Disease Classification Model Using Public Health Data

Kalp Shah

November 27, 2023

Aim: Develop a machine learning model for disease classification.

1 Introduction

By using machine learning algorithm we have developed an model used for healthcare system. A data set of various symptom to predict the disease,with the help of machine learning model we have classifier the disease on the base of symptoms using supervised learning model such as logistic regression and decision tree.

2 Methods

Uploading the data set to the program using pandas.performing the preprocesing operation on the data set and make it to proper feature for the machine learning model.After that developed logistic regression using training data and apply on test data.then evaluated the model using accuracy,precision,recall values.same process apply to decision tree model. compare both the model using graph.

3 Step by step process

- **step1:** import python library such as numpy,pandas,mathplotlib that are used for machine learning.

- **step2:** import the disease dataset using pd.read_csv method.
- **step3:** remove the missing values colounm by not selecting them.at the same time divide the dataset to attribute and target value.
- **step4:** convert the categrocial values to numeric values by performing one hot label encoding using pd.get_dummies.
- **step5:** split the data into training and test data using split_train_test method from sklearn.model_selection.
- **step6:** apply feature scaling using StandardScaler form sklearn.preproces
- **step7:** Developed logistic regression model on the training data by using LogisticRegression method form sklearn.linear_model.
- **step8:** Apply model on the test data.
- **step 9:** evaluation of model is done on the base of accuracy,precision,recall values that can be get by from sklearn.metrics import accuracy_score,precision_score,recall_score.
- **step10:** form step 7 to step 9 apply for decision tree model.
- **step11:** Comparing both model on the base of accuracy,precision,recall values using bar graph by mathplotlib library.

4 Result

accuracy=0.945
precision=0.969
recall=0.945

By seeing the graph,all the factor has same value for both.Hence we can apply any of them for the disease classification.

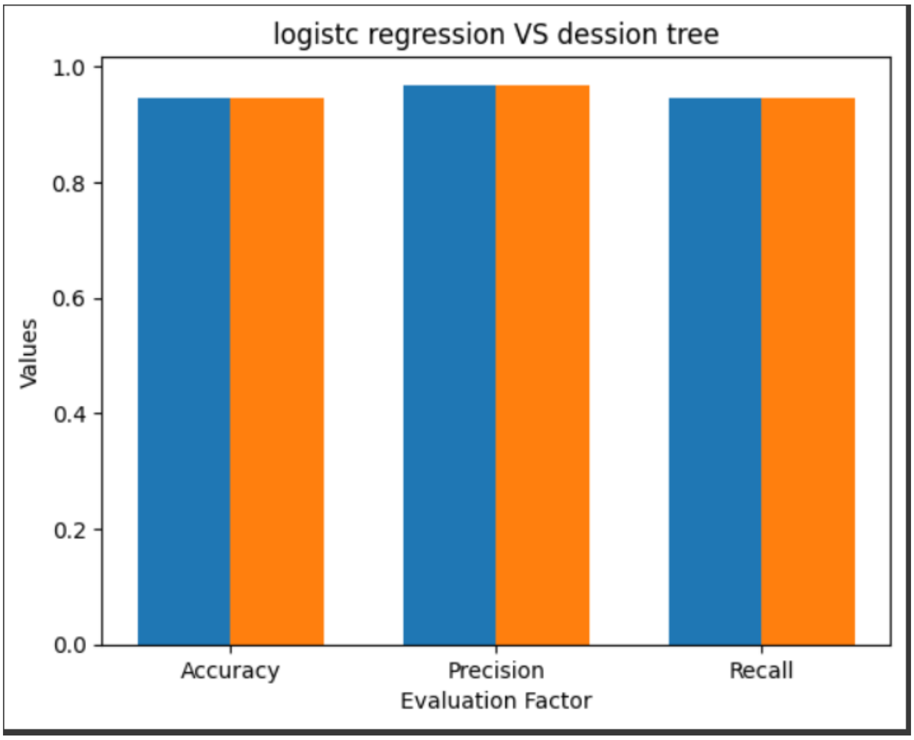

Figure 1: logistic regression vs decision tree

5 Tool & Technology

Programming languages: python.
Library: Numpy,Pandas,Mathplotlib,Scikit.
Dataset: dataset-symptoms.csv.
Platform: Google Colab.