Pallavi Das - Mastery 2: Prediction
5/17/2024

Mastery 2: Prediction

Section 1: Introduction

Lung cancer is one of the leading causes of cancer-related deaths worldwide. Early detection and prediction of lung cancer risk factors are critical for improving patient outcomes and survival rates. By analyzing various factors such as smoking habits, environmental exposures, and genetic conditions, it is possible to better understand the likelihood of developing lung cancer. The use of various prediction models can allow us to aid in the early diagnosis of the disease.

Dataset Overview

The dataset "Lung Cancer Survey" contains various features related to the risk factors and symptoms associated with lung cancer. It includes a mix of categorical and numerical variables, which provides a comprehensive basis for both regression and classification tasks. The target variable for the classification task is whether the individual has lung cancer or not. The columns in the dataset include:

- Gender: M(male), F(female)
- Age: Age of the patient
- Smoking: YES=2 , NO=1.
- Yellow fingers: YES=2 , NO=1.
- Anxiety: YES=2 , NO=1.
- Peer_pressure: YES=2 , NO=1.
- Chronic Disease: YES=2 , NO=1.
- Fatigue: YES=2 , NO=1.
- Allergy: YES=2 , NO=1.
- Wheezing: YES=2 , NO=1.
- Alcohol: YES=2 , NO=1.
- Coughing: YES=2 , NO=1.
- Shortness of Breath: YES=2 , NO=1.
- Swallowing Difficulty: YES=2 , NO=1.
- Chest pain: YES=2 , NO=1.
- Lung Cancer: YES , NO.

The dataset contains 309 entries, each representing a respondent with various recorded attributes related to lung cancer risk factors. The mix of categorical and numerical data types requires appropriate preprocessing steps, such as encoding categorical variables and scaling numerical features, to ensure effective model training and evaluation.

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
0	M	69	1	2	2	1	1	2	1	2	2	2	2	2	2	YES
1	M	74	2	1	1	1	2	2	2	1	1	1	2	2	2	YES
2	F	59	1	1	1	2	1	2	1	2	1	2	2	1	2	NO
3	M	63	2	2	2	1	1	1	1	1	2	1	1	2	2	NO
4	F	63	1	2	1	1	1	1	1	2	1	2	2	1	1	NO
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
304	F	56	1	1	1	2	2	2	1	1	2	2	2	2	1	YES
305	M	70	2	1	1	1	1	2	2	2	2	2	2	1	2	YES
306	M	58	2	1	1	1	1	1	2	2	2	2	1	1	2	YES
307	M	67	2	1	2	1	1	2	2	1	2	2	2	1	2	YES
308	M	62	1	1	1	2	1	2	2	2	2	1	1	2	1	YES

Objectives

Using the lung cancer survey dataset, the primary aim of this report is to enhance the early detection and prediction of lung cancer in patients. This involves conducting an in-depth analysis to identify key risk factors and symptoms associated with lung cancer, as well as developing predictive models to accurately identify at-risk individuals and gain insights into the likelihood of developing the disease. Specifically, the mastery aims to compare and evaluate the performance of different predictive models, such as Linear Regression and K-Nearest Neighbors, to ensure their reliability and accuracy. By analyzing these predictive models and key risk factors for lung cancer, this can aid in creating insights for healthcare professionals and support medical treatment approaches for lung cancer patients.

Section 2: Data Preprocessing

Before performing the prediction tasks, the categorical variables were encoded by one-hot encoding Gender and giving it binary indicators for Male and Female. Binary categorical variables (e.g., SMOKING, YELLOW_FINGERS, etc.) were also already numerically coded, with 2 for 'YES' and 1 for 'NO', and thus were left untouched. Additionally, the Age column was standardized using StandardScaler to ensure it contributes equally to the model training process. ColumnTransformer was used to apply the transformations of the one-hot encoding to categorical features and standard scaling to numerical features. The target variable LUNG_CANCER lastly was encoded to binary, with 1 for 'YES' and 0 for 'NO'.

	GENDER_F	GENDER_M	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN
0	0.0	1.0	0.771850	1.0	2.0	2.0	1.0	1.0	2.0	1.0	2.0	2.0	2.0	2.0	2.0	2.0
1	0.0	1.0	1.381829	2.0	1.0	1.0	1.0	2.0	2.0	2.0	1.0	1.0	1.0	2.0	2.0	2.0
2	1.0	0.0	-0.448107	1.0	1.0	1.0	2.0	1.0	2.0	1.0	2.0	1.0	2.0	2.0	1.0	2.0
3	0.0	1.0	0.039876	2.0	2.0	2.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	2.0	2.0
4	1.0	0.0	0.039876	1.0	2.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	2.0	2.0	1.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
304	1.0	0.0	-0.814095	1.0	1.0	1.0	2.0	2.0	2.0	1.0	1.0	2.0	2.0	2.0	2.0	1.0
305	0.0	1.0	0.893846	2.0	1.0	1.0	1.0	1.0	2.0	2.0	2.0	2.0	2.0	2.0	1.0	2.0
306	0.0	1.0	-0.570103	2.0	1.0	1.0	1.0	1.0	1.0	2.0	2.0	2.0	2.0	1.0	1.0	2.0
307	0.0	1.0	0.527859	2.0	1.0	2.0	1.0	1.0	2.0	2.0	1.0	2.0	2.0	2.0	1.0	2.0
308	0.0	1.0	-0.082120	1.0	1.0	1.0	2.0	1.0	2.0	2.0	2.0	2.0	1.0	1.0	2.0	1.0

