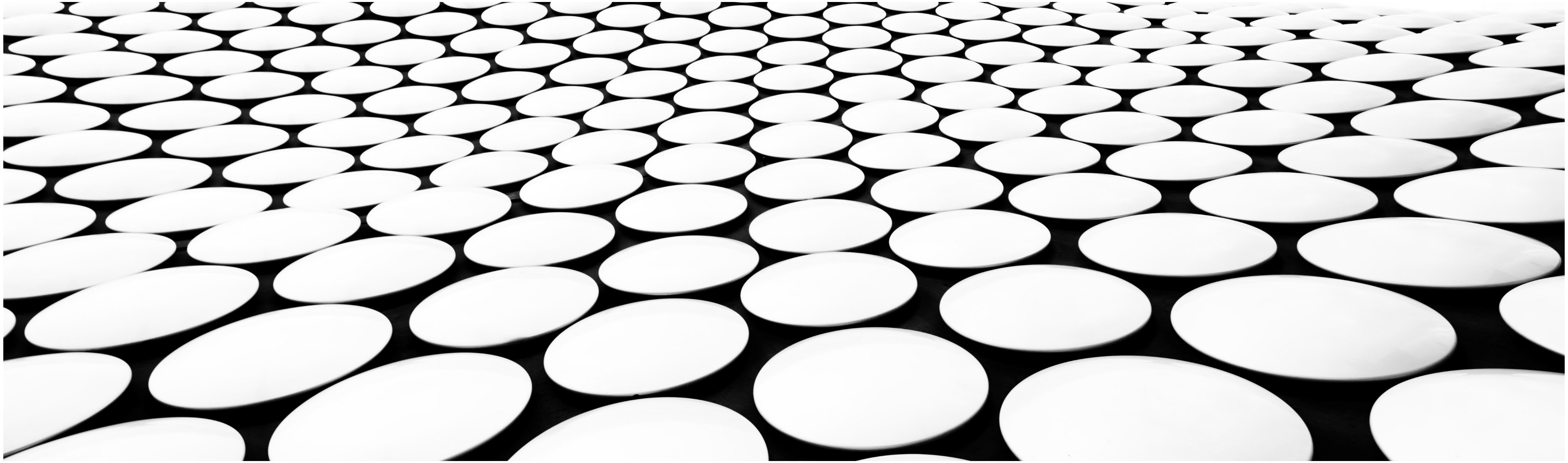
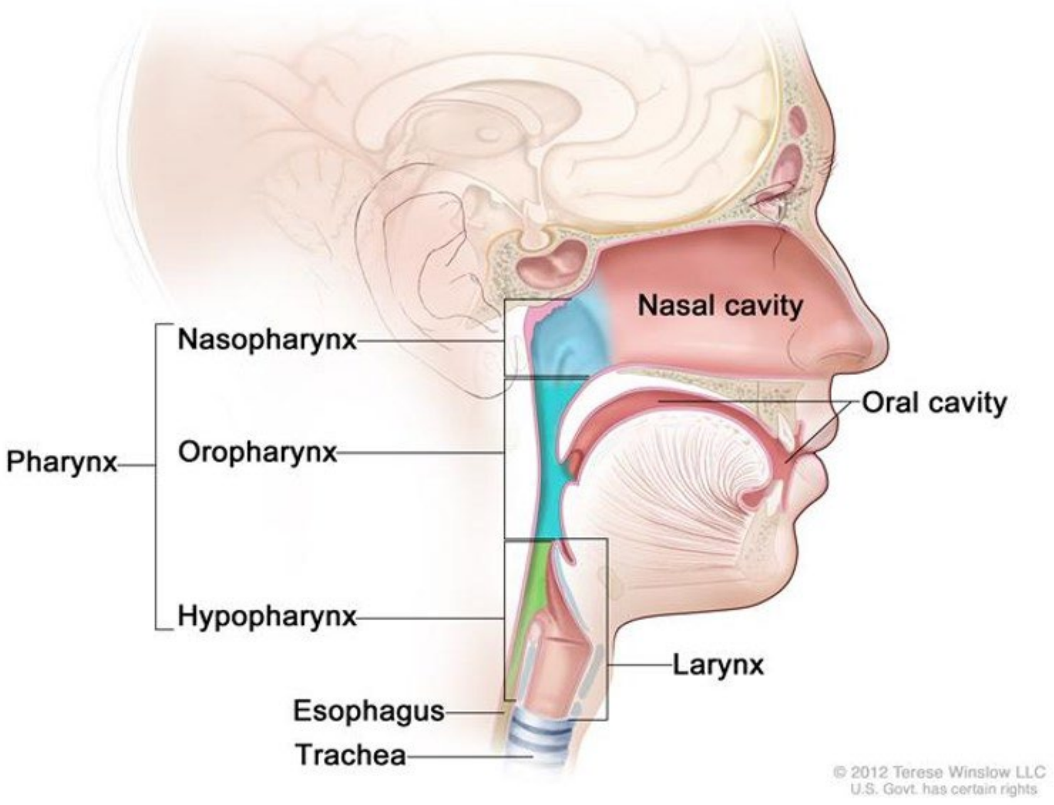
USING A MACHINE LEARNING STRATEGY TO PREDICT CANCER PATIENT OUTCOME BASED ON CLINICAL AND MOLECULAR FEATURES

