PROJECT TEAL

Jenny Fish

HELLO!

We are Project Teal

MEET THE PROJECT TEAL TEAM

Jenny Fish

(1) Big Picture

(2) Background

Isha Angadi

(3) Research

Adam Claudy

(4) Model

Samiha Khan

(5) Experiments

Sandeep Pvn

(6) Code

(8) Future Work

Mehar Chaturvedi

(7) Results

1

WHAT’S THE BIG PICTURE?

Let’s start with the high level overview

PROJECT OVERVIEW

- The **goal** is to build a model that accurately **predicts malignant or benign tumors**.
- Our **Base Model** is a research paper, which analyzed data to find out if someone is at **serious risk** of **Ovarian Cancer** based on their **49 biomarkers and non-biomarkers**.
- We sought to **improve the accuracy** of the base model by **tuning various hyperparameters**.

Jenny Fish

BACKGROUND: OVARIAN CANCER

Social Impact

2

Jenny Fish

21,000 Diagnoses

14,000 ***Deaths!***

Source: https://www.cdc.gov/cancer/ovarian/statistics/index.htm

OVARIAN CANCER IMPACTS

- Often **asymptomatic** until later stages (25% detected at Stage I)
  - Diagnosed early - 90% survival rate
- Later stages, **very low survival rate**
- **CA125, HE4, CEA** are common **biomarkers** associated with Ovarian Cancer
  - **CA125** considered a gold standard biomarker
  - Current diagnosis algorithm — **ROMA test** (based on CA125 and HE4)

Isha Angadi

RESEARCH 3

Ovarian Cancer Scientific Information

Isha Angadi

OVARIAN CANCER STUDY (PAPER)

"Using Machine Learning to Predict Ovarian Cancer" by Lu, Fan, et al.
Published: International Journal of Medical Informatics

Aim:

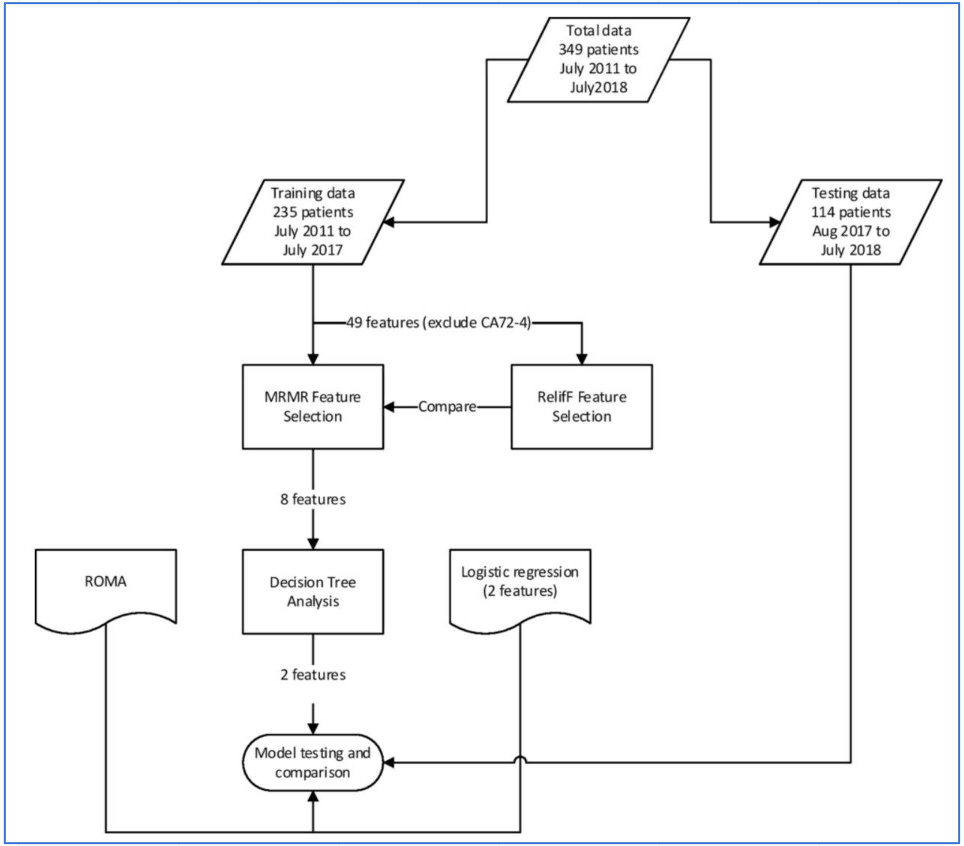
- To **improve the accuracy of early diagnosis and detection of ovarian cancer** using machine learning feature selection method — **MRMR** to build **decision tree**.

