Big Data

For the project you need to identify a suitable dataset, implement three different dimension reduction techniques and train four different machine learning models using the dimension-reduced data and the actual data.

Dataset

I chose the AIDS Virus Infection Prediction dataset available at: https://www.kaggle.com/datasets/aadarshvelu/aids-virus-infection-prediction/data .

Dataset contains healthcare statistics and categorical information about patients who have been diagnosed with AIDS. This dataset was initially published in 1996.

- Attribute Information :

time: time to failure or censoring

trt: treatment indicator (0 = ZDV only; 1 = ZDV + ddI, 2 = ZDV + Zal, 3 = ddI only)

age: age (yrs) at baseline

wtkg: weight (kg) at baseline

hemo: hemophilia (0=no, 1=yes)

homo: homosexual activity (0=no, 1=yes)

drugs: history of IV drug use (0=no, 1=yes)

karnof: Karnofsky score (on a scale of 0-100)

oprior: Non-ZDV antiretroviral therapy pre-175 (0=no, 1=yes)

z30: ZDV in the 30 days prior to 175 (0=no, 1=yes)

preanti: days pre-175 anti-retroviral therapy

race: race (0=White, 1=non-white)

gender: gender (0=F, 1=M)

str2: antiretroviral history (0=naive, 1=experienced)

strat: antiretroviral history stratification (1='Antiretroviral Naive',2='> 1 but <= 52 weeks of prior antiretroviral therapy',3='> 52 weeks)

symptom: symptomatic indicator (0=asymp, 1=symp)

treat: treatment indicator (0=ZDV only, 1=others)

offtrt: indicator of off-trt before 96+/-5 weeks (0=no,1=yes)

cd40: CD4 at baseline

cd420: CD4 at 20+/-5 weeks

cd80: CD8 at baseline

cd820: CD8 at 20+/-5 weeks

infected: is infected with AIDS (0=No, 1=Yes)

- Additional Variable Information :

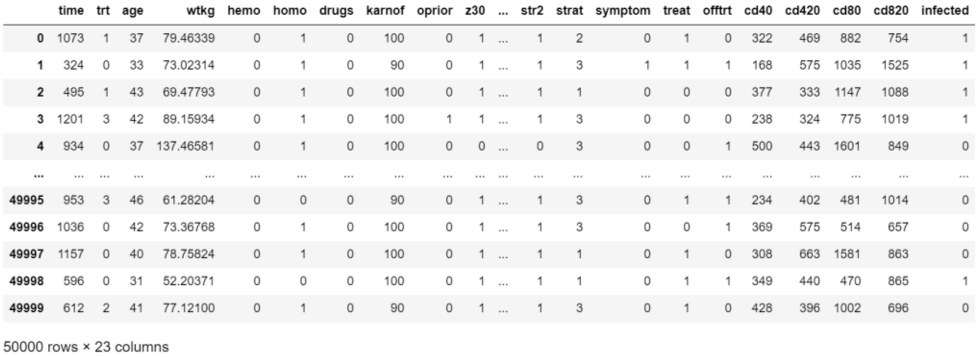
Personal information (age, weight, race, gender, sexual activity)

Medical history (hemophilia, history of IV drugs)

Treatment history (ZDV/non-ZDV treatment history)

Lab results (CD4/CD8 counts)

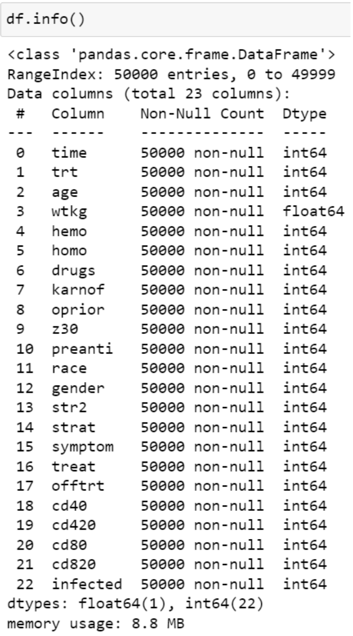
The dataset used has 50000 entries and 23 columns.



	time	trt	age	wtkg	hemo	homo	drugs	karnof	oprior	z30	...	str2	strat	symptom	treat	offtrt	cd40	cd420	cd80	cd820	infected
0	1073	1	37	79.46339	0	1	0	100	0	1	...	1	2	0	1	0	322	469	882	754	1
1	324	0	33	73.02314	0	1	0	90	0	1	...	1	3	1	1	1	168	575	1035	1525	1
2	495	1	43	69.47793	0	1	0	100	0	1	...	1	1	0	0	0	377	333	1147	1088	1
3	1201	3	42	89.15934	0	1	0	100	1	1	...	1	3	0	0	0	238	324	775	1019	1
4	934	0	37	137.46581	0	1	0	100	0	0	...	0	3	0	0	1	500	443	1601	849	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
49995	953	3	46	61.28204	0	0	0	90	0	1	...	1	3	0	1	1	234	402	481	1014	0
49996	1036	0	42	73.36768	0	1	0	100	0	1	...	1	3	0	0	1	369	575	514	657	0
49997	1157	0	40	78.75824	0	1	0	100	0	1	...	1	1	0	1	0	308	663	1581	863	0
49998	596	0	31	52.20371	0	0	0	100	0	1	...	1	1	0	1	1	349	440	470	865	1
49999	612	2	41	77.12100	0	1	0	90	0	1	...	1	3	0	1	0	428	396	1002	696	0

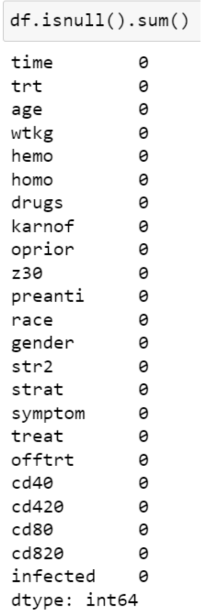
50000 rows × 23 columns

There are all int variables beside a float and there are no null values.



```
df.info()
```

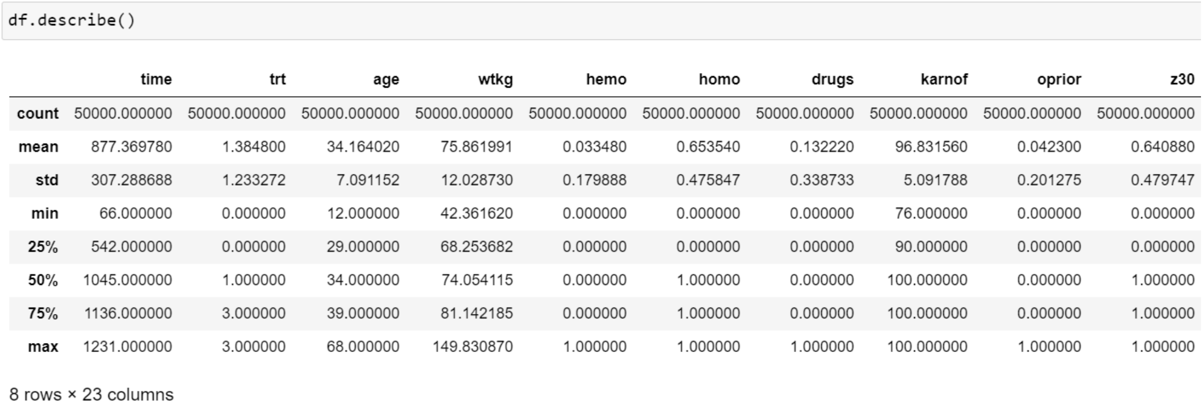
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 23 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   time      50000 non-null  int64  
 1   trt       50000 non-null  int64  
 2   age       50000 non-null  int64  
 3   wtkg      50000 non-null  float64
 4   hemo      50000 non-null  int64  
 5   homo      50000 non-null  int64  
 6   drugs     50000 non-null  int64  
 7   karnof    50000 non-null  int64  
 8   oprior    50000 non-null  int64  
 9   z30       50000 non-null  int64  
 10  preanti   50000 non-null  int64  
 11  race      50000 non-null  int64  
 12  gender    50000 non-null  int64  
 13  str2      50000 non-null  int64  
 14  strat     50000 non-null  int64  
 15  symptom   50000 non-null  int64  
 16  treat     50000 non-null  int64  
 17  offtrt    50000 non-null  int64  
 18  cd40      50000 non-null  int64  
 19  cd420     50000 non-null  int64  
 20  cd80      50000 non-null  int64  
 21  cd820     50000 non-null  int64  
 22  infected  50000 non-null  int64  
dtypes: float64(1), int64(22)
memory usage: 8.8 MB
```



```
df.isnull().sum()
```

```
time        0
trt         0
age         0
wtkg        0
hemo        0
homo        0
drugs       0
karnof      0
oprior      0
z30         0
preanti     0
race        0
gender      0
str2        0
strat       0
symptom     0
treat       0
offtrt      0
cd40        0
cd420       0
cd80        0
cd820       0
infected    0
dtype: int64
```

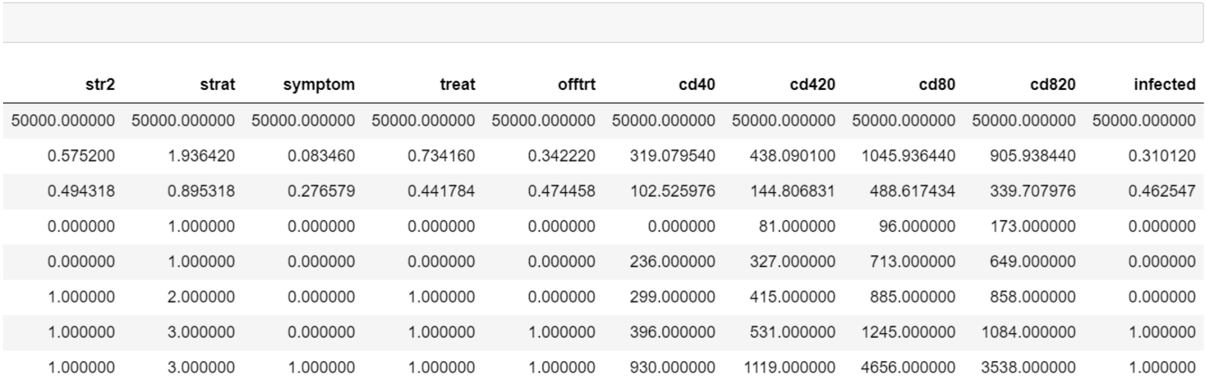
I used **`df.describe()`** `for` statistics that give a quick overview of the central tendency, dispersion, and shape of the distribution of the data in each column



