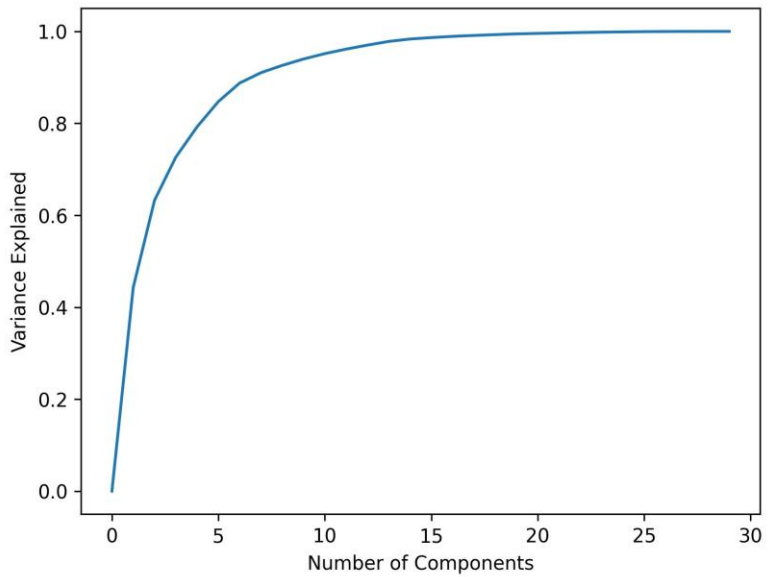
Examination of different numbers of principal components for PCA



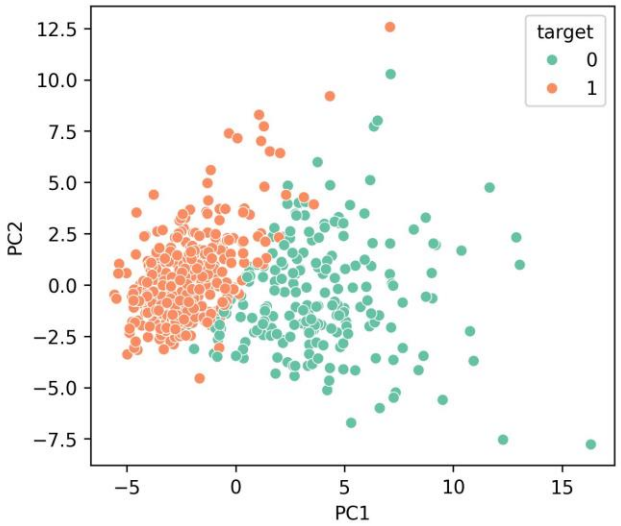
4, 5 or 6 principal components represent the raw data in a relatively high degree (explained variance is about 75-90%).

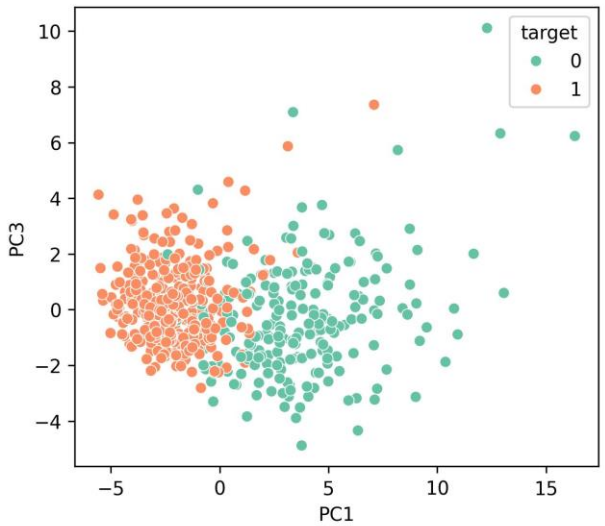
```
Number of principal components: 4
Total variance explained: 79.24 %

Number of principal components: 5
Total variance explained: 84.73 %

Number of principal components: 6
Total variance explained: 88.76 %
```

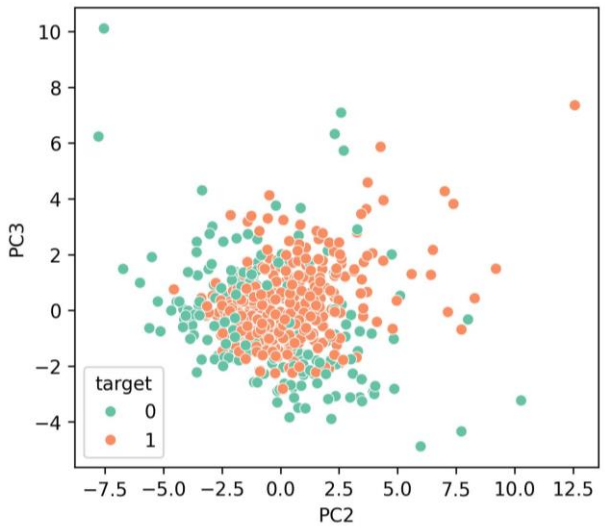
```
Variance explained by the different principal components:
PC1: 44.27 %
PC2: 18.97 %
PC3: 9.39 %
PC4: 6.6 %
PC5: 5.5 %
PC6: 4.02 %
```

