```
# do the same thing, but use scikitlearn randomforest classifier


!pip install scikit-learn==1.3.0 --upgrade
!pip install --upgrade xgboost
```

⇥ Requirement already satisfied: scikit-learn==1.3.0 in /usr/local/lib/python3.11/dist-packages (1.3.0)
    Requirement already satisfied: numpy>=1.17.3 in /usr/local/lib/python3.11/dist-packages (from scikit-l
    Requirement already satisfied: scipy>=1.5.0 in /usr/local/lib/python3.11/dist-packages (from scikit-le
    Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from scikit-l
    Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from s
    Requirement already satisfied: xgboost in /usr/local/lib/python3.11/dist-packages (2.1.4)
    Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (from xgboost) (1.26.4
    Requirement already satisfied: nvidia-nccl-cu12 in /usr/local/lib/python3.11/dist-packages (from xgboo
    Requirement already satisfied: scipy in /usr/local/lib/python3.11/dist-packages (from xgboost) (1.13.1

◀ ━━ ▶

```
#classify with cycle features including alignment
import pandas as pd
# import xgboost as xgb
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier as RFC
from sklearn.metrics import classification_report
import xgboost as xgb
from sklearn.metrics import confusion_matrix
from sklearn.metrics import roc_curve
import seaborn as sns
from matplotlib import pyplot as plt
import numpy as np
from IPython import get_ipython
from IPython.display import display
from sklearn.impute import SimpleImputer # Import SimpleImputer for imputation
import shap
shap.initjs()
```

⇥   ⟨js⟩

∨ Set up

```
df = pd.read_csv('/content/cycle_and_HMM_features_false_bw-3-0_dataset_48days.csv')


df.head()
```

	hub_id	pat_cat_map	cycle_min	cycle_max	cycle_median	cycle_mean	cycle_range	cycle_st
0	U303F6B17404145	PCOS	17	43	33.0	33.0625	26	7.35272
1	U2FC1C617263332	nonPCOS-nonBaseline	21	27	23.0	23.6000	6	2.40831
2	U2F65CA17170226	PCOS	27	45	36.0	35.3000	18	6.34297
3	U2F65AE17170162	Baseline	30	39	31.0	32.7500	9	4.27200
4	UD5E9E4230202	PCOS	29	29	29.0	29.0000	0	Na

```python
# LOOK AT LAUREN'S GITHUB FOR CODE

# try w xgboost
# try w subset of features
# explanatory tools to see which variables are important (SHAP values)


df = df.loc[df['pat_cat_map'].isin(['Baseline','PCOS'])]


df['label_01'] = df['pat_cat_map'].map({'Baseline':0, 'PCOS':1})
```

<ipython-input-792-1fe60784182b>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing
  df['label_01'] = df['pat_cat_map'].map({'Baseline':0, 'PCOS':1})

```python
df = df.replace(-np.inf, np.nan)


df.columns
```

```
Index(['hub_id', 'pat_cat_map', 'cycle_min', 'cycle_max', 'cycle_median',
       'cycle_mean', 'cycle_range', 'cycle_std', 'num_cycles',
       'viterbi_logprob_mean', 'viterbi_logprob_min', 'viterbi_logprob_max',
       'viterbi_logprob_std', 'viterbi_logprob_median',
       'complete_logprob_mean', 'complete_logprob_min', 'complete_logprob_max',
       'complete_logprob_std', 'complete_logprob_median', 'label_01'],
      dtype='object')
```

```python
HMM_features = [ 'viterbi_logprob_mean',
      'viterbi_logprob_min', 'viterbi_logprob_max', 'viterbi_logprob_std',
      'viterbi_logprob_median', 'complete_logprob_mean',
      'complete_logprob_min', 'complete_logprob_max', 'complete_logprob_std',
      'complete_logprob_median']
cycle_features = ['cycle_min', 'cycle_max', 'cycle_median',
      'cycle_mean', 'cycle_range', 'cycle_std']

target = 'label_01'
```

∨ All features

```
print('Performance with all features')

X_train_all, X_test_all, y_train_all, y_test_all = train_test_split(df[HMM_features+cycle_features], df[ta
                                                    shuffle=True, random_state=51)
```

⊋ Performance with all features

```
clf = xgb.XGBClassifier(random_state=51)
clf.fit(X_train_all, y_train_all)
y_pred_all = clf.predict(X_test_all)
y_score_all = clf.predict_proba(X_test_all)
print(confusion_matrix(y_test_all, y_pred_all, normalize='true'))
```

⊋ [[0.27142857 0.72857143]
   [0.09139785 0.90860215]]

```
print(classification_report(y_pred_all, y_test_all))
```

⊋

	precision	recall	f1-score	support
0	0.27	0.53	0.36	36
1	0.91	0.77	0.83	220
accuracy			0.73	256
macro avg	0.59	0.65	0.60	256
weighted avg	0.82	0.73	0.77	256

```
fpr_full, tpr_full, thresholds_full = roc_curve(y_test_all, y_score_all[:,1])#, pos_label='PCOS')
sns.lineplot(x=fpr_full, y=tpr_full, label='Cycle + HMM features - False Missingness 0-3 Days', errorbar=None
#plt.savefig('/content/drive/MyDrive/fall_research/feature distribution plots/xgb_full_features.pdf')
```
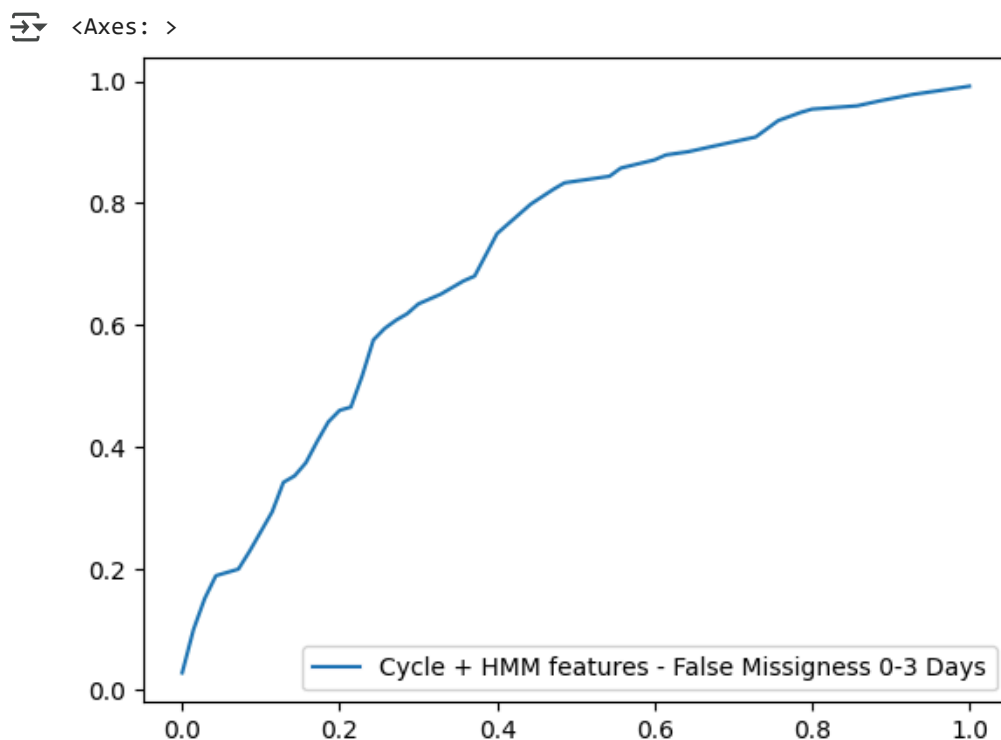
```
#overall accuracy:
print((y_pred_all==y_test_all).sum()/len(y_pred_all))
```

⇥ 0.734375

⌄ Cycle features only

```
#PERFORMANCE WITH CYCLE FEATURES ONLY
print('Performance with cycle features only')

X_train_cycle, X_test_cycle, y_train_cycle, y_test_cycle = train_test_split(df[cycle_features], df[target]
                                                            shuffle=True, random_state=51)
```

⇥ Performance with cycle features only

```
clf = xgb.XGBClassifier(random_state=51)
clf.fit(X_train_cycle, y_train_cycle)
y_pred_cycle = clf.predict(X_test_cycle)
y_score_cycle = clf.predict_proba(X_test_cycle)
print(confusion_matrix(y_test_cycle, y_pred_cycle, normalize='true'))
```