```
df.describe()
```

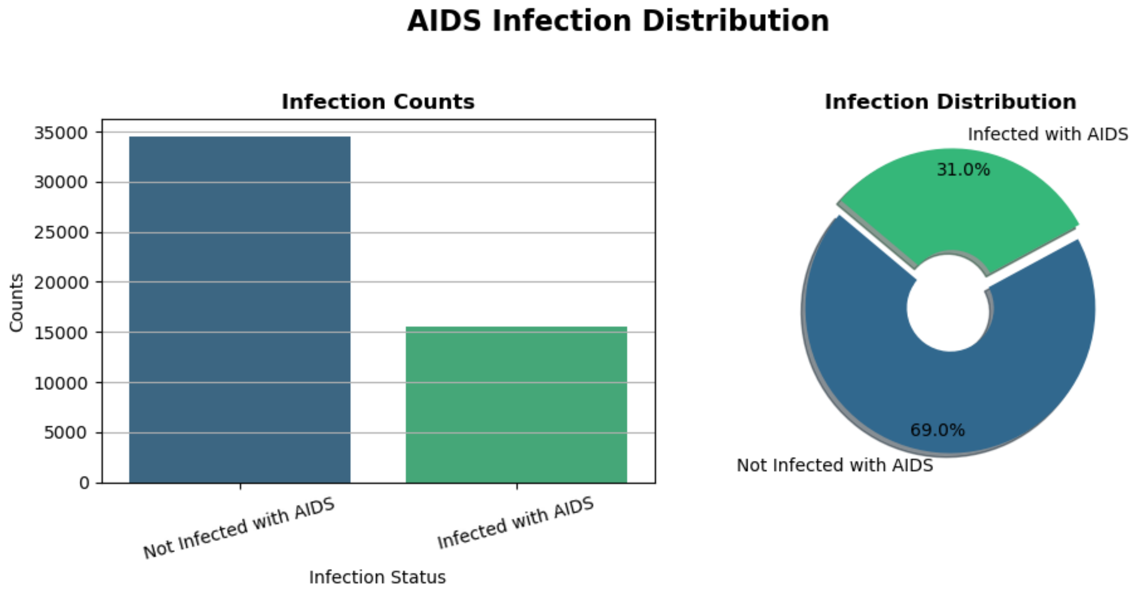
	time	trt	age	wtkg	hemo	homo	drugs	karnof	oprior	z30
count	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000
mean	877.369780	1.384800	34.164020	75.861991	0.033480	0.653540	0.132220	96.831560	0.042300	0.640880
std	307.288688	1.233272	7.091152	12.028730	0.179888	0.475847	0.338733	5.091788	0.201275	0.479747
min	66.000000	0.000000	12.000000	42.361620	0.000000	0.000000	0.000000	76.000000	0.000000	0.000000
25%	542.000000	0.000000	29.000000	68.253682	0.000000	0.000000	0.000000	90.000000	0.000000	0.000000
50%	1045.000000	1.000000	34.000000	74.054115	0.000000	1.000000	0.000000	100.000000	0.000000	1.000000
75%	1136.000000	3.000000	39.000000	81.142185	0.000000	1.000000	0.000000	100.000000	0.000000	1.000000
max	1231.000000	3.000000	68.000000	149.830870	1.000000	1.000000	1.000000	100.000000	1.000000	1.000000

8 rows × 23 columns



str2	strat	symptom	treat	offtrt	cd40	cd420	cd80	cd820	infected
50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000
0.575200	1.936420	0.083460	0.734160	0.342220	319.079540	438.090100	1045.936440	905.938440	0.310120
0.494318	0.895318	0.276579	0.441784	0.474458	102.525976	144.806831	488.617434	339.707976	0.462547
0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	81.000000	96.000000	173.000000	0.000000
0.000000	1.000000	0.000000	0.000000	0.000000	236.000000	327.000000	713.000000	649.000000	0.000000
1.000000	2.000000	0.000000	1.000000	0.000000	299.000000	415.000000	885.000000	858.000000	0.000000
1.000000	3.000000	0.000000	1.000000	1.000000	396.000000	531.000000	1245.000000	1084.000000	1.000000
1.000000	3.000000	1.000000	1.000000	1.000000	930.000000	1119.000000	4656.000000	3538.000000	1.000000

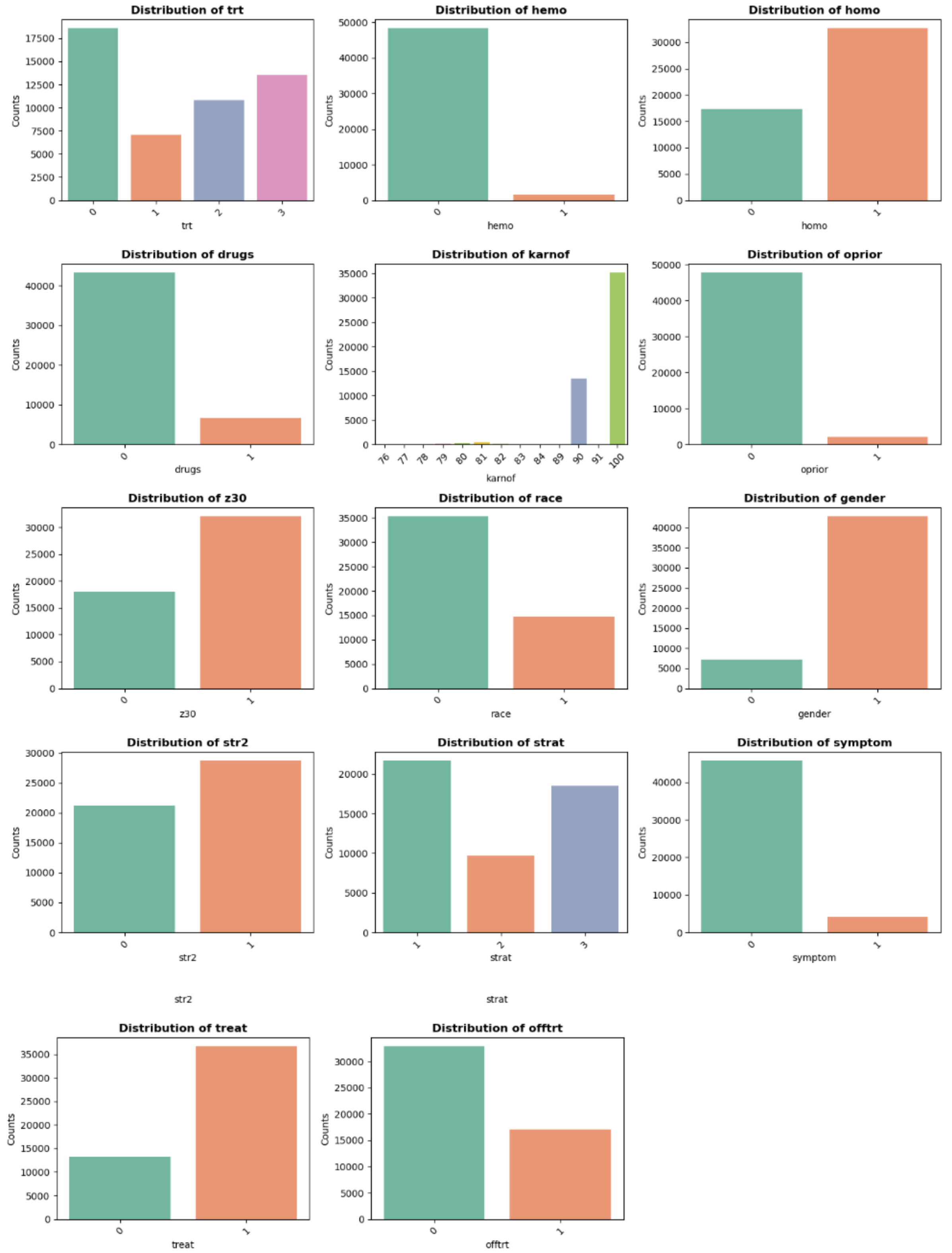
Categorical Variable



69% of people are infected in the dataset.

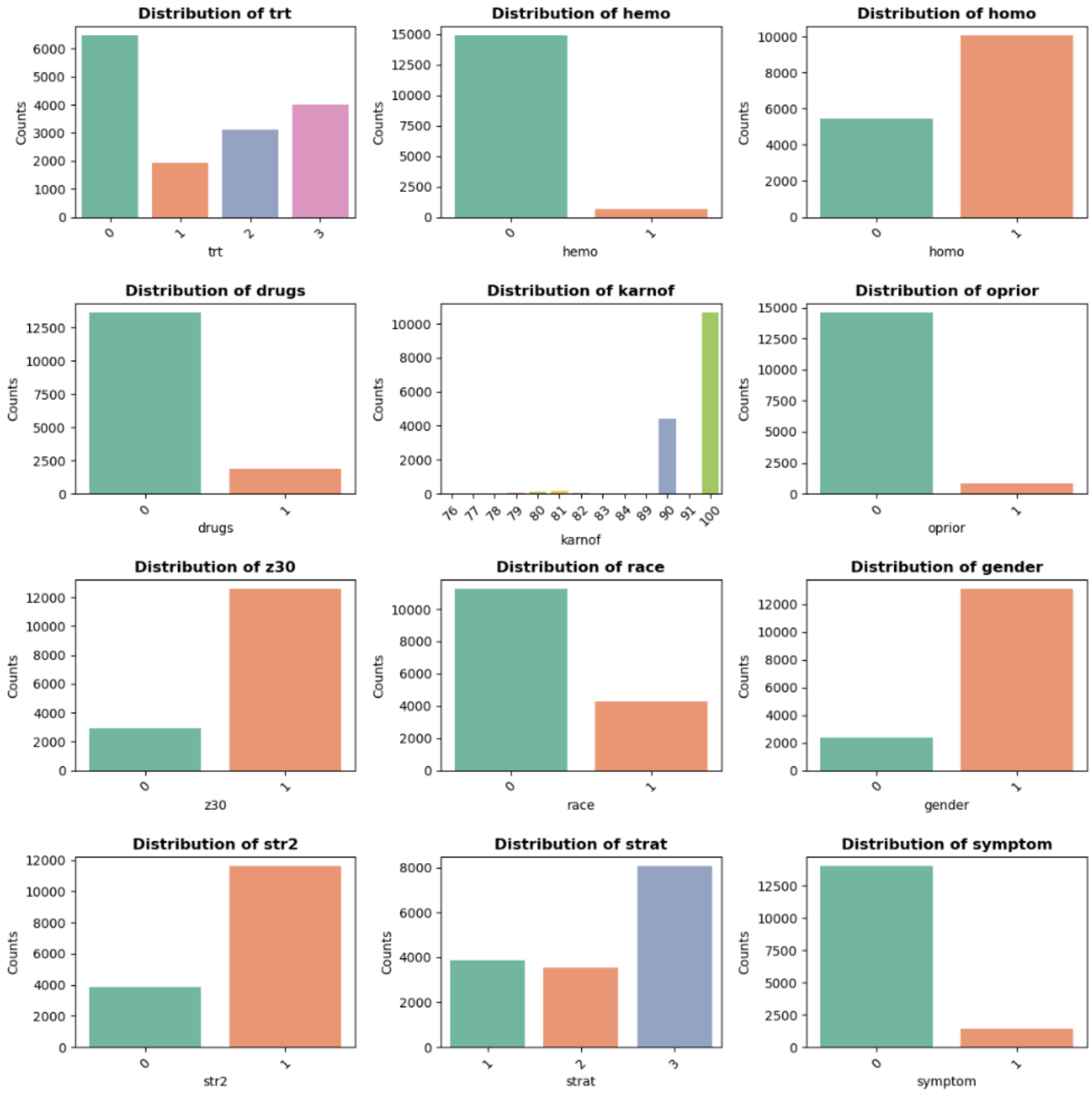
The dataset is unbalanced so I will duplicate some rows in order to have the classes more balanced.

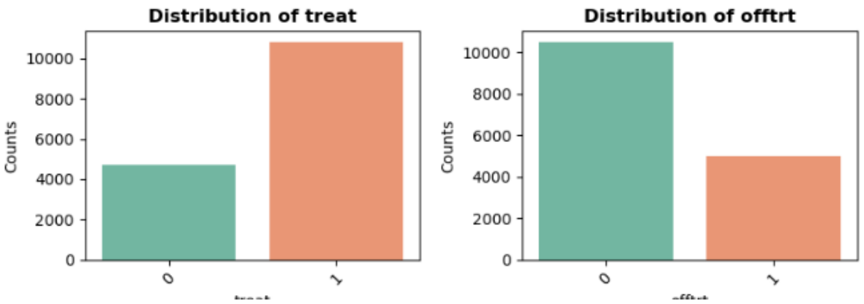
Overall distribution:



Distribution of trt
Counts
17500
15000
12500
10000
7500
5000
2500
0
0
1
2
3
trt
Distribution of hemo
Counts
50000
40000
30000
20000
10000
0
0
1
hemo
Distribution of homo
Counts
30000
25000
20000
15000
10000
5000
0
0
1
homo
Distribution of drugs
Counts
40000
30000
20000
10000
0
0
1
drugs
Distribution of karnof
Counts
35000
30000
25000
20000
15000
10000
5000
0
76
77
78
79
80
81
82
83
84
89
90
91
100
karnof
Distribution of oprior
Counts
50000
40000
30000
20000
10000
0
0
1
oprior
Distribution of z30
Counts
30000
25000
20000
15000
10000
5000
0
0
1
z30
Distribution of race
Counts
35000
30000
25000
20000
15000
10000
5000
0
0
1
race
Distribution of gender
Counts
40000
35000
30000
25000
20000
15000
10000
5000
0
0
1
gender
Distribution of str2
Counts
30000
25000
20000
15000
10000
5000
0
0
1
str2
Distribution of strat
Counts
20000
15000
10000
5000
0
1
2
3
strat
Distribution of symptom
Counts
40000
30000
20000
10000
0
0
1
symptom
str2
strat
Distribution of treat
Counts
35000
30000
25000
20000
15000
10000
5000
0
0
1
treat
Distribution of offtrt
Counts
30000
25000
20000
15000
10000
5000
0
0
1
offtrt

Filter when infected is 1





-Treatment Indicator (trt):

Most infected individuals received ZDV only (0).

A significant portion received ddI only (3).

Fewer individuals received combinations of ZDV with ddI (1) or Zal (2).

-Hemophilia (hemo):

Almost all infected individuals do not have hemophilia (0).

-Homosexual Activity (homo):

A large proportion of infected individuals have a history of homosexual activity (1).

-History of IV Drug Use (drugs):

The majority of infected individuals do not have a history of IV drug use (0).

-Karnofsky Score (karnof):

The Karnofsky score is concentrated at higher values, indicating better physical function among infected individuals.

-Non-ZDV Antiretroviral Therapy (oprior):

Most infected individuals have not received non-ZDV antiretroviral therapy before the study (0).

-ZDV in the 30 Days Prior to the Study (z30):

Most infected individuals received ZDV in the 30 days prior to the study (1).

-Race (race):

A significant proportion of infected individuals are non-white (1).

-Gender (gender):

Most infected individuals are male (1).

-Antiretroviral History (str2):

Most infected individuals have experience with antiretroviral therapy (1).

-Antiretroviral History Stratification (strat):

A majority of infected individuals have more than 52 weeks of prior antiretroviral therapy.

Fewer individuals have between 1 and 52 weeks of prior antiretroviral therapy.

Very few are antiretroviral naive.

-Symptomatic Indicator (symptom):

Most infected individuals are asymptomatic (0).

Treatment Indicator (treat):

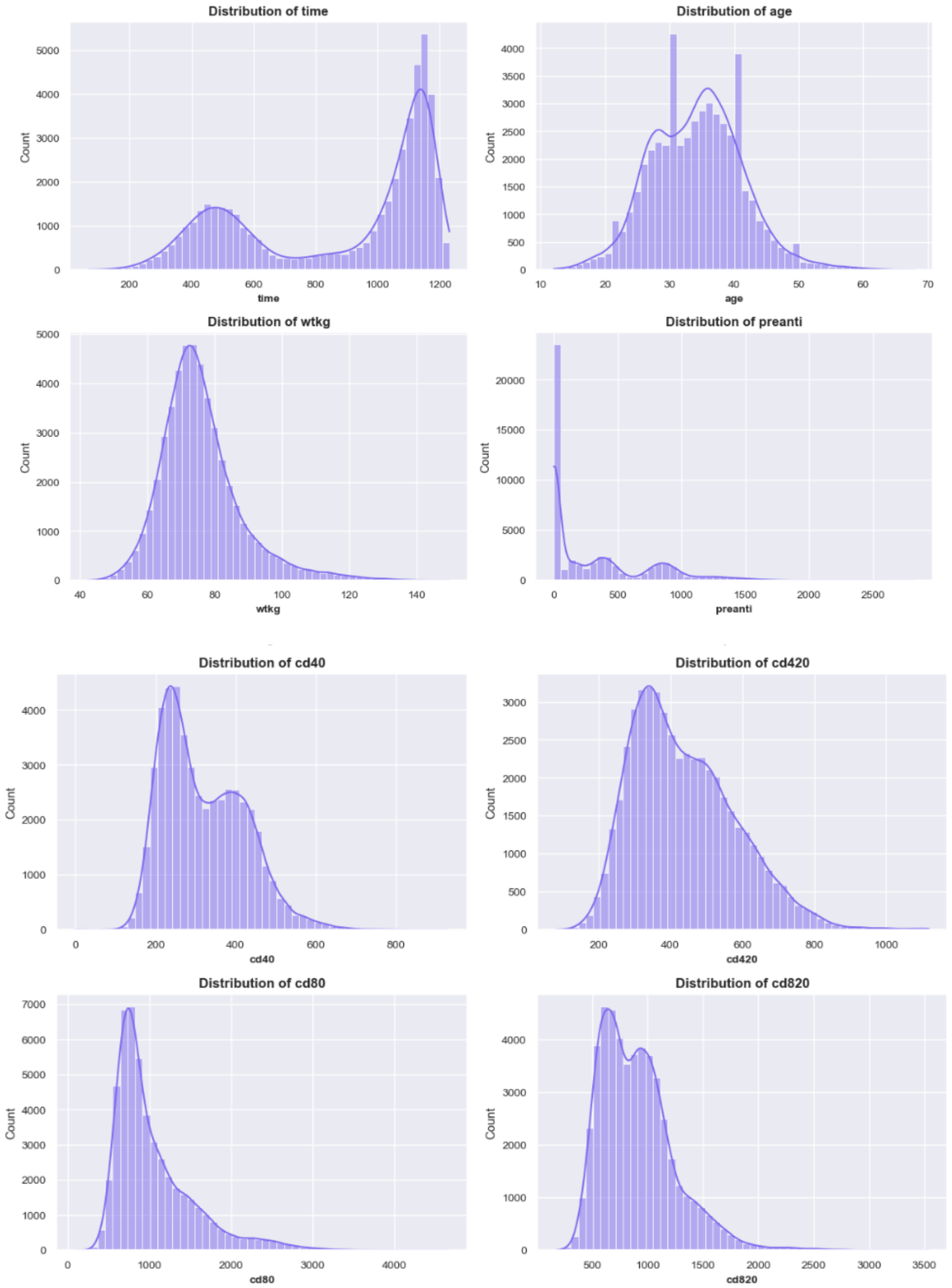
Most infected individuals received treatments other than ZDV only

-Indicator of Off-Treatment Before 96+/-5 Weeks (offtrt):

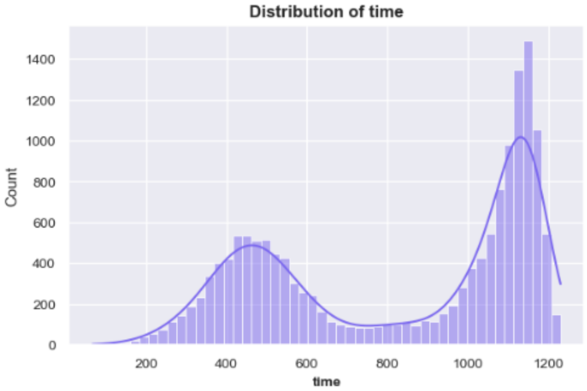
A significant proportion of infected individuals were not off-treatment before 96+/-5 weeks (0).

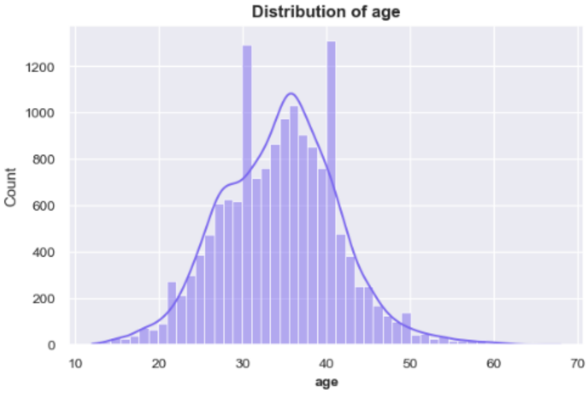
Numerical Variables

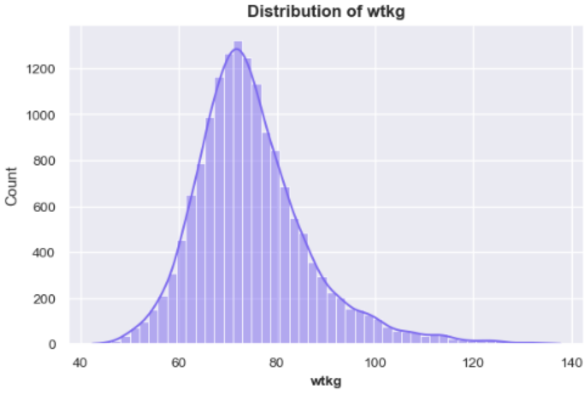
Overall values:

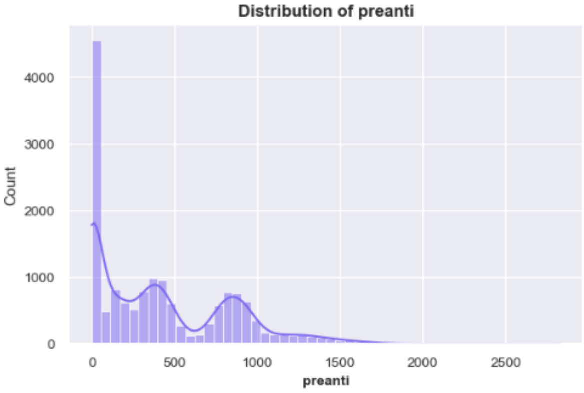


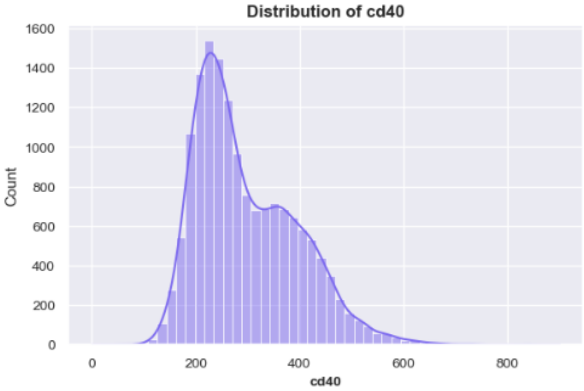
When infected is 1:

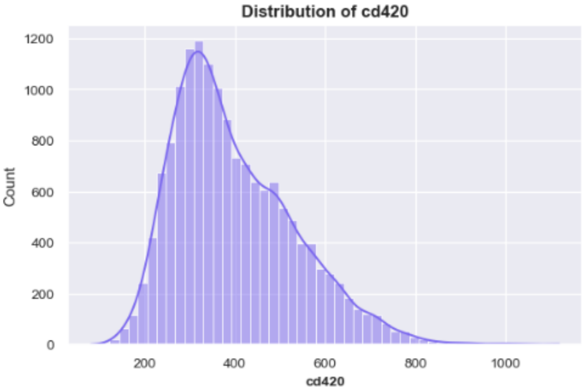