TOM BELBIN



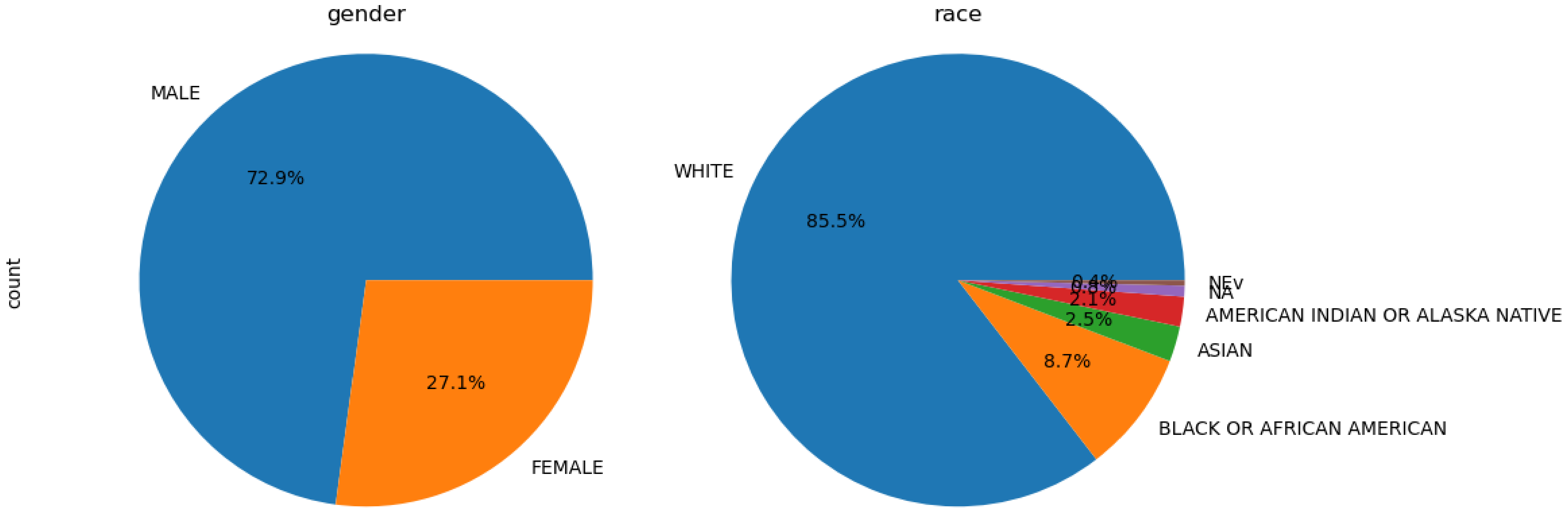
HEAD AND NECK CANCER



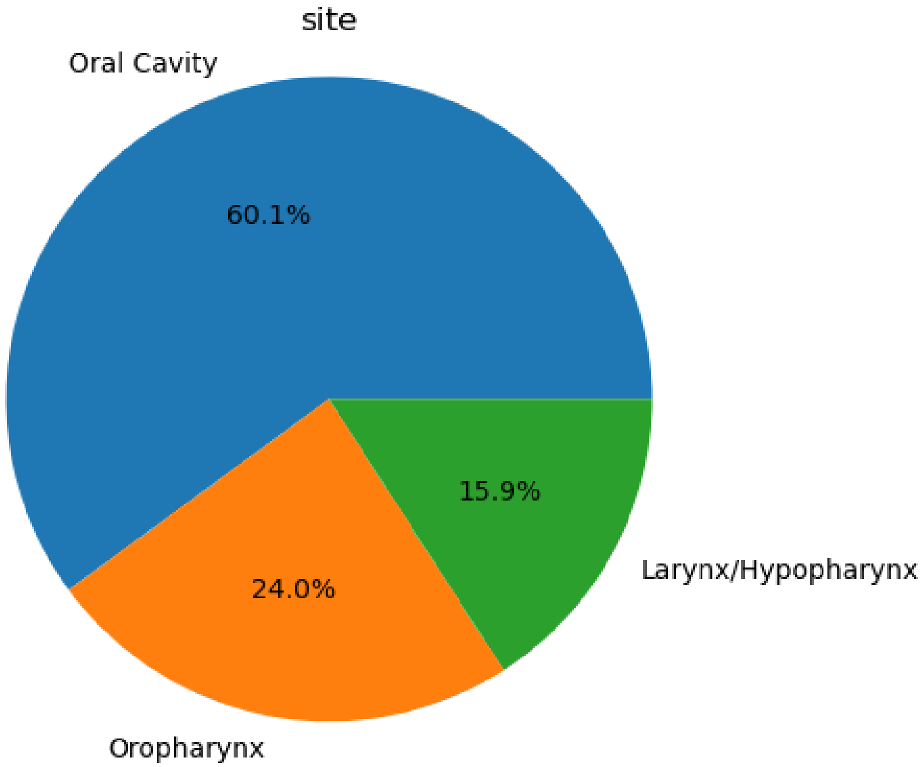
© 2012 Terese Winslow LLC
U.S. Govt. has certain rights

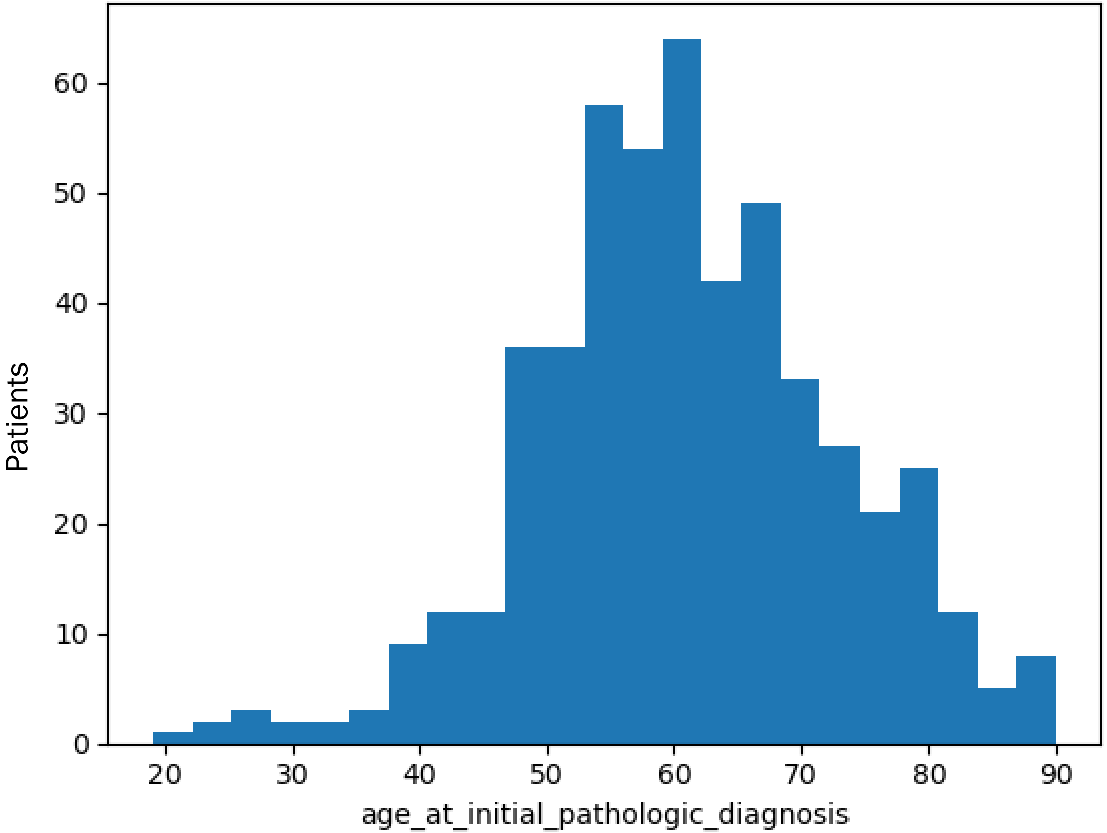
- Tobacco use, alcohol consumption, HPV infection
- Surgery, radiation therapy, chemotherapy
- Speech, swallowing
- Recurrence
- 5-yr survival
- Biomarkers?