309 rows × 16 columns

```
0      1
1      1
2      0
3      0
4      0
      ..
304    1
305    1
306    1
307    1
308    1
Name: LUNG_CANCER, Length: 309, dtype: int64
```

Section 3: Linear Regression Model

Setup/Methods

Before performing a linear regression model, the dataset was split into a training set and a validation set. Using a train-test split method ensured a random and reproducible split, with 50% of the data being allocated to the training set and the remaining 50% to the validation set. The training set size was 154 samples, and the validation set size was 155 samples.

A Linear Regression model through sklearn was initialized and trained on the training data. This training process was aimed to learn the relationship between the predictor variables and the target variable of the lung cancer diagnosis. The model training process was completed successfully, indicating that the model parameters were learned from the training data without significant convergence issues.

Model Results/Analysis

The trained Linear Regression model was used to make predictions on the validation set. The performance of the model was evaluated using two key metrics using the scikit-learn library, being the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE):

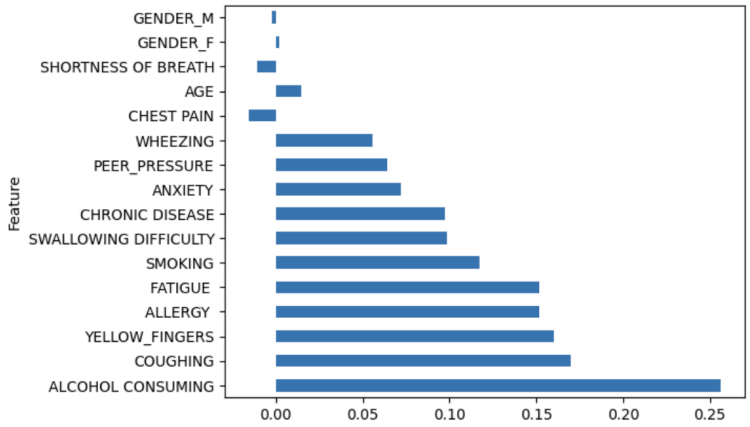
- **Root Mean Squared Error (RMSE):** 0.284
- **Mean Absolute Error (MAE):** 0.204

The Root Mean Squared Error (RMSE) was identified to be 0.284, indicating that, on average, the predicted values deviate from the actual values by about 0.284 units. Given that the lung cancer diagnosis is binary (0 for NO, 1 for YES), this RMSE value suggests that the model is making reasonably accurate predictions in distinguishing between patients with and without lung cancer. However, there is room for improvement to obtain a lower RMSE. The Mean Absolute Error (MAE) was identified to be 0.204, indicating that, on average, the model's predictions deviate from the actual values by about 0.204 units. For the binary target variable, this means that the model's predictions are quite close to the actual values, with an average error of slightly more than 0.2. This relatively low MAE value suggests that the model is performing well and making accurate predictions for the presence or absence of lung cancer.

Loading Plot

A **loading plot** was also created to visualize the coefficients of the linear regression model. This plot helps to understand the impact of each feature on the predictions for lung cancer.

```
Root Mean Squared Error (RMSE): 0.2842291899554955
Mean Absolute Error (MAE): 0.20410250134402544
[151]: <Axes: ylabel='Feature'>
```



Based on the loading plot, the Alcohol Consuming feature has the highest positive coefficient, suggesting a strong association between alcohol consumption and the likelihood of being predicted to have lung cancer. Coughing, Fatigue, Yellow Fingers, and Allergy also have considerable positive coefficients, indicating that patients experiencing these symptoms are more likely to be predicted as having lung cancer. Smoking has a positive coefficient, reinforcing its role as a significant predictor of lung cancer. However, the coefficient is smaller compared to alcohol consumption and coughing, suggesting other factors may have a stronger direct association in this model. The Gender (Gender_M and Gender_F) features show very small coefficients, indicating that gender differences might not be as significant in this model's predictions. However, the negative coefficient for Gender_M suggests that male patients are slightly less likely to be predicted as having lung cancer compared to the baseline of female patients.

Section 4: K-Nearest Neighbors Regression Model

Setup/Methods

Similar to the Linear Regression model, the dataset is split into a training set and a validation set, ensuring a random and reproducible split, with 50% of the data allocated to the training set and the remaining 50% to the validation set.