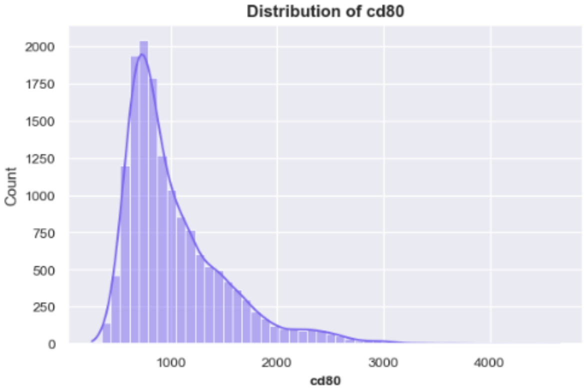


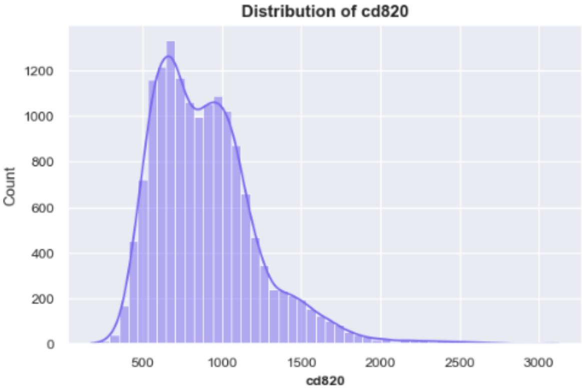












-Time (time):

There are two peaks in the distribution: one around 400-600 days and another around 1000-1200 days.

This suggests that the time to failure or censoring has two distinct groups, possibly indicating different stages or responses to treatment.

-Age (age):

The age distribution is roughly normal, centered around 30-40 years.

Most infected individuals are in their 30s, with fewer individuals at younger and older ages.

-Weight (wtkg):

The weight distribution is also roughly normal, centered around 70-80 kg.

There is a wide range of weights, indicating diversity in the physical health of the individuals.

-Pre-Anti-Retroviral Therapy Days (preanti):

Most individuals have a preanti value close to zero, indicating no or minimal pre-study antiretroviral therapy.

A small number of individuals have higher values, indicating prior extensive therapy.

-CD4 Count at Baseline (cd40):

The CD4 count at baseline is skewed towards lower values, with most individuals having counts between 100-400.

This indicates a weakened immune system in most infected individuals at baseline.

-CD4 Count at 20+/-5 Weeks (cd420):

Similar to the baseline CD4 count, the distribution is centered around 300-400, with fewer individuals having higher counts.

This indicates a general trend of CD4 count remaining within this range post-treatment.

-CD8 Count at Baseline (cd80):

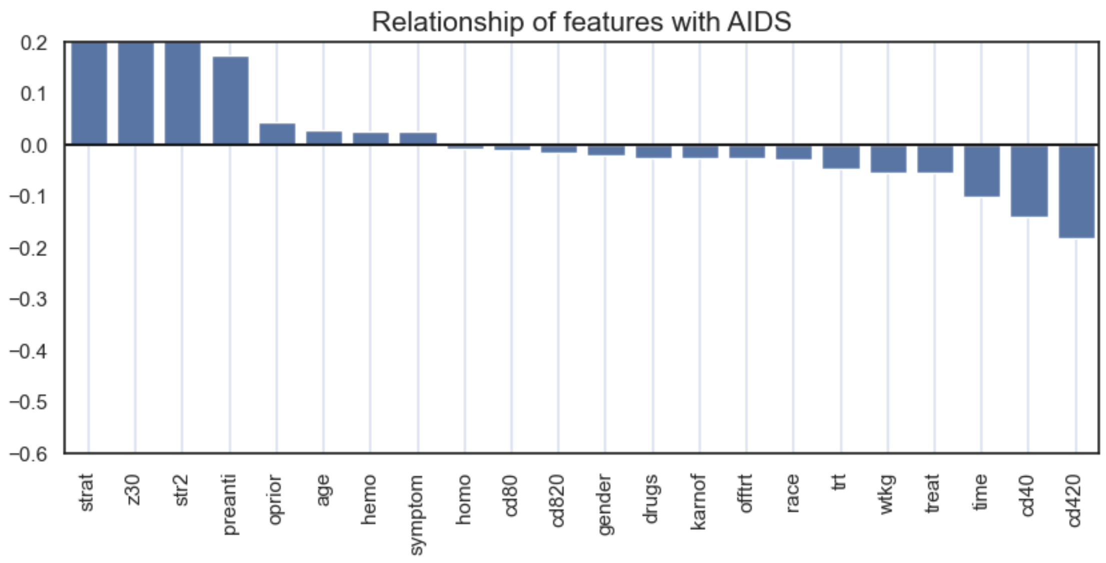
The CD8 count at baseline is also skewed, with most individuals having counts between 500-2000.

This suggests a broad range of immune response among individuals.

-CD8 Count at 20+/-5 Weeks (cd820):

The distribution is similar to the baseline CD8 count, centered around 500-1500.

This suggests that the CD8 count remains relatively stable post-treatment.



Positive Correlations:

strat: Higher values of stratification are positively correlated with being infected.

z30: Use of ZDV in the 30 days prior to the study is positively correlated with infection.

str2: Experience with antiretroviral therapy is positively correlated with infection.

preanti: More days of pre-study antiretroviral therapy are positively correlated with infection.

Negative Correlations:

cd420: Lower CD4 count at 20+/-5 weeks is strongly negatively correlated with infection.

cd40: Lower CD4 count at baseline is also strongly negatively correlated with infection.

time: Shorter time to failure or censoring is negatively correlated with infection.

treat: Treatment indicator is negatively correlated with infection.

wtkg: Lower weight is negatively correlated with infection.

trt: Specific treatment regimens are negatively correlated with infection.

offtrt: Being off-treatment before 96+/-5 weeks is negatively correlated with infection.

race: Non-white race is slightly negatively correlated with infection.

karnof: Lower Karnofsky score is slightly negatively correlated with infection.

drugs: History of IV drug use is slightly negatively correlated with infection.

gender: Gender is slightly negatively correlated with infection.

cd820: Lower CD8 count at 20+/-5 weeks is negatively correlated with infection.

cd80: Lower CD8 count at baseline is negatively correlated with infection.

homo: Homosexual activity is slightly negatively correlated with infection.

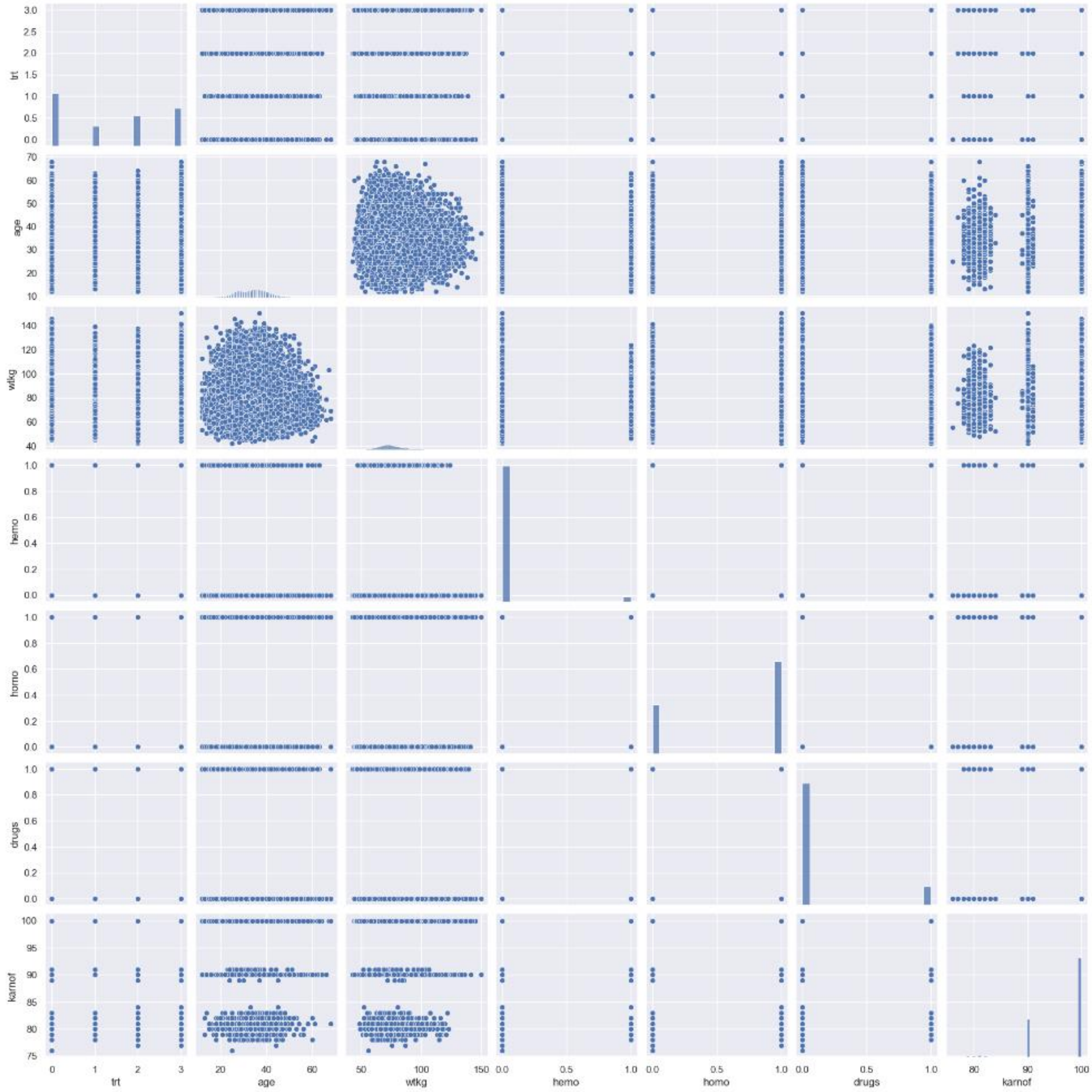
symptom: Being symptomatic is slightly negatively correlated with infection.

hemo: Hemophilia is slightly negatively correlated with infection.

age: Lower age is slightly negatively correlated with infection.

oprior: Non-ZDV antiretroviral therapy pre-study is slightly negatively correlated with infection.

Visual representation of relationships between pairs of features



Age vs. Weight (wtkg):

There appears to be a normal distribution for weight around 70-80 kg.

Age is centered around 30-40 years, indicating most participants are in this age group.

Karnofsky Score (karnof):

The Karnofsky score is heavily skewed towards higher values, indicating better physical function among participants.

Binary Features (hemo, homo, drugs, gender):

Most features like hemophilia (hemo), homosexual activity (homo), IV drug use (drugs), and gender (gender) are binary, showing clear separation between 0 and 1 values.

Treatment Indicator (trt):

The treatment indicator shows discrete values, indicating different treatment groups (ZDV only, ZDV + ddI, etc.).

Correlations and Clusters:

Some scatter plots show distinct clusters indicating possible correlations, especially among continuous variables like age and weight.

The binary variables show distinct separation but no clear correlation patterns in scatter plots.

Preprocess:

Drop the less relevant columns based on the correlation analysis:

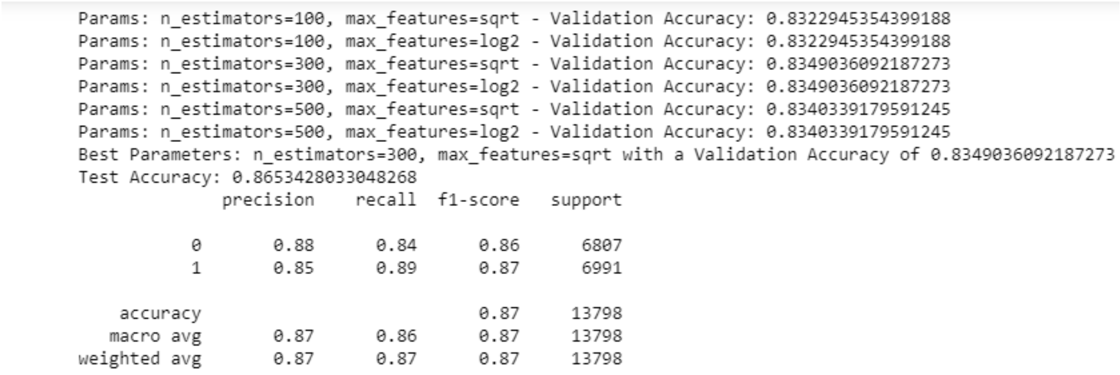
```
columns_to_drop = ['oprior', 'age', 'hemo', 'symptom', 'homo', 'cd80', 'cd820', 'gender', 'drugs', 'karnof', 'offtrt', 'race']
```

Normalizes the feature data using MinMaxScaler.

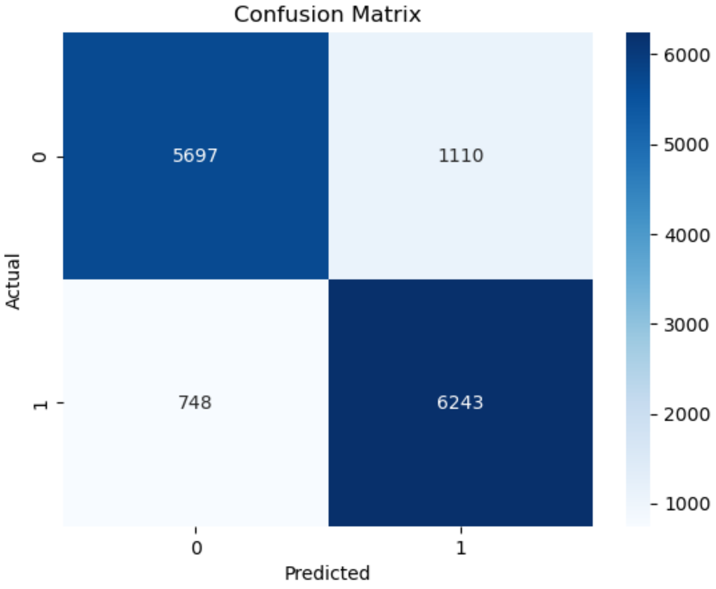
ML Models:

Random Forest

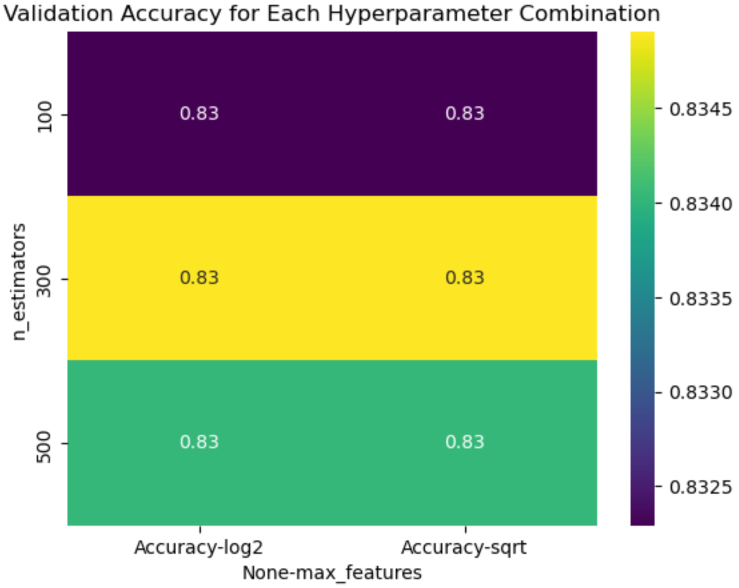
```
param_grid = {
    'n_estimators': [100, 300, 500],
    'max_features': ['sqrt', 'log2']
}
```