EXPLORATORY ANALYSIS OF THE TCGA HNSCC COHORT

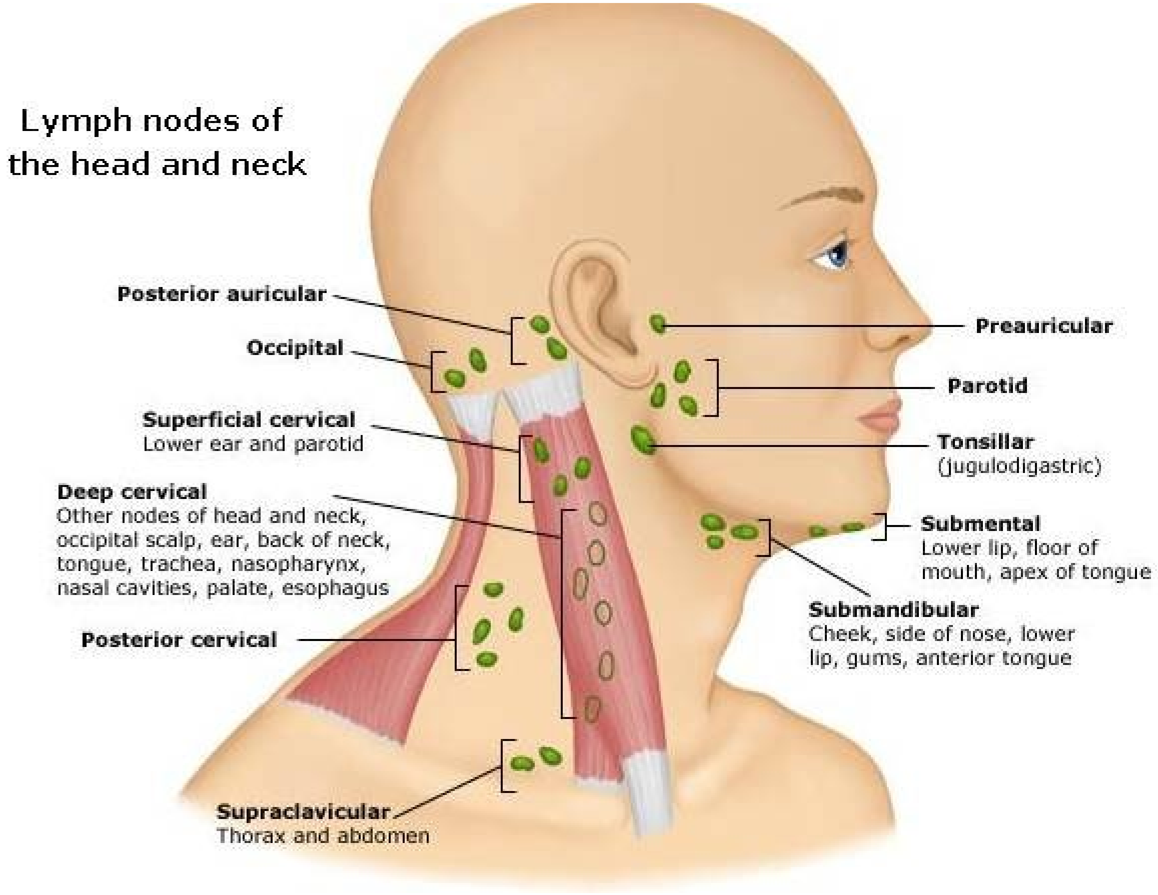


EXPLORATORY ANALYSIS OF THE TCGA HNSCC COHORT





EXPLORATORY ANALYSIS OF THE TCGA HNSCC COHORT


Lymph nodes of the head and neck
Posterior auricular
Occipital
Superficial cervical
Lower ear and parotid
Deep cervical
Other nodes of head and neck, occipital scalp, ear, back of neck, tongue, trachea, nasopharynx, nasal cavities, palate, esophagus
Posterior cervical
Supraclavicular
Thorax and abdomen
Preauricular
Parotid
Tonsillar
(jugulodigastric)
Submental
Lower lip, floor of mouth, apex of tongue
Submandibular
Cheek, side of nose, lower lip, gums, anterior tongue

