Data Preparation

Data Collecting and EDA

After merging data from 4 files from two different sources we have one large dataframe with the shape (1706, 17818): 17810 columns with gene expression level, and 8 info columns.

GTEx - 54670 columns with genes + 2 info columns (Sex, Age), 578 patients
Then we add 6 info columns for having opportunity of joining the TCGA – so 8 info cols in total.

1. **Age** – we replace values as: {'20-29': 25, '30-39': 35, …} because we have only ranges in the GTEx dataset with normal samples.
2. Comparing total expression levels by Sex and Age (ranges) shows similar results in both cases – no differences were found.

TCGA - 20501 columns with genes + 8 info columns, 1128 patients.
Info columns: Sex, Age, Dataset, Sample, Histilogy, Location, TimeSurv, EventSurv.
Additional information: here we have 1018 tumor patients and 110 normal patients,
673 male and 455 female,
576 from LUAD (59 norm) and 552 from LUSC (51 norm)
772 – alive and 322 - dead

1. **Age** – here we have an exact age but for joining with the GTEx we also transform the data by putting them in the ranges as higher. Maybe it's worth considering only this TCGA dataset without replacing it for further age research.
2. **Sex** - we replace values as: {'male': 1, 'female': 2} because these values are in GTEx.
3. Comparing total expression levels by Sex and Age (ranges) shows a similar result in both cases – no differences were found.

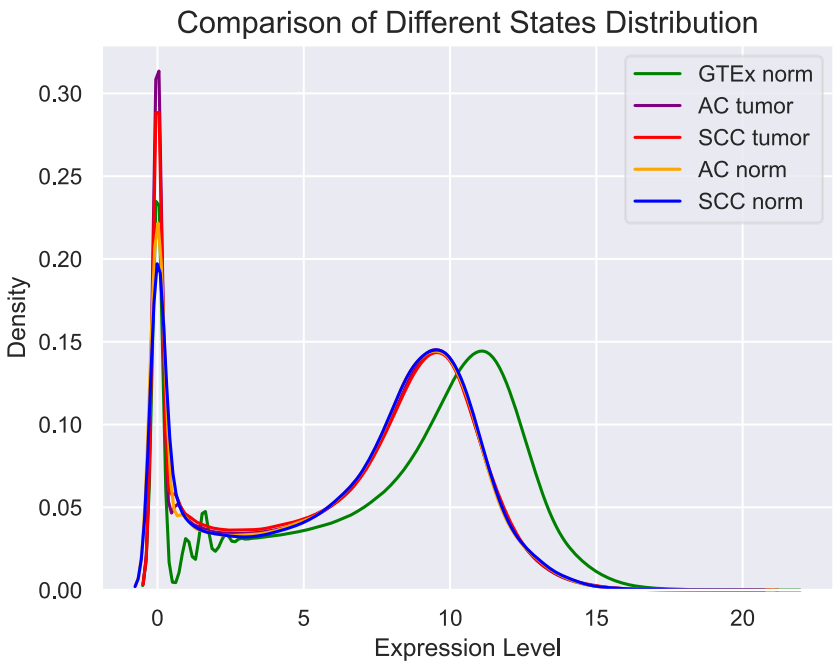
GTEx + TCGA

1. We create a **new feature STATE** which contains useful info from 2 cols (Dataset, Sample), and drop these 2.
   So all normal patients from three datasets will have the State: norm – in total 688 = 578+110;
   AC tumor patients from LUAD dataset will have the State: AC – in total 517 = 576-59;
   SCC tumor patients from LUSC dataset will have the State: SCC – in total 501 = 552-51.

After merging genes that are present in both sources we get a new common dataset (df3) with:
17817 columns (17810 genes, 7 info) and 1706 rows for patients.

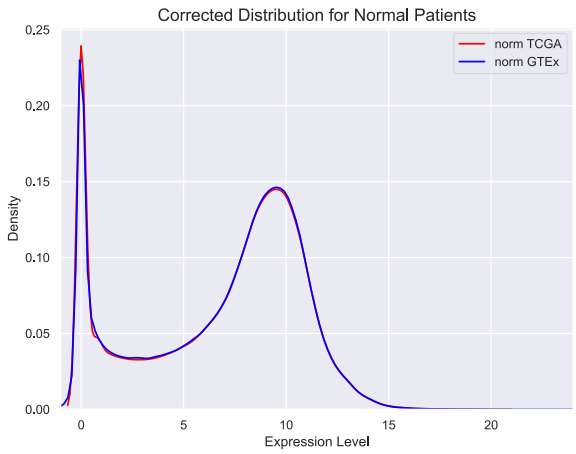
2. Merge validity check because we have two different sources of data.

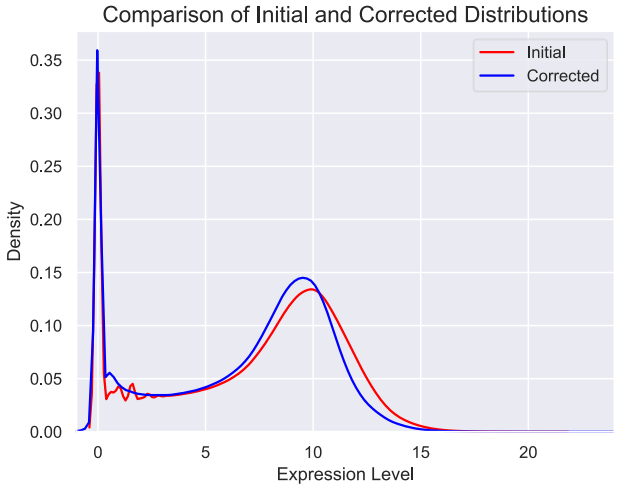
Here we see that data distributions from different sources are shifted. In GTEx we have only normal patients, while in TCGA we have normal and tumor patients. We don't know about differences in tumor_normal, so we take only normal patients from TCGA and correct the GTEx distribution as:



GTEx_new = (GTEx – avg(GTEx)) * std(TCGA)/std(GTEx) + avg(TCGA)

So we get:



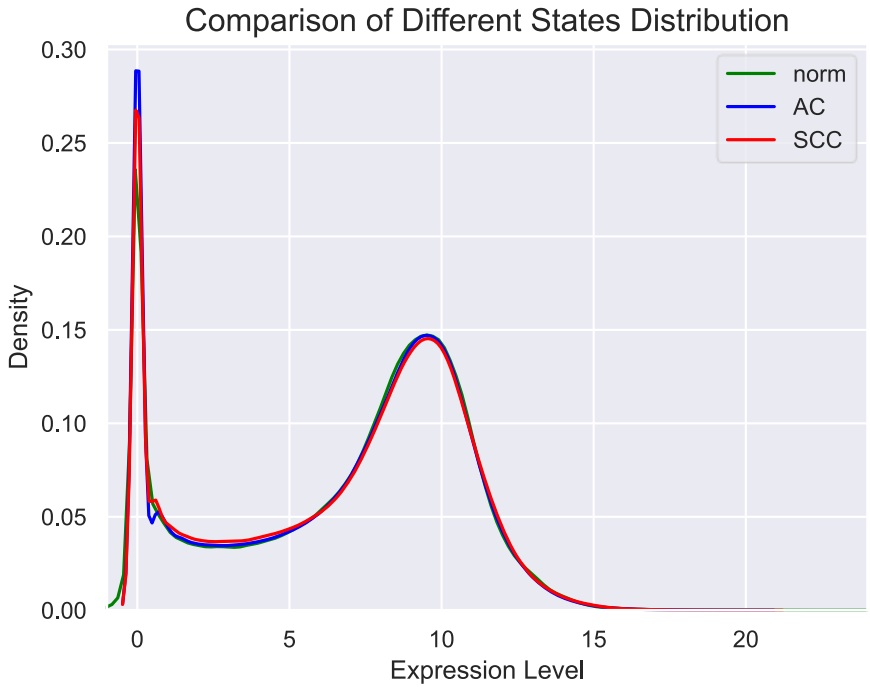


3. **NaN values** – we investigate the result dataset for detecting missing values and fill them by 0 because this is the result of calculations in the previous step.
4. **Null_genes** – we removed all columns with ONLY ZERO values, there were 217 such.

So now we have dataset (df4) with: 17600 columns ((17810-2+1-217) genes, 7 info) and 1706 rows for patients.

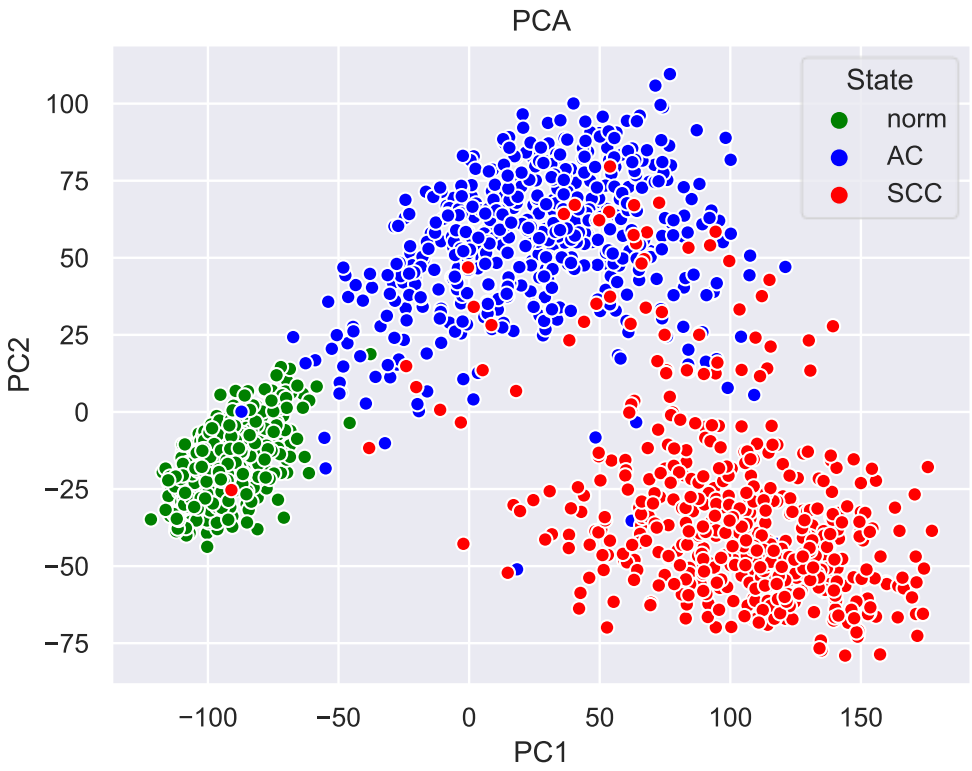
We save the result as *'L_Prep_df4.csv'* file.

Finally, for the State investigations, we save the dataset with the shape (1706, 17594): 17593 genes and 1 info (State) and save it as a **pickle object 'L_Prep_State'**.