```
Params: n_estimators=100, max_features=sqrt - Validation Accuracy: 0.8322945354399188
Params: n_estimators=100, max_features=log2 - Validation Accuracy: 0.8322945354399188
Params: n_estimators=300, max_features=sqrt - Validation Accuracy: 0.8349036092187273
Params: n_estimators=300, max_features=log2 - Validation Accuracy: 0.8349036092187273
Params: n_estimators=500, max_features=sqrt - Validation Accuracy: 0.8340339179591245
Params: n_estimators=500, max_features=log2 - Validation Accuracy: 0.8340339179591245
Best Parameters: n_estimators=300, max_features=sqrt with a Validation Accuracy of 0.8349036092187273
Test Accuracy: 0.8653428033048268
              precision    recall  f1-score   support

           0       0.88      0.84      0.86      6807
           1       0.85      0.89      0.87      6991

    accuracy                           0.87     13798
   macro avg       0.87      0.86      0.87     13798
weighted avg       0.87      0.87      0.87     13798
```

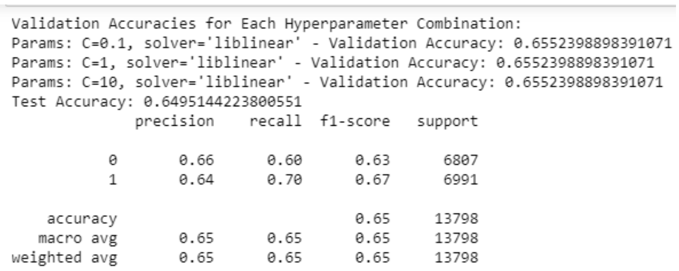


```
Negative Predictive Value (NPV): 0.8839410395655547
```

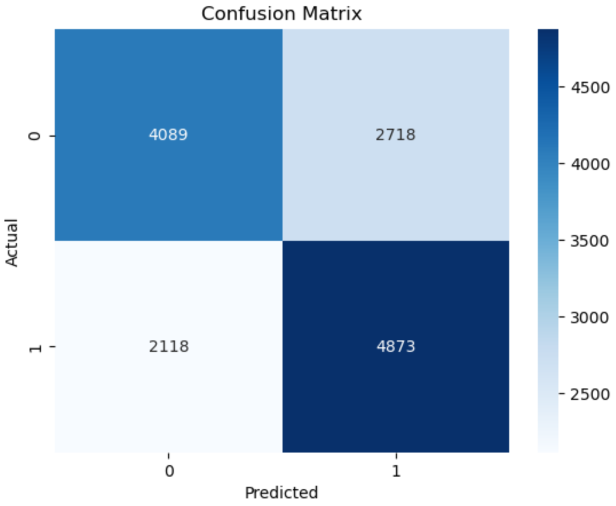


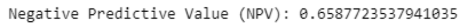
Logistic Regression

```
param_grid = {
    'C': [0.1, 1, 10],
    'solver': ['liblinear']
}
```

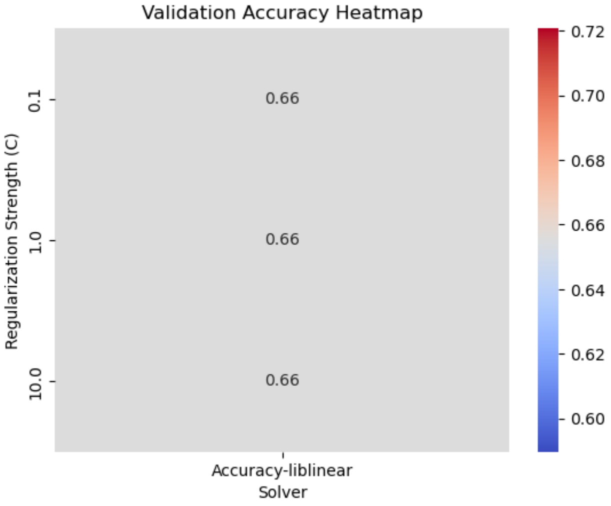


```
Validation Accuracies for Each Hyperparameter Combination:
Params: C=0.1, solver='liblinear' - Validation Accuracy: 0.6552398898391071
Params: C=1, solver='liblinear' - Validation Accuracy: 0.6552398898391071
Params: C=10, solver='liblinear' - Validation Accuracy: 0.6552398898391071
Test Accuracy: 0.6495144223800551
              precision    recall  f1-score   support

           0       0.66      0.60      0.63      6807
           1       0.64      0.70      0.67      6991

    accuracy                           0.65     13798
   macro avg       0.65      0.65      0.65     13798
weighted avg       0.65      0.65      0.65     13798
```



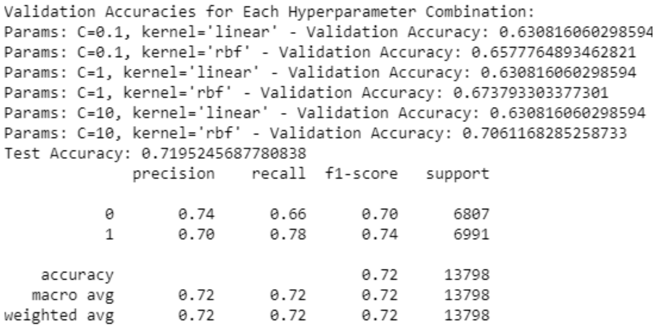


```
Negative Predictive Value (NPV): 0.6587723537941035
```

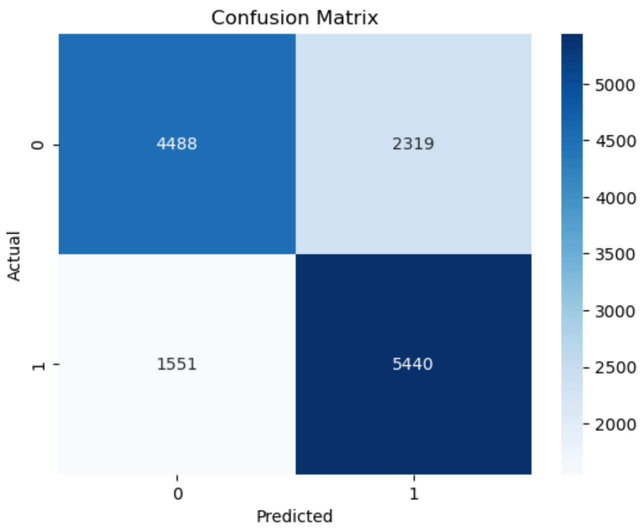


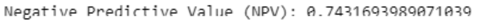
SVM

```
param_grid = {
    'C': [0.1, 1, 10],
    'kernel': ['linear', 'rbf']
}
```

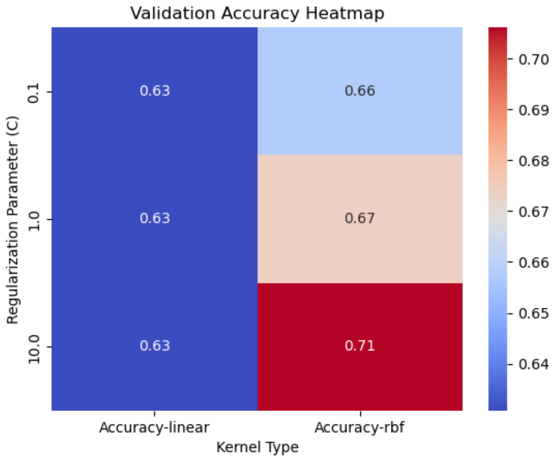


```
Validation Accuracies for Each Hyperparameter Combination:
Params: C=0.1, kernel='linear' - Validation Accuracy: 0.630816060298594
Params: C=0.1, kernel='rbf' - Validation Accuracy: 0.6577764893462821
Params: C=1, kernel='linear' - Validation Accuracy: 0.630816060298594
Params: C=1, kernel='rbf' - Validation Accuracy: 0.673793303377301
Params: C=10, kernel='linear' - Validation Accuracy: 0.630816060298594
Params: C=10, kernel='rbf' - Validation Accuracy: 0.7061168285258733
Test Accuracy: 0.7195245687780838
              precision    recall  f1-score   support

           0       0.74      0.66      0.70      6807
           1       0.70      0.78      0.74      6991

    accuracy                           0.72     13798
   macro avg       0.72      0.72      0.72     13798
weighted avg       0.72      0.72      0.72     13798
```



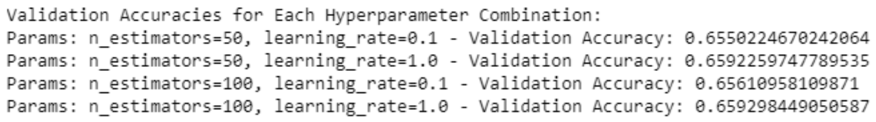


```
Negative Predictive Value (NPV): 0.7431693989071039
```

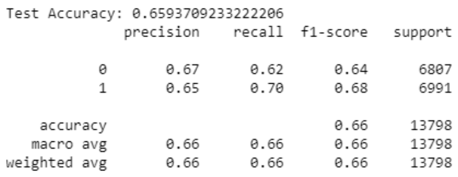


AdaBoost(decision trees with a depth of one)

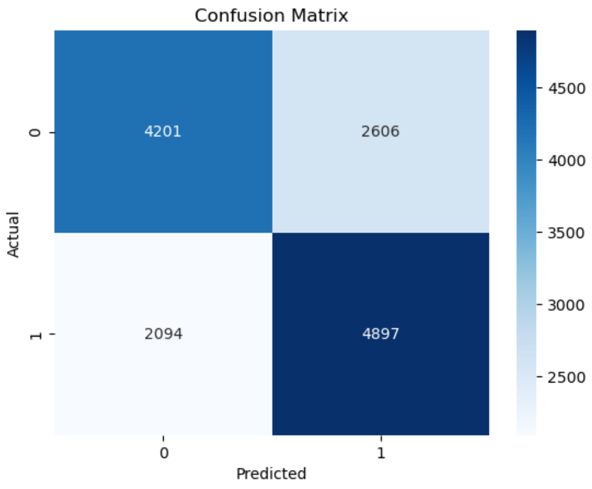
```
param_grid = {
    'n_estimators': [50, 100],
    'learning_rate': [0.1, 1.0]
}
```

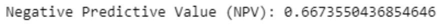