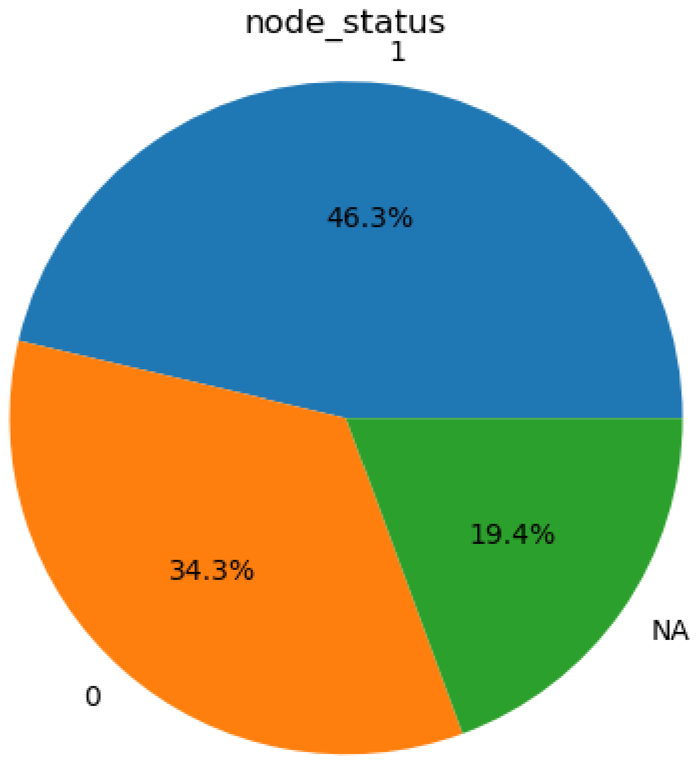
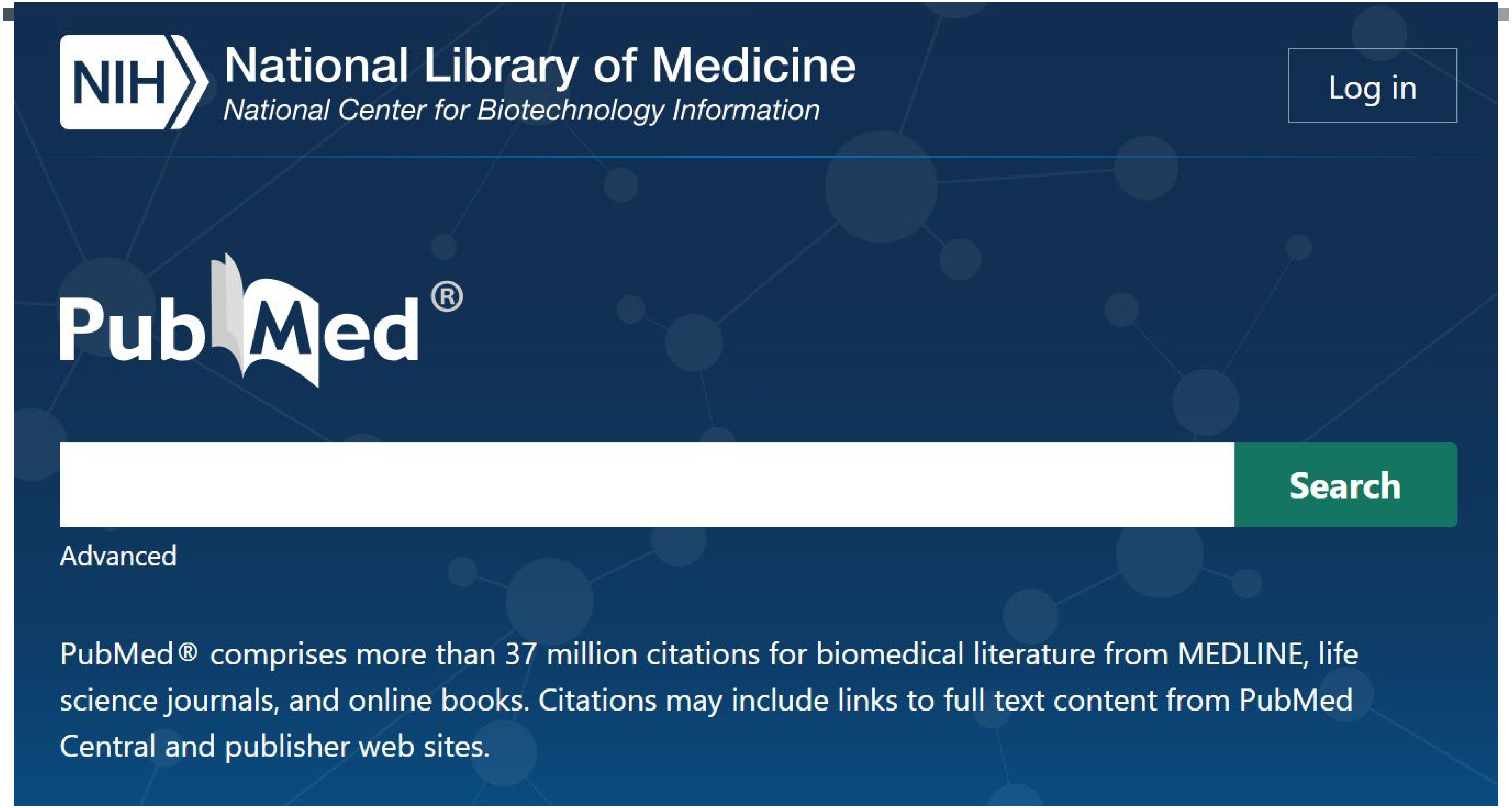


Table 1. HNSCC patient characteristics by tumor site

	Oral Cavity		Oropharynx		Larynx	
	N=310		N=82		N=124	
	N	%	N	%	N	%
Gender						
Male	206	66.5	69	84.1	101	81.5
Female	104	33.5	13	15.9	23	18.5
Race						
White	266	85.8	76	92.7	99	79.8
Black or African American	20	6.5	6	7.3	19	15.3
Asian	10	3.2	0	0	1	0.8
American Indian or Alaska Native	1	0.3	0	0	1	0.8
Information not available	13	4.2	0	0	4	3.2
Ethnicity						
Hispanic/Latino	16	5.2	3	3.7	5	4
Non-Hispanic/Latino	269	86.8	76	92.7	109	87.9
Information not available	25	8.1	3	3.7	10	8.1
Smoking						
Ever smoker	215	69.4	56	68.3	113	91.1
Lifelong non-smoker	86	27.7	25	30.5	8	6.5
Information not available	9	2.9	1	1.2	3	2.4
HPV Status						
HPV +	31	10	56	68.3	11	8.9
HPV -	278	89.7	26	31.7	112	90.3
Indeterminate	1	0.3	0	0	1	0.8
Vital Status						
Alive	197	63.5	69	84.1	82	66.1
Deceased	113	36.5	13	15.9	42	33.9
Nodal Status						
Positive	147	47.4	34	41.5	58	46.8
Negative	119	38.4	16	19.5	42	33.9
Information not available	44	14.2	32	39	24	19.4
Pathologic Tumor Stage						
Stage I	19	6.1	4	4.9	2	1.6
Stage II	56	18.1	10	12.2	12	9.7
Stage III	52	16.8	9	11.0	14	11.3
Stage IV	160	51.6	28	34.1	79	63.7
Information not available	23	7.4	31	37.8	17	13.7