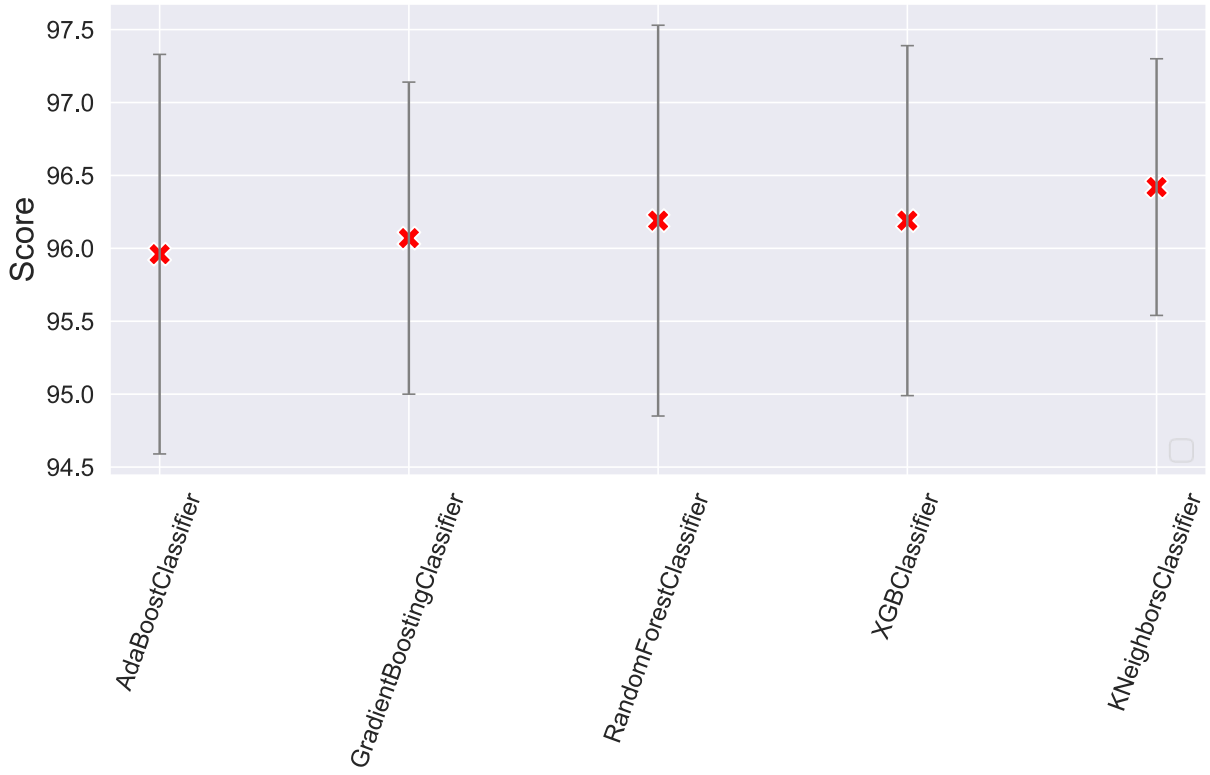


688 - norm, 517 - AC, 501 - SCC

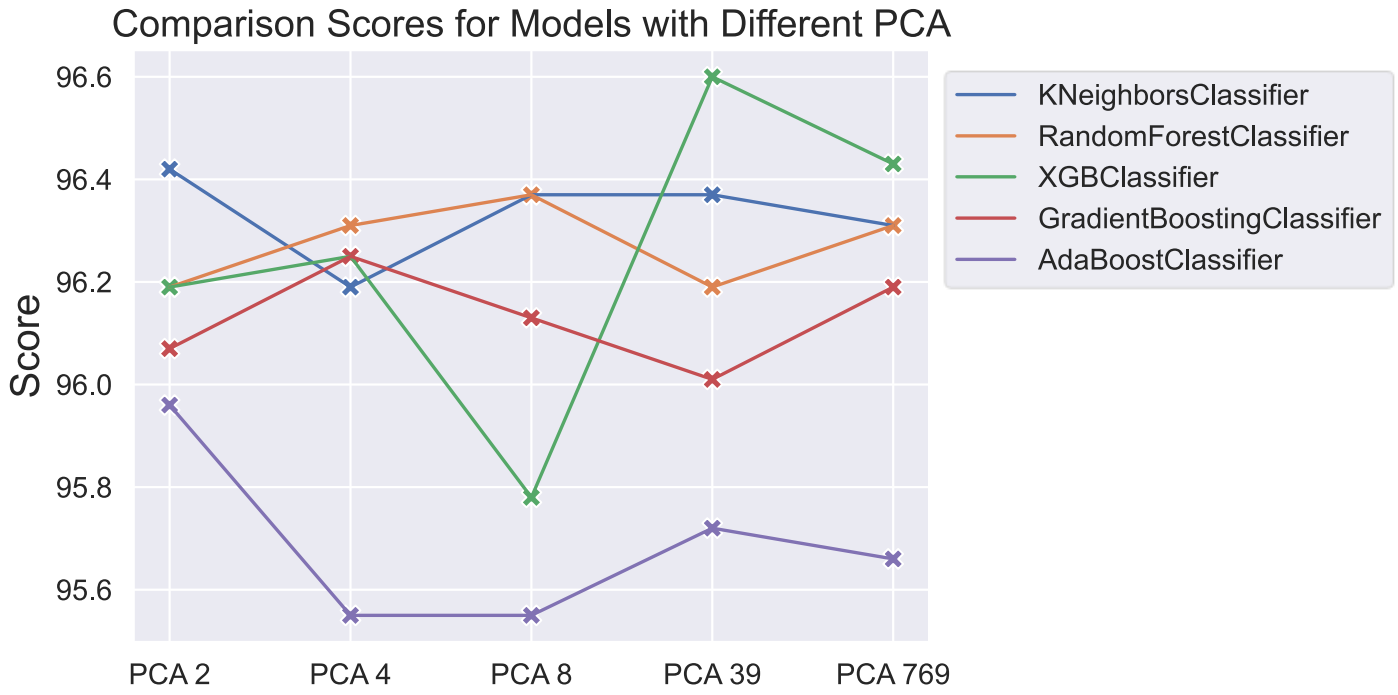
Principal Component Analysis



We see 3 groups here as it is, and there is a pretty good separation between norm and tumor, and worse between AC and SCC. Further, we use the best methods for modeling with cross-validation (cv=5) and receive scores indicated on the graph.