Visualization of the principal components (up until 3)

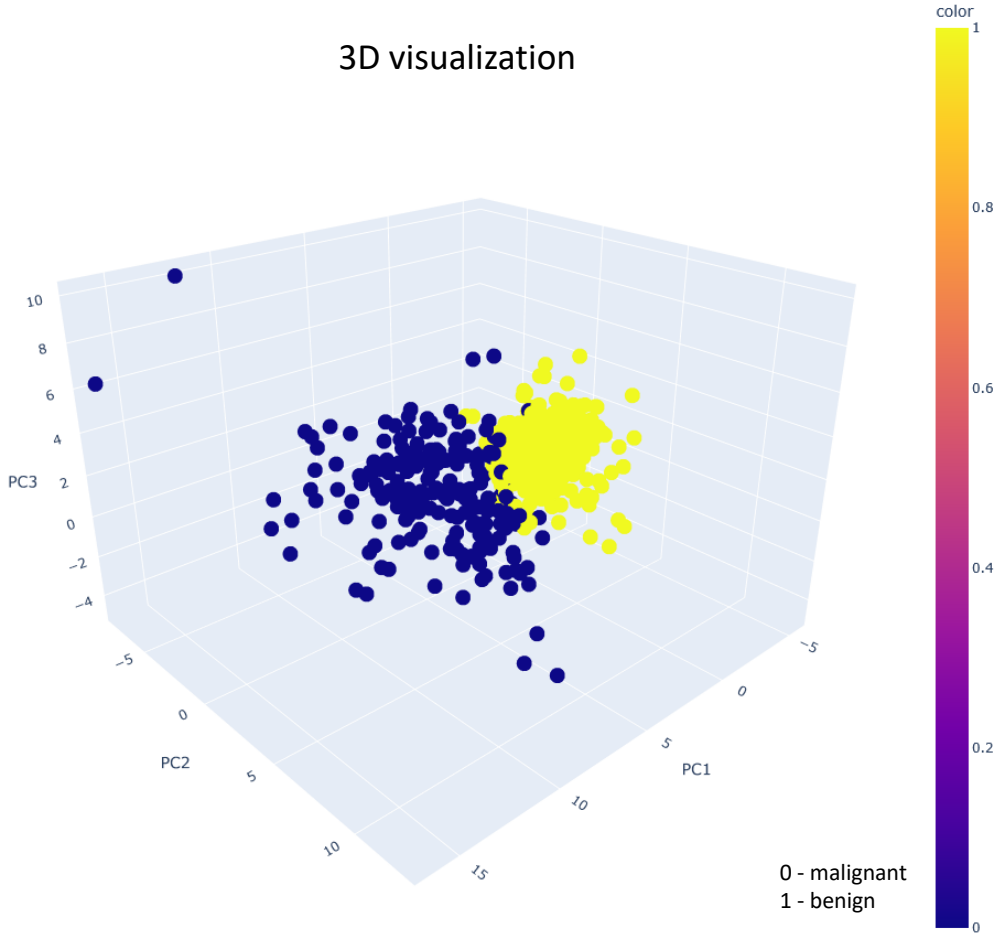
2D visualization







3D visualization



0 - malignant
1 - benign

Comparison of different ML models – all features vs. PCA

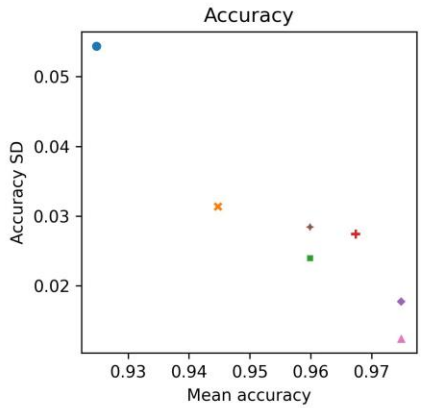
(Mean and SD values are calculated from the results of cross-validation)