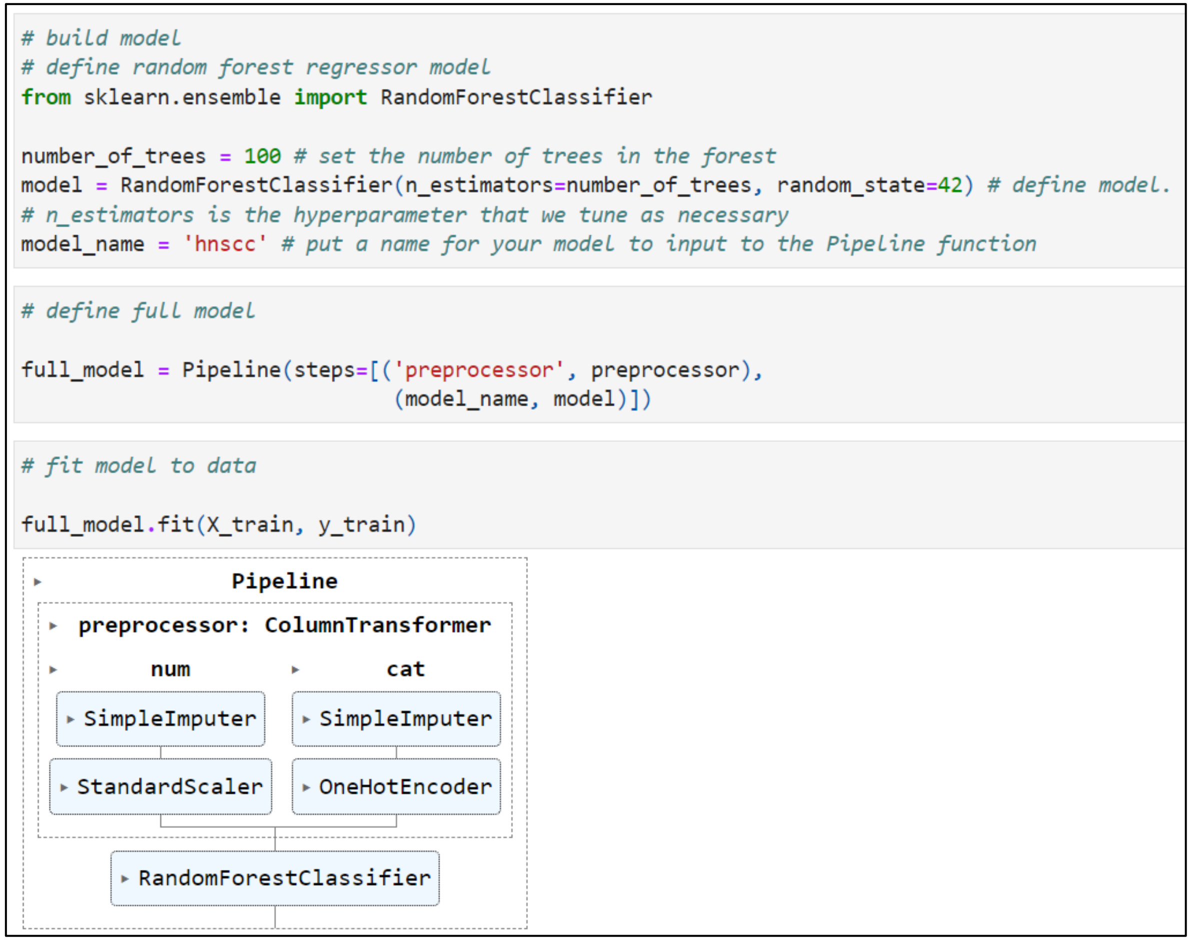
National Library of Medicine

National Center for Biotechnology Information

PubMed®

PubMed® comprises more than 37 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites.

RANDOM FOREST CLASSIFIER FOR NODAL STATUS PREDICTION



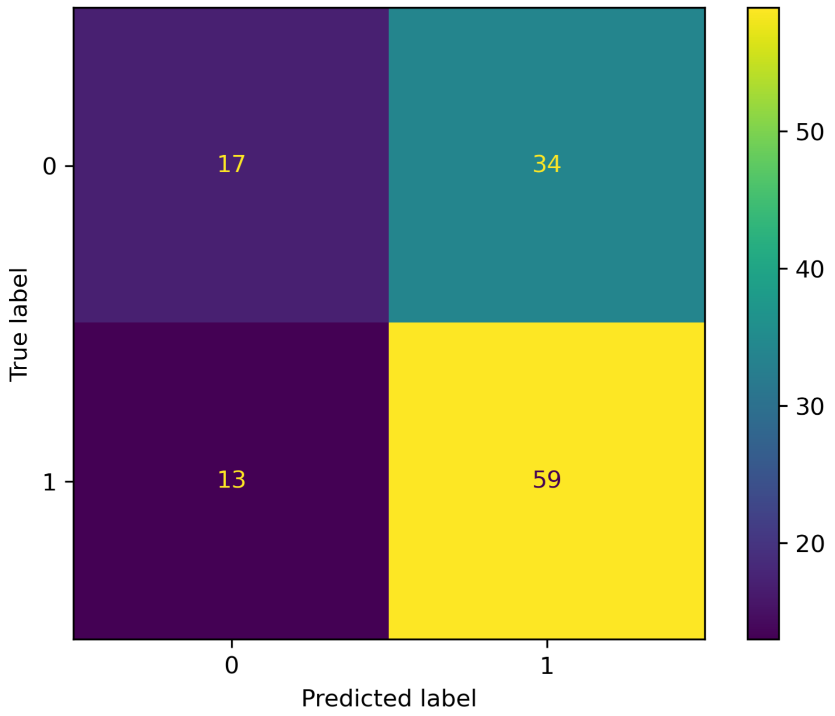
```
# build model
# define random forest regressor model
from sklearn.ensemble import RandomForestClassifier

number_of_trees = 100 # set the number of trees in the forest
model = RandomForestClassifier(n_estimators=number_of_trees, random_state=42) # define model.
# n_estimators is the hyperparameter that we tune as necessary
model_name = 'hnscc' # put a name for your model to input to the Pipeline function
```

```
# define full model

full_model = Pipeline(steps=[('preprocessor', preprocessor),
                             (model_name, model)])
```