So, we have the similar results generally where **accuracy score** is **≈ 96 %** - the same in all models. It's interesting that if we take all 769 components, which give 95% reproducibility, we receive the same result.



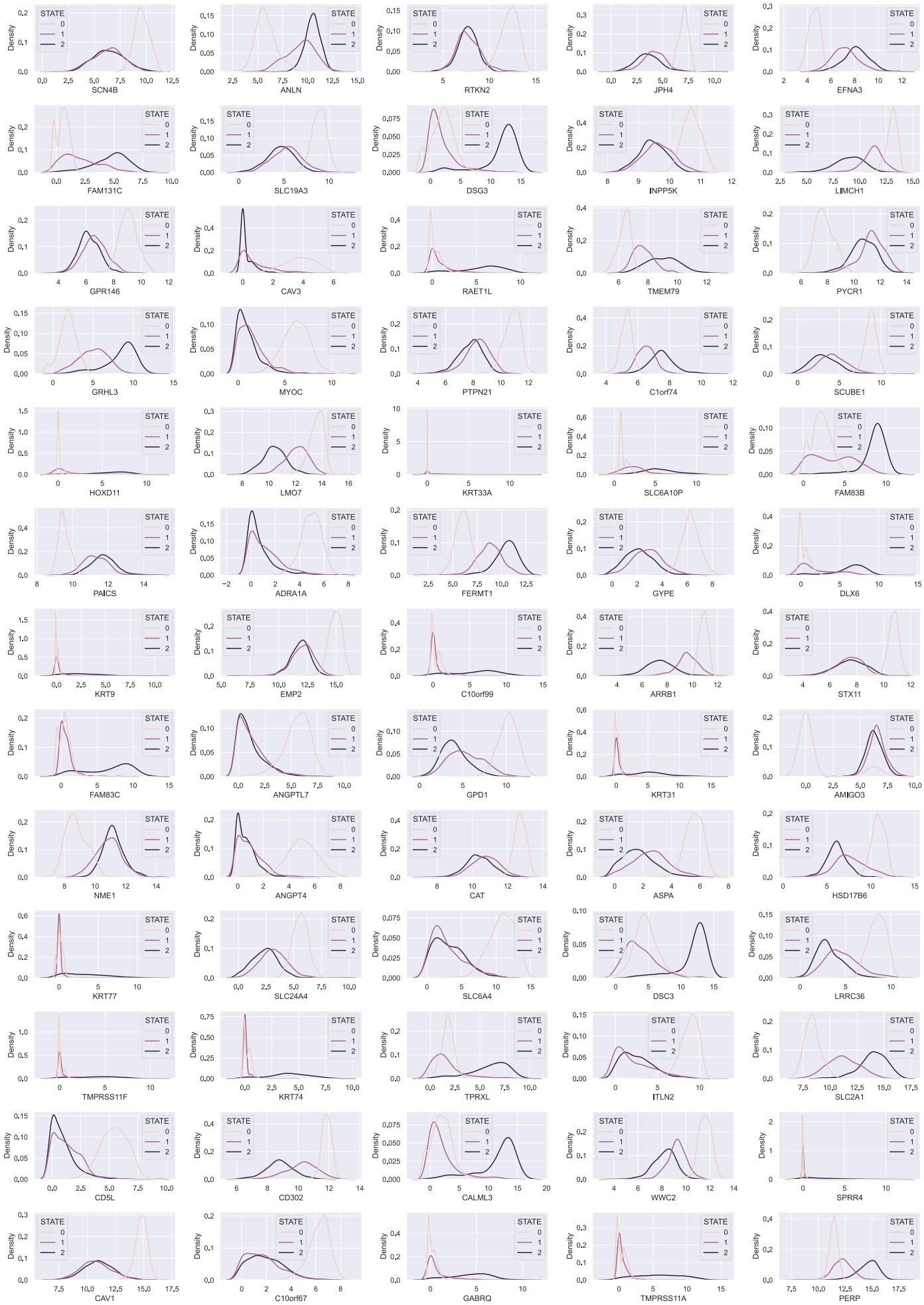
So, the model with two principal components is satisfactory.

Feature Selection

We create two variables X (all 17593 genes) and y (State).

1. Detect constant features with VarianceThreshold = 0.01 and remove them. 17593-271=17322.
2. Detect correlated features with threshold = 0.9 and remove them from X: 17322-583=16739.
3. Selecting features using Mutual Information (MI) gained from Classification, select the 30 best features and save them as best_MI_State. Check them on correlation just in case, so now we have best_MI_State.
4. Selecting features using ANOVA Statistical Analysis. We can't use the Chi2 test because we have negative values in genes. So we use f_statistic and also select 30 best features and save them as best_ANOVA_State.
5. Selecting features using Feature Importance, select select 15 best features, and save them as best_TreesClass.

