function
7	thyroid_function_3cat
8	thyroid_function_2cat
9	physical_examination
10	adenopathy
11	pathology
12	focality
13	risk
14	t
15	t_4cat
16	n
17	m
18	stage
19	stage_4cat
20	stage_2cat
21	response
22	response_3cat
23	recurrence

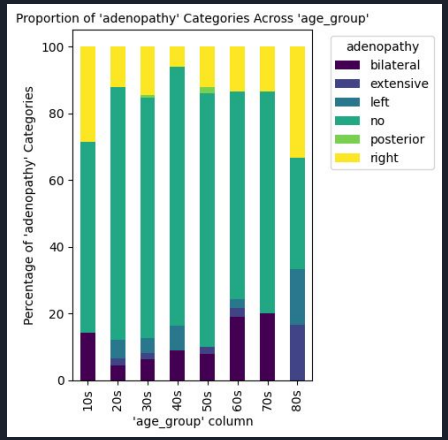
Exploratory Data Analysis (EDA)

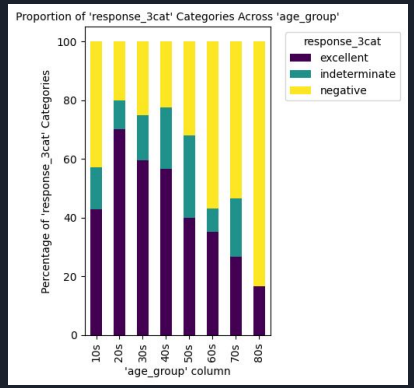




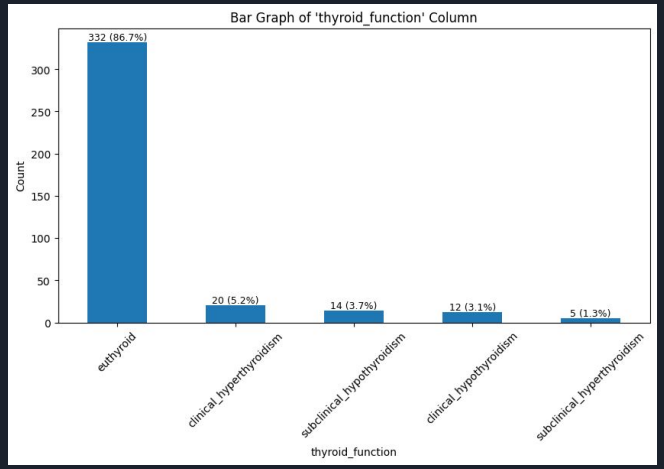


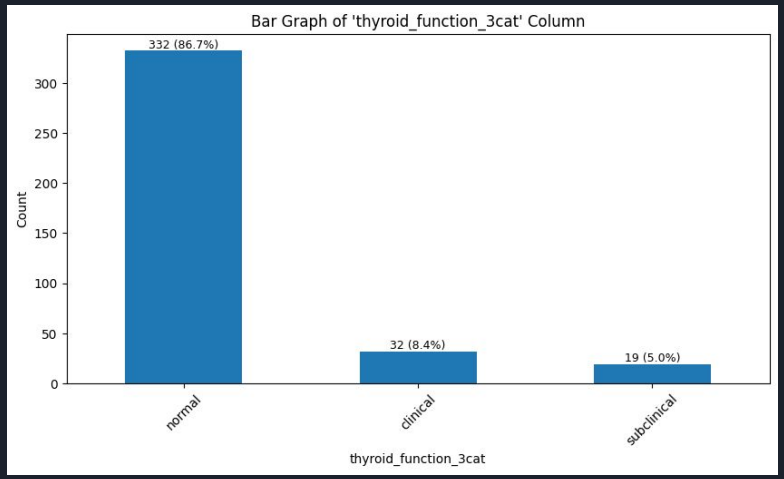


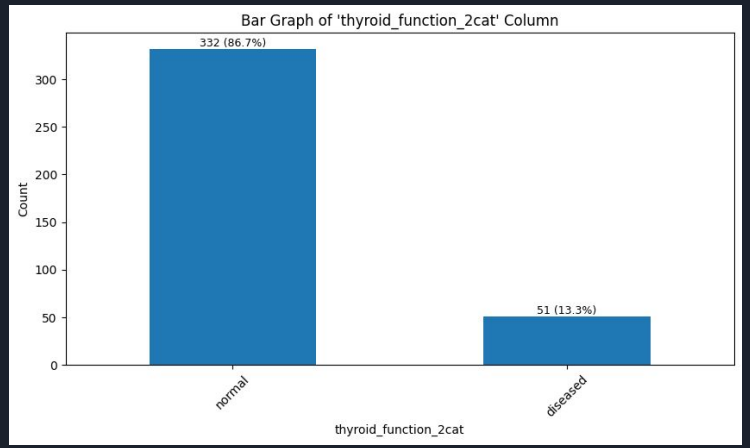




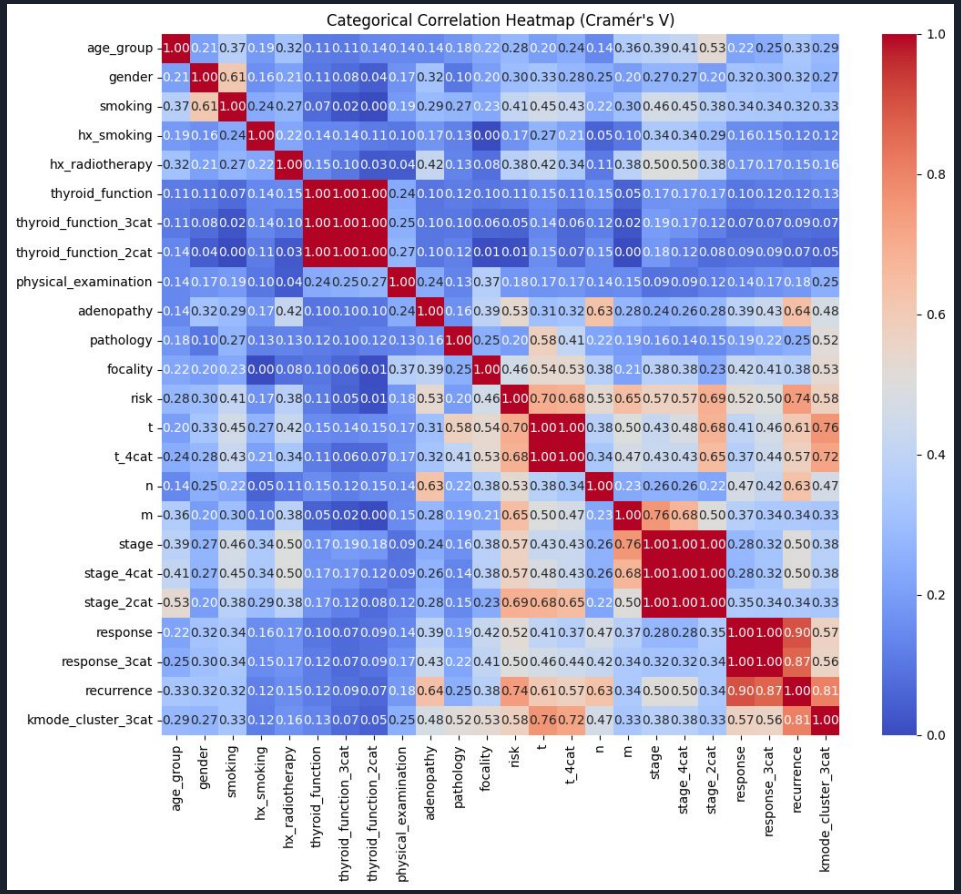
Exploratory Data Analysis (EDA)