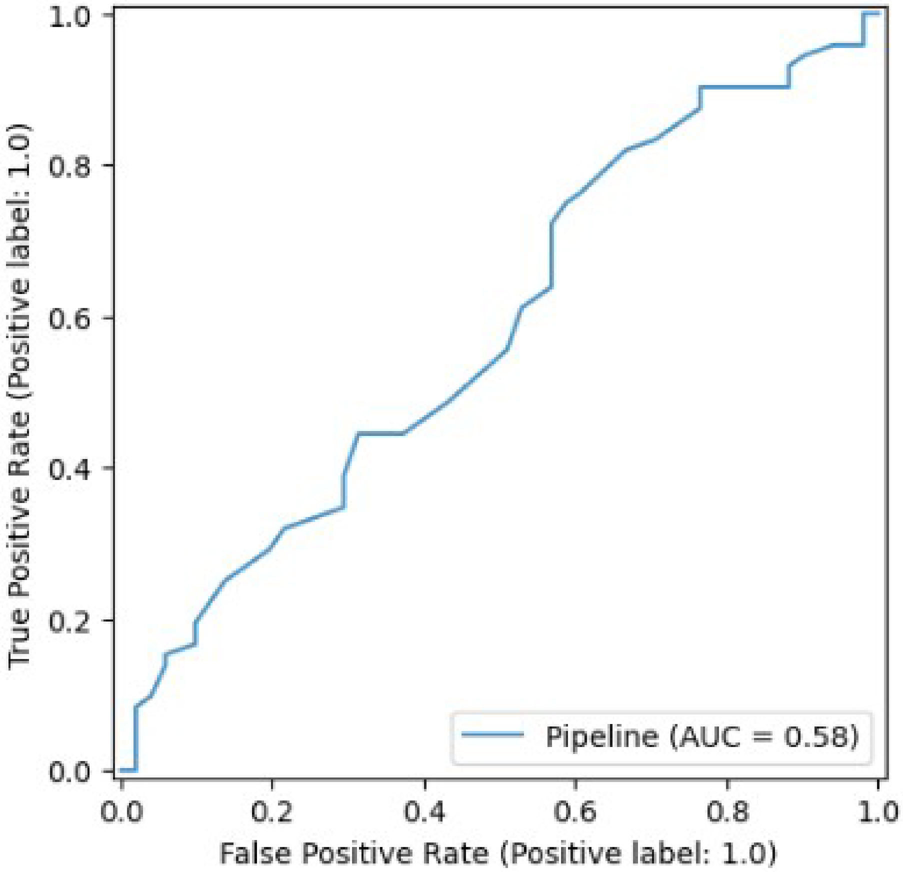
```
# fit model to data

full_model.fit(X_train, y_train)
```

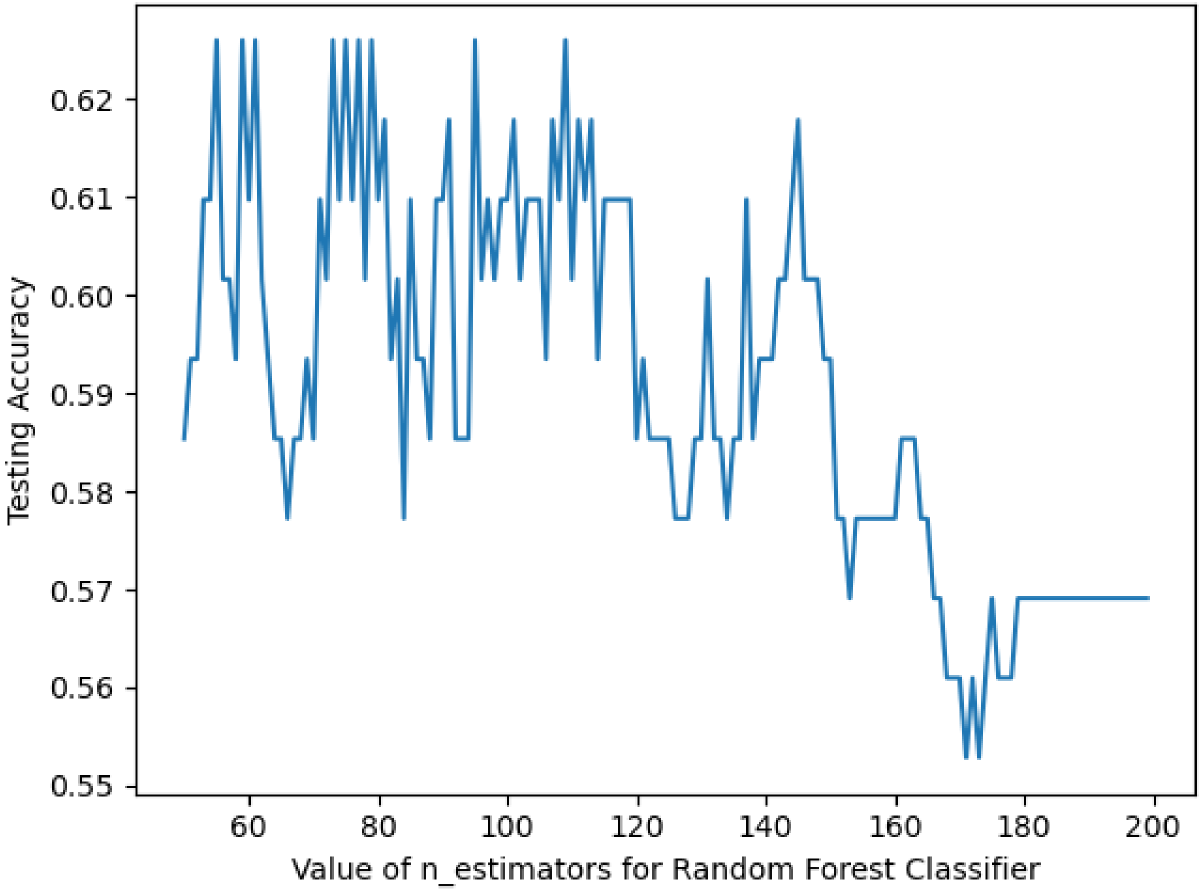
```
Pipeline
  preprocessor: ColumnTransformer
    num                 cat
    SimpleImputer       SimpleImputer
    StandardScaler      OneHotEncoder
  RandomForestClassifier
```