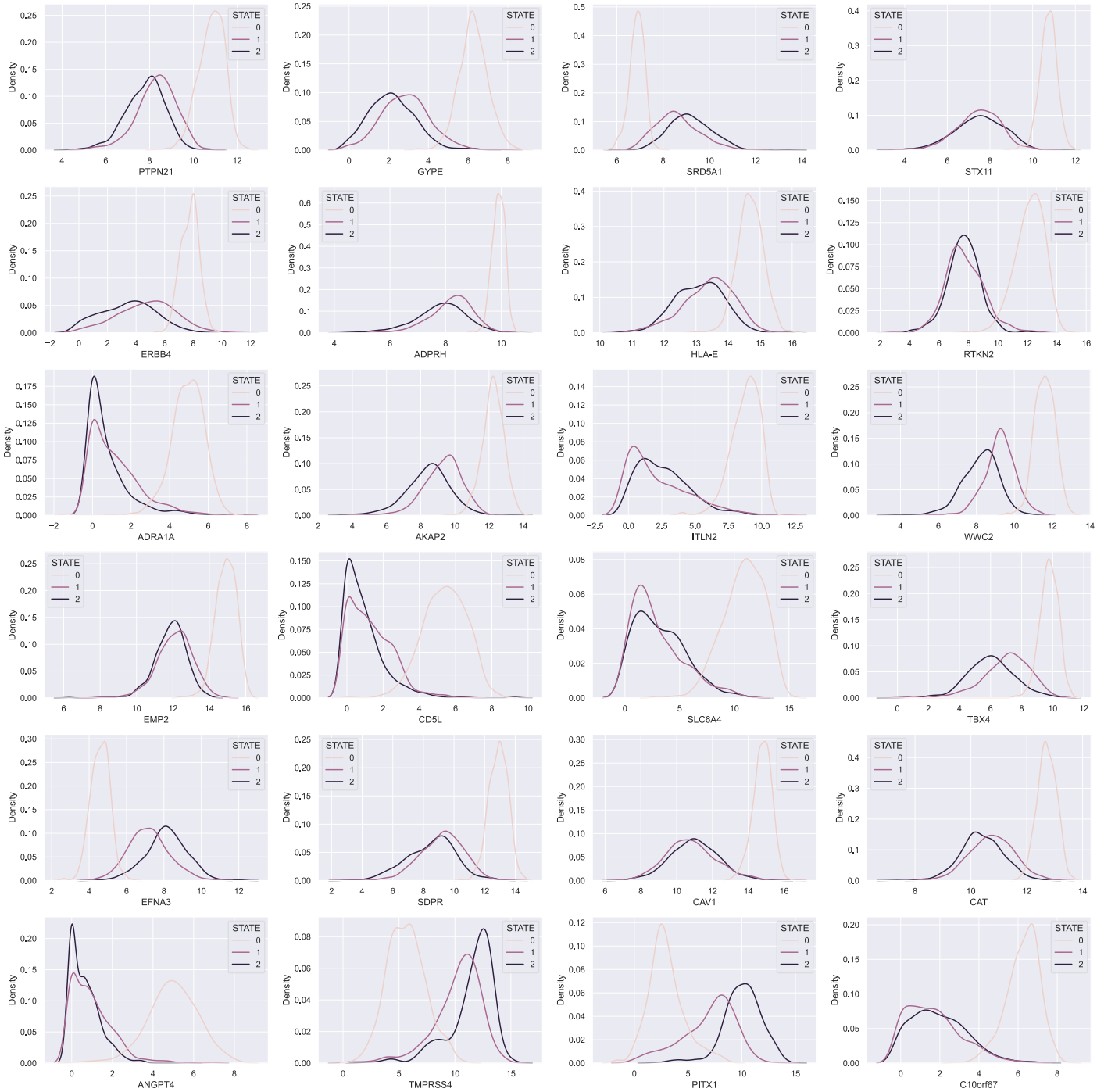
So we have 65 features that we save as standard_features:



We have no genes that can be good predictors for norm, AC, SCC simultaneously.

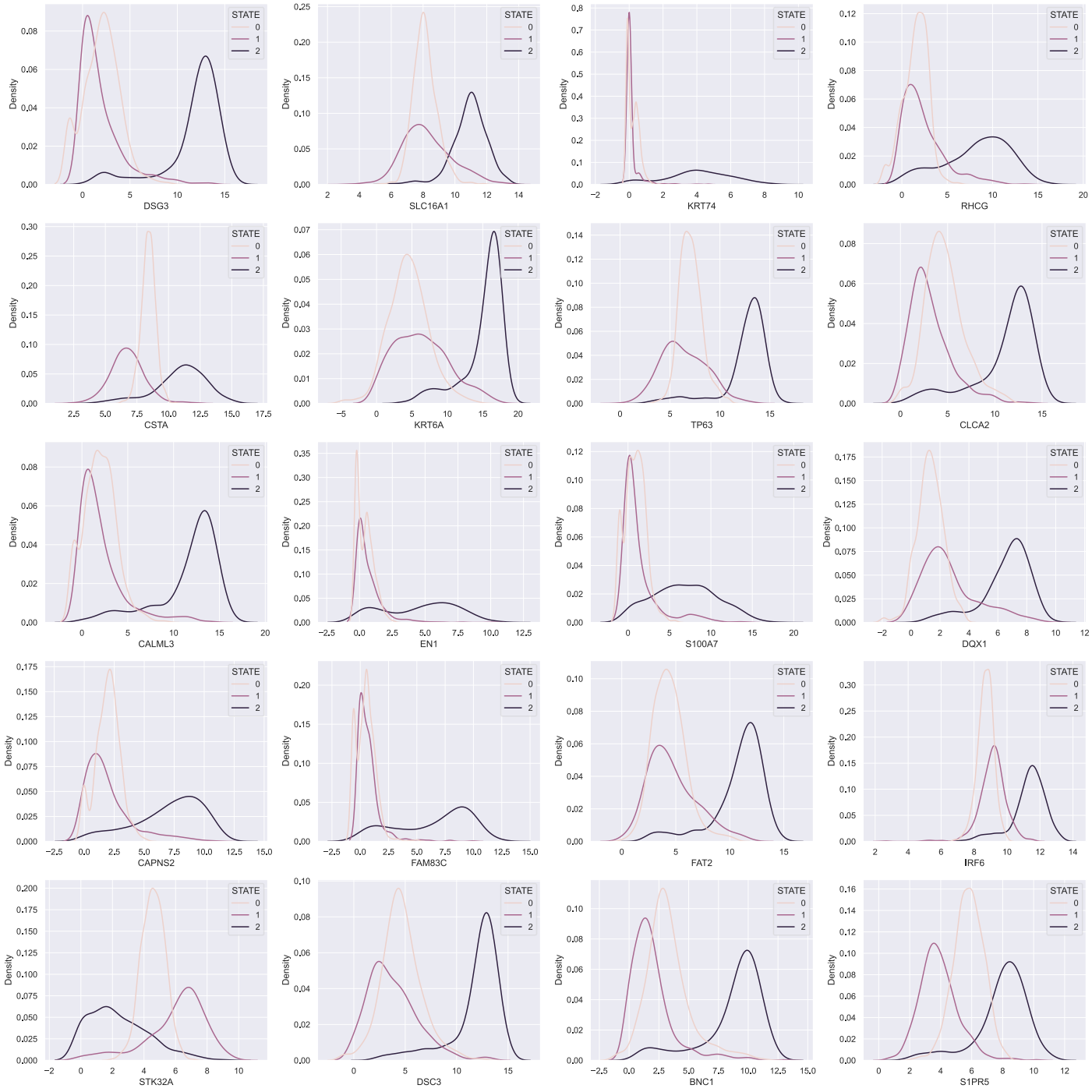
We divide genes for 2 groups which are the best for detecting:

