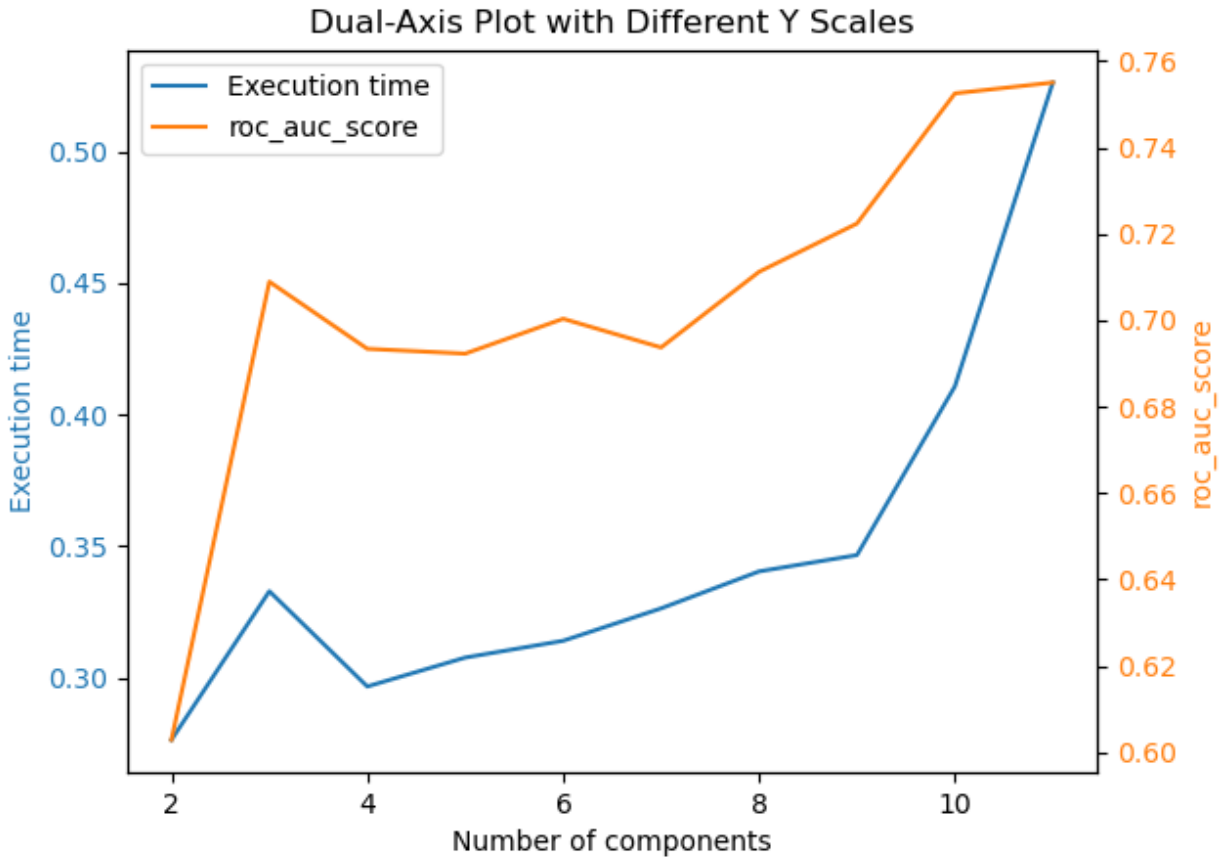
- Get dataset from Kaggle (any tabular dataset you want)
- Make simple classifier / regressor on the dataset
- Reasonably reduce dataset dimensionality
- Plot explained variance
- Explain chosen number of components
- Retrain the same classifier / regressor on the dataset with reduced dimensionality
- Compare accuracies / MSEs and speed of the two approaches (with and without dimensionality reduction)

Мабуть вибрав не найоптимальніший датасет для цієї задачі - датасет якості вина.

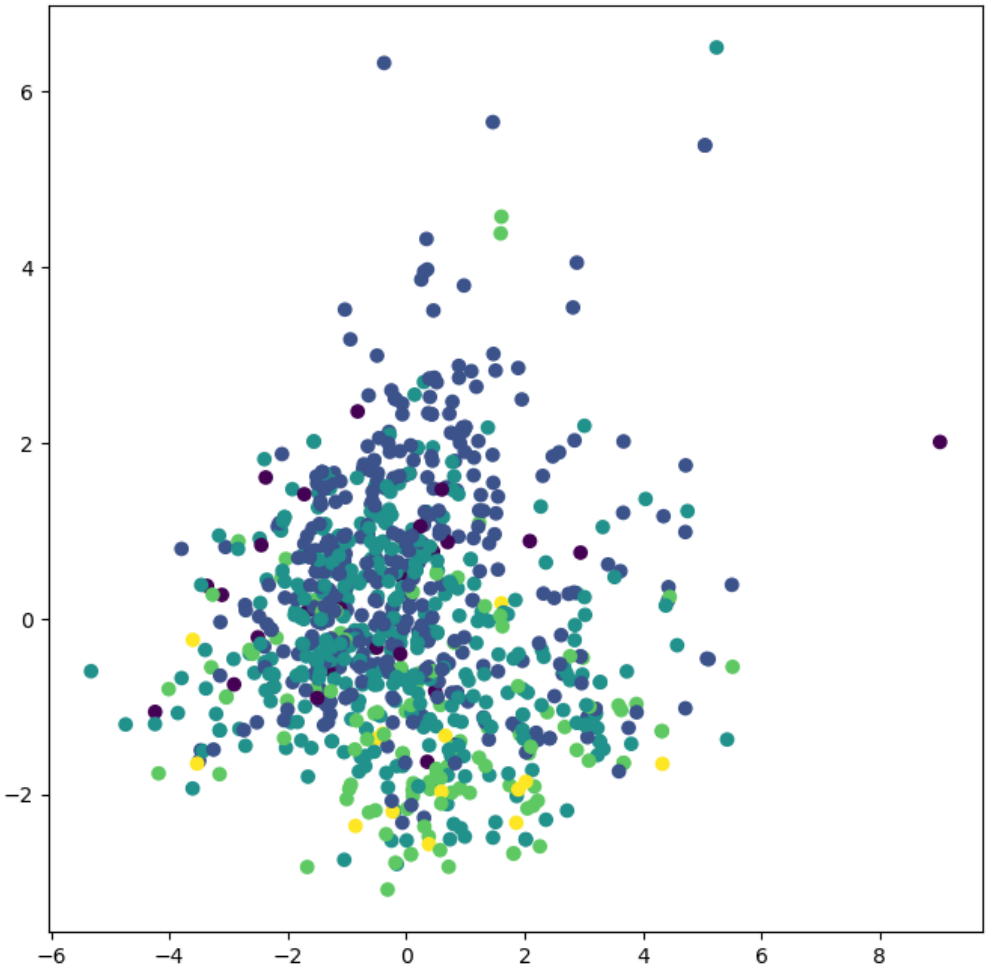
Побудував графік зміни часу виконання в секундах та результату метрики roc_auc_score в залежності від розмірності датасету.