```
Validation Accuracies for Each Hyperparameter Combination:
Params: n_estimators=50, learning_rate=0.1 - Validation Accuracy: 0.6550224670242064
Params: n_estimators=50, learning_rate=1.0 - Validation Accuracy: 0.6592259747789535
Params: n_estimators=100, learning_rate=0.1 - Validation Accuracy: 0.65610958109871
Params: n_estimators=100, learning_rate=1.0 - Validation Accuracy: 0.659298449050587
```

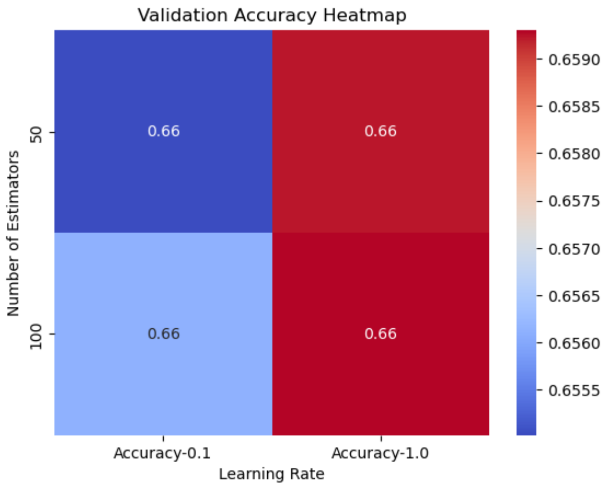


```
Test Accuracy: 0.6593709233222206
              precision    recall  f1-score   support

           0       0.67      0.62      0.64      6807
           1       0.65      0.70      0.68      6991

    accuracy                           0.66     13798
   macro avg       0.66      0.66      0.66     13798
weighted avg       0.66      0.66      0.66     13798
```





```
Negative Predictive Value (NPV): 0.6673550436854646
```



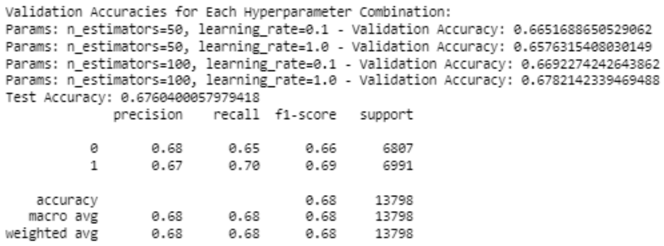
CatBoost

param_grid = {

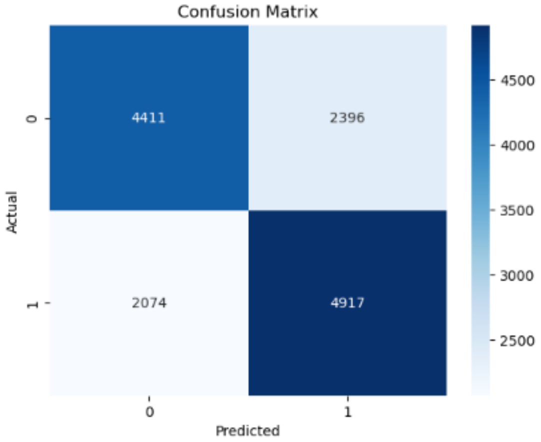
'n_estimators': [50, 100],

'learning_rate': [0.1, 1.0]

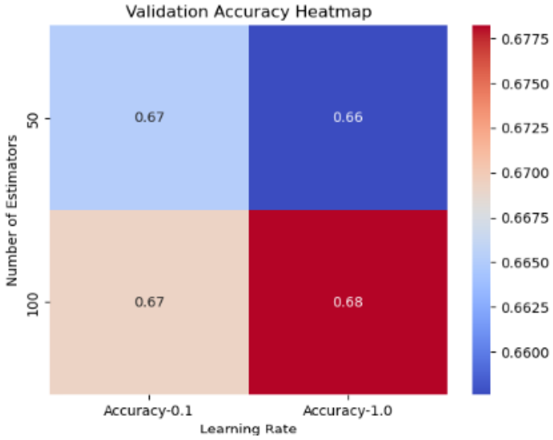
}



```
Validation Accuracies for Each Hyperparameter Combination:
Params: n_estimators=50, learning_rate=0.1 - Validation Accuracy: 0.6651688650529062
Params: n_estimators=50, learning_rate=1.0 - Validation Accuracy: 0.6576315408030149
Params: n_estimators=100, learning_rate=0.1 - Validation Accuracy: 0.6692274242643862
Params: n_estimators=100, learning_rate=1.0 - Validation Accuracy: 0.6782142339469488
Test Accuracy: 0.6760400057979418
              precision    recall  f1-score   support

           0       0.68      0.65      0.66      6807
           1       0.67      0.70      0.69      6991

    accuracy                           0.68     13798
   macro avg       0.68      0.68      0.68     13798
weighted avg       0.68      0.68      0.68     13798
```

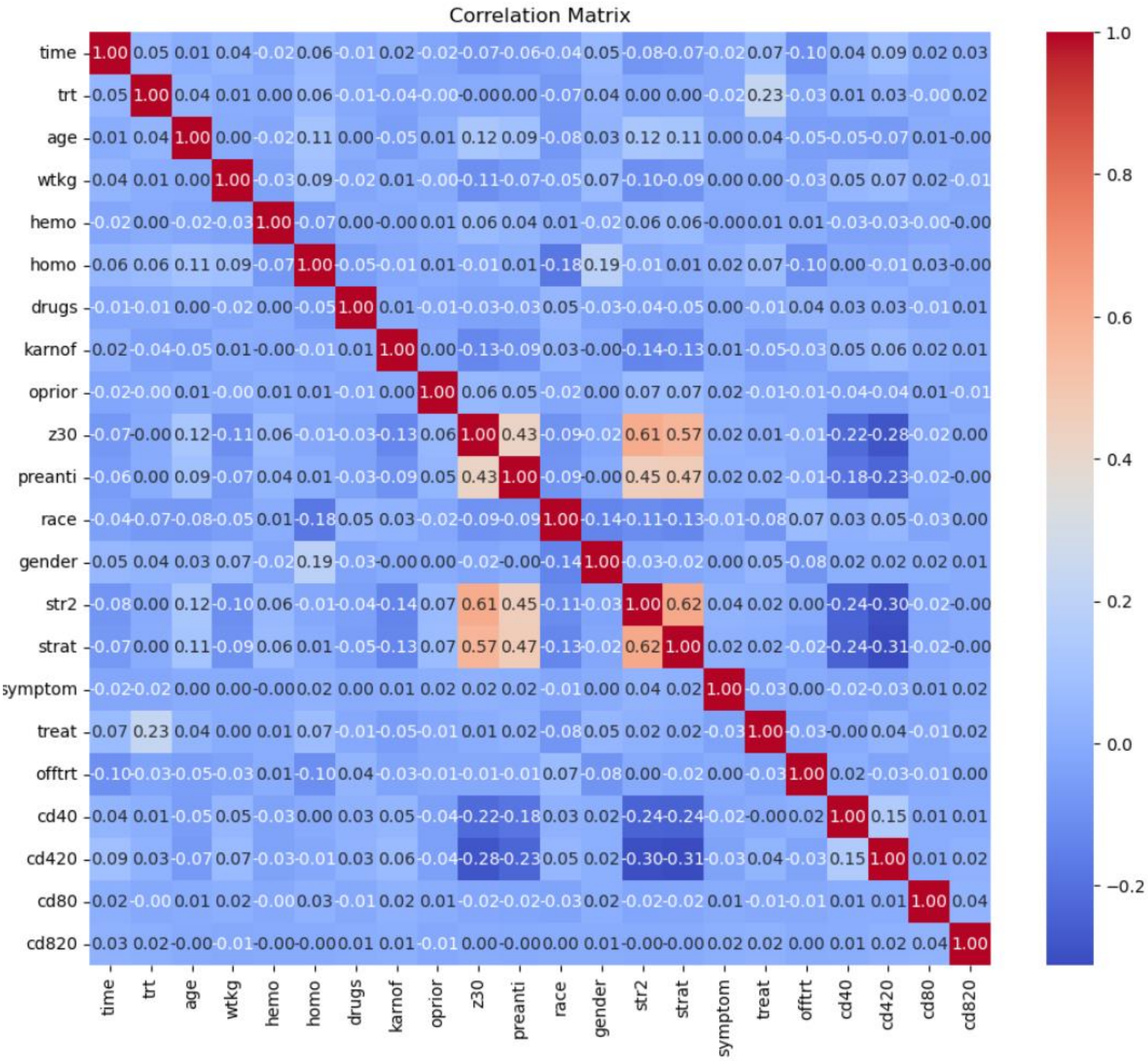


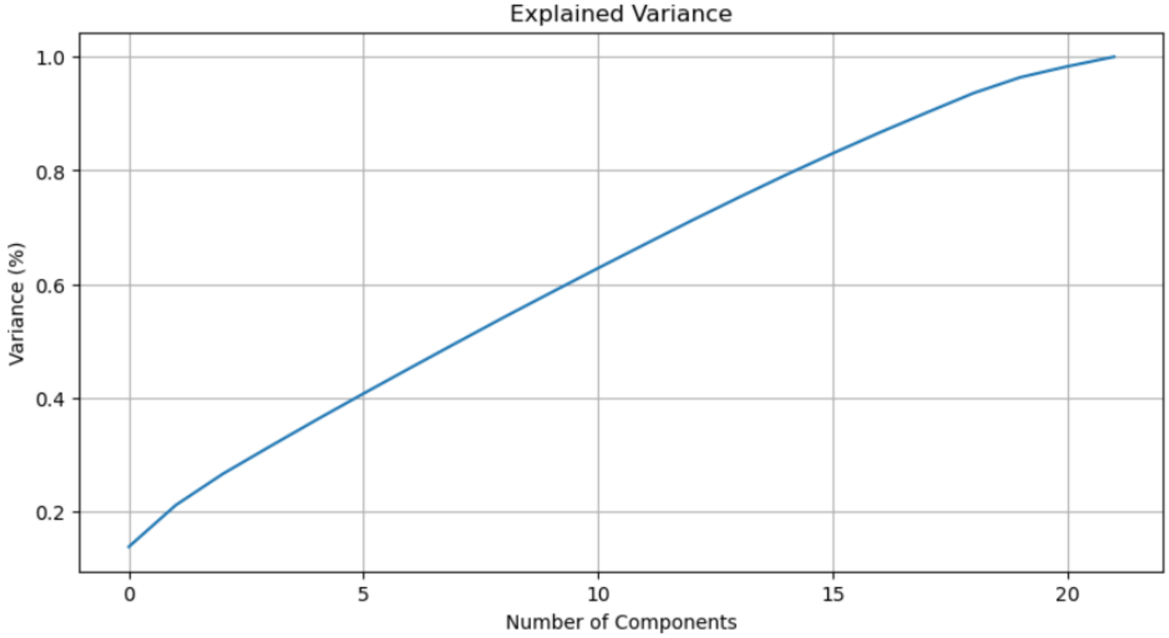
```
Negative Predictive Value (NPV): 0.6801850424055512
```



Dimension reduction

PCA

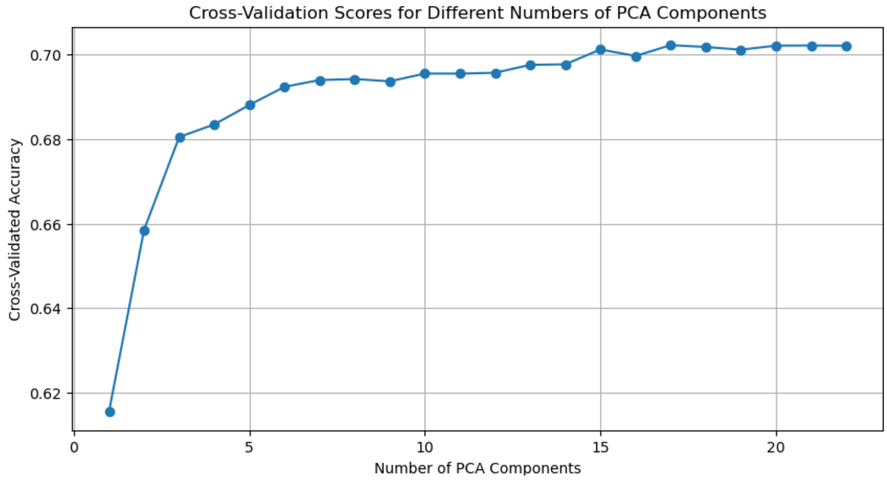




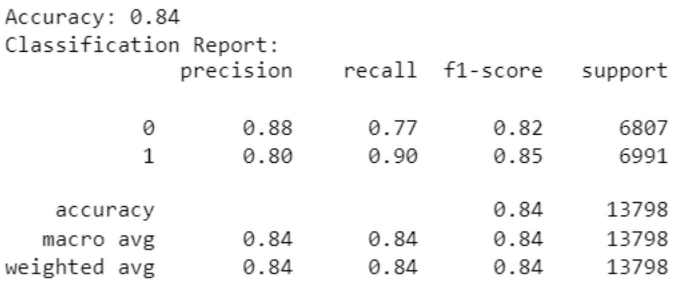
```
Number of components selected: 18
```

Random Forest

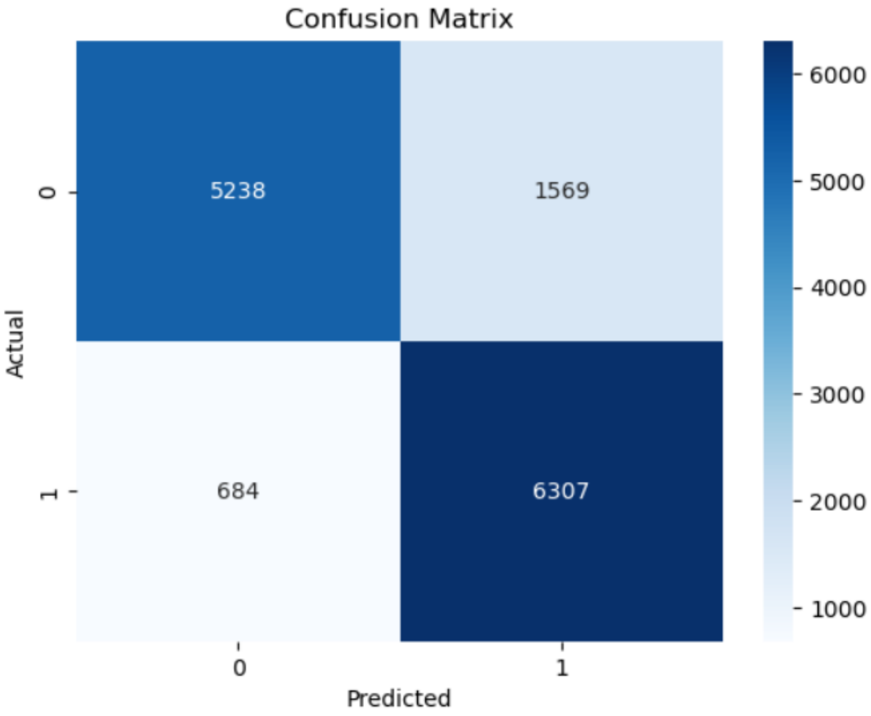
CV=5

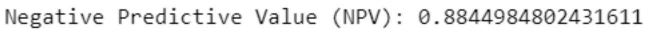


With 7 comp



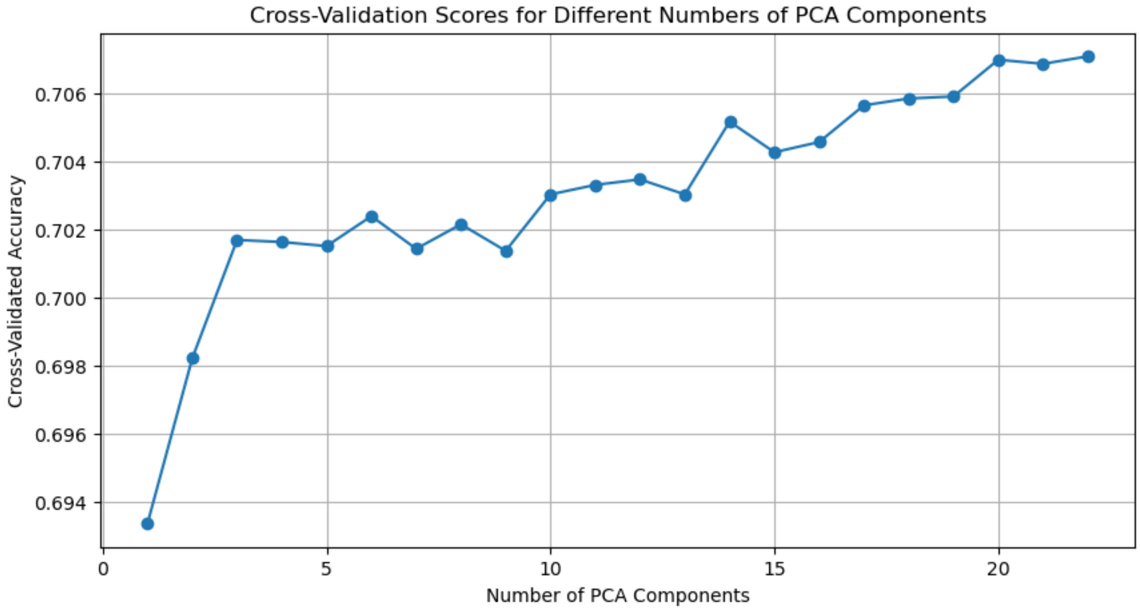
```
Accuracy: 0.84
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.77      0.82      6807
           1       0.80      0.90      0.85      6991