⇥ [[0.14285714 0.85714286]
 [0.11290323 0.88709677]]

```
print(classification_report(y_pred_cycle, y_test_cycle))
```

⇥

	precision	recall	f1-score	support
0	0.14	0.32	0.20	31
1	0.89	0.73	0.80	225

accuracy			0.68	256
macro avg	0.51	0.53	0.50	256
weighted avg	0.80	0.68	0.73	256

```
fpr_cycle, tpr_cycle, thresholds_cycle = roc_curve(y_test_cycle, y_score_cycle[:,1])#, pos_label='PCOS')
sns.lineplot(x=fpr_cycle, y=tpr_cycle, label='Cycle features only', errorbar=None)
#plt.savefig('/content/drive/MyDrive/fall_research/feature distribution plots/xgb_cycle_features_only.pdf'
```

⊶ <Axes: >
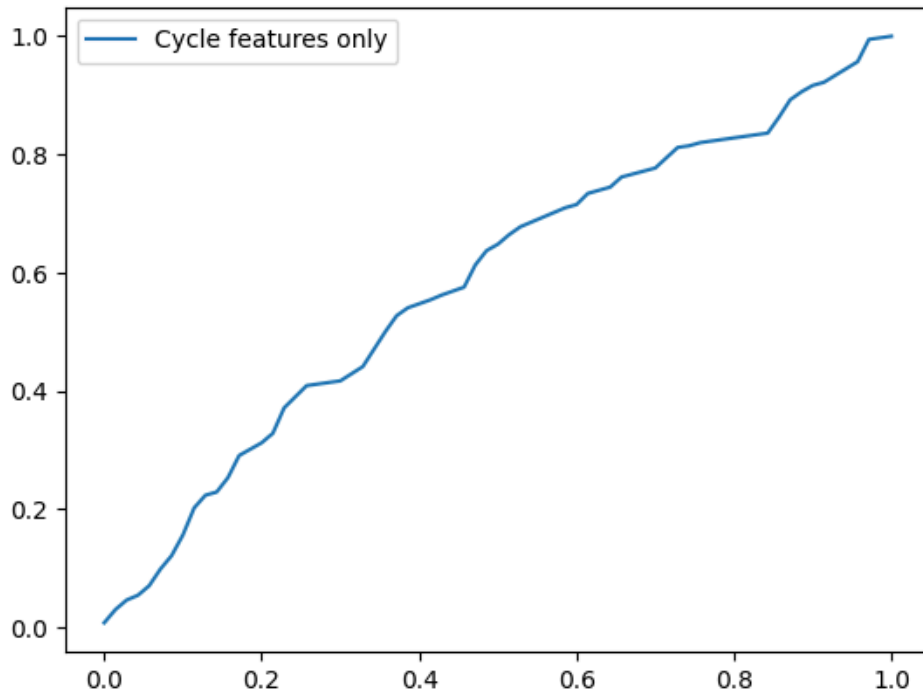

```
#overall accuracy:
print((y_pred_cycle==y_test_cycle).sum()/len(y_pred_cycle))
```

⊶ 0.68359375

⌄ HMM Features only

```
#PERFORMANCE WITH HMM FEATURES ONLY
print('Performance with HMM features only')

X_train_hmm, X_test_hmm, y_train_hmm, y_test_hmm = train_test_split(df[HMM_features], df[target],
                                                    shuffle=True, random_state=51)
```

⊶ Performance with HMM features only

```
# Impute missing values using SimpleImputer
imputer = SimpleImputer(strategy='mean') # Replace 'mean' with other strategies if needed
X_train_hmm = imputer.fit_transform(X_train_hmm)
X_test_hmm = imputer.transform(X_test_hmm)
```

```
clf = RFC(random_state=101)
clf.fit(X_train_hmm, y_train_hmm)
y_pred_hmm = clf.predict(X_test_hmm)
y_score_hmm = clf.predict_proba(X_test_hmm)
print(confusion_matrix(y_test_hmm, y_pred_hmm, normalize='true'))
fpr_hmm, tpr_hmm, thresholds_hmm = roc_curve(y_test_hmm, y_score_hmm[:,1])#, pos_label='PCOS')
sns.lineplot(x=fpr_hmm, y=tpr_hmm, label='HMM features only', errorbar=None)
#plt.savefig('/content/drive/MyDrive/fall_research/feature distribution plots/xgb_hmm_features_only.pdf')
```

⮑ [[0.24285714 0.75714286]
     [0.09139785 0.90860215]]
    <Axes: >
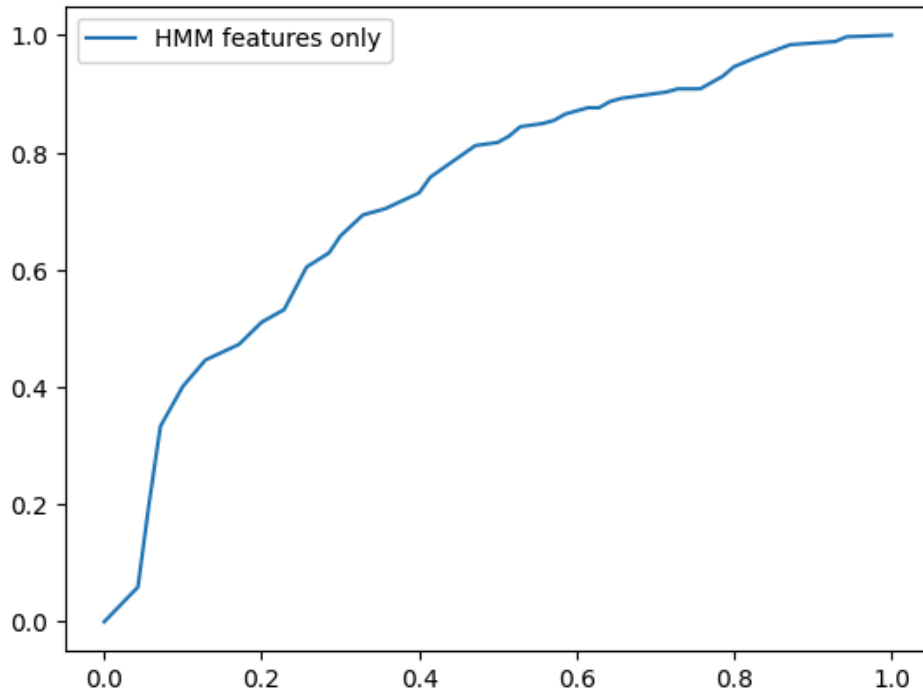

```
print(classification_report(y_pred_cycle, y_test_cycle))
```

⮑

	precision	recall	f1-score	support
0	0.14	0.32	0.20	31
1	0.89	0.73	0.80	225
accuracy			0.68	256
macro avg	0.51	0.53	0.50	256
weighted avg	0.80	0.68	0.73	256

```
#overall accuracy:
print((y_pred_cycle==y_test_cycle).sum()/len(y_pred_cycle))
#fpr_algn, tpr_algn, thresholds_algn = roc_curve(y_test, -1*X_test, pos_label='PCOS')
#sns.lineplot(x=fpr_algn, y=tpr_algn, label='HMM features only', errorbar=None)
```

⮑ 0.68359375

```
#make kdeplots of all features
for feature in HMM_features+cycle_features:
    sns.kdeplot(data=df, x=feature, hue='pat_cat_map', common_norm=False)
```

```
#plt.savefig('/content/drive/MyDrive/fall_research/feature distribution plots/xgb_kdeplots_feature_dis
plt.clf()
```