Отримав наступні значення:

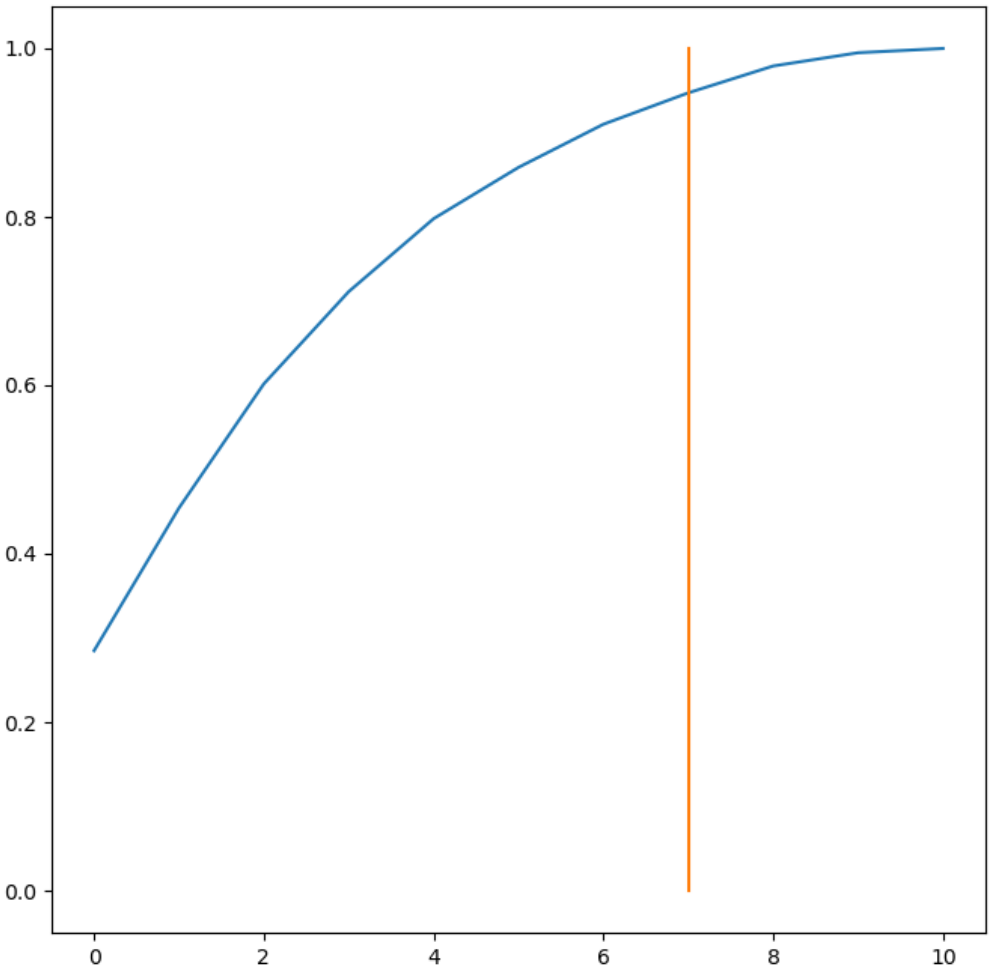


Як видно після 7 компоненту час виконання скрипта різко зростає, при втраті всього 0.04 значення roc_auc_score

Невисокий рівень roc_auc_score пояснюється збалансованістю датасету, що чітко видно з графіку розподілу класів:



Як видно з графіку комулятивних сум **explained_variance_ratio_** загалом на 7 компоненті можна досягти збереження 90% всієї інформації датасету



Враховуючи час виконання скрипта, оптимальне значення збереження точності (91%) та некритичність датасету з точки зору “нашкодити комусь” виглядає що раціональним рішенням буде зменшити перелік компонент до 7

[**0.28482615 0.16946752 0.14746712 0.10951631 0.08661631 0.06092188**
0.05121262 0.03726945 0.0318998 0.01573143 0.0050714] 1.0