    accuracy                           0.84     13798
   macro avg       0.84      0.84      0.84     13798
weighted avg       0.84      0.84      0.84     13798
```



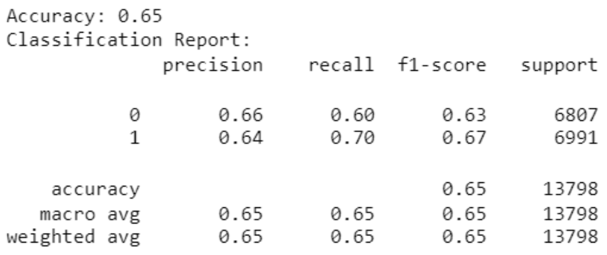


```
Negative Predictive Value (NPV): 0.8844984802431611
```

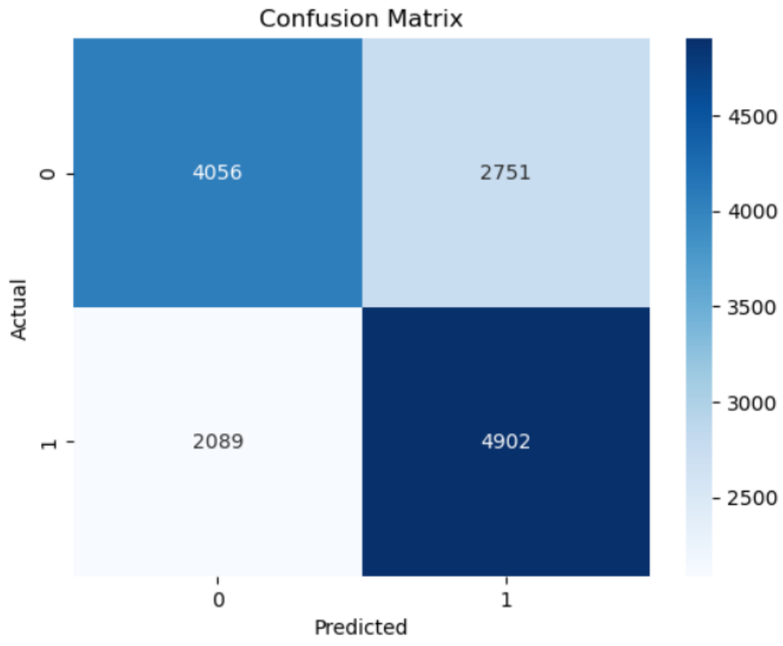
Log Reg



With 3 components

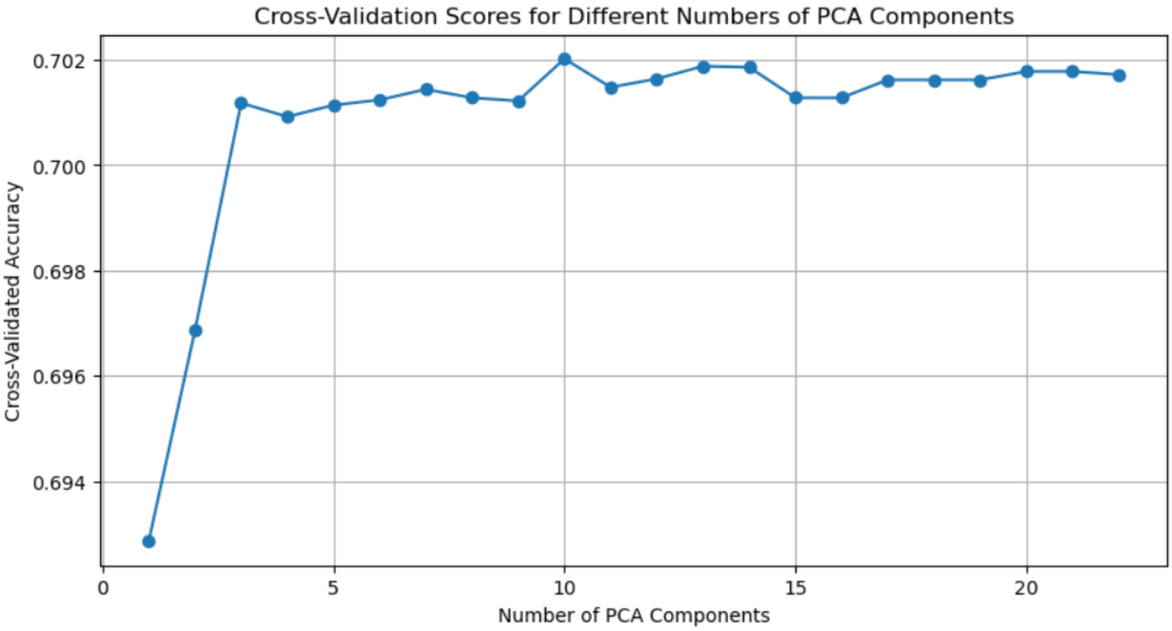


```
Accuracy: 0.65
Classification Report:
              precision    recall  f1-score   support

           0       0.66      0.60      0.63      6807
           1       0.64      0.70      0.67      6991

    accuracy                           0.65     13798
   macro avg       0.65      0.65      0.65     13798
weighted avg       0.65      0.65      0.65     13798
```

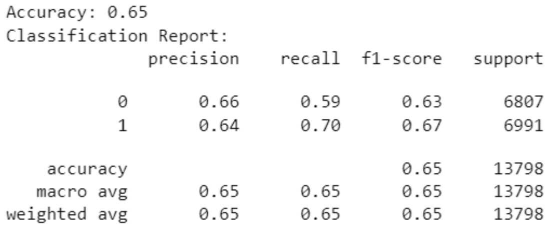


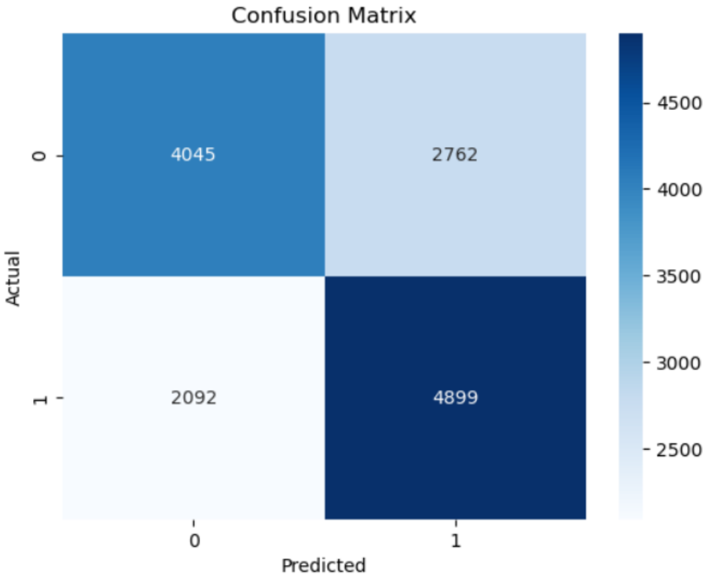
```
Negative Predictive Value (NPV): 0.6600488201790073
```

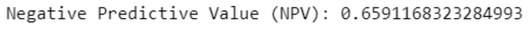
AdaBoost



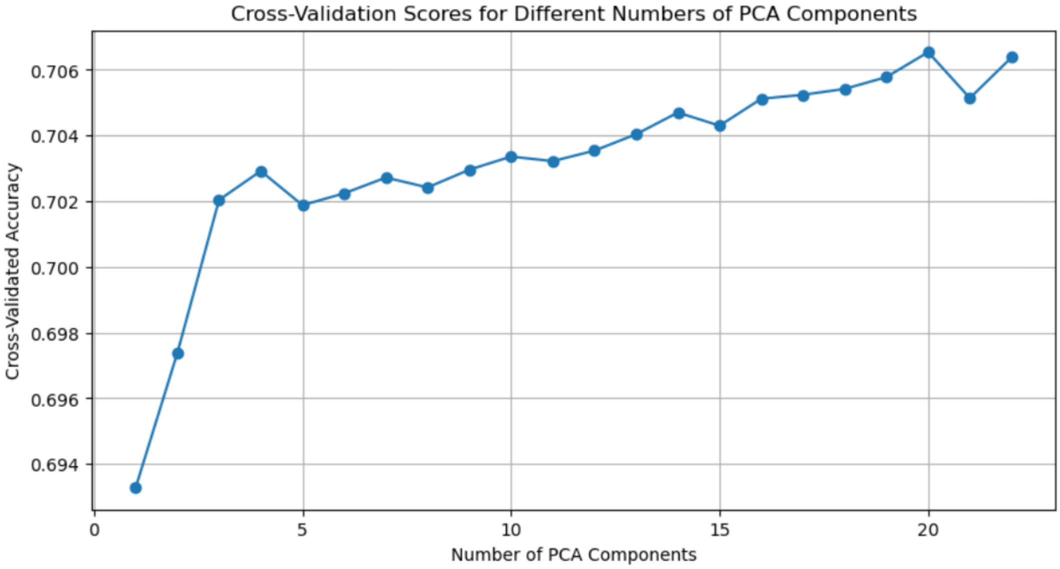
With 3 components


```
Accuracy: 0.65
Classification Report:
              precision    recall  f1-score   support

           0       0.66      0.59      0.63      6807
           1       0.64      0.70      0.67      6991

    accuracy                           0.65     13798
   macro avg       0.65      0.65      0.65     13798
weighted avg       0.65      0.65      0.65     13798
```

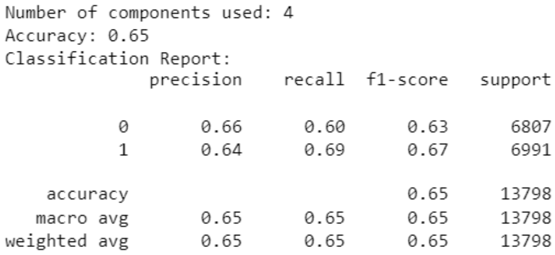



```
Negative Predictive Value (NPV): 0.6591168323284993
```

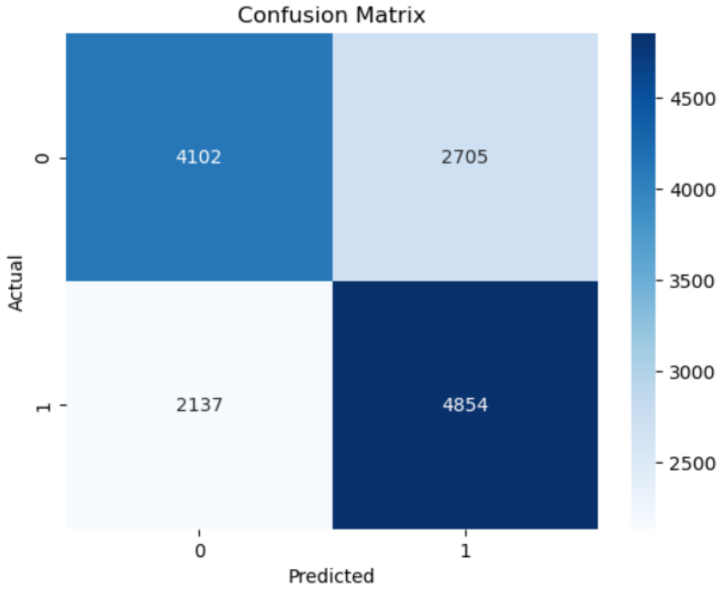
CatBoost

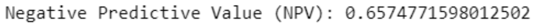


With 4 components



```
Number of components used: 4
Accuracy: 0.65
Classification Report:
              precision    recall  f1-score   support