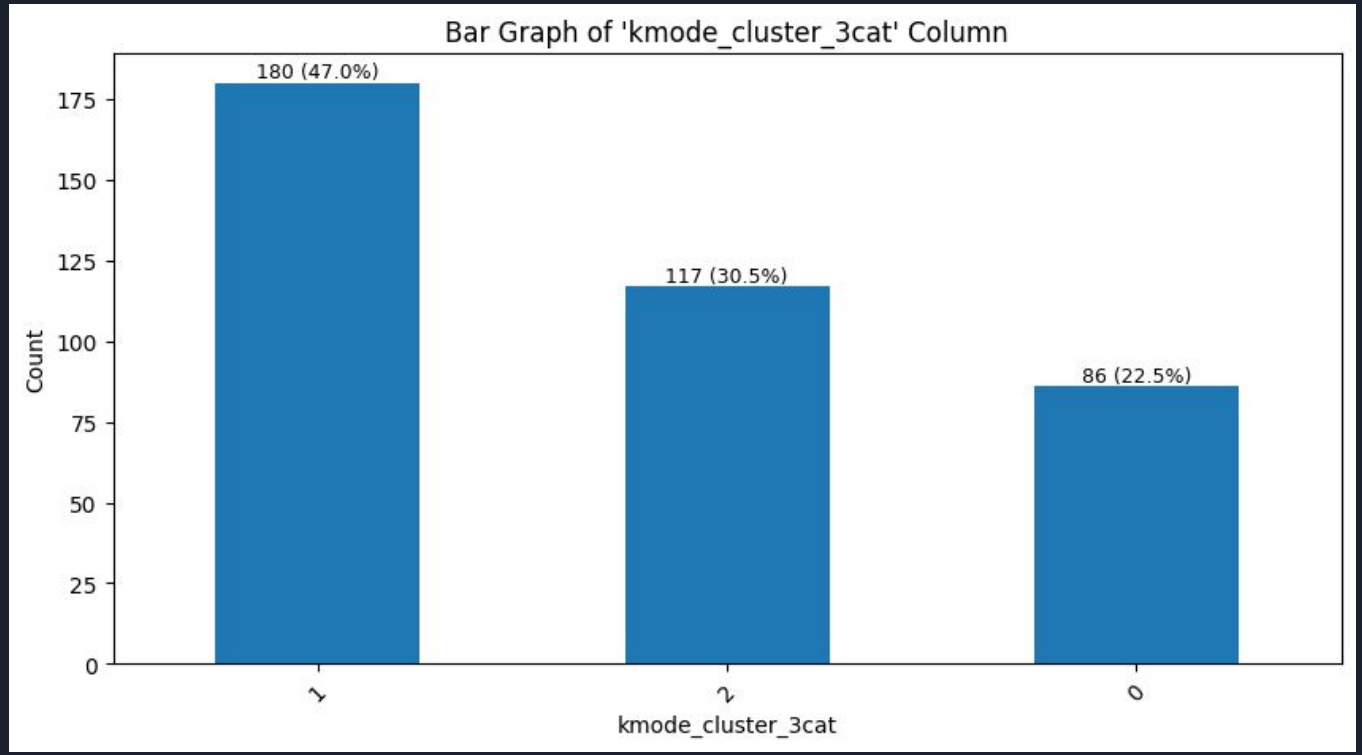




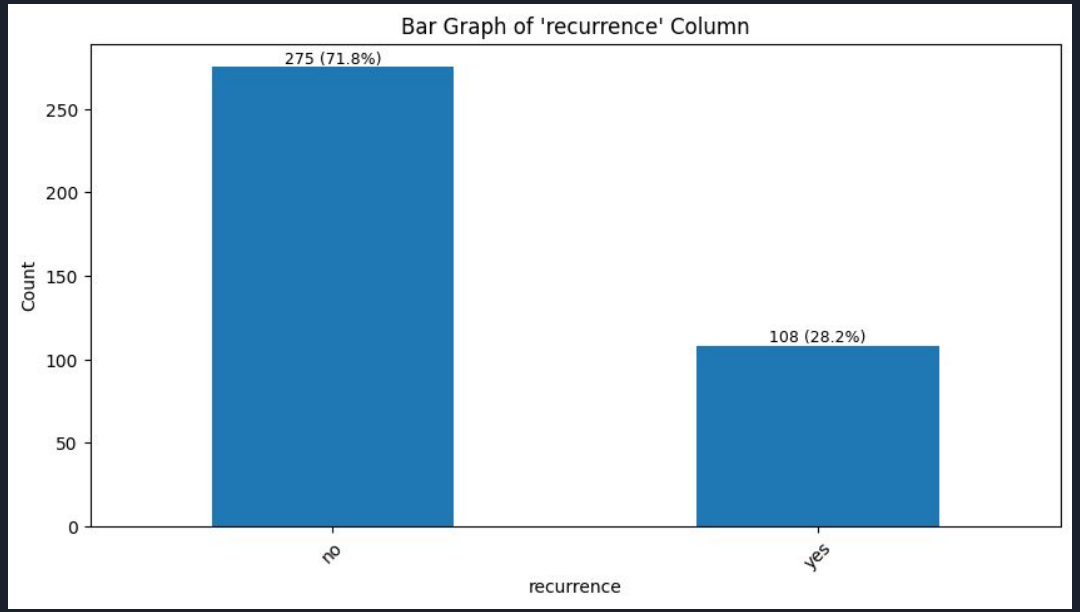
Exploratory Data Analysis (EDA)



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?