⌲ <Figure size 640x480 with 0 Axes>

⌄ ROC Curves

```
# put 3 ROC curves on one axis (cycle, hmm, all)


# # Create subplots
# fig, axes = plt.subplots(1, 3, figsize=(15, 5))  # 1 row, 3 columns

# Plot Cycle + HMM features
sns.lineplot(x=fpr_full, y=tpr_full, label='Cycle + HMM features', errorbar=None)
# axes[0].set_title("Cycle + HMM ROC Curve")

# Plot Cycle features only
sns.lineplot(x=fpr_cycle, y=tpr_cycle,  label='Cycle features only', errorbar=None)
# axes[1].set_title("Cycle Only ROC Curve")

# Plot HMM features only
sns.lineplot(x=fpr_hmm, y=tpr_hmm, label='HMM features only', errorbar=None)
# axes[2].set_title("HMM Only ROC Curve")

# Adjust layout
# plt.tight_layout()
plt.show()
#plt.savefig('/content/drive/MyDrive/fall_research/feature distribution plots/xgb_roc_curves.pdf')
```
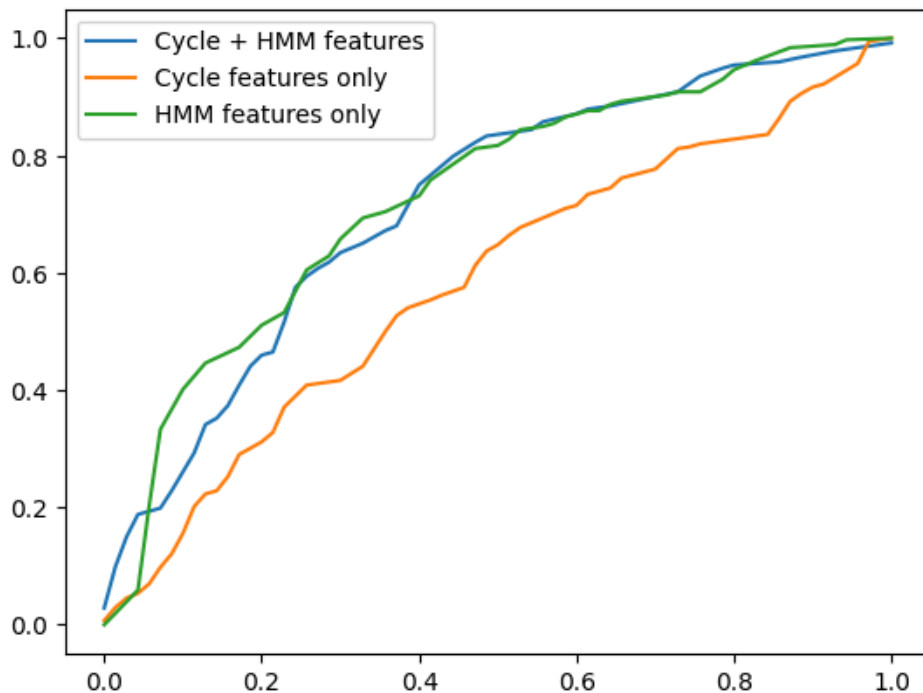
⌲

⌄

use HMM features and take one out to see if any features are important (leave one out version)

```
HMM_features = ['viterbi_logprob_mean',
      'viterbi_logprob_min', 'viterbi_logprob_max', 'viterbi_logprob_std',
      'viterbi_logprob_median', 'complete_logprob_mean',
      'complete_logprob_min', 'complete_logprob_max', 'complete_logprob_std',
      'complete_logprob_median']
```

∨ without viterbi_logprob_mean

```
HMM_features = [
      'viterbi_logprob_min', 'viterbi_logprob_max', 'viterbi_logprob_std',
      'viterbi_logprob_median', 'complete_logprob_mean',
      'complete_logprob_min', 'complete_logprob_max', 'complete_logprob_std',
      'complete_logprob_median']


print('Performance with HMM features _without_viterbi_logprob_mean ')

X_train_without_viterbi_logprob_mean, X_test_without_viterbi_logprob_mean, y_train_without_viterbi_logprob_r
                                          shuffle=True, random_state=51)
```

⤷ Performance with HMM features _without_viterbi_logprob_mean
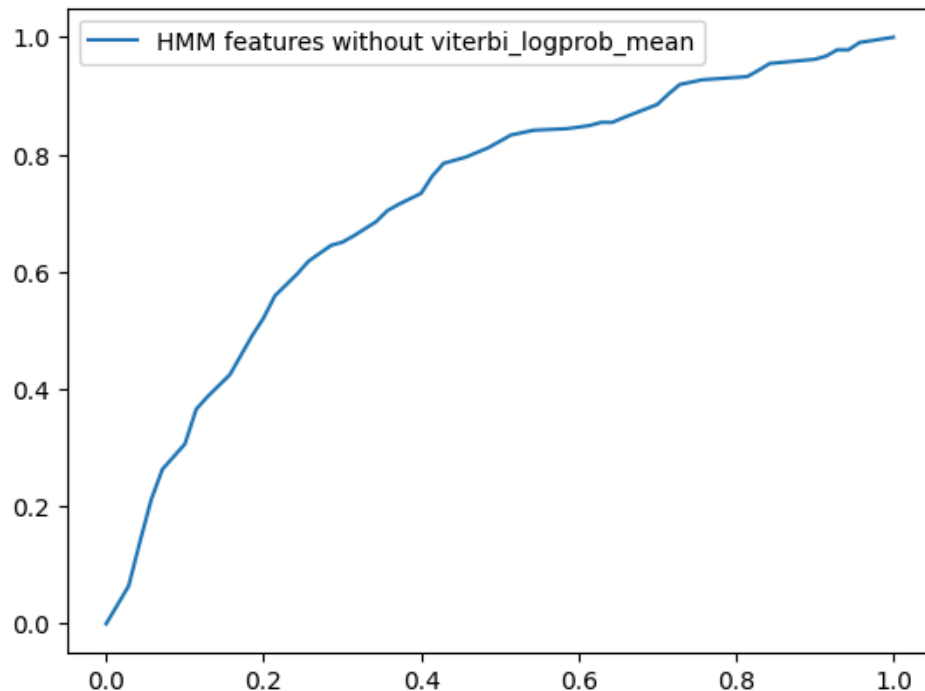
```
# Impute missing values using SimpleImputer
imputer = SimpleImputer(strategy='mean') # Replace 'mean' with other strategies if needed
X_train_without_viterbi_logprob_mean = imputer.fit_transform(X_train_without_viterbi_logprob_mean)
X_test_without_viterbi_logprob_mean = imputer.transform(X_test_without_viterbi_logprob_mean)


clf = RFC(random_state=101)
clf.fit(X_train_without_viterbi_logprob_mean, y_train_without_viterbi_logprob_mean)
y_pred_without_viterbi_logprob_mean = clf.predict(X_test_without_viterbi_logprob_mean)
y_score_without_viterbi_logprob_mean = clf.predict_proba(X_test_without_viterbi_logprob_mean)
print(confusion_matrix(y_test_without_viterbi_logprob_mean, y_pred_without_viterbi_logprob_mean, normalize=
fpr_without_viterbi_logprob_mean, tpr_without_viterbi_logprob_mean, thresholds_without_viterbi_logprob_mean
sns.lineplot(x=fpr_without_viterbi_logprob_mean, y=tpr_without_viterbi_logprob_mean, label='HMM features wi
#plt.savefig('/content/drive/MyDrive/fall_research/feature distribution plots/viterbi adjusted plots/xgb_wi
```

```
[[0.24285714 0.75714286]
 [0.07526882 0.92473118]]
<Axes: >
```



```
print(classification_report(y_pred_without_viterbi_logprob_mean, y_test_without_viterbi_logprob_mean))
```

```
              precision    recall  f1-score   support

           0       0.24      0.55      0.34        31
           1       0.92      0.76      0.84       225

    accuracy                           0.74       256
   macro avg       0.58      0.66      0.59       256
weighted avg       0.84      0.74      0.78       256
```

```
#overall accuracy:
print((y_pred_without_viterbi_logprob_mean==y_test_without_viterbi_logprob_mean).sum()/len(y_pred_without_
#fpr_algn, tpr_algn, thresholds_algn = roc_curve(y_test, -1*X_test, pos_label='PCOS')
#sns.lineplot(x=fpr_algn, y=tpr_algn, label='HMM features only', errorbar=None)
```

```
0.73828125
```

∨ without viterbi_logprob_min

```
HMM_features = ['viterbi_logprob_mean',
       'viterbi_logprob_max', 'viterbi_logprob_std',
       'viterbi_logprob_median', 'complete_logprob_mean',
       'complete_logprob_min', 'complete_logprob_max', 'complete_logprob_std',
       'complete_logprob_median']
```

```
print('Performance with HMM features _without_viterbi_logprob_min ')
```

```
X_train_without_viterbi_logprob_min, X_test_without_viterbi_logprob_min, y_train_without_viterbi_logprob_m
                                        shuffle=True, random_state=51)
```