1. norm_tumor:



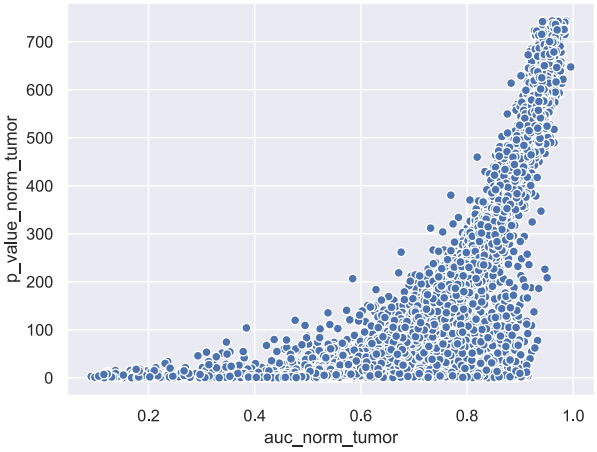
So we have 24 genes from standard methods in order to train the model for detecting norm_tumor.

2. AC_SCC:

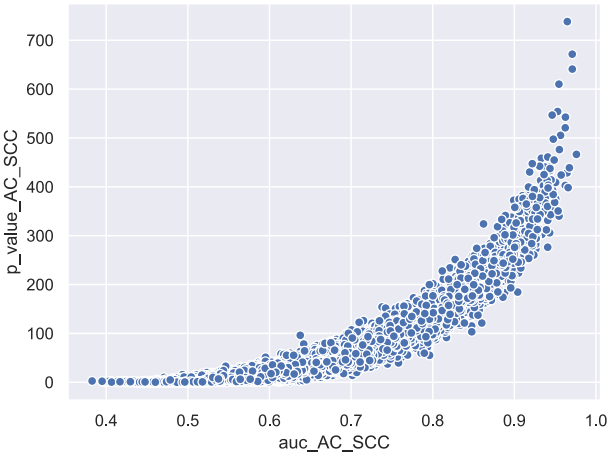


Here we have 20 genes from standard methods in order to train the model for detecting AC_SCC.