Hyperparameter tuning is crucial for KNN regression to find the optimal number of neighbors (k). GridSearchCV was used to perform a search over a specified parameter grid, using **cross-validation** to evaluate model performance. The optimal k value was selected based on the cross-validation results. A parameter grid with k values ranging from 1 to 20 was defined, and 5-fold cross-validation within GridSearchCV was used, essentially splitting the training data into 5 subsets, and training and validating the model 5 times. With this process, the optimal k value came out to be **k = 8**.

Next, the KNN Regression model was initialized and trained using the training data and the optimal hyperparameters identified from the tuning process. The model was trained on the training set using the best k value of 8. The trained KNN Regression model was then used to make predictions on the validation set.

Model Results/Analysis

The performance of the trained KNN model was evaluated using two key metrics from the scikit-learn library, being the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE):

- **Root Mean Squared Error (RMSE):** 0.294
- **Mean Absolute Error (MAE):** 0.143

First, The Root Mean Squared Error (RMSE) was identified to be 0.294, indicating that, on average, the predicted values deviate from the actual values by about 0.294 units. Given that the lung cancer diagnosis is binary (0 for NO, 1 for YES), this RMSE value suggests that the KNN Regression model is making reasonably accurate predictions in distinguishing between patients with and without lung cancer. The Mean Absolute Error (MAE) was identified to be 0.143, indicating that, on average, the model's predictions deviate from the actual values by about 0.143 units. For the binary target variable, this indicates that the model's predictions are relatively close to the actual values, with an average error of less than 0.15. This low MAE value suggests that the KNN Regression model is performing well and making accurate predictions for the presence or absence of lung cancer.

While both models are effective in predicting lung cancer, the KNN Regression model based on k=8 provides more precise predictions on average, as indicated by its lower MAE, but the Linear Regression model, with a lower RMSE, suggests better consistency in its predictions. For practical purposes in a medical context, where predictions are more consistent, the Linear Regression model might be the better model for diagnosing lung cancer.

Section 5: K-Nearest Neighbors Classification Model

Setup/Methods

Similar to the regression models, the dataset was split into a training set and a validation set, ensuring a random and reproducible split, with 50% of the data allocated to the training set and the remaining 50% to the validation set. Additionally, **Hyperparameter tuning** is crucial for KNN regression to find the optimal number of neighbors (k). GridSearchCV was used to perform a search over a specified parameter grid, using cross-validation to evaluate model performance. The optimal k value was selected based on the cross-validation results. A parameter grid with k values ranging from 1 to 20 was defined, and 5-fold cross-validation within GridSearchCV was used, essentially splitting the training data into 5 subsets, and training and validating the model 5 times. With this process, the optimal k value came out to be **k = 5**. With the optimal k value of 5 identified, the KNN Classification model was initialized and trained using the training data. The trained KNN Classification model was then used to make predictions on the validation set.

Model Results/Analysis

The trained model was then used to make predictions on the validation set and evaluate the performance of the model using metrics like accuracy, precision, recall, f1-score, sensitivity, and specificity:

- **Accuracy:** 0.8516
- **Precision:** 0.8929
- **Recall:** 0.9398
- **F1-Score:** 0.9158
- **Sensitivity:** 0.9398
- **Specificity:** 0.3182

The **accuracy** of the model is 0.8516, meaning that out of all the predictions made by the model, 85.16% of them were correct. This indicates that the model performs well overall in distinguishing between individuals with and without lung cancer. The **precision** was 0.8929, indicating that when the model predicts a positive case of lung cancer, it is correct 89.29% of the time. This metric highlights the model's ability to avoid false positives. The **recall** metric is 0.9398, meaning the model successfully identifies 93.98% of the true positive cases with lung cancer. The high **f1-score** of 0.9158 provides a balance between precision and recall, and reflects both the model's accuracy in predicting positives and its ability to identify actual positive cases. The **sensitivity** metric was 0.9398, confirming that the model identifies 93.98% of actual positives. Lastly, the **specificity** metric was 0.3182, indicating that only 31.82% of actual negative cases of those without lung cancer are correctly identified by the model. The relatively low specificity suggests that while the model is good at identifying positive lung cancer cases, it has a higher rate of incorrectly predicting lung cancer for some patients who don't have the disease. Overall, this KNN Classification model has high recall and precision in identifying true lung cancer cases.

Section 6: Overall Results/Conclusion

In conclusion, both the Linear Regression and KNN regression models showed strong performance in predicting lung cancer, with the KNN model demonstrating more precise predictions as it had a lower MAE and the linear regression model showing better consistency as it had a lower RMSE. The KNN classification model, with an optimal k value of 5, achieved high accuracy and precision in identifying lung cancer cases. However, its low specificity indicates a higher rate of false positives. While the KNN regression model offers precise predictions, the linear regression model's consistency makes it a reliable choice for diagnosing lung cancer and thus the more preferred model. The high recall and precision of the KNN classification model make it an effective tool for early lung cancer detection, but the specificity needs to be improved to reduce false positives. Overall, this mastery highlights the strengths and potential improvements for each model, providing valuable insights for the early detection and accurate prediction of lung cancer within patients.