⮡ Performance with HMM features _without_viterbi_logprob_min

```
# Impute missing values using SimpleImputer
imputer = SimpleImputer(strategy='mean') # Replace 'mean' with other strategies if needed
X_train_without_viterbi_logprob_min = imputer.fit_transform(X_train_without_viterbi_logprob_min)
X_test_without_viterbi_logprob_min = imputer.transform(X_test_without_viterbi_logprob_min)
```

```
clf = RFC(random_state=101)
clf.fit(X_train_without_viterbi_logprob_min, y_train_without_viterbi_logprob_min)
y_pred_without_viterbi_logprob_min = clf.predict(X_test_without_viterbi_logprob_min)
y_score_without_viterbi_logprob_min = clf.predict_proba(X_test_without_viterbi_logprob_min)
print(confusion_matrix(y_test_without_viterbi_logprob_min, y_pred_without_viterbi_logprob_min, normalize='
fpr_without_viterbi_logprob_min, tpr_without_viterbi_logprob_min, thresholds_without_viterbi_logprob_min =
sns.lineplot(x=fpr_without_viterbi_logprob_min, y=tpr_without_viterbi_logprob_min, label='HMM features wit
#plt.savefig('/content/drive/MyDrive/fall_research/feature distribution plots/viterbi adjusted plots/xgb_w
```

⮡ [[0.25714286 0.74285714]
    [0.08064516 0.91935484]]
   <Axes: >
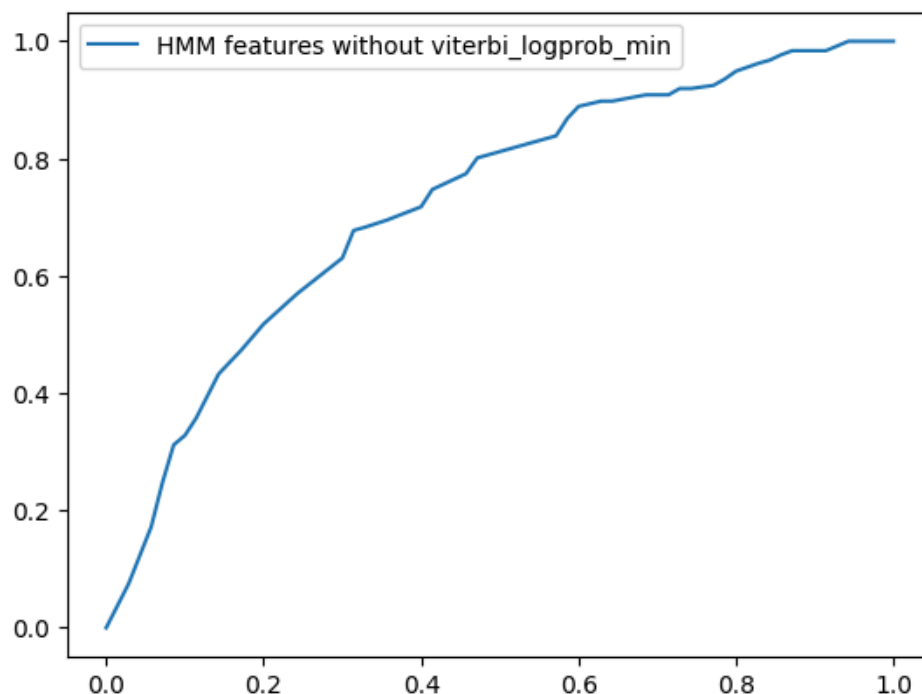

```
print(classification_report(y_pred_without_viterbi_logprob_min, y_test_without_viterbi_logprob_min))
```

⮡

	precision	recall	f1-score	support
0	0.26	0.55	0.35	33
1	0.92	0.77	0.84	223
accuracy			0.74	256
macro avg	0.59	0.66	0.59	256
weighted avg	0.83	0.74	0.77	256

```
#overall accuracy:
print((y_pred_without_viterbi_logprob_min==y_test_without_viterbi_logprob_min).sum()/len(y_pred_without_vi
#fpr_algn, tpr_algn, thresholds_algn = roc_curve(y_test, -1*X_test, pos_label='PCOS')
#sns.lineplot(x=fpr_algn, y=tpr_algn, label='HMM features only', errorbar=None)
```

⤳ 0.73828125

⌄ without viterbi_logprob_max

```
HMM_features = ['viterbi_logprob_mean',
        'viterbi_logprob_min', 'viterbi_logprob_std',
        'viterbi_logprob_median', 'complete_logprob_mean',
        'complete_logprob_min', 'complete_logprob_max', 'complete_logprob_std',
        'complete_logprob_median']


print('Performance with HMM features _without_viterbi_logprob_max ')

X_train_without_viterbi_logprob_max, X_test_without_viterbi_logprob_max, y_train_without_viterbi_logprob_ma:
                                              shuffle=True, random_state=51)
```

⤳ Performance with HMM features _without_viterbi_logprob_max

```
# Impute missing values using SimpleImputer
imputer = SimpleImputer(strategy='mean') # Replace 'mean' with other strategies if needed
X_train_without_viterbi_logprob_max = imputer.fit_transform(X_train_without_viterbi_logprob_max)
X_test_without_viterbi_logprob_max = imputer.transform(X_test_without_viterbi_logprob_max)


clf = RFC(random_state=101)
clf.fit(X_train_without_viterbi_logprob_max, y_train_without_viterbi_logprob_max)
y_pred_without_viterbi_logprob_max = clf.predict(X_test_without_viterbi_logprob_max)
y_score_without_viterbi_logprob_max = clf.predict_proba(X_test_without_viterbi_logprob_max)
print(confusion_matrix(y_test_without_viterbi_logprob_max, y_pred_without_viterbi_logprob_max, normalize='t1
fpr_without_viterbi_logprob_max, tpr_without_viterbi_logprob_max, thresholds_without_viterbi_logprob_max = 1
sns.lineplot(x=fpr_without_viterbi_logprob_max, y=tpr_without_viterbi_logprob_max, label='HMM features with
#plt.savefig('/content/drive/MyDrive/fall_research/feature distribution plots/viterbi adjusted plots/xgb_wit
```
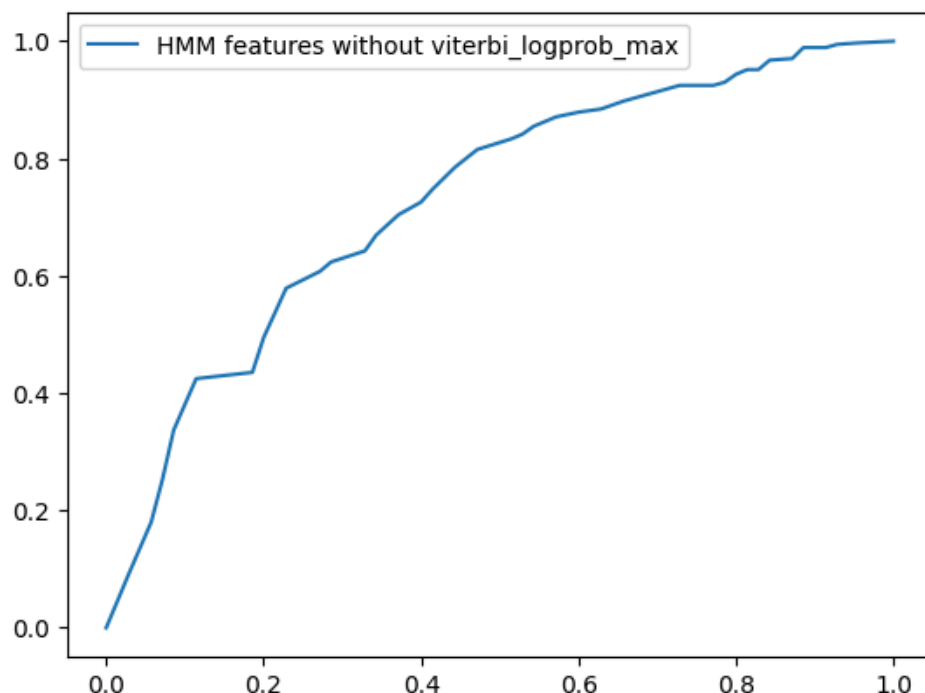
```
[[0.27142857 0.72857143]
 [0.07526882 0.92473118]]
<Axes: >
```



```
print(classification_report(y_pred_without_viterbi_logprob_max, y_test_without_viterbi_logprob_max))
```

	precision	recall	f1-score	support
0	0.27	0.58	0.37	33
1	0.92	0.77	0.84	223
accuracy			0.75	256
macro avg	0.60	0.67	0.61	256
weighted avg	0.84	0.75	0.78	256

```
#overall accuracy:
print((y_pred_without_viterbi_logprob_max==y_test_without_viterbi_logprob_max).sum()/len(y_pred_without_vi
#fpr_algn, tpr_algn, thresholds_algn = roc_curve(y_test, -1*X_test, pos_label='PCOS')
#sns.lineplot(x=fpr_algn, y=tpr_algn, label='HMM features only', errorbar=None)
```

```
0.74609375
```