Data:

- **171 OC patients** and **178 BOT patients, 49 features**
- **Train/Test split — 235/114 values**

Source: *https://www.sciencedirect.com/science/article/pii/S1386505620302781*

Isha Angadi



"Using Machine Learning to Predict Ovarian Cancer" Process

Isha Angadi

OVARIAN CANCER STUDY (PAPER)

"Using Machine Learning to Predict Ovarian Cancer" by Lu, Fan, et al.
Published: International Journal of Medical Informatics

Procedure:

- Handling **missing data**
- Using **MRMR feature reduction**,
- Building a **decision tree model**.
  - Performing **cross validation**.
  - Produce **confusion matrix** and **accuracies**.

Results:

- **CEA and HE4** have the most significant prediction power when it comes to the classification of ovarian cancer vs the benign ovarian tumors.

Source: https://www.sciencedirect.com/science/article/pii/S1386505620302781

Adam Claudy

4 BUILDING OUR MODEL

Comparing Research Model with Our’s

Adam Claudy

PROJECT PIPELINE

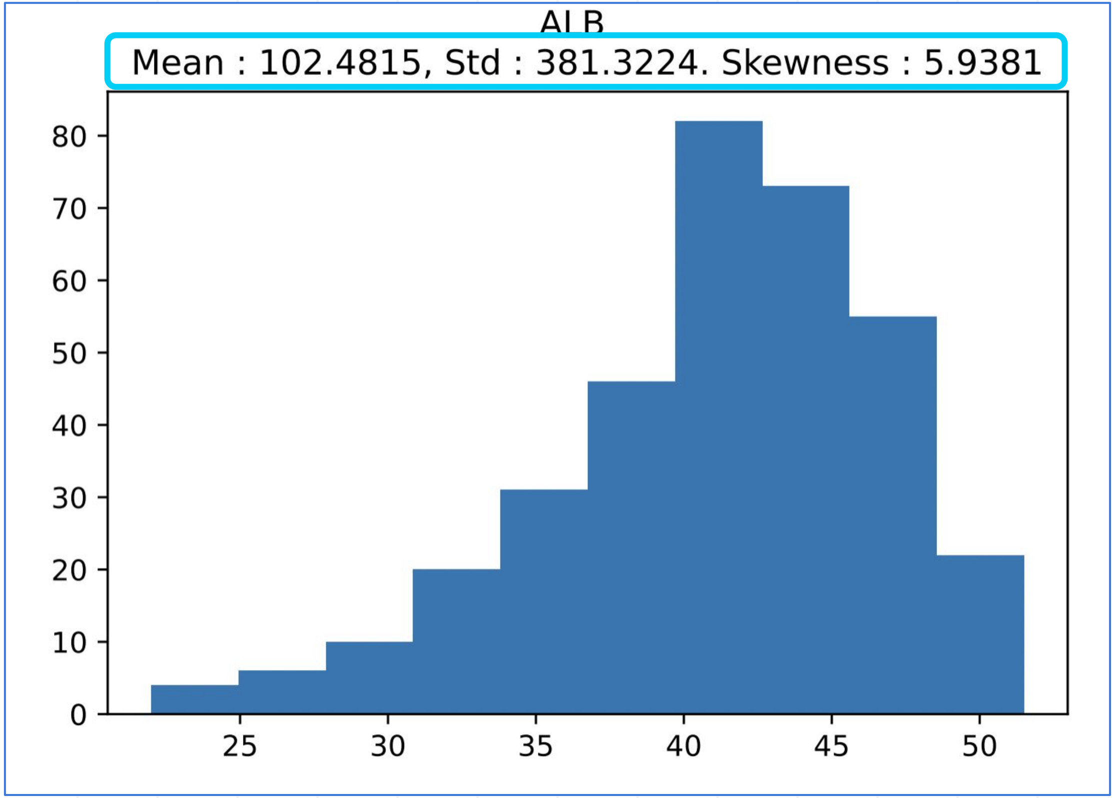
01	DATA PREPROCESSING	• Handling object data type • Handling columns with a given missing rate tolerance. • Handling missing data
02	DECISION TREE MODEL	• Impurity measure • Depth
03	EXPERIMENTS	• 3 Model Versions
04	RESULTS	• Visualizations • Confusion Matrices • Metrics

Adam Claudy

DATA PREPROCESSING

- Convert all feature columns into numeric form.
- Data is missing at random (MAR)
- Remove columns which exceed the specified missing rate tolerance. (25%, 50%)
  - **2 biomarkers removed** (CA72-4, NEU)
- Impute NAs with mean, median or mode.

