TRAINING MACHINE LEARNING MODELS TO PREDICT THE LIKELIHOOD OF RECURRENCE OF THYROID CANCER

DATALAB PYTHON PROJECT 2025

Written and Presented By
Bisong Enow Njang-Ebob Rene

TABLE OF CONTENT

- Summary of the project
- Definition of terms
- Visualization of the dataset
- Comparison of the different machine models
- Report and Results
- Conclusion

Summary

“Thyroid cancer is a type of cancer that affects the thyroid gland. The thyroid gland produces hormones that regulate, metabolism, growth and development. Building machine learning models that can quickly predict the presence of this cancer will help in quick administration of treatment and follow up of patients with the cancer.”

Definition of some terms

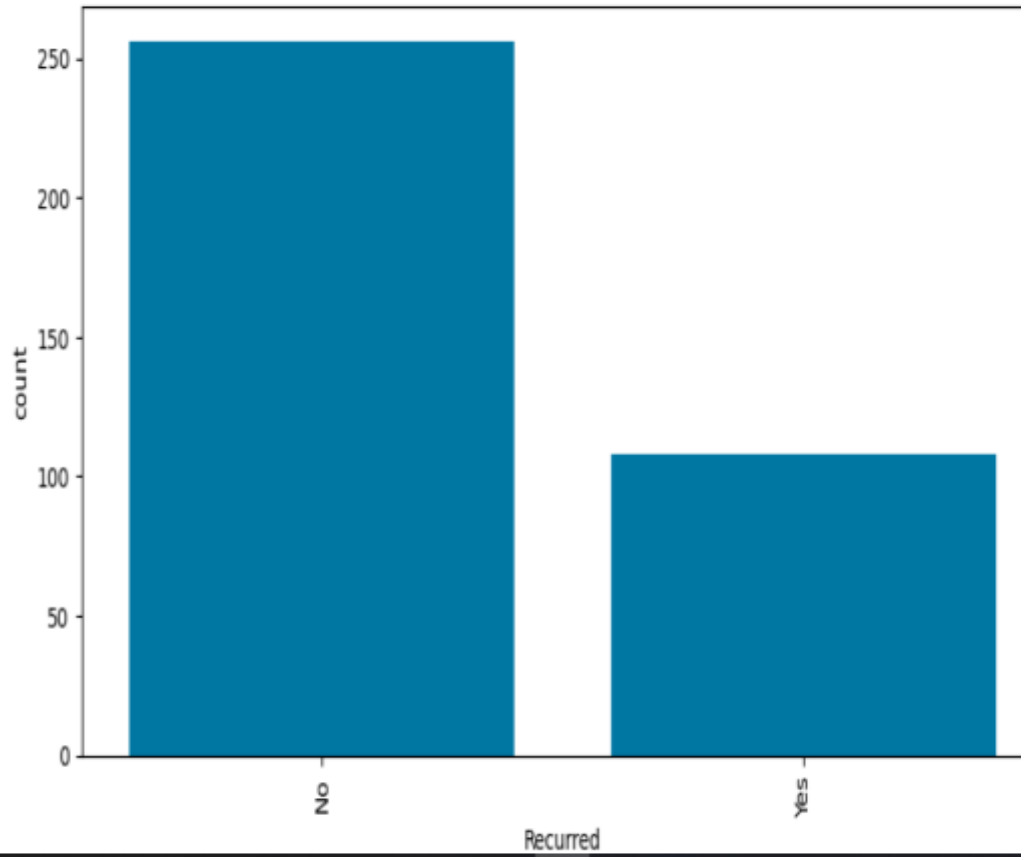
- Thyroid cancer: Thyroid cancer is a type of cancer that affects the thyroid gland.
- Adenopathy : This is a disease that cause enlargement of the lymph nodes.
- Pathology: This is the study of disease