6. Specific laboratory technique:



auc > 0.99 and -log(p_value) > 650



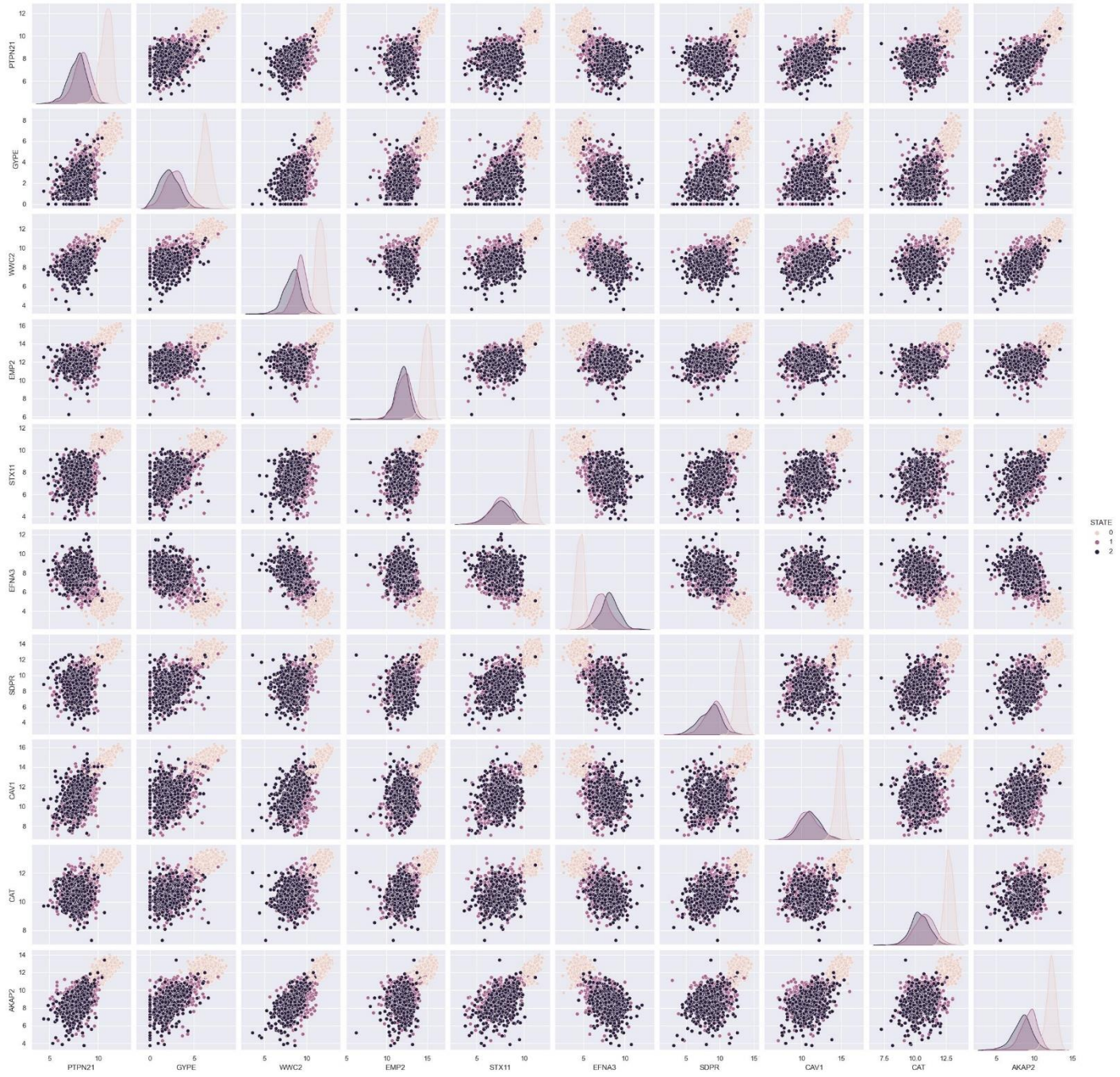
auc > 0.925 and -log(p_value) > 430

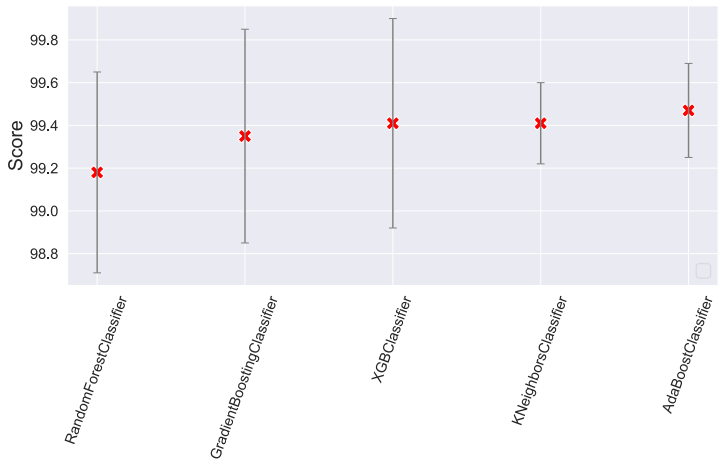
Finally, we create 4 lists of features (norm_tumor, AC_SCC) which are the intersections and unions of the best features from these techniques:

final_norm_tumor list consists of 10 features and final_AC_SCC – 12 features;

union_norm_tumor list consists of 40 features and union_AC_SCC – 32 features.