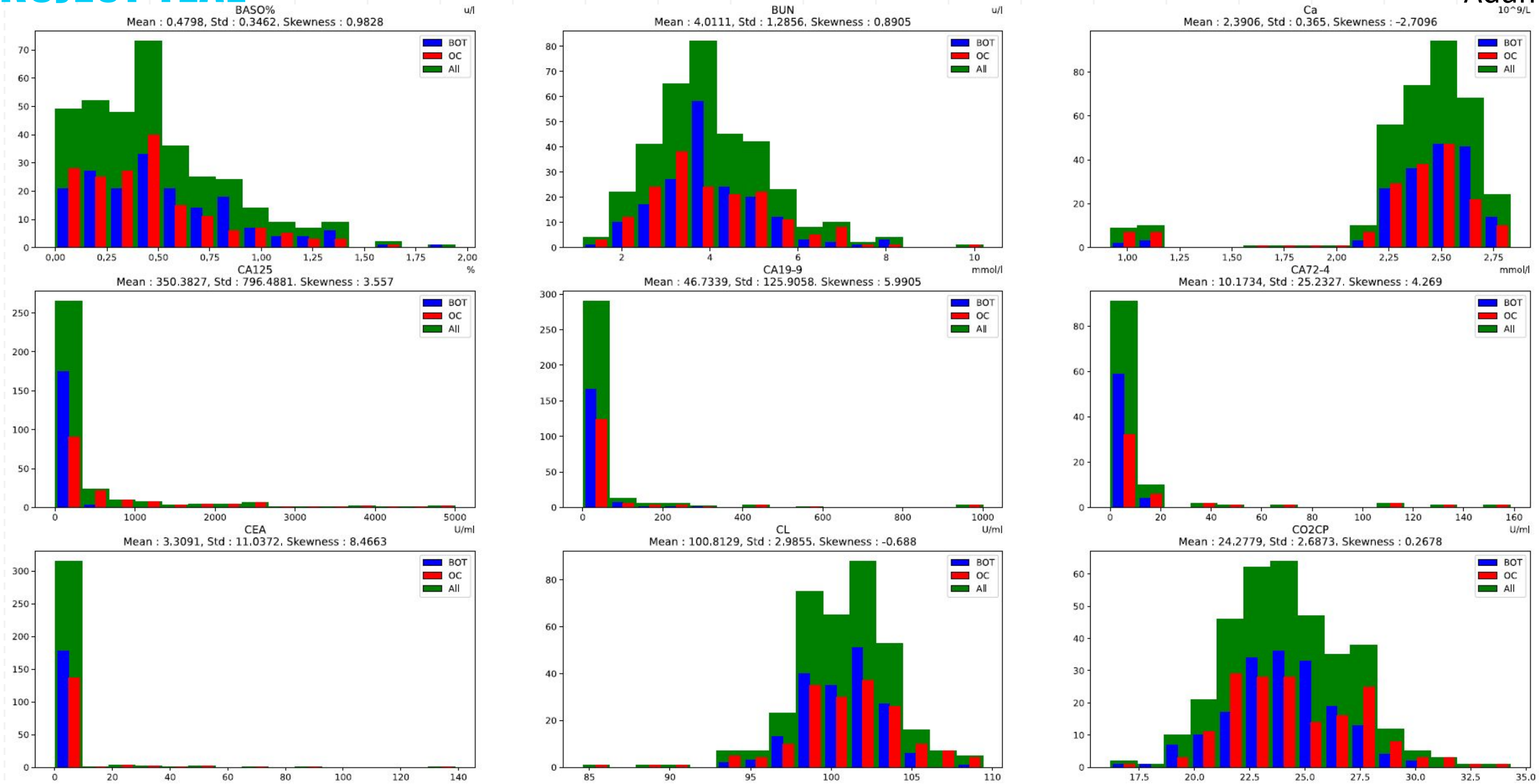
Adam Claudy



SHOWING DATA SKEWNESS - MEAN VS MEDIAN

PROJECT TEAL

Adam Claudy



Features Histogram

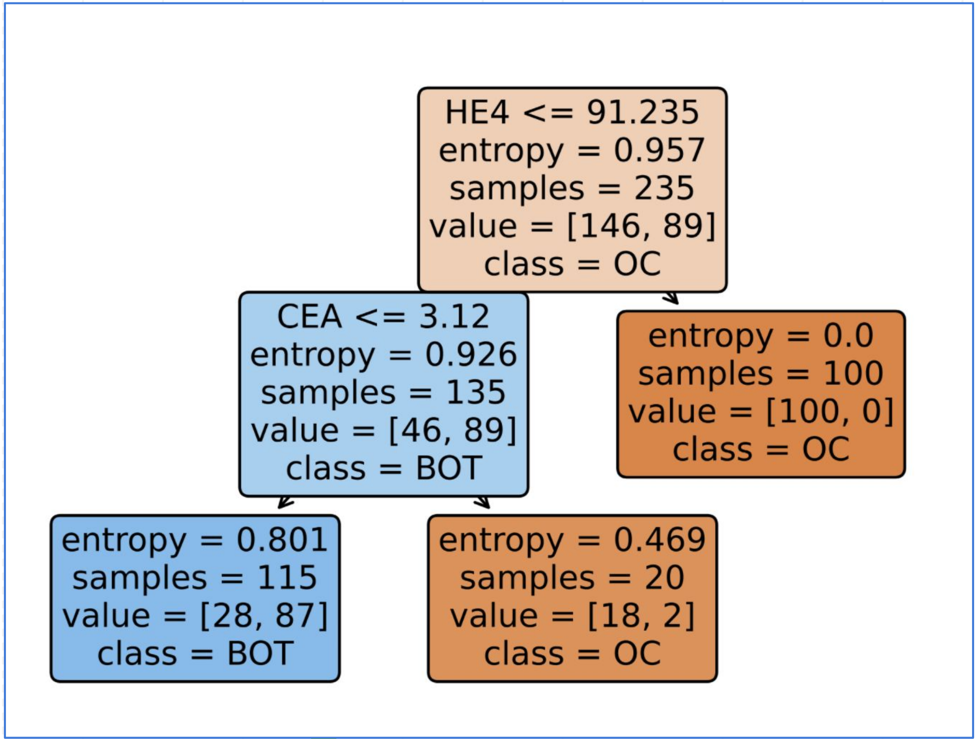
FEATURE SELECTION

Why do we need feature selection?