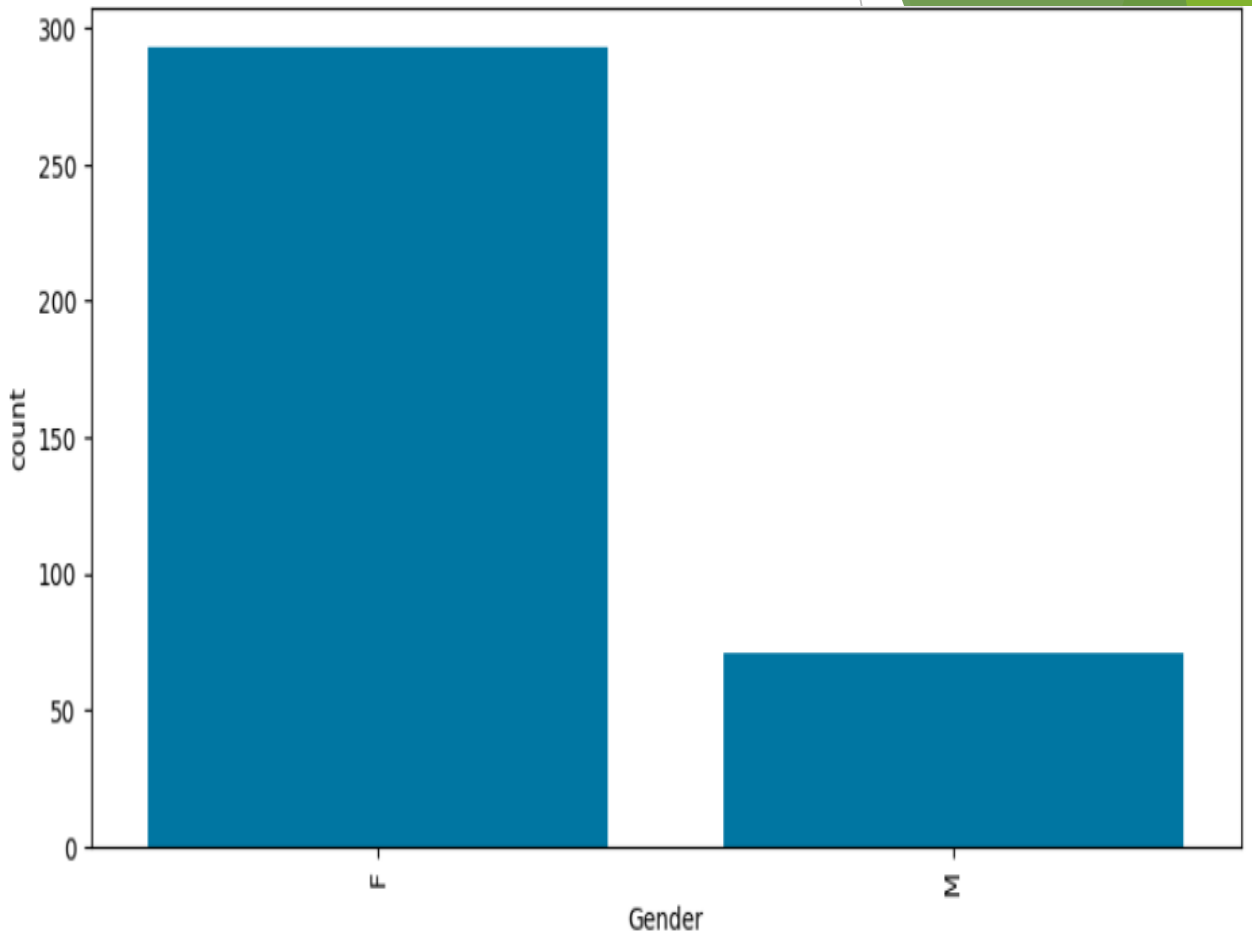
Looking at the dataset

OBSERVATIONS:

