Mousumi Deb
Data scientist
Machine learning

Machine Learning Project Report

Introduction:

Machine learning is a branch of artificial intelligence that focuses on developing algorithms and models that can learn relationships from data and make predictions based on that. In this project, I will be using the mnist dataset and machine learning models to predict data. The mnist dataset is a well-known dataset used to evaluate the performance of machine learning algorithms and is often referred to as the "hello world" of machine learning.

Question:

In this dataset which model performs better to recognize the digits image and gives the best accuracy score, and which model is accurate? So, for this testing, several models and exploring the dataset need to find the accuracy score, R 2, precision, recall, RMSE and MAE.

Data Description:

The mnist dataset is a well-known dataset used to evaluate the performance of machine learning algorithms. It consists of 60,000 hand-written digits for training and 10,000 hand-written digits for testing. Each digit is a grayscale image of size 28*28 pixels, giving a total of 784 pixels per image. Each pixel is represented as a value between 0-255, where 0 corresponds to white and 255 corresponds to black.

(EDA)Analysis of the dataset:

Load the mnist dataset by using fetch_openml, define the data and target key as X, y variable, and print it as DESCR(describing the dataset). Plotting the frequency distribution of each digit class for this uses a bar plot to visualize. Then plot a random sample of images from each digit class with a specified size of 10 * 10 and then loops through each digit class and select the random images from the corresponding dataset. Creating the data frame to see the features' names in X, y and see the statistical value of the dataset by describe(), info() function. Use the unique () function to see how many unique values are in the dataset. checking the pixel value of some of the columns for this using the group by (), agg() function. Checking the null value in the datasets for data cleaning use IsNull (),sum(), isna() functions. There is no missing value so visualize some digits from the dataset.

Using the boxplot to detect outliers in the dataset. After checking the box plot there is no outliers in the datasets so now implement the clustering algorithms Standardsacaler() and

then print the mean and standard deviation of the normalized pixel values and check again the shape of the scaled data.

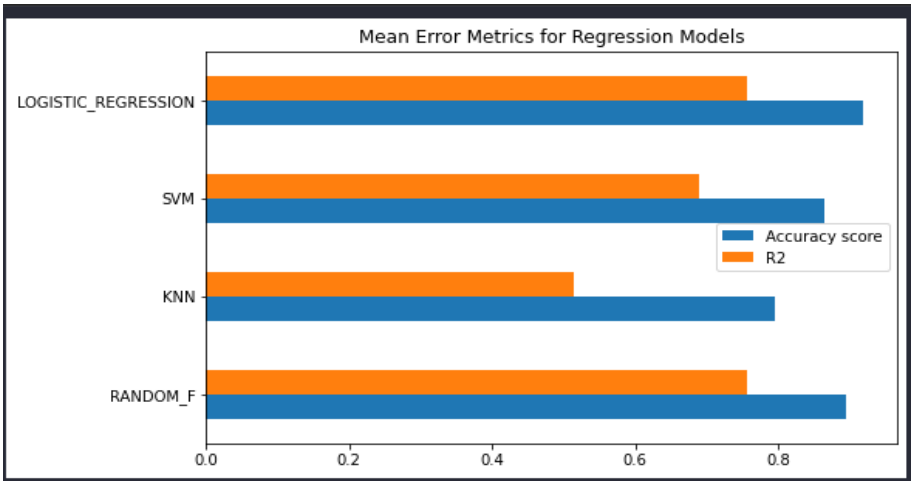
Now implementing the PCA algorithm to scaled data and use n_components to transform data into principal components. And the plot of the PCA data into a scatter plot and visualized in 3d.

For modeling the dataset first I split the data into train and test. And using only 1000 data points to speed up the model training.

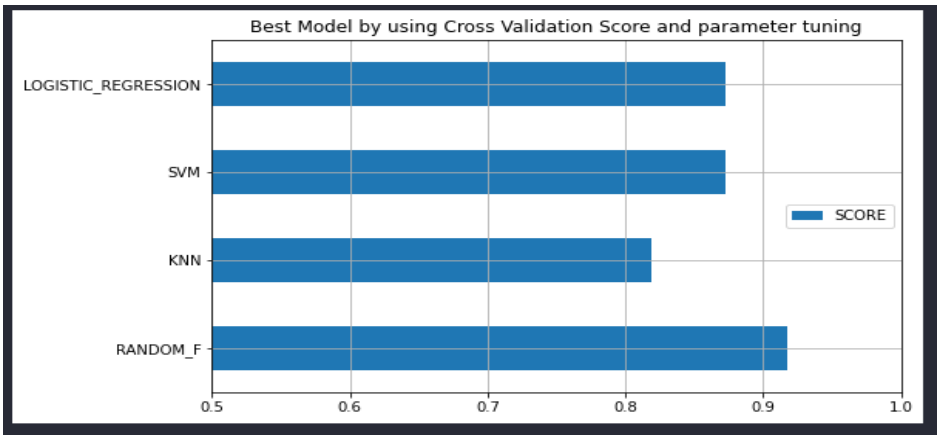
Method and models:

In this project, I used **RandomForestClassifiers** with 100 decision trees and trained the data using fit() function. then evaluate the model and computes root mean square error (RMSE), mean absolute error(MAE), and R- square (R^2).

Also testing the logistic regression model, KNeighborsClassifiers and SVM model then plot the model evaluation metrics for R^2 and accuracy score.



Then testing random forest, KNN, SVM, and logistic regression with k-cross-validation and evaluation scores using cross_val_score() and then plotting all models with cross-validation scores



For evaluating machine learning use precision as a performance metric and it is useful to measure how many positive predictions can be made by the model.

Result and analysis

Model	**Precision**	**recall**	**Accuracy score**
Logistic regression	0.89	0.90	0.89
Random Forest	0.92	0.92	0.92
KNN	0.807	0.803	0.795
SVM	0.87	0.86	.0.87

Analysis:

I got the best accuracy score by using random forest classifiers with 100 trees, training the classifiers on the training set, and evaluating the classifiers' accuracy on the test data. The hyperparameter n_estimator = 100 for 1000 mnist datapoint worked well and achieved 0.92 accuracy score on the testing dataset.

Conclusion

In conclusion of my project, when I analyzed the data I got that the datasets are no missing values and no outliers as well. When I checked in the t-sne plot properly I checked the digit 2 value is scattered. In the machine learning model, I have used 4 models t.ex logistic regression, KNN, SVM, and Random forest. All of them work very well for predicting the digit but among them, random forest s the best accuracy, and it trains well. So the best model is RandomForestClassifiers with a precision of 0.92.

Further potential development:

Further potential development could have been used till ex: convolutional neural network and can be used ensemble methods for combining multiple classifiers to improve accuracy.

Short statement

As it is a large dataset, it was hard to work whole data set, it takes a long time to execute so then I take less data to speed up my work. Also, I find it hard when I try to data dimensions with t-sne.

I think I should have to get VG because I tried to analyze the dataset elaborately with pca and t-sne and plotted the dataset and tried 4 models to train.