- Base Model reduced features using Minimum Redundancy - Maximum Relevance (MRMR) (from 48 to 8).
- Experiment using all features to test if feature selection is required.

Adam Claudy

DECISION TREE MODEL



Hyperparameters

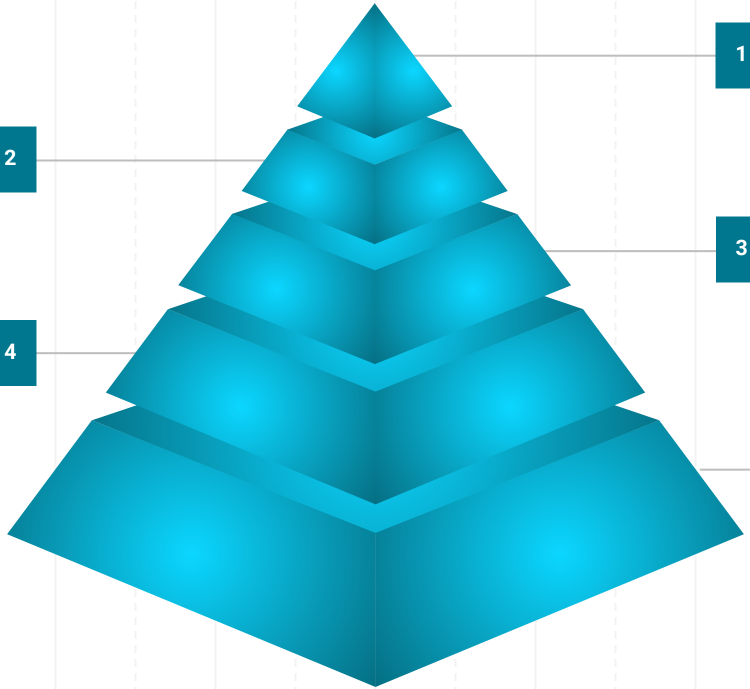
- Impurity Measure
  - Gini
  - Entropy
- Depth of tree