⌄ without viterbi_logprob_std

```
HMM_features = ['viterbi_logprob_mean',
       'viterbi_logprob_min', 'viterbi_logprob_max',
       'viterbi_logprob_median', 'complete_logprob_mean',
       'complete_logprob_min', 'complete_logprob_max', 'complete_logprob_std',
       'complete_logprob_median']
```

```
print('Performance with HMM features _without_viterbi_logprob_std ')
```

```
X_train_without_viterbi_logprob_std, X_test_without_viterbi_logprob_std, y_train_without_viterbi_logprob_st
```

Performance with HMM features _without_viterbi_logprob_std

```
# Impute missing values using SimpleImputer
imputer = SimpleImputer(strategy='mean') # Replace 'mean' with other strategies if needed
X_train_without_viterbi_logprob_std = imputer.fit_transform(X_train_without_viterbi_logprob_std)
X_test_without_viterbi_logprob_std = imputer.transform(X_test_without_viterbi_logprob_std)
```

```
clf = RFC(random_state=101)
clf.fit(X_train_without_viterbi_logprob_std, y_train_without_viterbi_logprob_std)
y_pred_without_viterbi_logprob_std = clf.predict(X_test_without_viterbi_logprob_std)
y_score_without_viterbi_logprob_std = clf.predict_proba(X_test_without_viterbi_logprob_std)
print(confusion_matrix(y_test_without_viterbi_logprob_std, y_pred_without_viterbi_logprob_std, normalize='
fpr_without_viterbi_logprob_std, tpr_without_viterbi_logprob_std, thresholds_without_viterbi_logprob_std =
sns.lineplot(x=fpr_without_viterbi_logprob_std, y=tpr_without_viterbi_logprob_std, label='HMM features wit
#plt.savefig('/content/drive/MyDrive/fall_research/feature distribution plots/viterbi adjusted plots/xgb_w
```

```
[[0.24285714 0.75714286]
 [0.08064516 0.91935484]]
<Axes: >
```
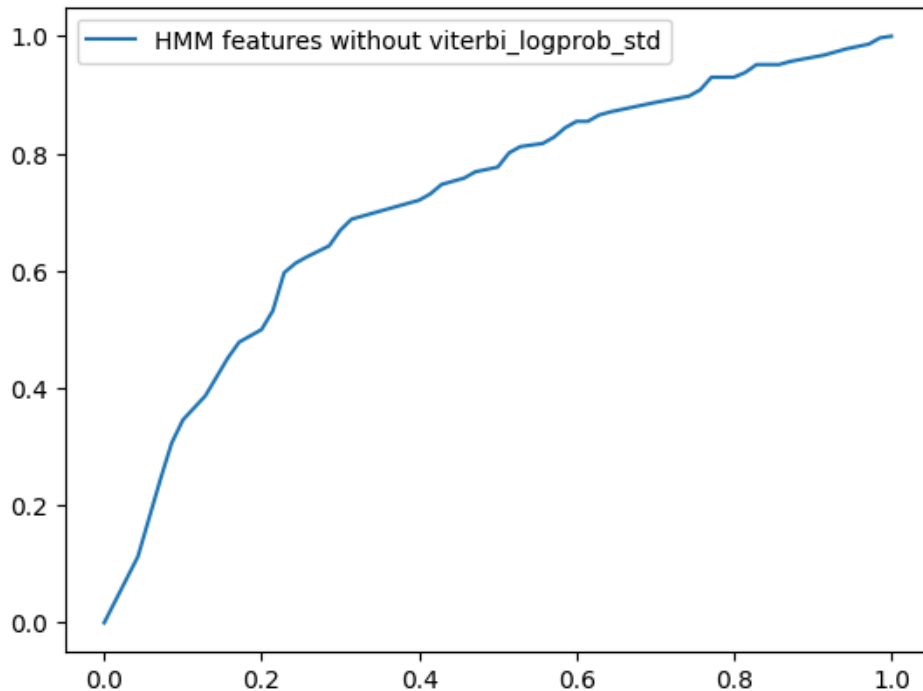


```
print(classification_report(y_pred_without_viterbi_logprob_std, y_test_without_viterbi_logprob_std))
```

	precision	recall	f1-score	support
0	0.24	0.53	0.33	32
1	0.92	0.76	0.83	224
accuracy			0.73	256
macro avg	0.58	0.65	0.58	256
weighted avg	0.83	0.73	0.77	256

```
#overall accuracy:
print((y_pred_without_viterbi_logprob_std==y_test_without_viterbi_logprob_std).sum()/len(y_pred_without_vi
```

```
#fpr_algn, tpr_algn, thresholds_algn = roc_curve(y_test, -1*X_test, pos_label='PCOS')
#sns.lineplot(x=fpr_algn, y=tpr_algn, label='HMM features only', errorbar=None)
```

⮕ 0.734375

˅ without viterbi_logprob_median

```
HMM_features = ['viterbi_logprob_mean',
       'viterbi_logprob_min', 'viterbi_logprob_max', 'viterbi_logprob_std',
       'complete_logprob_mean',
       'complete_logprob_min', 'complete_logprob_max', 'complete_logprob_std',
       'complete_logprob_median']


print('Performance with HMM features _without_viterbi_logprob_median ')

X_train_without_viterbi_logprob_median, X_test_without_viterbi_logprob_median, y_train_without_viterbi_log
                                             shuffle=True, random_state=51)
```

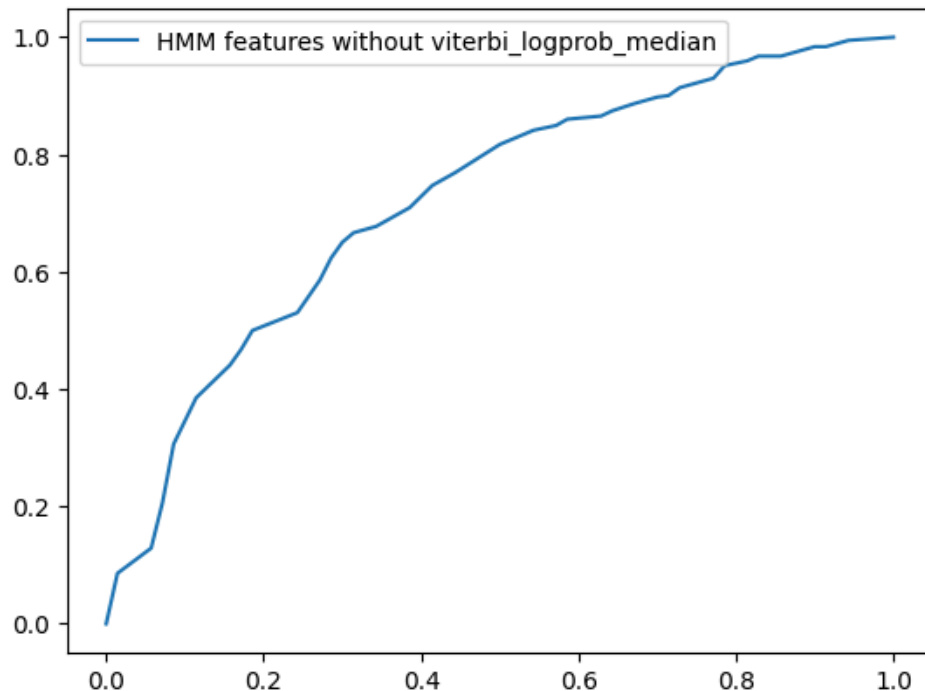⮕ Performance with HMM features _without_viterbi_logprob_median

```
# Impute missing values using SimpleImputer
imputer = SimpleImputer(strategy='mean') # Replace 'mean' with other strategies if needed
X_train_without_viterbi_logprob_median = imputer.fit_transform(X_train_without_viterbi_logprob_median)
X_test_without_viterbi_logprob_median = imputer.transform(X_test_without_viterbi_logprob_median)


clf = RFC(random_state=101)
clf.fit(X_train_without_viterbi_logprob_median, y_train_without_viterbi_logprob_median)
y_pred_without_viterbi_logprob_median = clf.predict(X_test_without_viterbi_logprob_median)
y_score_without_viterbi_logprob_median = clf.predict_proba(X_test_without_viterbi_logprob_median)
print(confusion_matrix(y_test_without_viterbi_logprob_median, y_pred_without_viterbi_logprob_median, norma
fpr_without_viterbi_logprob_median, tpr_without_viterbi_logprob_median, thresholds_without_viterbi_logprob
sns.lineplot(x=fpr_without_viterbi_logprob_median, y=tpr_without_viterbi_logprob_median, label='HMM featur
#plt.savefig('/content/drive/MyDrive/fall_research/feature distribution plots/viterbi adjusted plots/xgb_w
```

```
[[0.27142857 0.72857143]
 [0.08602151 0.91397849]]
<Axes: >
```



```python
print(classification_report(y_pred_without_viterbi_logprob_median, y_test_without_viterbi_logprob_median))
```

```
              precision    recall  f1-score   support

           0       0.27      0.54      0.36        35
           1       0.91      0.77      0.84       221

    accuracy                           0.74       256
   macro avg       0.59      0.66      0.60       256
weighted avg       0.83      0.74      0.77       256
```

```python
#overall accuracy:
print((y_pred_without_viterbi_logprob_median==y_test_without_viterbi_logprob_median).sum()/len(y_pred_with
#fpr_algn, tpr_algn, thresholds_algn = roc_curve(y_test, -1*X_test, pos_label='PCOS')
#sns.lineplot(x=fpr_algn, y=tpr_algn, label='HMM features only', errorbar=None)
```

```
0.73828125
```