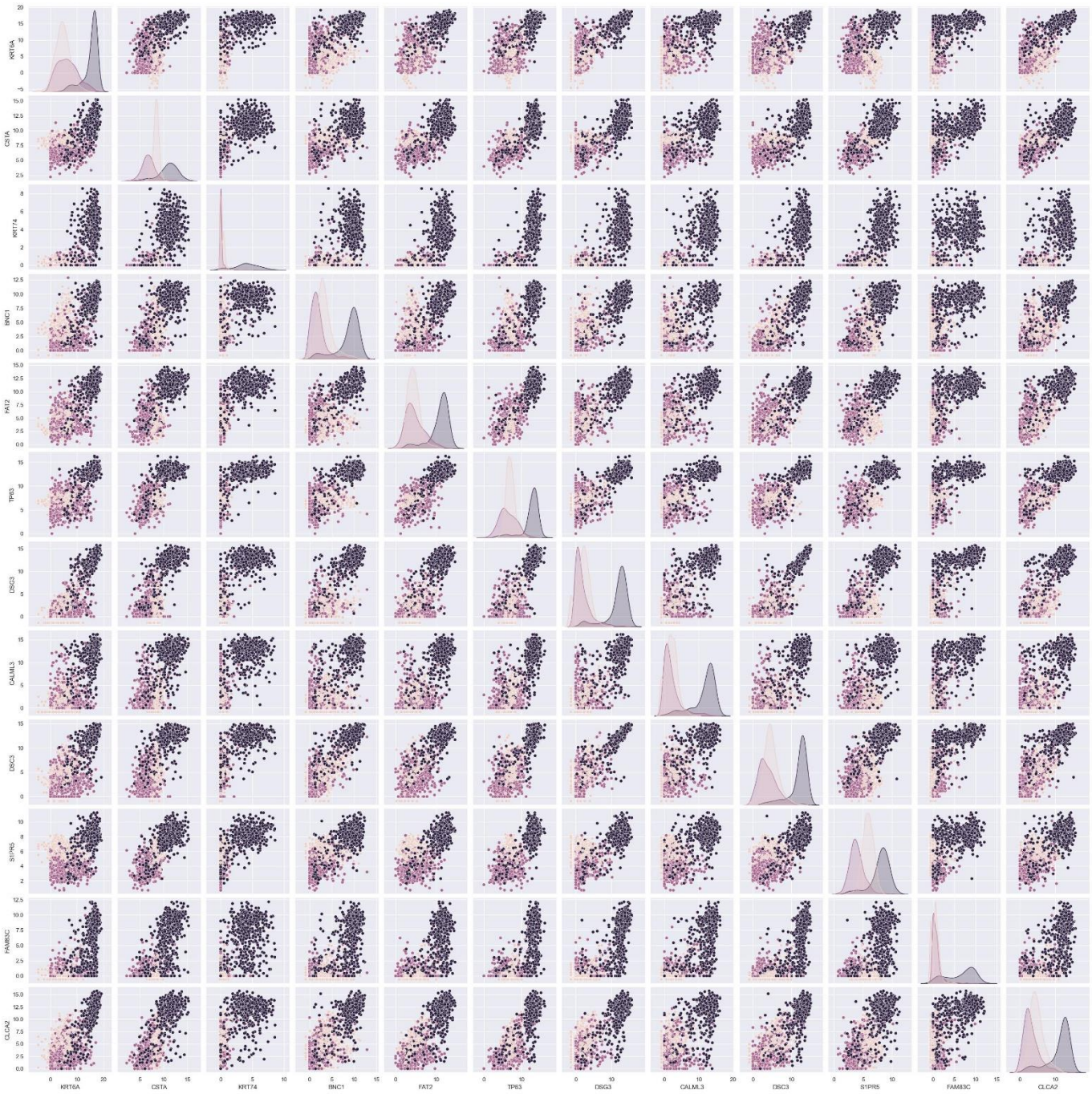
Modeling norm_tumor

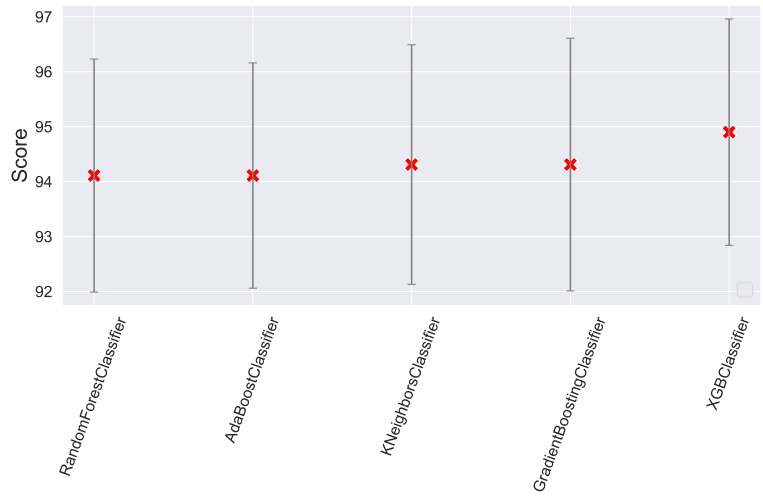




So with these 10 features for prediction, we have an **accuracy score ≈99 %** for all the models.

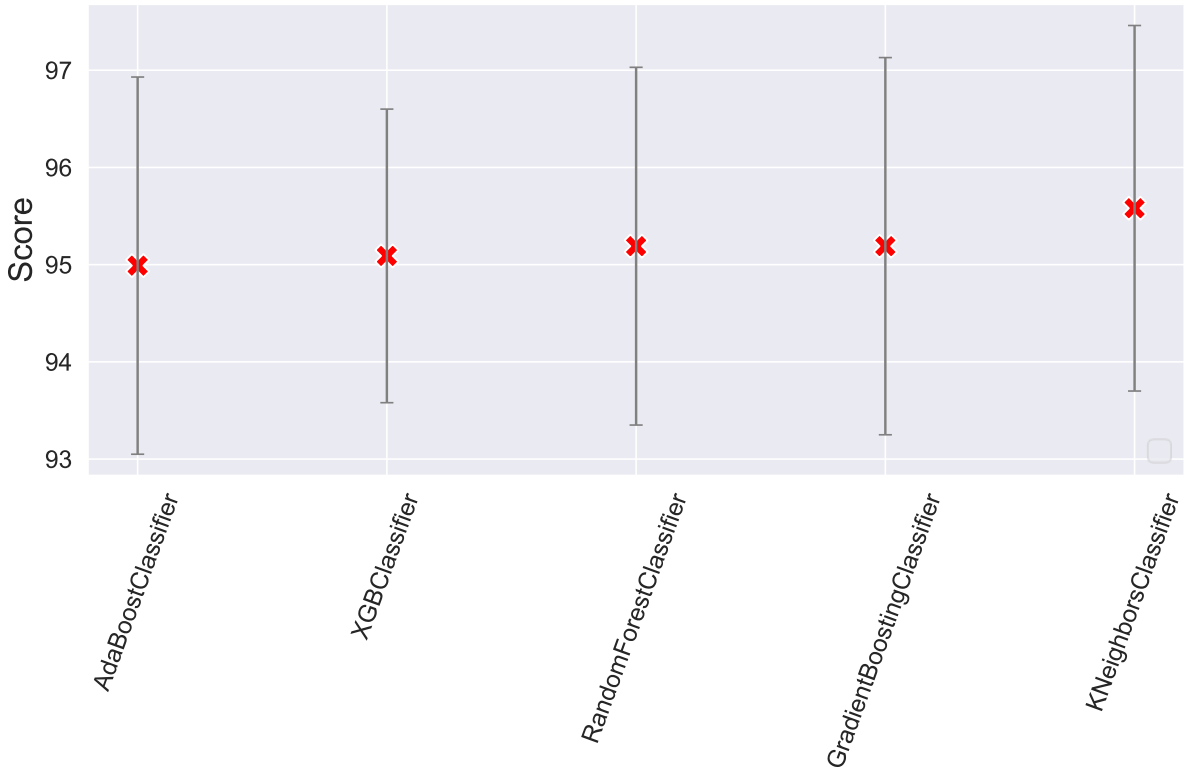
Modeling AC_SCC





So with the selected 12 features for prediction, we have **accuracy scores ≈94÷95 %** for all the models.

When we use all 32 features which were determined in FeatureSelection we have:



The **accuracy scores** are about **95÷96 %**.

As we see all these models show the results in the range of 1% approximately. And in total we can predict the tumor with an accuracy of more than 99%, and AC_SCC – 95%.