Risk Factors for COVID-19 Hospitalization

A **proposal** for Introduction to Data Mining Project work

Identification of the risk factors responsible for hospitalization due to COVID-19 infections is crucial for monitoring of patients before they develop a serious illness outcome. Several studies have been conducted to determine the risk factors for COVID infection hospitalization and death in different settings. In this project various machine learning (ML) models will be used for the classification of a publicly available datasets to determine the risk factors of hospitalization due to COVID-19 infection. The data was collected for the research aimed to characterize early symptoms, exposures, comorbidities, and other risk factors associated with hospitalization between December 2020 and June 2021. The data is contributed by the John Hopkins Bloomberg school of public health and is freely available on Human Data Exchange (HDX) website. It has 519 entries in 57 columns.

A comparative performance analysis of different classification algorithms will be used to analyze a set of risk factors associated with hospitalization due to COVID-19 infection. The overall process includes data pre-processing, explanatory data analysis and data visualization, algorithm selection for classification and feature predictions, model training and testing and model evaluation. Accuracy of the machine learning models will be used to evaluate the models and to compare models each other.

Risk Factors for COVID-19 Hospitalization

Introduction to Data Mining Project work

Introduction

Identifying and characterizing risk factors for severe COVID-19 is critical for identifying individuals who may benefit from increased monitoring before serious illness outcome. Risk factors for severe illness from COVID-19 with strong evidence supported by systematic reviews and meta-analyses include cancer, cerebrovascular disease, chronic kidney disease, chronic obstructive pulmonary disease, diabetes mellitus, cardiovascular disease, obesity, pregnancy, and smoking. Demographic risk factors, including older age and male sex, have also been associated with poor prognosis (2).

In this project various machine learning (ML) models are used for the classification of a publicly available datasets to determine the risk factors of hospitalization due to COVID-19 infection. The data was collected for the research aimed to characterize early symptoms, exposures, comorbidities, and other risk factors associated with hospitalization and death from COVID-19 in South Sudan (SSD) and Democratic Republic of Congo (DRC) between December 2020 and June 2021.

Description of the Data set

The data is a publicly available data from Human Data Exchange (HDX) website, provided by the Johns Hopkins Bloomberg School of Public Health. It is a deidentified dataset used for analysis presented in *"Risk Factors for Hospitalization and Death from COVID-19: A Prospective Cohort Study in South Sudan and Eastern Democratic Republic of the Congo*" by Leidman et al. The url link for the data is (humdata.org).

Data preprocessing / EDA

The collected data has 519 entries in 57 columns. 55 features (risk factors) and two outcome variables were included; hospitalization of the patient and whether the patient deceased or not if

hospitalized. For this project, only one of the outcome variables is used; whether the patient would be hospitalized or not based on the risk factors after having a positive covid test.

Python 3.9 language libraries are used for data processing.

Grouping/minimizing the features

For the sake of analysis in this project, among the 55 features in the data, 11 features that have potential effect on the outcome (risk factors that potentially affect the outcome of covid infection) has been selected. Some features were summed up (grouped) and included in one column, in this case the summary column of specific factors is included in the analysis; for example, for symptom, the column for 'any symptom' sum-up all different kinds of symptoms listed in separate column. Same principle is used for having any history of chronic disease. Features about the 'source of exposure' were dropped, as all cases are covid positive cases source of exposure doesn't really affect the infection outcome. The following features are included for analysis.

Selected variables

- Age by categories, gender any other infection during the test, body mass index, fever, high blood pressure, history of chronic disease, low oxygen level during enrollment, sex, smoking, any symptoms, and test reason are included as features.

- Hospitalization status (whether hospitalized or not) is used as an outcome variable.

Data cleaning

The following data cleanings are performed before using the data to train and test models.