Accuracy: 0.62
Precision: 0.63
Recall: 0.82

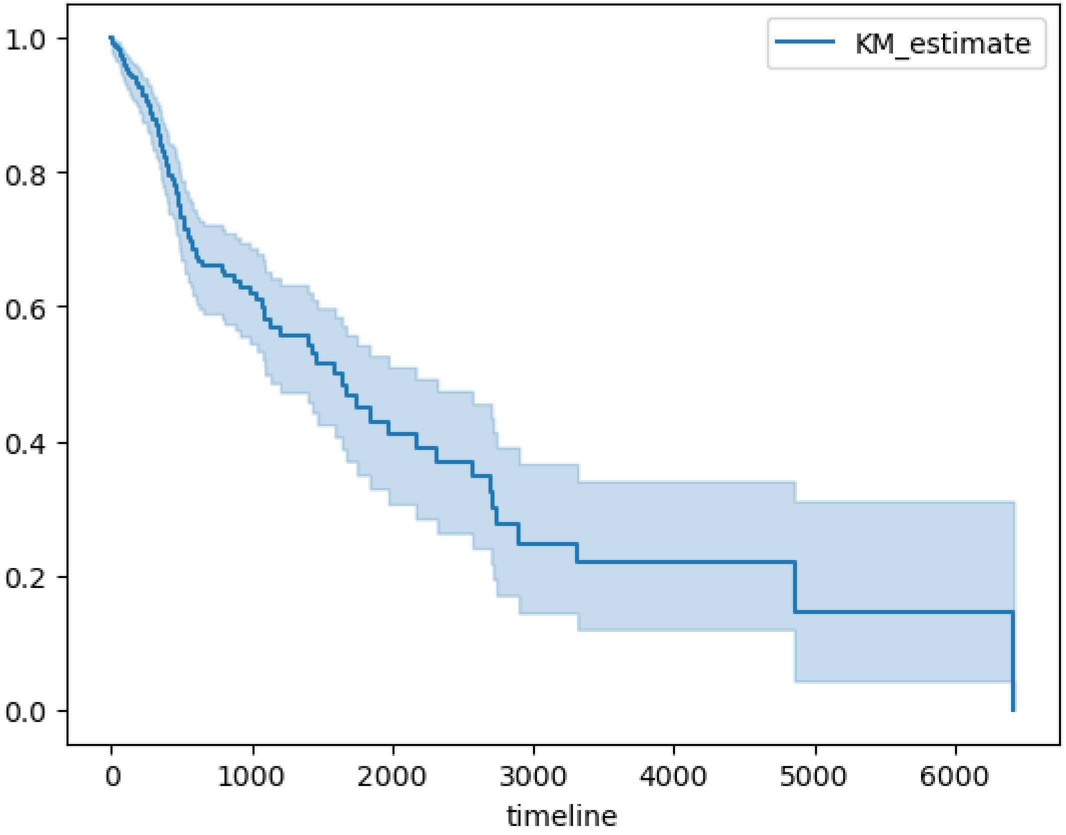


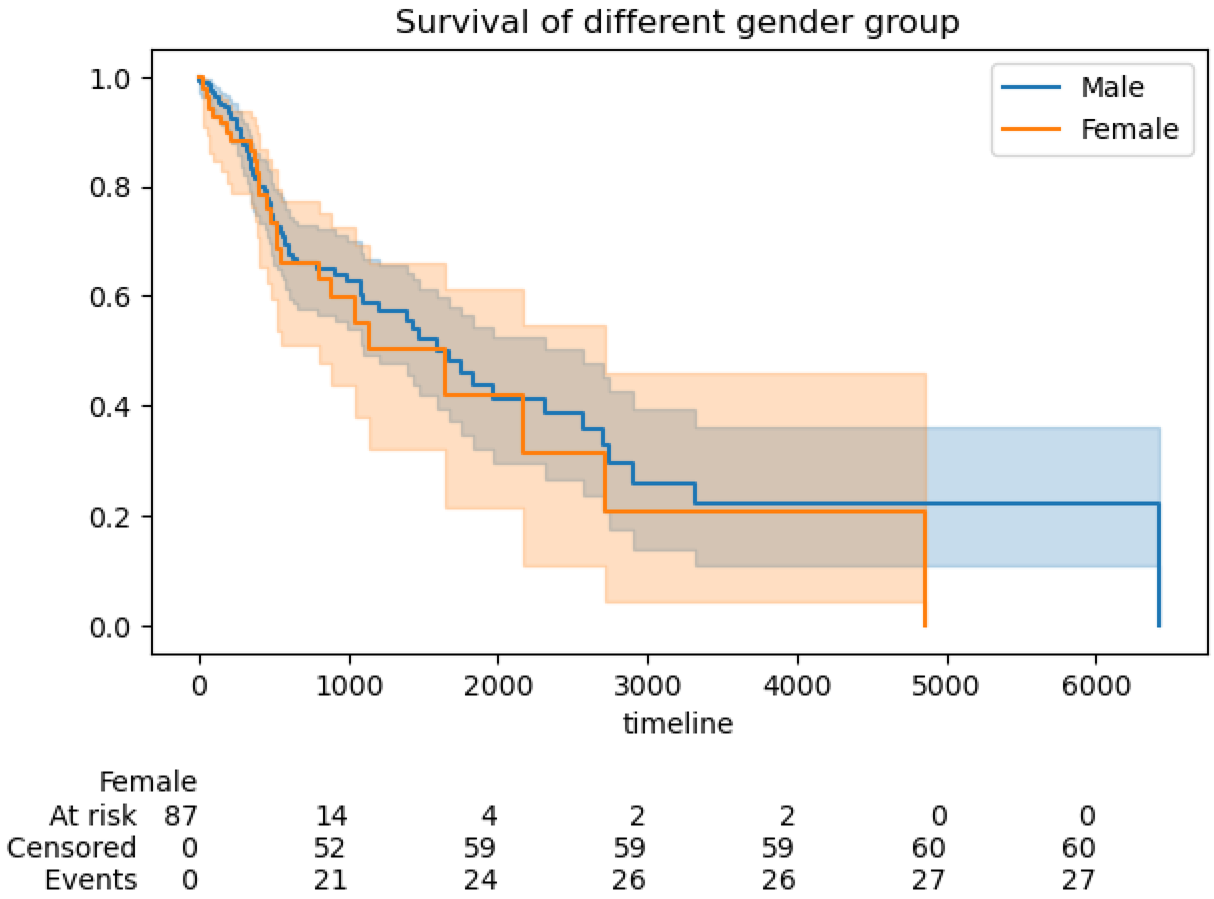
EXPANDING THE NUMBER OF TREES DID NOT IMPROVE PERFORMANCE OF THE NODAL METASTASIS CLASSIFIER





