Classification and Clustering

Classification:

I chose to work with two classification datasets. These are: The Titanic Dataset and Heart Disease Dataset. I took a cross_val_score before and after hyperparameter tuning.

Titanic:

The purpose of this dataset is to predict if a person was going to survive the titanic disaster or not. In my preprocessed dataset I base this on these attributes: Ticket class, sex, age group, fare, embarked and family member count.

I chose to use these three algorithms: XGBoost due to its popularity in the ML field, Random Forest due to it being a classic and Gaussian Naïve Bayes due to it being shown in class.

I got the best accuracy with XGBoost with Hyperparameter tuning. The accuracy measured with skleanrs accuracy_score function was 82.06%.

Heart Disease:

The purpose of this dataset is to predict if a person is more or less likely going to have a heart attack. The dataset needs no preprocessing and therefor contains every attribute in the original dataset which can be found here: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

I chose to use the same three algorithms as in the other classification problem to better be able to compare the algorithms across different datasets.

I got the best accuracy with Random Forest with Hyperparameter tuning. The accuracy was 82.35%.

Clustering:

I chose to work with one clustering dataset: Airline satisfaction dataset. This is originally a classification dataset, so I had to remove the ground truth. The dataset consists of 22 columns where airline passengers have given a score on 22 different metrices. The goal of this dataset is to cluster satisfied customers together and un-satisfied customers together.

I chose to use these three algorithms: Kmeans due to it being a standard clustering algorithm for newbies, Birch due to it being good with large datasets, and Agglomerative due to it being shown in class.

I got the best accuracy measured with silhouette_score with Agglomerative clustering. The accuracy was 81%.

Final Reflections:

I managed to improve score on every model after the hyperparameter tuning. Some had barely any difference while others had quite a decent change.

I also noticed that GridSearch combined with cross_val_score takes a long time, even on a beefy computer.

For the Birch and Agglomerative clustering I had to reduce the dataset with 90 000 rows.