- Identify unique values of each column
- Rename longer column names by shorter and easier to use names

- Missing values:
    - Check the null values in each column
    - Drop columns with significant number of Nan
    - A column for ‘anyinfectious’ has 223 of 519 null values
    - Replace categorical data with the most frequent value in that column
    - Check the null values again in each column
- Change categorical values with numerical values
    - Create a dictionary of categorical values and numeric values to replace the categorical values by number.
- Feature selection – removing features with low variance using sklearn.feature_selction.
    - No feature was eligible to be dropped by using the variance threshold feature selection method.
- Check the distribution of the outcome variables and the correlation between all variables.

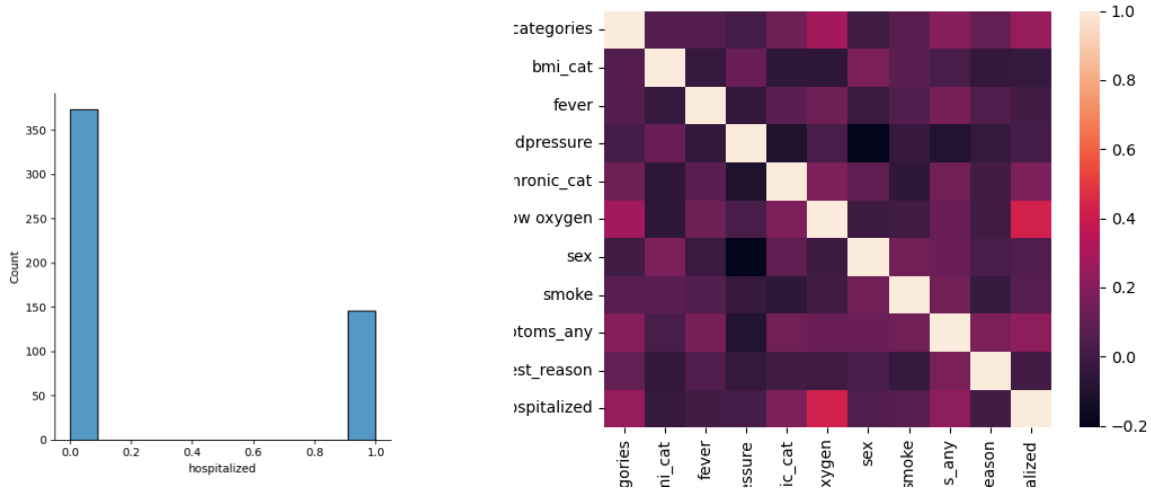

Figure 1. Distribution plot of the outcome variables (left), and heatmap of the correlation between variables (right).

Model Training and Testing

In this study, the following machine learning models are used for the classification analysis.

- SVM (Kernel = linear)
- Naïve Bayes
- Decision Tree

- Logistic regression
- Knn, and
- Random Forest

The steps used to train and test the machine learning models are: -

- Split the dataset into features and outcomes
- Splitting the dataset into train and test
    - Split data for training and testing (0.7:0.3) with some random state.
- Standardize the Variables
- Train the model
- perform prediction
- Metrix Evaluations: Confusion matrix, accuracy

Results:

Model Evaluation

The classification models are evaluated with classification metrics: accuracy score, classification report, and confusion matrix. The classification report includes precision, recall, and F1-score. Accuracy tells us how close a measured value is to the real one. Precision determines how close a measured value is to the real one. Recall or sensitivity defines the total number of positives (actual) returned by the machine learning model. **Accuracy** is used to compare the models used here.

```
lmconfusion matrix:
[[109   2]
 [ 27  18]]
lm Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.98      0.88       111
           1       0.90      0.40      0.55        45

    accuracy                           0.81       156
   macro avg       0.85      0.69      0.72       156
weighted avg       0.83      0.81      0.79       156
```

Logistic regression model, Accuracy= 81%

```
knn confusion matrix:
[[97 14]
 [26 19]]
knn Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.87      0.83       111
           1       0.58      0.42      0.49        45

    accuracy                           0.74       156
   macro avg       0.68      0.65      0.66       156
weighted avg       0.73      0.74      0.73       156
```