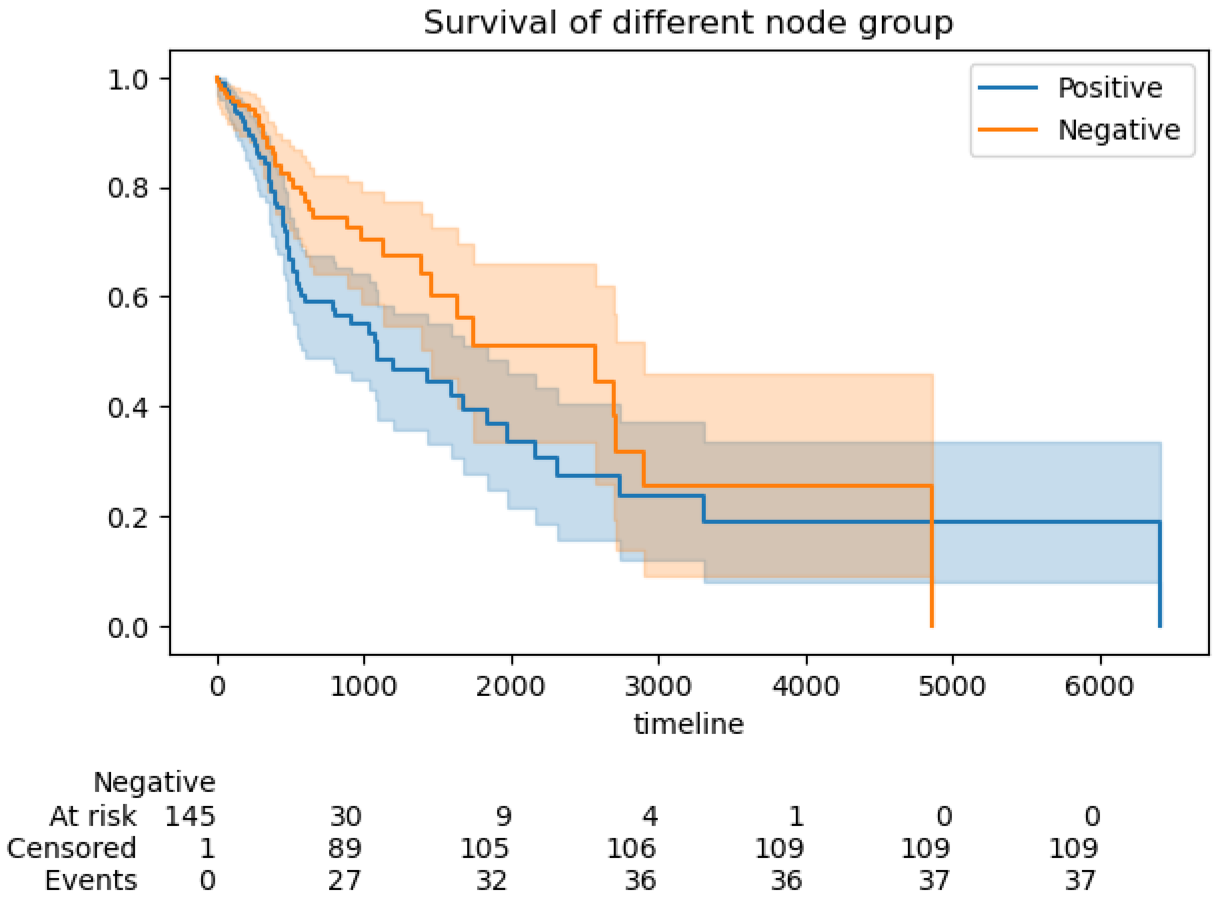
WHAT ABOUT PATIENT SURVIVAL?





Female							
At risk	87	14	4	2	2	0	0
Censored	0	52	59	59	59	60	60
Events	0	21	24	26	26	27	27

NODAL STATUS ARE SIGNIFICANTLY ASSOCIATED WITH POOR OUTCOME



Survival of different node group

Negative

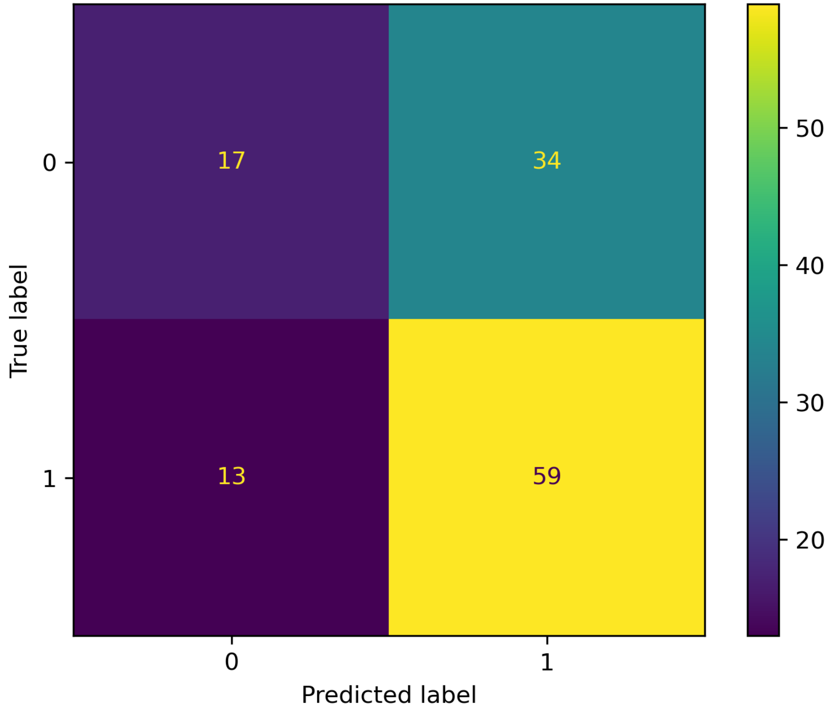
At risk	145	30	9	4	1	0	0
Censored	1	89	105	106	109	109	109
Events	0	27	32	36	36	37	37

- Nodal metastasis
- Age
- History of other malignancy

FUTURE DIRECTIONS

- Re-try RF classifier with new features (much high feature set, other molecular measurements, etc.)
- Repeat survival analysis with many more features
- Is node-negative REALLY node-negative? Is RF classifier calling them node-positive for a reason?
- Try other ML approaches

Accuracy:	**0.62**
Precision:	**0.63**
Recall:	**0.82**


True label
Predicted label