           0       0.66      0.60      0.63      6807
           1       0.64      0.69      0.67      6991

    accuracy                           0.65     13798
   macro avg       0.65      0.65      0.65     13798
weighted avg       0.65      0.65      0.65     13798
```

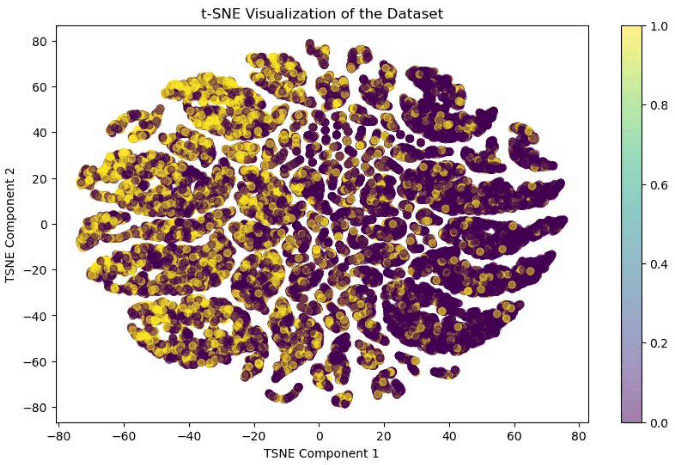




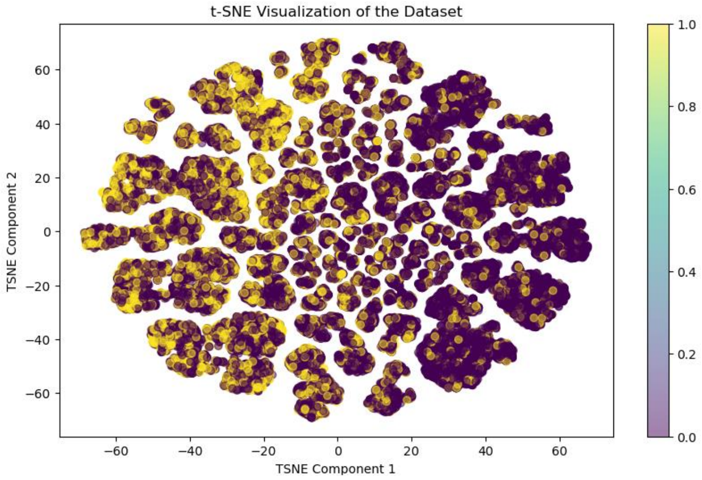
```
Negative Predictive Value (NPV): 0.6574771598012502
```

T-SNE

Apply on PCA with 5 n_components:

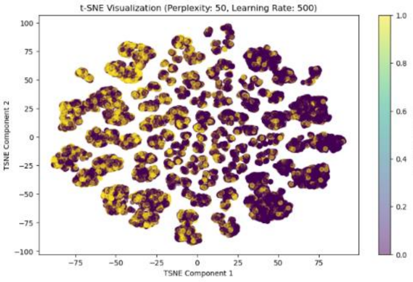


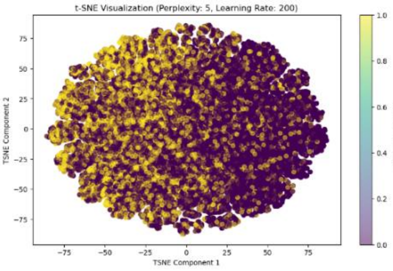
Without PCA

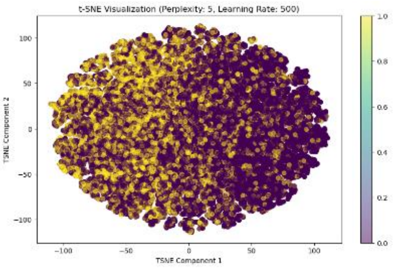


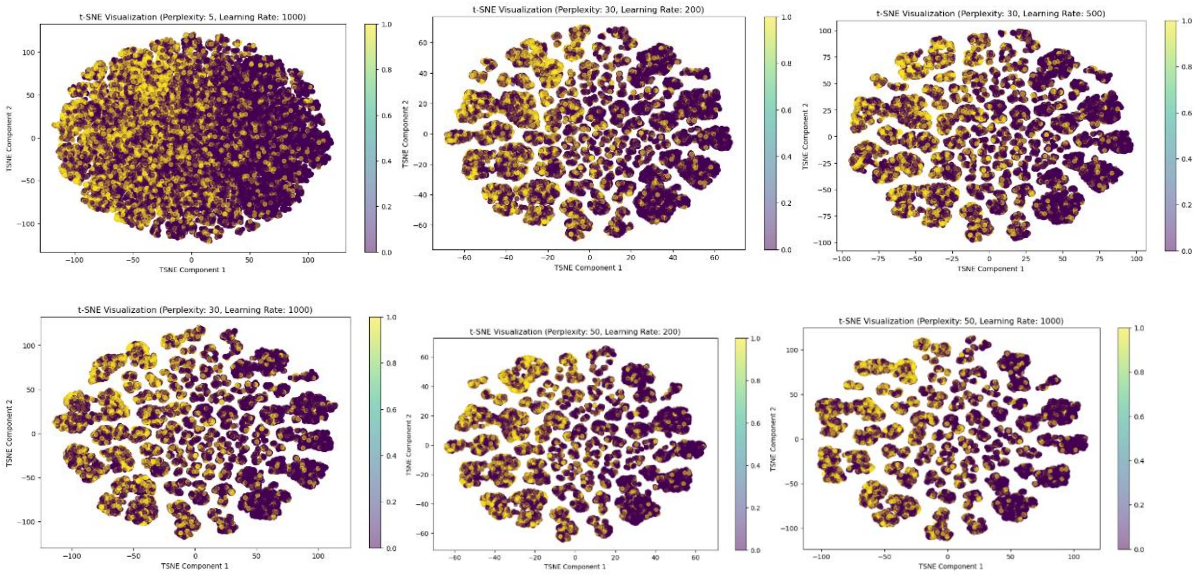
perplexity_values = [5, 30, 50]

learning_rates = [200, 500, 1000]



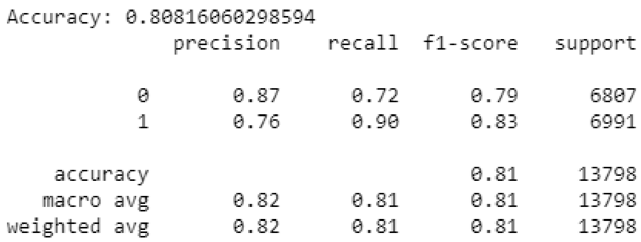




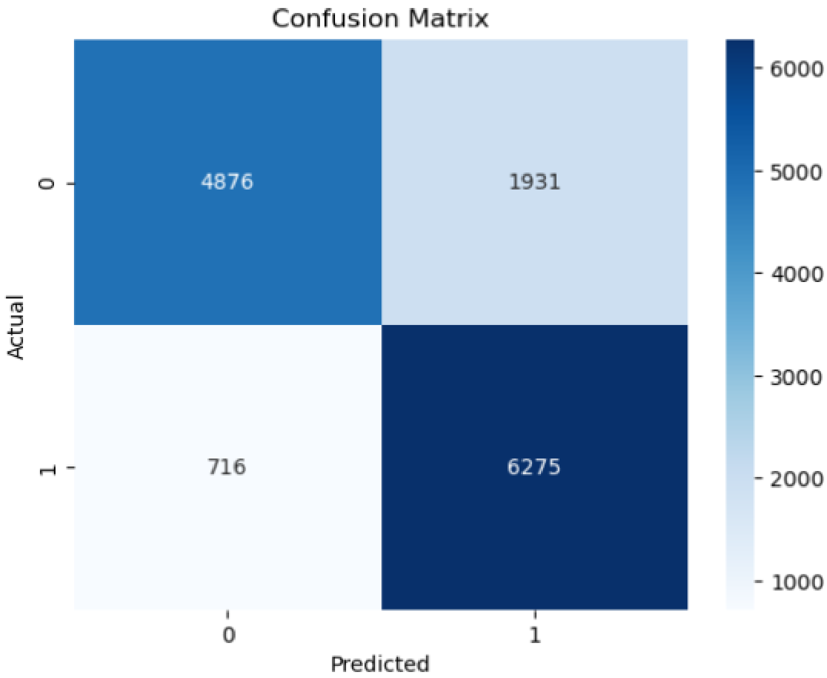


Perplexity 50, Learning Rate 500 :The clusters are more distinct and well-formed, which indicates good global structure capture. The higher perplexity value allows the model to consider a broader context, which seems beneficial.

Random Forest

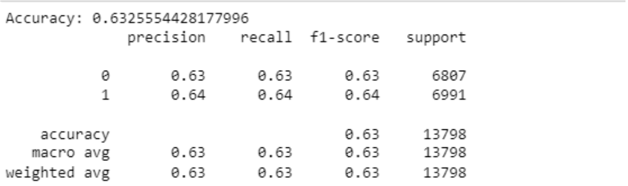


```
Accuracy: 0.80816060298594
              precision    recall  f1-score   support

           0       0.87      0.72      0.79      6807
           1       0.76      0.90      0.83      6991

    accuracy                           0.81     13798
   macro avg       0.82      0.81      0.81     13798
weighted avg       0.82      0.81      0.81     13798
```

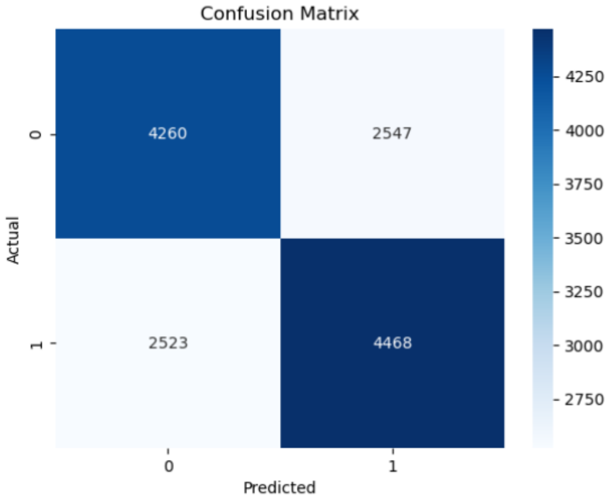


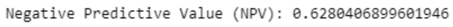
```
Negative Predictive Value (NPV): 0.8719599427753935
```

Log Reg



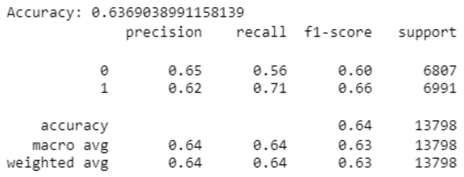
```
Accuracy: 0.6325554428177996
              precision    recall  f1-score   support

           0       0.63      0.63      0.63      6807
           1       0.64      0.64      0.64      6991

    accuracy                           0.63     13798
   macro avg       0.63      0.63      0.63     13798
weighted avg       0.63      0.63      0.63     13798
```



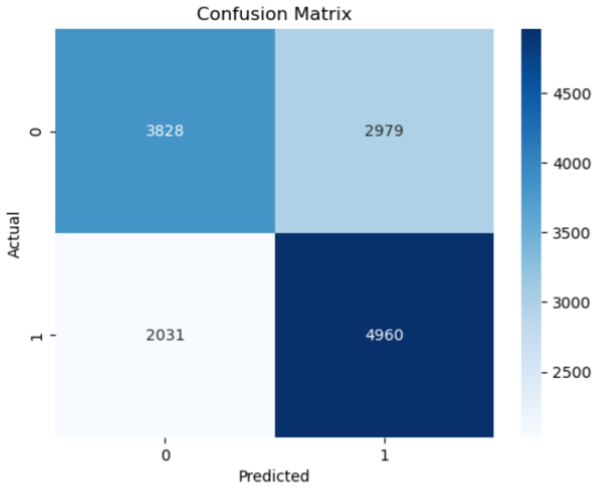


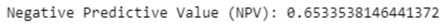
```
Negative Predictive Value (NPV): 0.6280406899601946
```

AdaBoost



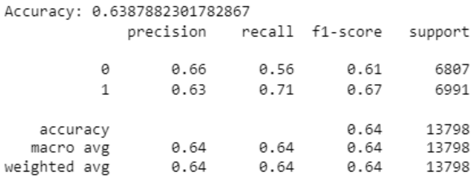
```
Accuracy: 0.6369038991158139
              precision    recall  f1-score   support

           0       0.65      0.56      0.60      6807
           1       0.62      0.71      0.66      6991

    accuracy                           0.64     13798
   macro avg       0.64      0.64      0.63     13798
weighted avg       0.64      0.64      0.63     13798
```



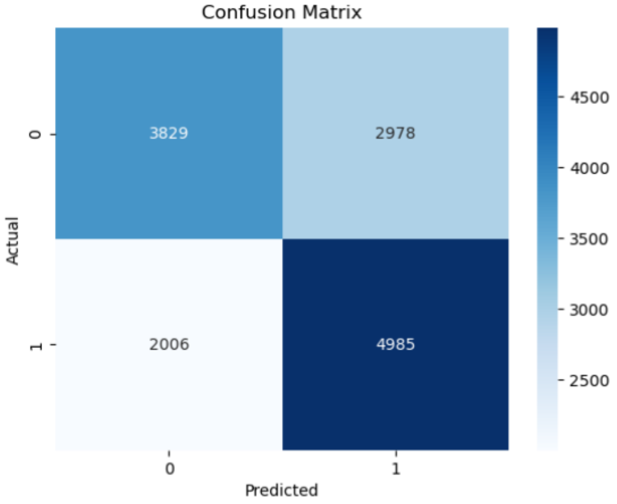


```
Negative Predictive Value (NPV): 0.6533538146441372
```

CatBoost



```
Accuracy: 0.6387882301782867
              precision    recall  f1-score   support

           0       0.66      0.56      0.61      6807
           1       0.63      0.71      0.67      6991

    accuracy                           0.64     13798
   macro avg       0.64      0.64      0.64     13798
weighted avg       0.64      0.64      0.64     13798
```

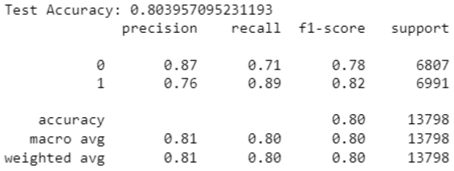


```
Negative Predictive Value (NPV): 0.6562125107112253
```

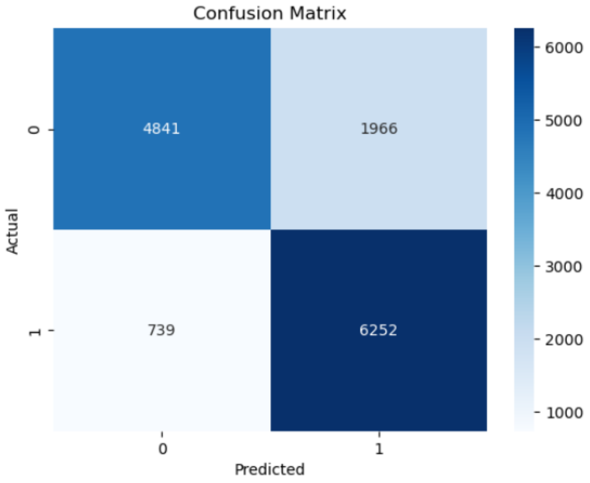
LDA

n_components=1

Random Forest

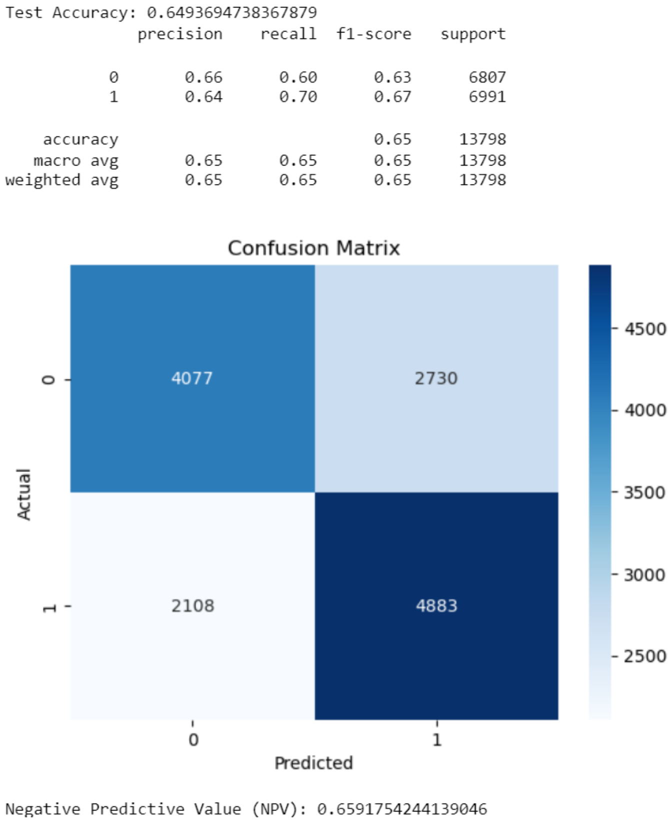