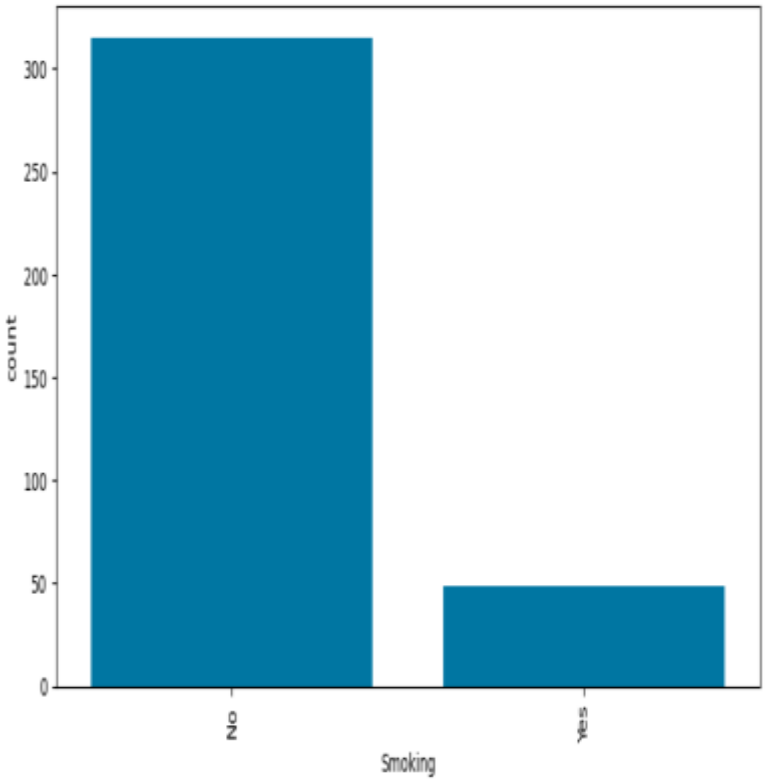
- The shape of the dataset is 383, 17. implying that the dataset has 17 columns and 383 rows.
- Looking at the info, the dataset has 16 categorical columns and I numerical column.
- Stage is in roman numerals and must be changed to Arabic or integer.
- Statistical summary shows that the maximum age is ~82 years, minimum age ~15years, the mean age ~45
- No nulls in the data, 19 duplicated values.