Samiha Khan

EXPERIMENTS 5

Samiha Khan

EXPERIMENT VARIATIONS



1. **Stratified k-cross Validation :** True or False
2. **Feature Selection :** 8 (selected by MRMR) or all features
3. **Shuffle Data :** True or False
4. **Imputation of Missing Data :** Mean, Mode, or Median
5. **DT Impurity :** gini or entropy

EXPERIMENT OUTPUTS

- Confusion Matrix
- Specificity
  Sensitivity
- PPV
- NPV
- Overall Accuracy
- F1 Score
- Mean Stratified Cross Validation Accuracy
- Teal Score

Sandeep Pvn

6 CODE

Metrics Insight and Code

CODE

Jupyter Notebook:

https://colab.research.google.com/drive/12dhDfeTJQj8NSfUnsfsw06HlpoqOjQcy#scrollTo=00rR7B5NwI2J

Sandeep Pvn

METRICS

Confusion Matrix

Actual / Predict	BOT	OC
BOT	TP	FP
OC	FN	TN

Objective : To reduce FP

- We take 2 metrics, specificity and precision into account.
- We combine the metrics into one score, the Teal score.

$$Precision = \frac{TP}{TP + FP}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Teal\ Score = \frac{1}{1 + \frac{FP}{2}\left[\frac{1}{TN} + \frac{1}{TP}\right]}$$

Mehar Chaturvedi

7 RESULTS

Model Results

Mehar Chaturvedi

RESULTS

- Why Stratified k-cross validation is required?
- Feature selection
- Why Shuffling is required?
  - What do we mean by shuffling?
- Which impute method is better and why?

CLASS IMBALANCE!!
Overfitting for OC class

Data

Training -
37.8% BOT class
62.2 % OC class

Testing -
78.2 % BOT class
21.8 % OC class

RESULTS : Confusion Matrix (Shuffle)

Mehar Chaturvedi

Before Shuffling : Paper

Actual / Predicted	BOT	OC
BOT	80	0
OC	9	**25**

Before Shuffling : Teal

Actual / Predicted	BOT	OC
BOT	76	0
OC	13	**25**

After Shuffling : Paper

Actual / Predicted	BOT	OC
BOT	43	13
OC	8	**50**

After Shuffling : Teal

Actual / Predicted	BOT	OC
BOT	47	13
OC	4	**50**

RESULTS : Confusion Matrix (Impute Methods)

**After Stratified-K-Cross Validation and Shuffling*

Mehar Chaturvedi

Mean

Actual / Predicted	BOT	OC
BOT	47	13
OC	4	50

Median

Actual / Predicted	BOT	OC
BOT	43	13
OC	8	50

Mode

Actual / Predicted	BOT	OC
BOT	45	16
OC	6	47

Actual / Predicted	Mean	Median	Mode
Teal Score	0.9798	0.9788	0.9787
Precision	0.783	0.768	0.738
Specificity	0.794	0.794	0.746