```
Test Accuracy: 0.803957095231193
              precision    recall  f1-score   support

           0       0.87      0.71      0.78      6807
           1       0.76      0.89      0.82      6991

    accuracy                           0.80     13798
   macro avg       0.81      0.80      0.80     13798
weighted avg       0.81      0.80      0.80     13798
```



```
Negative Predictive Value (NPV): 0.8675627240143369
```

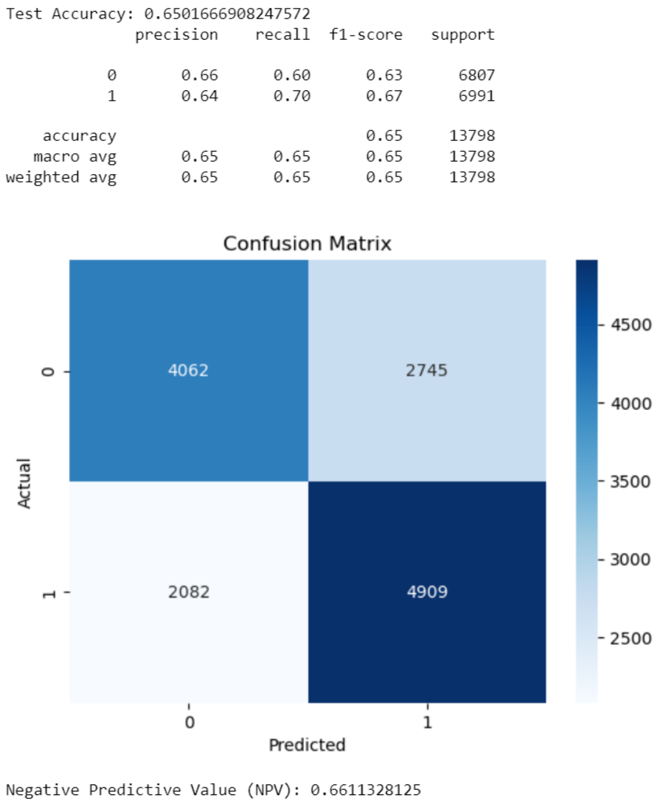
Log reg



```
Test Accuracy: 0.6493694738367879
              precision    recall  f1-score   support

           0       0.66      0.60      0.63      6807
           1       0.64      0.70      0.67      6991

    accuracy                           0.65     13798
   macro avg       0.65      0.65      0.65     13798
weighted avg       0.65      0.65      0.65     13798
```

```
Negative Predictive Value (NPV): 0.6591754244139046
```

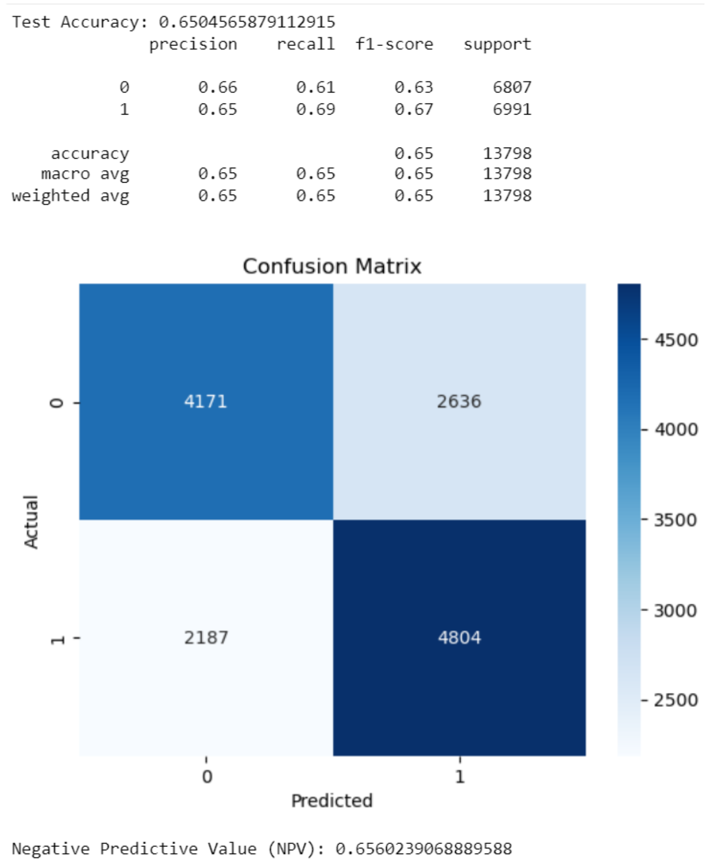
AdaBoost



```
Test Accuracy: 0.6501666908247572
              precision    recall  f1-score   support

           0       0.66      0.60      0.63      6807
           1       0.64      0.70      0.67      6991

    accuracy                           0.65     13798
   macro avg       0.65      0.65      0.65     13798
weighted avg       0.65      0.65      0.65     13798
```

```
Negative Predictive Value (NPV): 0.6611328125
```

CatBoost



```
Test Accuracy: 0.6504565879112915
              precision    recall  f1-score   support

           0       0.66      0.61      0.63      6807
           1       0.65      0.69      0.67      6991

    accuracy                           0.65     13798
   macro avg       0.65      0.65      0.65     13798
weighted avg       0.65      0.65      0.65     13798
```

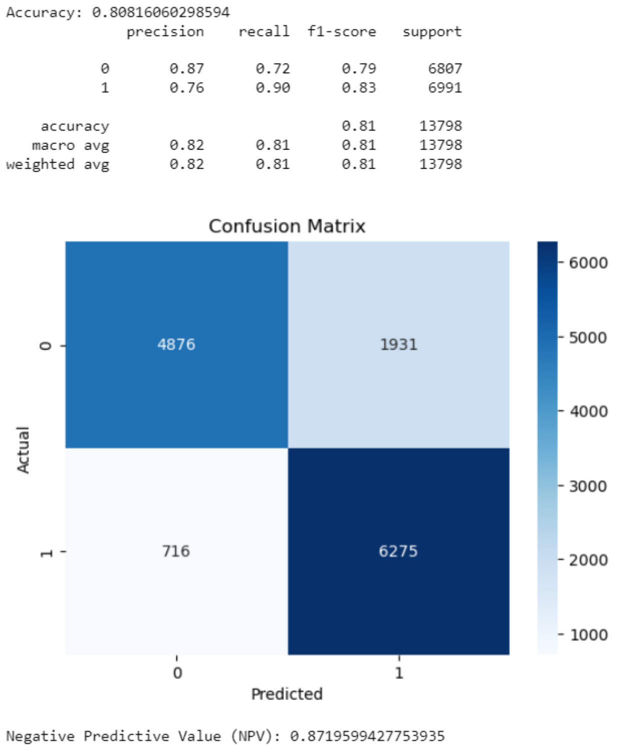
Confusion Matrix

Actual \ Predicted	0	1
0	4171	2636
1	2187	4804

```
Negative Predictive Value (NPV): 0.6560239068889588
```

LDA

Random Forest



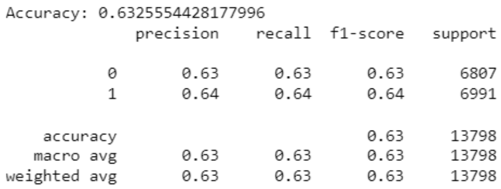
```
Accuracy: 0.80816060298594
              precision    recall  f1-score   support

           0       0.87      0.72      0.79      6807
           1       0.76      0.90      0.83      6991

    accuracy                           0.81     13798
   macro avg       0.82      0.81      0.81     13798
weighted avg       0.82      0.81      0.81     13798
```

Confusion Matrix

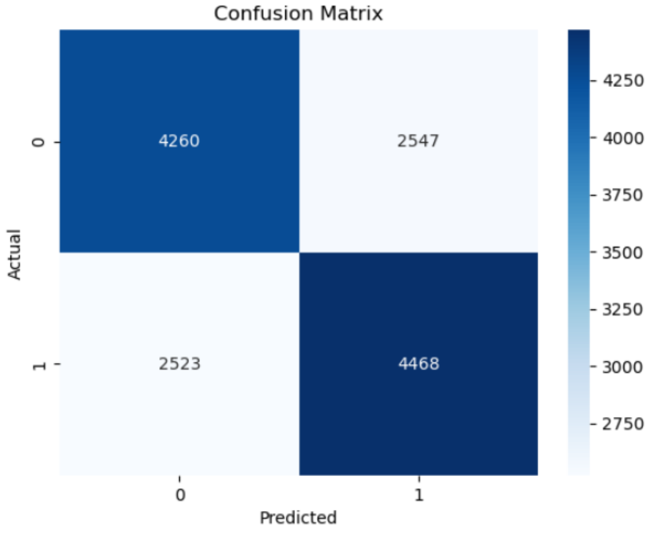
Actual \ Predicted	0	1
0	4876	1931
1	716	6275

```
Negative Predictive Value (NPV): 0.8719599427753935
```

Log Reg

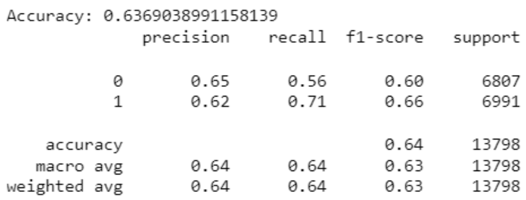


```
Accuracy: 0.6325554428177996
              precision    recall  f1-score   support

           0       0.63      0.63      0.63      6807
           1       0.64      0.64      0.64      6991

    accuracy                           0.63     13798
   macro avg       0.63      0.63      0.63     13798
weighted avg       0.63      0.63      0.63     13798
```

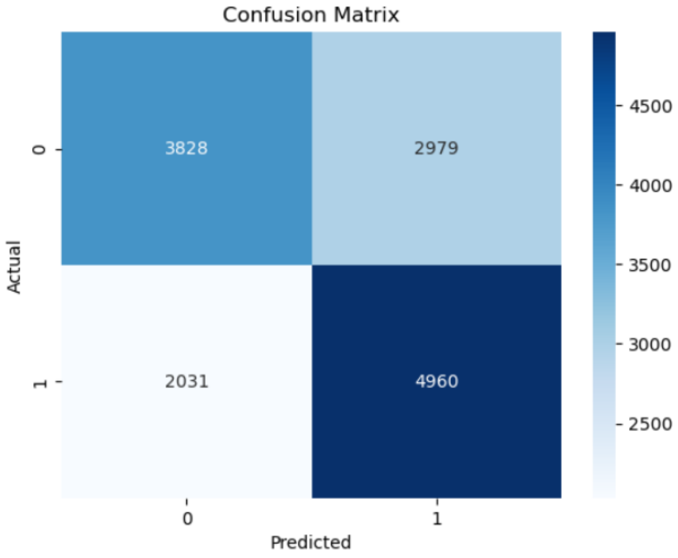


```
Negative Predictive Value (NPV): 0.6280406899601946
```

AdaBoost

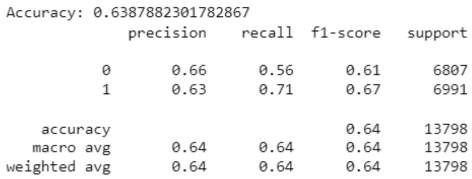


```
Accuracy: 0.6369038991158139
              precision    recall  f1-score   support

           0       0.65      0.56      0.60      6807
           1       0.62      0.71      0.66      6991

    accuracy                           0.64     13798
   macro avg       0.64      0.64      0.63     13798
weighted avg       0.64      0.64      0.63     13798
```

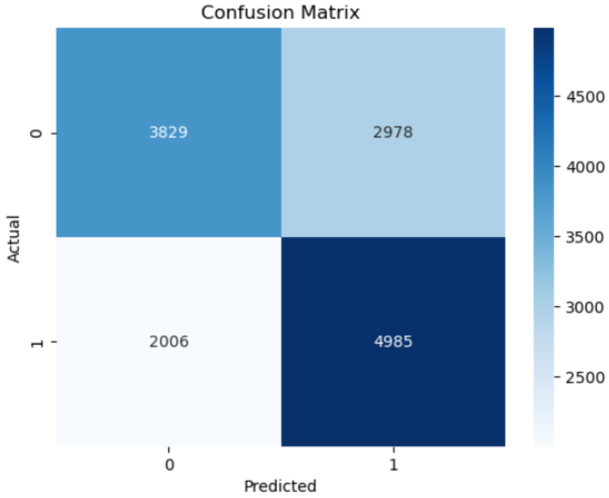


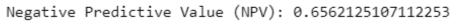
```
Negative Predictive Value (NPV): 0.6533538146441372
```

CatBoost



```
Accuracy: 0.6387882301782867
              precision    recall  f1-score   support

           0       0.66      0.56      0.61      6807
           1       0.63      0.71      0.67      6991

    accuracy                           0.64     13798
   macro avg       0.64      0.64      0.64     13798
weighted avg       0.64      0.64      0.64     13798
```





```
Negative Predictive Value (NPV): 0.6562125107112253
```

Conclusions:

Accuracy:

	Dataset	PCA	T-SNE	LDA
RF	0.86	0.84	0.80	0.80
Log Reg	0.64	0.65	0.63	0.64
AdaBoost	0.65	0.65	0.63	0.65
CatBoost	0.67	0.65	0.63	0.65

Negative predictive values:

	Dataset	PCA	T-SNE	LDA
RF	0.88	0.88	0.87	0.86
Log Reg	0.65	0.66	0.62	0.65
AdaBoost	0.66	0.65	0.65	0.66
CatBoost	0.68	0.65	0.65	0.65

In my case with Random Forest I got the best values with 0.86 accuracy and 0.88 negative predicted values.

With dimension reduction, the result are slightly worse or even the same in some cases. PCA has the best results then it is LDA and the last it would be T-SNE.