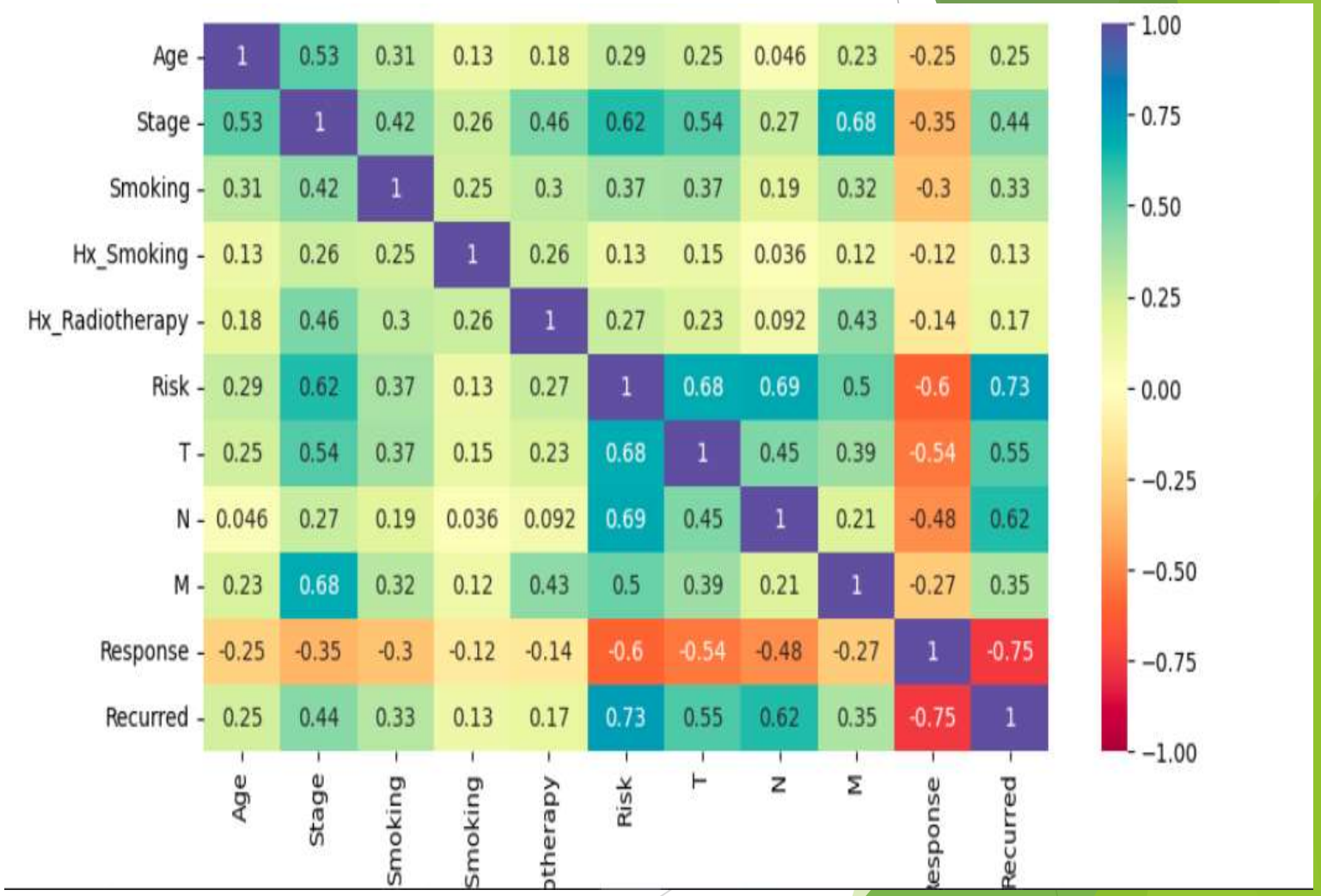
Visualization of the dataset





CONTINUE….


300
250
200
150
100
50
0
count
No
Yes
Smoking


Age 1 0.53 0.31 0.13 0.18 0.29 0.25 0.046 0.23 -0.25 0.25
Stage 0.53 1 0.42 0.26 0.46 0.62 0.54 0.27 0.68 -0.35 0.44
Smoking 0.31 0.42 1 0.25 0.3 0.37 0.37 0.19 0.32 -0.3 0.33
Hx_Smoking 0.13 0.26 0.25 1 0.26 0.13 0.15 0.036 0.12 -0.12 0.13
Hx_Radiotherapy 0.18 0.46 0.3 0.26 1 0.27 0.23 0.092 0.43 -0.14 0.17
Risk 0.29 0.62 0.37 0.13 0.27 1 0.68 0.69 0.5 -0.6 0.73
T 0.25 0.54 0.37 0.15 0.23 0.68 1 0.45 0.39 -0.54 0.55
N 0.046 0.27 0.19 0.036 0.092 0.69 0.45 1 0.21 -0.48 0.62
M 0.23 0.68 0.32 0.12 0.43 0.5 0.39 0.21 1 -0.27 0.35
Response -0.25 -0.35 -0.3 -0.12 -0.14 -0.6 -0.54 -0.48 -0.27 1 -0.75
Recurred 0.25 0.44 0.33 0.13 0.17 0.73 0.55 0.62 0.35 -0.75 1
Age
Stage
Smoking
Smoking
Risk
T
N
M
Recurred
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Machine learning models

model	accuracy
Logistic Regression	0.90
Naïve Bayes	0.93
KNN	0.96
RandomForest	0.96
Decision Tree	0.95
Gradient boost	0.95
Adaboost	0.93
Xgboost	0.93

Hyperparameter tuning

- **Adaboost model**

accuracy	recall	precision	F1 SCORE
0.94	0.93	0.88	0.91

Gradient Boost classifier

accuracy	recall	precision	F1 SCORE
0.97	0.92	0.97	0.95

OBSERVATIONS

- After looking at the most important features, it was observed that the most important feature is RESPONSE followed by Risk.

Conclusion

After tuning the models, the best model is Adaboost with an accuracy of 0.94 followed by Naïve Bayes with an accuracy of 0.93 as the other models can be over fitted.

RECOMMENDATION:

- Consider the risk, stage and response as they are the most important features that are related to recurred of thyroid cancer.
- More information is needed to investigate the reoccurrence of thyroid cancer in the patients.

Thank you