Mehar Chaturvedi

RESULTS

Experiment	TP	FP	FN	TN	Sensitivity (Recall)	Specificity	Positive Predictive Value	Negative Predictive Value	F1 score	Accuracy	Teal score	Tree depth
Teal_MRMR_features__gini__mean_	**80**	**0**	**9**	**25**	**0.899**	**1.000**	**1.000**	**0.735**	**0.947**	**0.921**	**1.000**	**2**
Teal_MRMR_features__gini__mean__stratified_k_cross	80	0	9	25	0.899	1.000	1.000	0.735	0.947	0.921	1.000	2
Teal_all_features__gini__mean_	76	0	13	25	0.854	1.000	1.000	0.658	0.921	0.886	1.000	4
Teal_all_features__gini__mean__stratified_k_cross	76	0	13	25	0.854	1.000	1.000	0.658	0.921	0.886	1.000	4
Teal_all_features__entropy__median_	70	0	19	25	0.787	1.000	1.000	0.568	0.881	0.833	1.000	3
Teal_all_features__entropy__median__stratified_k_cross	70	0	19	25	0.787	1.000	1.000	0.568	0.881	0.833	1.000	3
Teal_MRMR_features__entropy__median_	45	0	44	25	0.506	1.000	1.000	0.362	0.672	0.614	1.000	6
Teal_MRMR_features__entropy__median__stratified_k_cross	45	0	44	25	0.506	1.000	1.000	0.362	0.672	0.614	1.000	6
Teal_all_features__gini__mode_	28	1	61	24	0.315	0.960	0.966	0.282	0.475	0.456	0.963	6
Teal_all_features__gini__mode__stratified_k_cross	28	1	61	24	0.315	0.960	0.966	0.282	0.475	0.456	0.963	6
Teal_MRMR_features__entropy__mean_	81	2	8	23	0.910	0.920	0.976	0.742	0.942	0.912	0.947	2
Teal_MRMR_features__entropy__mean__stratified_k_cross	**81**	**2**	**8**	**23**	**0.910**	**0.920**	**0.976**	**0.742**	**0.942**	**0.912**	**0.947**	**2**
Teal_all_features__entropy__mean_	77	2	12	23	0.865	0.920	0.975	0.657	0.917	0.877	0.947	4
Teal_all_features__entropy__mean__stratified_k_cross	77	2	12	23	0.865	0.920	0.975	0.657	0.917	0.877	0.947	4
Teal_MRMR_features__gini__median_	66	2	23	23	0.742	0.920	0.971	0.500	0.841	0.781	0.945	1
Teal_all_features__gini__median_	66	2	23	23	0.742	0.920	0.971	0.500	0.841	0.781	0.945	1
Teal_MRMR_features__gini__median__stratified_k_cross	66	2	23	23	0.742	0.920	0.971	0.500	0.841	0.781	0.945	1
Teal_all_features__gini__median__stratified_k_cross	66	2	23	23	0.742	0.920	0.971	0.500	0.841	0.781	0.945	1
Teal_MRMR_features__gini__mode_	57	3	32	22	0.640	0.880	0.950	0.407	0.765	0.693	0.914	7
Teal_MRMR_features__gini__mode__stratified_k_cross	57	3	32	22	0.640	0.880	0.950	0.407	0.765	0.693	0.914	7
Teal_MRMR_features__entropy__mode_	53	3	36	22	0.596	0.880	0.946	0.379	0.731	0.658	0.912	3
Teal_all_features__entropy__mode_	53	3	36	22	0.596	0.880	0.946	0.379	0.731	0.658	0.912	3
Teal_MRMR_features__entropy__mode__stratified_k_cross	53	3	36	22	0.596	0.880	0.946	0.379	0.731	0.658	0.912	3
Teal_all_features__entropy__mode__stratified_k_cross	53	3	36	22	0.596	0.880	0.946	0.379	0.731	0.658	0.912	3
Teal_MRMR_features__entropy__mean__shuffle	47	13	4	50	0.922	0.794	0.783	0.926	0.847	0.851	0.788	2

Mehar Chaturvedi

RESULTS METRICS

PAPER

Experiment
Teal_MRMR_features__gini__mean_

TP	FP	FN	TN	Sensitivity (Recall)	Specificity	Positive Predictive Value	Negative Predictive Value	F1 score	Accuracy	Teal score	Tree depth
80	0	9	25	0.899	1.000	1.000	0.735	0.947	0.921	1.000	2
81	2	8	23	0.910	0.920	0.976	0.742	0.942	0.912	0.947	2