without complete_logprob_mean

```python
HMM_features = ['viterbi_logprob_mean',
       'viterbi_logprob_min', 'viterbi_logprob_max', 'viterbi_logprob_std',
       'viterbi_logprob_median',
       'complete_logprob_min', 'complete_logprob_max', 'complete_logprob_std',
       'complete_logprob_median']
```

```python
print('Performance with HMM features _without_complete_logprob_mean ')
```

```
X_train_without_complete_logprob_mean, X_test_without_complete_logprob_mean, y_train_without_complete_logp
                                      shuffle=True, random_state=51)
```

Performance with HMM features _without_complete_logprob_mean

```
# Impute missing values using SimpleImputer
imputer = SimpleImputer(strategy='mean') # Replace 'mean' with other strategies if needed
X_train_without_complete_logprob_mean = imputer.fit_transform(X_train_without_complete_logprob_mean)
X_test_without_complete_logprob_mean = imputer.transform(X_test_without_complete_logprob_mean)


clf = RFC(random_state=101)
clf.fit(X_train_without_complete_logprob_mean, y_train_without_complete_logprob_mean)
y_pred_without_complete_logprob_mean = clf.predict(X_test_without_complete_logprob_mean)
y_score_without_complete_logprob_mean = clf.predict_proba(X_test_without_complete_logprob_mean)
print(confusion_matrix(y_test_without_complete_logprob_mean, y_pred_without_complete_logprob_mean, normali
fpr_without_complete_logprob_mean, tpr_without_complete_logprob_mean, thresholds_without_complete_logprob_
sns.lineplot(x=fpr_without_complete_logprob_mean, y=tpr_without_complete_logprob_mean, label='HMM features
#plt.savefig('/content/drive/MyDrive/fall_research/feature distribution plots/viterbi adjusted plots/xgb_w
```

```
[[0.24285714 0.75714286]
 [0.09677419 0.90322581]]
<Axes: >
```
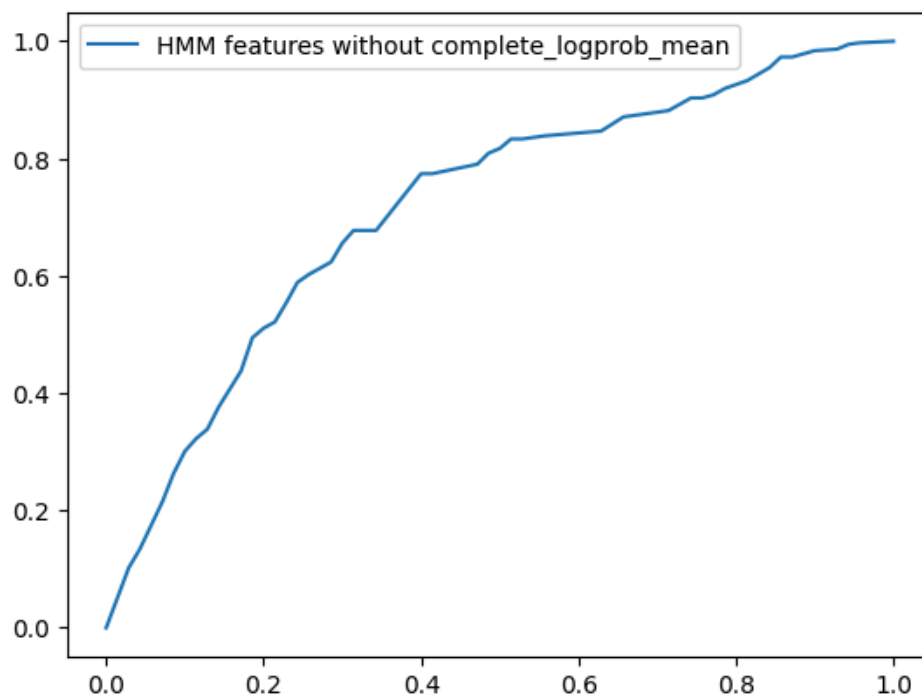


```
print(classification_report(y_pred_without_complete_logprob_mean, y_test_without_complete_logprob_mean))
```

```
              precision    recall  f1-score   support

           0       0.24      0.49      0.32        35
           1       0.90      0.76      0.83       221

    accuracy                           0.72       256
   macro avg       0.57      0.62      0.57       256
weighted avg       0.81      0.72      0.76       256
```

```
#overall accuracy:
print((y_pred_without_complete_logprob_mean==y_test_without_complete_logprob_mean).sum()/len(y_pred_withou
#fpr_algn, tpr_algn, thresholds_algn = roc_curve(y_test, -1*X_test, pos_label='PCOS')
#sns.lineplot(x=fpr_algn, y=tpr_algn, label='HMM features only', errorbar=None)
```

⮞ 0.72265625

⌄ without complete_logprob_min

```
HMM_features = ['viterbi_logprob_mean',
        'viterbi_logprob_min', 'viterbi_logprob_max', 'viterbi_logprob_std',
        'viterbi_logprob_median', 'complete_logprob_mean',
        'complete_logprob_max', 'complete_logprob_std',
        'complete_logprob_median']


print('Performance with HMM features _without_complete_logprob_min ')

X_train_without_complete_logprob_min, X_test_without_complete_logprob_min, y_train_without_complete_logpro
                                            shuffle=True, random_state=51)
```