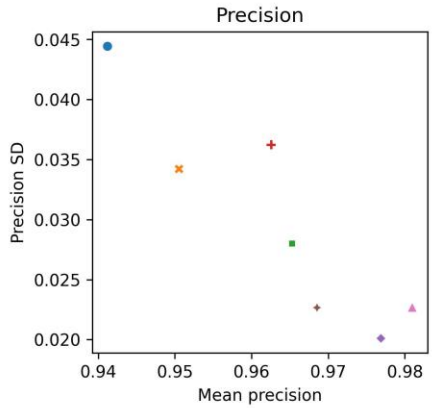
All features

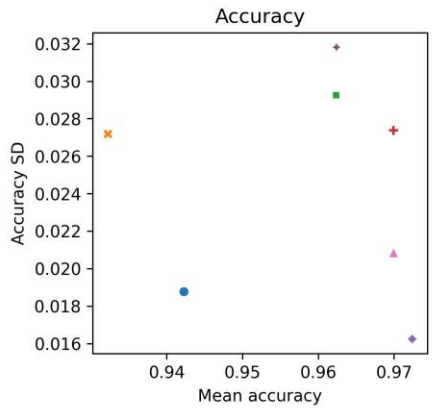
model	test_accuracy	test_precision	test_recall	test_roc_auc	test_accuracy_std	test_precision_std	test_recall_std	test_roc_auc_std
DecisionTreeClassifier(random_state=101)	0.924842	0.941213	0.940863	0.919167	0.054332	0.044429	0.053629	0.057396
GaussianNB()	0.944778	0.950542	0.964392	0.988062	0.031348	0.034211	0.035562	0.010236
GradientBoostingClassifier(random_state=101)	0.959905	0.965298	0.972314	0.991070	0.023950	0.027995	0.017629	0.008550
KNeighborsClassifier()	0.967405	0.962576	0.988157	0.986734	0.027409	0.036229	0.010812	0.011232
LogisticRegression(random_state=101)	0.974873	0.976886	0.984157	0.997555	0.017735	0.020100	0.021696	0.002215
RandomForestClassifier(random_state=101)	0.959905	0.968543	0.968392	0.988142	0.028425	0.022660	0.029752	0.012297
SVC(random_state=101)	0.974905	0.980967	0.980157	0.997558	0.012461	0.022692	0.019806	0.002460

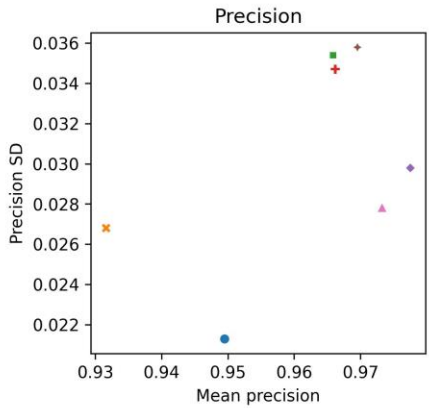
5 principal components from PCA

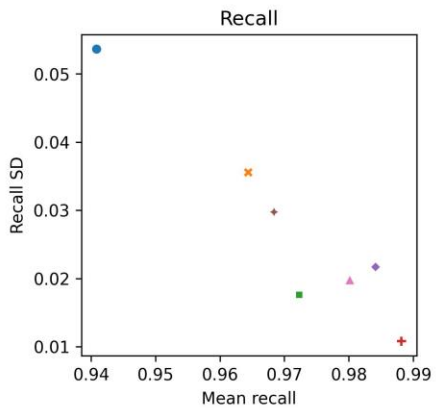
model	test_accuracy	test_precision	test_recall	test_roc_auc	test_accuracy_std	test_precision_std	test_recall_std	test_roc_auc_std
DecisionTreeClassifier(random_state=101)	0.942278	0.949520	0.960392	0.935713	0.018771	0.021285	0.024260	0.020899
GaussianNB()	0.932247	0.931642	0.964314	0.985779	0.027185	0.026799	0.025870	0.009961
GradientBoostingClassifier(random_state=101)	0.962373	0.965861	0.976157	0.988239	0.029248	0.035397	0.021768	0.016684
KNeighborsClassifier()	0.969937	0.966204	0.988157	0.988204	0.027372	0.034704	0.010812	0.010556
LogisticRegression(random_state=101)	0.972405	0.977533	0.980157	0.997693	0.016250	0.029790	0.019806	0.002461
RandomForestClassifier(random_state=101)	0.962405	0.969547	0.972157	0.988992	0.031807	0.035794	0.026747	0.013278
SVC(random_state=101)	0.969937	0.973264	0.980235	0.996886	0.020850	0.027827	0.013867	0.003464

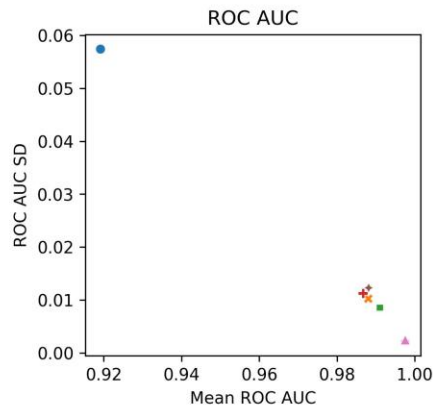












Training the models with 5 principal components from PCA provided the highest overall performance from the tested options. Compared to using all features, a slight improvement can be observed in each case with PCA and standard deviations tend to be lesser. In general, model performances following PCA are as good as the results of feature selection.