Titanic Survivor Prediction Project

Project Summary

The code above is a Python script that builds and tests a logistic regression model on the Titanic dataset. The script uses the pandas library to read in CSV files containing training and testing data, and preprocesses the data by replacing missing values, converting categorical variables into numerical ones, and filling in age and fare values with medians.

The preprocess() function takes two dataframes, targetDF and sourceDF, and modifies targetDF based on values in sourceDF. The function replaces male and female sex values with 0 and 1, respectively, and fills in missing age and fare values with the median. Additionally, the function fixes two errors in the data by filling in missing embarked values with the mode, converting categorical embarked values into numerical ones, and converting all not-available embarked values to 2.

The buildAndTestModel() function reads in the training data, preprocesses it using the preprocess() function, and builds a logistic regression model using the LogisticRegression() function from scikit-learn's linear model module. The function selects certain input columns and output columns from the preprocessed data and uses cross-validation to calculate the average accuracy score of the model. The function then reads in the testing data, preprocesses it using the preprocess() function, predicts the survival outcomes for the passengers in the testing data using the built logistic regression model, and saves the predictions to a CSV file for submission to Kaggle.

Overall, the program is designed to predict the survival of passengers on the Titanic using a real-life data set. To accomplish this, the program implements the K-Nearest-Neighbor Algorithm by creating testing and training data sets for k-fold cross validations and 1-NN algorithms using Python libraries and custom Python classes. The key focus of the program is on preprocessing the data, avoiding data leakage, and making models from pre-defined classes. The program takes into account a variety of attributes, such as Room Numbers, Floor, Patch, Cabin, and Age, to predict whether a passenger survived the Titanic sinking. Overall, the program

provides a practical example of how to use machine learning algorithms to analyze real-life data sets and make accurate predictions.