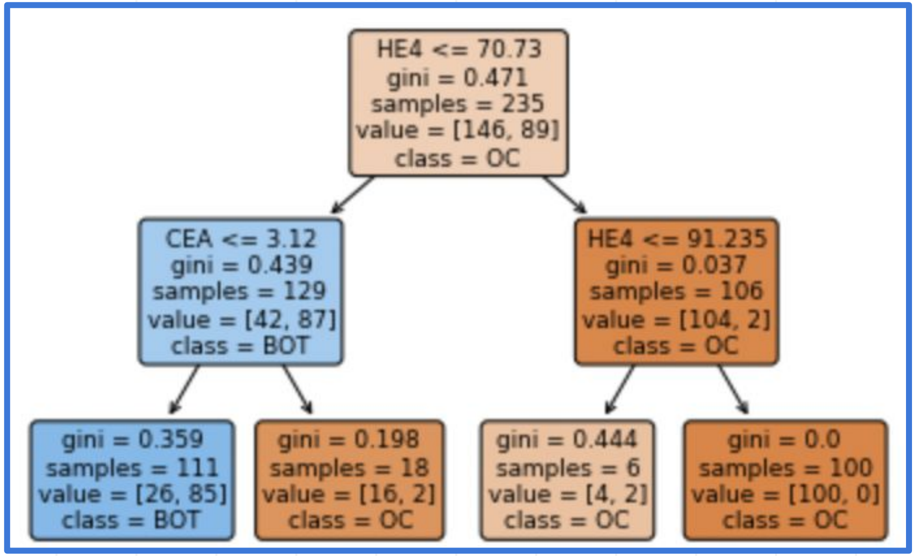
TEAL

Teal_MRMR_features__entropy__mean__stratified_k_cross

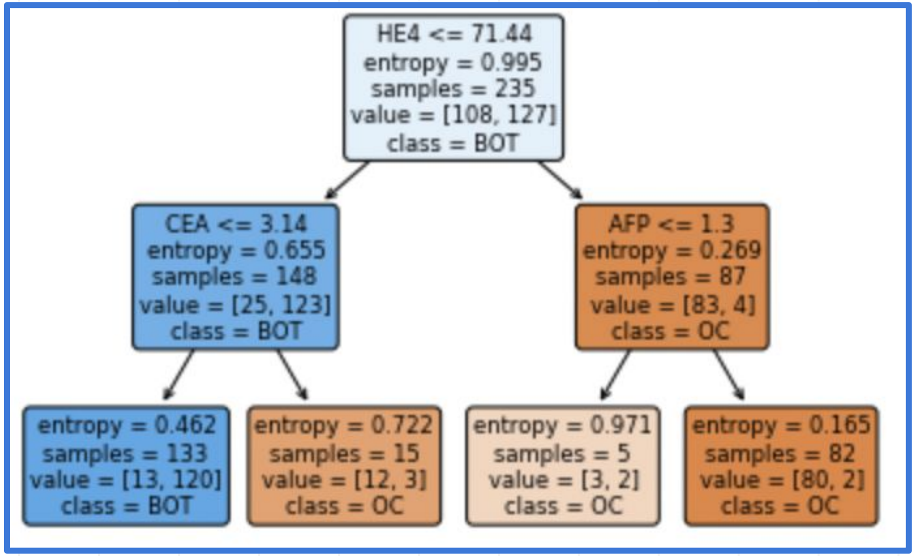
Mehar Chaturvedi

DECISION TREE COMPARISON

PAPER



TEAL

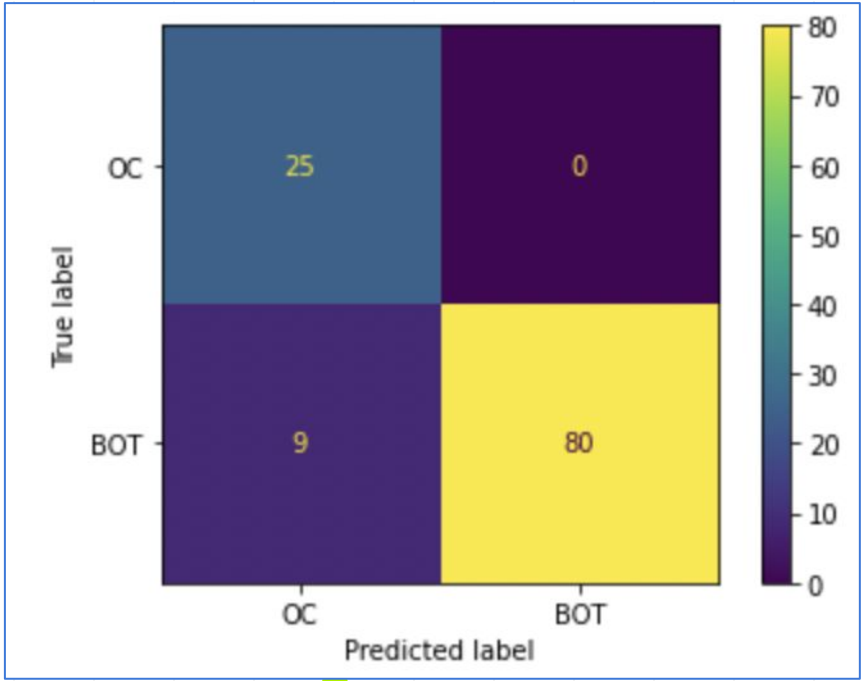


- The tree achieves the best mean cross-validation accuracy 87.65957 +/-4.73852 % on training dataset

