HW2.1_Titanic Report
Name & Surname: Atena Jafari Parsa
Course: AIN3001 - Machine Learning

Introduction

The Titanic Machine Learning competition on Kaggle has a goal of predicting the survival of passengers based on various attributes such as age, gender, class, and embarkation location. The task is to build a machine learning model to predict whether a passenger survived or not during the tragic event of the Titanic disaster.

The dataset is divided into a training set, used to train our model, and a test set where predictions are to be made. The features include both numerical and categorical attributes, and careful preprocessing is required to handle missing values and convert categorical data into a format suitable for machine learning algorithms.

Our approach involves the creation of preprocessing pipelines for numerical and categorical features, utilizing tools such as SimpleImputer for handling missing values, StandardScaler for standardizing numerical features, and OrdinalEncoder for encoding categorical features. The ultimate goal is to develop a robust and accurate predictive model that generalizes well to new, unseen data.

Three libraries were added to the beginning of the code:

```
from sklearn.pipeline import Pipeline#This library is essential for
constructing a sequence of data processing steps, #facilitating a clean and
organized workflow in machine learning pipelines.
from sklearn.impute import SimpleImputer#This library is used to handle missing
values in the dataset by replacing them #with a specified strategy, such as the
median or most frequent value.
from sklearn.preprocessing import StandardScaler, OrdinalEncoder#StandardScaler
library  is necessary to standardize #numerical features, ensuring they have a
mean of 0 and a standard deviation of 1, which is often important for machine
#learning algorithms. OrdinalEncoder library  is employed to encode categorical
features into numerical values, making #them suitable for input into machine
learning models.
```

It is necessary to establish two essential preprocessing pipelines for the Titanic dataset. The numerical pipeline addresses missing values in numerical features using the median as a replacement and standardizes the features. The categorical pipeline handles missing values in categorical features by replacing them with the most frequent

values and then encodes these categorical variables into numerical representations using ordinal encoding.

The added libraries for the prediction task and their purpose:

```
from sklearn.svm import SVC#This library imports the Support Vector Classifier as it is a ML model commonly used #for classification tasks.

from sklearn.neighbors import KNeighborsClassifier#Imports the k-Nearest Neighbors classifier, a simple and effective #algorithm for classification based on nearest neighbors.

from sklearn.ensemble import RandomForestClassifier#Imports the Random Forest classifier, an ensemble learning method #that combines multiple decision trees.

from sklearn.model_selection import cross_val_score#Used for cross-validation, providing an efficient way to assess a #model's performance by splitting the data into multiple subsets and evaluating the model on each subset.

from sklearn.metrics import accuracy_score#The metric used to measure the accuracy of a classification model by  #comparing the predicted labels to the true labels.
```

The accuracy results:

SVC Average Accuracy: 0.8204494382022471

KNN Average Accuracy: 0.7957553058676654

Random Forest Average Accuracy: 0.8149063670411986

Best Classifier: SVC with an average accuracy of 0.8204494382022471

Support Vector Classifier (SVC):

As SVC is effective in high-dimensional spaces, it is also powerful in capturing complex relationship among the attributes. SVC is sensitive to dataset characteristics and, as the Titanic dataset has features that align well with the strengths of SVC (e.g., clear separation between classes), it performs better.