Report for Assignment 1

Muntashir Bin Solaiman

Student id - 22021468

Contents

Task 1

a) Before preprocessing the dataset had 26215 rows and 10 columns. The variable types were integer for income, age and hours per week. Object was the variable type for work class, education, marital status, occupation, relationship, race and sex. The dataset contained duplicated rows. In 1396 and 1401 Work class and occupation, respectively, had no values.

b) The dataset had 26215 rows and 10 columns. Isna().sum() method was invoked in answer (a). Dropna() method was invoked on the dataset. The shape of the dataset was printed. 1396 and 1401 rows of Work class and occupation, respectively, had missing values. The rest had no missing values. The rows with missing values were removed. The dataset had 24814 rows, and 10 columns post processing.

c) The dataset had 24814 rows and 10 columns. In answer (a) duplicated().any() method was invoked on the dataset. Dropna() was invoked. The shape was printed. The result of duplicated().any() was true. The duplicated rows were removed. The dataset had 21537 rows, and 10 columns post processing.

d) Work class column had Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked as values. Education had Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th and Preschool as values. Marital status had Married, NotMarried, Separate and Widowed as values. Occupation had Tech-support, Craft-repair, Other-service, Sales, Exec-managerial,Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical,Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv and Armed-Forces as values. Relationship had Wife, Own-child, Husband, Not-in-family, Other-relative and Unmarried as values. Race had White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other and Black as values. Sex had Male and Female as values.

   Replace() method was invoked on the sex column. The parameters were 0 for male and 1 for female. Head() method was invoked. Replace() method was invoked on the education column. Preschool had 0, 1st-4th had 1, 5th-6th had 2, 7th-8th had 3,9th had 4, 10th had 5, 11th had 6, 12th had 7, HS-grad had 8, Some-college had 9, Assoc-voc had 10, Assoc-acdm had 11, Assoc-acdm had 12, Bachelors had 13,Masters' had 14, Prof-school had 15 and Doctorate had 16 as parameters. The head() method was invoked. Get_dummies() was used on the rest of the categorical variables. Head() was invoked again. The shape was printed.

The sex column had 0 and 1 as values. The values for education ranged from 0 to 16. For each of the possible values a categorical variable had, a column was created. The datatype was bool. This indicated whether the specific value was present for that row. This was the result of get_dummies().

There were 21537 rows and 36 columns post processing.

e) The values were converted into NumPy arrays. The first column was declared as the target variable. This was the column we were trying to predict. From the second column to the last were defined as input variables. Train_split_test() method was called. The parameters were the input variables, target variables, test size of 0.1 and random state of 1. The method returned four values. 10% of the data was reserved for testing. 80% for training. The data was split into input training variables, target training variable, input testing variables and target testing variables.

f) The array had values of various range. The minmaxscaler() and fit() method were used. A normalized dataset was created. The dataset's values were transformed to a range from 0 to 1. The fit() method identified the maximum and minimum value. The minmaxscaler() changed the values within the range of 0 and 1. The transform() method was called on the normalized dataset twice. The first parameter was input training data. The second parameter was input testing data. The minimum and maximum values from mimaxscaler() and fit() were used to scale the input training set and input test set.

Task 2

Logistic regression was a classification algorithm. It predicted a binary outcome. It used a sigmoid function. The function converted the input values to 0 or 1. The gradient descent algorithm was used to adjust the weights of the input.

Support vector machine (SVM) was a supervised machine learning algorithm. A hyperplane was used. The hyperplane separated the data into classes. The hyperplane's margins were maximally increased. If the data was linearly non- separable then different kernels were used.

A 10-fold cross validation was defined. It was passed to cross_val_score() method. The first parameter was different. In one instance it was logistic regression model. In another instance it was SVM model. The accuracy score for Logistic regression model and SVM were 0.8073043619880025 was 0.8042087060931156 respectively. The score for logistic regression was slightly higher.

The parameters for logistic regression were penalty, c, and solver. The penalties were l1 and l2. C were 1 and 10. Solver were saga and liblinear. Max iter was 150. All possible combinations were tried through grid search. The best score was 0.8079752364559732.

The parameters for SVM were kernel, C, degree, gamma. Kernal had linear and poly. C was either 1 or 10. Degree was either 3 or 8. Gamma was auto and scale. All possible combinations were tried through grid search. The best score was 0.8079752364559732.

The accuracy score for both models were very high. The accuracy score for SVM increased by 0.0037665303628576. The accuracy score for logistic regression increased by 0.0006708744679707.

The optimized Logistic regression model scored 0.7859795728876509. The best parameters were {'C': 1, 'penalty': 'l2', 'solver': 'liblinear'}. The optimized SVM scored 0.7887650882079852. The best parameters were {'C': 1, 'penalty': 'l2', 'solver': 'liblinear'}. SVM performed slightly better. Their score differed by 0.0027855153203343.

Task 3

K means clustering was an unsupervised learning algorithm. It classified data into clusters. The data that had had the lowest distance with a cluster was placed in it. After each placement, the centroid of the cluster was updated. The process continued until all data was processed. I chose 2 clusters. This represented the binary output of the target variable. The income was higher than 50 000 or lower than that. The first cluster had 5781 samples. The second cluster had 13602 samples. Row 3378 and row 16016 best represented cluster 1 and cluster 2 respectively. They were closest to the centroids of each cluster. Both rows differed in age, education, work class local government, work class private. Row 3378 had age of 43, education as some college, work class local government as true and work class private as false. Row 16016 had age as 69, education as 5th to 6th grade, work class local government as false and work class private as True. For all the columns they were similar. Both did not earn less than 50 000. The accuracy score for K means clustering was 0.3867223769730734. The accuracy score for logistic regression was 0.7887650882079852. The accuracy for svm score was 0.7859795728876509. This meant that logistic regression predicted income the best. Its performance was followed closely by SVM. K means clustering did not predict the income well.