⮞ Performance with HMM features _without_complete_logprob_min

```
# Impute missing values using SimpleImputer
imputer = SimpleImputer(strategy='mean') # Replace 'mean' with other strategies if needed
X_train_without_complete_logprob_min = imputer.fit_transform(X_train_without_complete_logprob_min)
X_test_without_complete_logprob_min = imputer.transform(X_test_without_complete_logprob_min)


clf = RFC(random_state=101)
clf.fit(X_train_without_complete_logprob_min, y_train_without_complete_logprob_min)
y_pred_without_complete_logprob_min = clf.predict(X_test_without_complete_logprob_min)
y_score_without_complete_logprob_min = clf.predict_proba(X_test_without_complete_logprob_min)
print(confusion_matrix(y_test_without_complete_logprob_min, y_pred_without_complete_logprob_min, normalize
fpr_without_complete_logprob_min, tpr_without_complete_logprob_min, thresholds_without_complete_logprob_mi
sns.lineplot(x=fpr_without_complete_logprob_min, y=tpr_without_complete_logprob_min, label='HMM features w
#plt.savefig('/content/drive/MyDrive/fall_research/feature distribution plots/viterbi adjusted plots/xgb_w
```

```
[[0.27142857 0.72857143]
 [0.07526882 0.92473118]]
<Axes: >
```
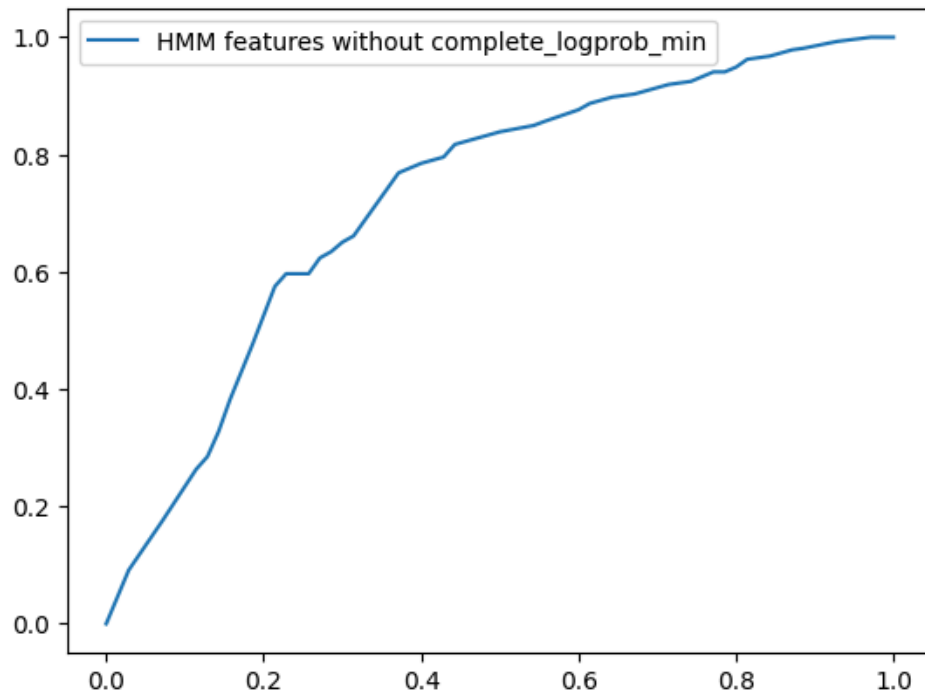


```
print(classification_report(y_pred_without_complete_logprob_min, y_test_without_complete_logprob_min))
```

	precision	recall	f1-score	support
0	0.27	0.58	0.37	33
1	0.92	0.77	0.84	223
accuracy			0.75	256
macro avg	0.60	0.67	0.61	256
weighted avg	0.84	0.75	0.78	256

```
#overall accuracy:
print((y_pred_without_complete_logprob_min==y_test_without_complete_logprob_min).sum()/len(y_pred_without_
#fpr_algn, tpr_algn, thresholds_algn = roc_curve(y_test, -1*X_test, pos_label='PCOS')
#sns.lineplot(x=fpr_algn, y=tpr_algn, label='HMM features only', errorbar=None)
```

```
0.74609375
```

⌄ without complete_logprob_max

```
HMM_features = ['viterbi_logprob_mean',
    'viterbi_logprob_min', 'viterbi_logprob_max', 'viterbi_logprob_std',
    'viterbi_logprob_median', 'complete_logprob_mean',
    'complete_logprob_min', 'complete_logprob_std',
    'complete_logprob_median']
```

```
print('Performance with HMM features _without_complete_logprob_max ')
```
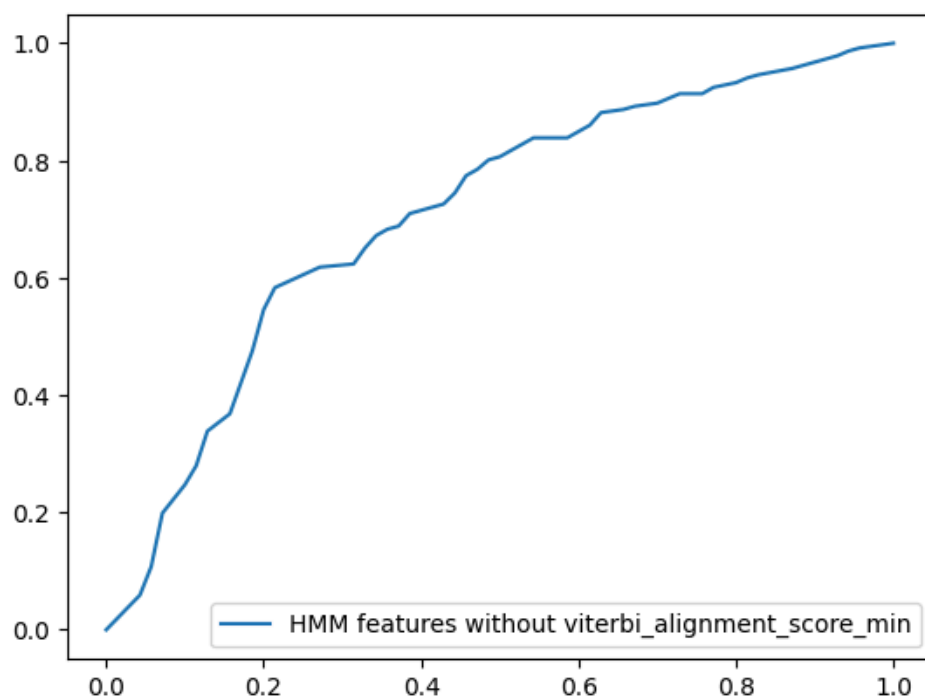
```
X_train_without_complete_logprob_max, X_test_without_complete_logprob_max, y_train_without_complete_logpro
                                                shuffle=True, random_state=51)
```

⇥ Performance with HMM features _without_complete_logprob_max

```
# Impute missing values using SimpleImputer
imputer = SimpleImputer(strategy='mean') # Replace 'mean' with other strategies if needed
X_train_without_complete_logprob_max = imputer.fit_transform(X_train_without_complete_logprob_max)
X_test_without_complete_logprob_max = imputer.transform(X_test_without_complete_logprob_max)
```

```
clf = RFC(random_state=101)
clf.fit(X_train_without_complete_logprob_max, y_train_without_complete_logprob_max)
y_pred_without_complete_logprob_max = clf.predict(X_test_without_complete_logprob_max)
y_score_without_complete_logprob_max = clf.predict_proba(X_test_without_complete_logprob_max)
print(confusion_matrix(y_test_without_complete_logprob_max, y_pred_without_complete_logprob_max, normalize
fpr_without_complete_logprob_max, tpr_without_complete_logprob_max, thresholds_without_complete_logprob_ma
sns.lineplot(x=fpr_without_complete_logprob_max, y=tpr_without_complete_logprob_max, label='HMM features w
#plt.savefig('/content/drive/MyDrive/fall_research/feature distribution plots/viterbi adjusted plots/xgb_w
```

⇥ [[0.27142857 0.72857143]
   [0.08602151 0.91397849]]
   <Axes: >


```
print(classification_report(y_pred_without_complete_logprob_max, y_test_without_complete_logprob_max))
```

⇥

	precision	recall	f1-score	support
0	0.27	0.54	0.36	35
1	0.91	0.77	0.84	221
accuracy			0.74	256
macro avg	0.59	0.66	0.60	256
weighted avg	0.83	0.74	0.77	256

```python
#overall accuracy:
print((y_pred_without_complete_logprob_max==y_test_without_complete_logprob_max).sum()/len(y_pred_without_
#fpr_algn, tpr_algn, thresholds_algn = roc_curve(y_test, -1*X_test, pos_label='PCOS')
#sns.lineplot(x=fpr_algn, y=tpr_algn, label='HMM features only', errorbar=None)
```

    0.73828125

without complete_logprob_std

```python
HMM_features = ['viterbi_logprob_mean',
        'viterbi_logprob_min', 'viterbi_logprob_max', 'viterbi_logprob_std',
        'viterbi_logprob_median', 'complete_logprob_mean',
        'complete_logprob_min', 'complete_logprob_max',
        'complete_logprob_median']

print('Performance with HMM features _without_complete_logprob_std ')

X_train_without_complete_logprob_std, X_test_without_complete_logprob_std, y_train_without_complete_logpro
                                                shuffle=True, random_state=51)
```