K-nearest neighbors, Accuracy= 74%

```
SVM confusion matrix:
[[108   3]
 [ 26  19]]
SVM Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.97      0.88       111
           1       0.86      0.42      0.57        45

    accuracy                           0.81       156
   macro avg       0.83      0.70      0.72       156
weighted avg       0.82      0.81      0.79       156
```

Support vector machine, Accuracy= 81%

```
dt confusion matrix:
[[94 17]
 [23 22]]
dt Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.85      0.82       111
           1       0.56      0.49      0.52        45

    accuracy                           0.74       156
   macro avg       0.68      0.67      0.67       156
weighted avg       0.73      0.74      0.74       156
```

Decision tree, Accuracy = 74%

```
rf confusion matrix:
[[99 12]
 [22 23]]
rf Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.89      0.85       111
           1       0.66      0.51      0.57        45

    accuracy                           0.78       156
   macro avg       0.74      0.70      0.71       156
weighted avg       0.77      0.78      0.77       156
```

Random forest, Accuracy 78%

```
NB confusion matrix:
[[101  10]
 [ 20  25]]
NB Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.91      0.87       111
           1       0.71      0.56      0.63        45

    accuracy                           0.81       156
   macro avg       0.77      0.73      0.75       156
weighted avg       0.80      0.81      0.80       156
```

Naïve Bayes, Accuracy 81%

As shown above in the classification report, logistic regression, support vector machine and Gaussian Naïve Bayes have an accuracy of 81%, better than the other three, random forest 78%, decision tree 78 % and KNN model 74%.

Optimizing models

- Use voting classifier (Hard Voting) to optimize the accuracy
    - Voting is an ensemble method that combines the performances of multiple models to make predictions. Hard voting using all models combined gave an accuracy of 79%.
    - Hard voting using the three better models, logistic regression, support vector machine and Gaussian NB is also 78%.
- K-fold cross validation (K = 5)
    - Applying K-fold cross validation on the voting classifier gave the following accuracies:
        - With K=5; Accuracy, mean (stdv) = 0.784 (0.056)

        - With K=10 Accuracy, mean (stdv) = 0.784 (0.068)
- To better optimize the models, after performing correlation between the variables, features with very low correlation were dropped and hard voting is run again.
    - Hard voting accuracy 79%
- Optimize the random forest by hyperparameter, increase the estimator from 100 to 200.
    - Accuracy =79%, increasing the estimator to 200 increases the accuracy only from 78% to 79%
- Grid search to update SVM,
    - After optimizing the SVM using grid search the accuracy of the SVM is still 81%.

Summary and Conclusion

Overall, in this project logistic regression (0.81), SVM (0.81) and random forest (0.81) has better accuracy of prediction than the other models. Parameter and hyperparameter optimization techniques used; K-fold cross validation, grid search in svm, increasing estimator in random forest, hard voting and dropping features with low correlation, all didn't optimize the accuracy of the models significantly. Further data preprocessing, and feature selections may help to train a better model to determine the risk factors for hospitalization due to COVID infection.

Lessons learnt from the project

- Data preprocessing and EDA are the most important part of data mining
- Choosing machine learning model suitable for the data type and data problem is crucial

Reference:

1. Data source: Risk Factors for Hospitalization and Death from COVID-19 in Humanitarian Settings - Humanitarian Data Exchange (humdata.org)
2. Leidman, Eva et al. "Risk factors for hospitalization and death from COVID-19: a prospective cohort study in South Sudan and Eastern Democratic Republic of the Congo." *BMJ open* vol. 12,5 e060639. 18 May. 2022, doi:10.1136/bmjopen-2021-060639
3. Chatterjee A, Gerdes MW, Martinez SG. Identification of Risk Factors Associated with Obesity and Overweight-A Machine Learning Overview. Sensors (Basel). 2020 May 11;20(9):2734. doi: 10.3390/s20092734. PMID: 32403349; PMCID: PMC7248873.