Prediction of Death Situation of Covid

Introduction: Covid-19 pandemic had caused significant global impact, with millions of confirmed cases and deaths reported worldwide. The situation was evolving and changing rapidly, with many countries implementing measures such as social distancing, mask-wearing, and vaccination campaigns to try to slow the spread of the virus.Many researchers and organizations are using machine learning models to analyze Covid-19 data and make predictions about the spread of the virus and the potential impact on populations. These models can use various data sources, including medical records, demographic information, and social media data, to make predictions about the spread and impact of Covid-19.

Objective of the work: To Predict whether death will happen based on features like sex,age,asthma,diabetes,hypertension,..etc

Methodology:

This dataset contains an enormous number of anonymized patient-related information including pre-conditions. The raw dataset consists of 21 unique features and 1,048,576 unique patients. In the Boolean features, 1 means "yes" and 2 means "no". values as 97 and 99 are missing data

1. The data contains 1048575 entries with 21 columns
2. There are No missing values,null values
3. We have some features that we expect them to have just 2 unique values but we see that these features have 3 or 4 unique values. For example the feature "PNEUMONIA" has 3 unique values (1,2,99) 99 represents NaN values. Hence we will just take the rows that includes 1 and 2 values.
4. In the "DATE_DIED" column, we have 971633 "9999-99-99" values which represent alive patients so this feature is taken as a "DEATH" that includes whether the patient died or not.
5. In "INTUBED" and "ICU" features there are too many missing values so i will drop them. Also we don't need the "DATE_DIED" column anymore because we used this feature as a "DEATH" feature.
6. Data is visualized using hist-plot Feature Age is plotted
7. Of all these features only few important features are selected based on the correlation with crystal structure using a heat map.
8. The data is split into training and testing using sk-learn library and thereafter scaling is done.Logistic Regression Classifier is used as a model for classification.
9. Logistic Regression is a statistical method used to analyze the relationship between a dependent variable and one or more independent variables. It is used to model the probability of a binary outcome, that is, an outcome that can take only one of two values (e.g., yes/no, true/false, etc.

Results and Discussions: The model gave an accuracy of around 93% which means that the model performed well.A confusion matrix is plotted.Logistic regression is a powerful statistical method for modeling the relationship between a binary outcome and one or more independent variables. Its methodology involves data preparation, model selection, model fitting, and model evaluation. Logistic regression is widely used in various fields, including medicine, finance, marketing, and social sciences.

Predictions made by machine learning models should always be considered in conjunction with expert analysis and guidance from public health officials and medical professionals. Ultimately, human judgment and decision-making are essential in determining the best course of action in response to the Covid-19 pandemic.

Conclusion: We got good accuracy with Logistic Regression.But it can mislead us so we have to check the other metrics.When we look at the F1 Score it says that we predicted the patients who survived well but we can't say the same thing for dead patients.

References:

1. https://www.kaggle.com/code/yasirakyzl/covid-19-ml-model-90-accuracy/notebook