   Performance with HMM features _without_complete_logprob_std

```python
# Impute missing values using SimpleImputer
imputer = SimpleImputer(strategy='mean') # Replace 'mean' with other strategies if needed
X_train_without_complete_logprob_std = imputer.fit_transform(X_train_without_complete_logprob_std)
X_test_without_complete_logprob_std = imputer.transform(X_test_without_complete_logprob_std)


clf = RFC(random_state=101)
clf.fit(X_train_without_complete_logprob_std, y_train_without_complete_logprob_std)
y_pred_without_complete_logprob_std = clf.predict(X_test_without_complete_logprob_std)
y_score_without_complete_logprob_std = clf.predict_proba(X_test_without_complete_logprob_std)
print(confusion_matrix(y_test_without_complete_logprob_std, y_pred_without_complete_logprob_std, normalize
fpr_without_complete_logprob_std, tpr_without_complete_logprob_std, thresholds_without_complete_logprob_st
sns.lineplot(x=fpr_without_complete_logprob_std, y=tpr_without_complete_logprob_std, label='HMM features w
#plt.savefig('/content/drive/MyDrive/fall_research/feature distribution plots/viterbi adjusted plots/xgb_w
```

```
[[0.24285714 0.75714286]
 [0.07526882 0.92473118]]
<Axes: >
```
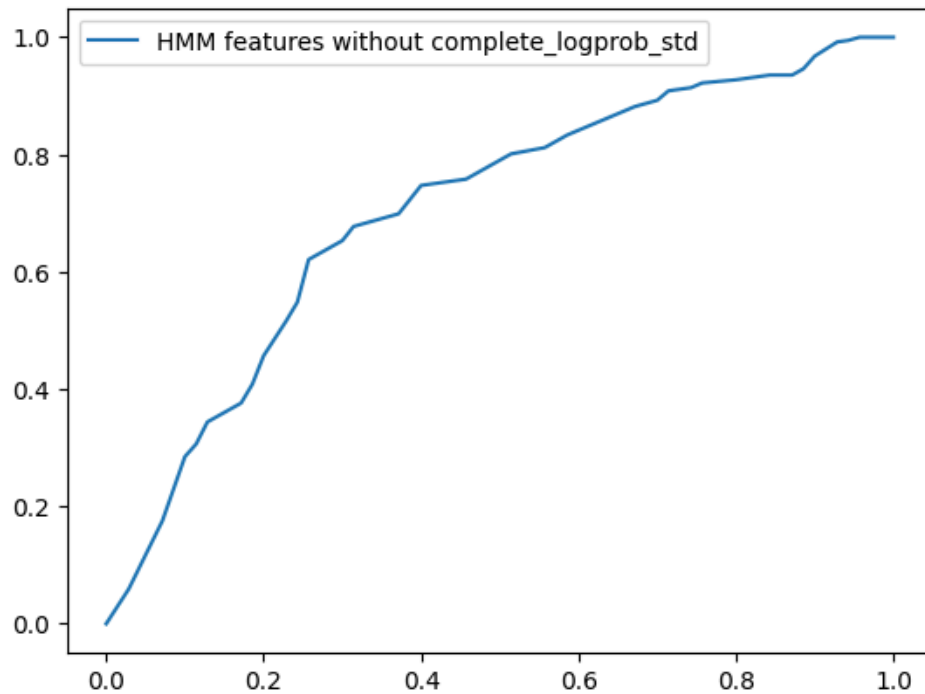


```
print(classification_report(y_pred_without_complete_logprob_std, y_test_without_complete_logprob_std))
```

	precision	recall	f1-score	support
0	0.24	0.55	0.34	31
1	0.92	0.76	0.84	225
accuracy			0.74	256
macro avg	0.58	0.66	0.59	256
weighted avg	0.84	0.74	0.78	256

```
#overall accuracy:
print((y_pred_without_complete_logprob_std==y_test_without_complete_logprob_std).sum()/len(y_pred_without_
#fpr_algn, tpr_algn, thresholds_algn = roc_curve(y_test, -1*X_test, pos_label='PCOS')
#sns.lineplot(x=fpr_algn, y=tpr_algn, label='HMM features only', errorbar=None)
```

```
0.73828125
```

⌄ without complete_logprob_median

```
HMM_features = ['viterbi_logprob_mean',
       'viterbi_logprob_min', 'viterbi_logprob_max', 'viterbi_logprob_std',
       'viterbi_logprob_median', 'complete_logprob_mean',
       'complete_logprob_min', 'complete_logprob_max', 'complete_logprob_std']
```
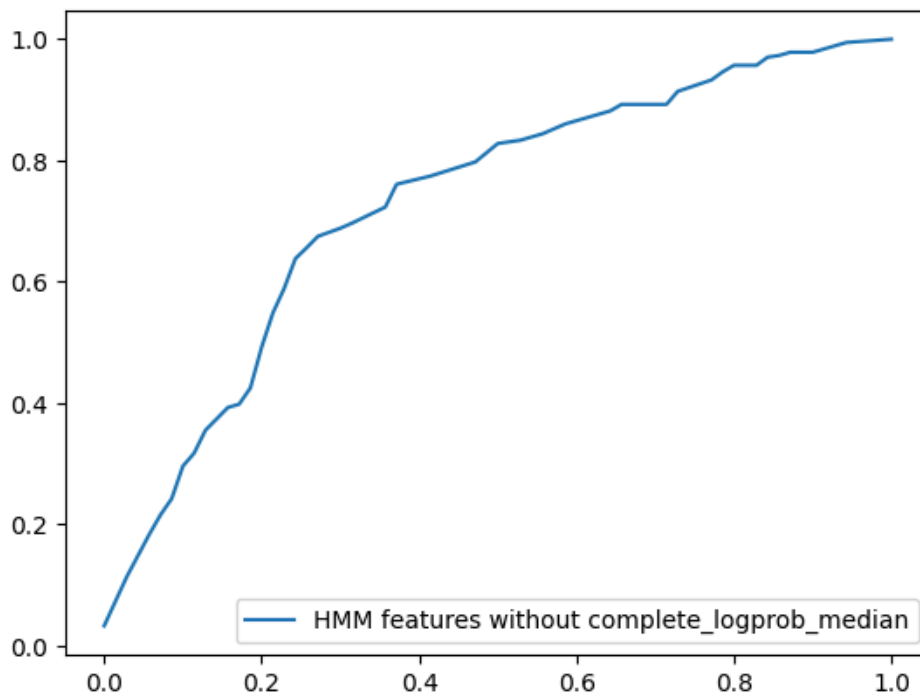
```
print('Performance with HMM features _without_viterbi_logprob_median ')

X train without viterbi logprob median, X test without viterbi logprob median, y train without viterbi log
# Impute missing values using SimpleImputer
imputer = SimpleImputer(strategy='mean') # Replace 'mean' with other strategies if needed
X_train_without_viterbi_logprob_median = imputer.fit_transform(X_train_without_viterbi_logprob_median)
X_test_without_viterbi_logprob_median = imputer.transform(X_test_without_viterbi_logprob_median)


clf = RFC(random_state=101)
clf.fit(X_train_without_viterbi_logprob_median, y_train_without_viterbi_logprob_median)
y_pred_without_viterbi_logprob_median = clf.predict(X_test_without_viterbi_logprob_median)
y_score_without_viterbi_logprob_median = clf.predict_proba(X_test_without_viterbi_logprob_median)
print(confusion_matrix(y_test_without_viterbi_logprob_median, y_pred_without_viterbi_logprob_median, norma
fpr_without_viterbi_logprob_median, tpr_without_viterbi_logprob_median, thresholds_without_viterbi_logprob
sns.lineplot(x=fpr_without_viterbi_logprob_median, y=tpr_without_viterbi_logprob_median, label='HMM featur
#plt.savefig('/content/drive/MyDrive/fall_research/feature distribution plots/viterbi adjusted plots/xgb_w
```

```
[[0.22857143 0.77142857]
 [0.06989247 0.93010753]]
<Axes: >
```



```
print(classification_report(y_pred_without_viterbi_logprob_median, y_test_without_viterbi_logprob_median))
```

```
              precision    recall  f1-score   support

           0       0.23      0.55      0.32        29
           1       0.93      0.76      0.84       227

    accuracy                           0.74       256
   macro avg       0.58      0.66      0.58       256
weighted avg       0.85      0.74      0.78       256
```

```
#overall accuracy:
print((y_pred_without_viterbi_logprob_median==y_test_without_viterbi_logprob_median).sum()/len(y_pred_with
```