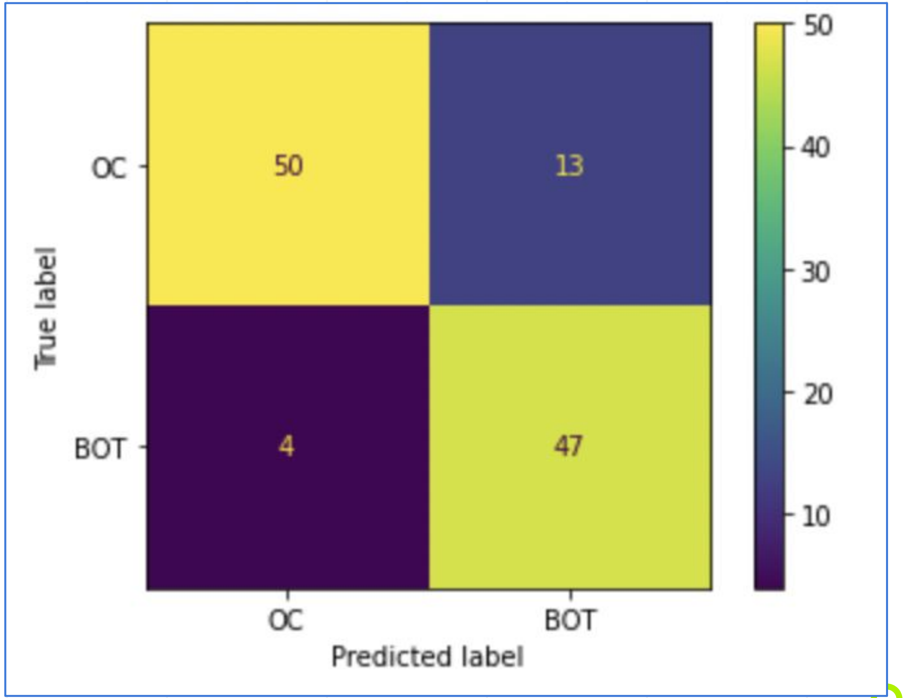
Mehar Chaturvedi

CONFUSION MATRIX COMPARISON

PAPER



TEAL



Sandeep Pvn

8 FUTURE WORK

Neural Networks

Sandeep Pvn

FUTURE WORK AND SUGGESTIONS

- **Customization:** run the model on any generalized data set
    - implement customizing imputing techniques for each column
    - Try to obtain and use genetic data
- **Gini vs. Entropy**
- **Grid search:** Increase code efficiency and compute the optimum values of hyperparameters.
- **Neural Network:** Running the model through a neural network to improve the accuracies.
- **Analyse and predict** if and when BOT converts to OC
    - Change is system. Need Time Series Data

Sandeep Pvn

FUTURE WORK AND SUGGESTIONS

01	Customization	• Run the model on any generalized data set • Implement customizing imputing techniques for each column
02	Genetic Data	• Try to obtain and use genetic data
03	Grid search & Pipelining	• Increase code efficiency and compute the optimum values of hyperparameters. • Use pipelining to speed up
04	Neural Network	• Running the model through a neural network to improve the accuracies.
05	BOT to OC	• Analyse and predict if and when BOT converts to OC • Change is system. Need Time Series Data

9 QUESTIONS FOR PROF. PREM

QUESTIONS FOR REFLECTION

- When we shuffle our data, it is making a very big difference — Why does shuffling makes such a big difference with our results?
- Why is mean giving a better result than median and mode?

THANKS!

Any questions?

SOURCES

- Lu, M., Fan, Z., Xu, B., Chen, L., Zheng, X., Li, J., Znati, T., Mi, Q. and Jiang, J., 2021. Using machine learning to predict ovarian cancer. https://www.sciencedirect.com/science/article/pii/S1386505620302781