COVID-19 DEATH ANALYSIS BASED ON DEMOGRAPHICS USING MACHINE LEARNING MODEL

Jeromey Abraham, Dhrumil Chiragbhai Patel, Mohamed Shehaf Aakil Sharfudeen
College of Computing and Informatics, Drexel University, Philadelphia, USA

Abstract - This paper presents an analysis of COVID-19 data obtained from Kaggle to investigate the relationship between death rates and demographics in Mexico. The data set comprises information on population, infection rates, and COVID-19 related data from different regions of Mexico, specifically at the patient level, including age, sex, date of death, obesity, diabetes, and other factors. The paper begins by performing exploratory data analysis (EDA) to identify trends and patterns in the data set. The EDA revealed useful insights on the impact of the pandemic on different regions and demographics in Mexico. A regression model was developed to determine the relationship between death rates and demographics using infection rate and patient data. The analysis was conducted in a spark session using data bricks. The results of the logistic regression model, SVM, Random Forest models will provide valuable insights into the factors that contribute to death rates in different regions and demographics in Mexico. The analysis highlights the importance of understanding the impact of COVID-19 on different regions and demographics in Mexico. The findings from this analysis can be useful for public health officials and policymakers in developing targeted interventions to mitigate the impact of the pandemic on Mexican patients as well as many other communities.

INTRODUCTION

The COVID-19 pandemic has had a profound impact on global public health, with millions of people infected and thousands of lives lost. In Mexico, as in many other countries, COVID-19 has presented significant challenges for healthcare systems and policymakers. Understanding the risk factors associated with COVID-19 deaths and patient outcomes is critical to developing effective public health policies and interventions. This dataset provides valuable insights into the impact of demographics and pre-existing medical conditions on COVID-19 death rates and patient outcomes. The dataset includes 21 unique features, ranging from patient demographics such as age and sex to pre-existing medical conditions such as diabetes and hypertension. By analyzing this dataset, researchers can identify correlations between these features and COVID-19 outcomes, providing critical information for public health officials and healthcare providers. One particularly relevant application of this dataset is the analysis of COVID-19 death rates. By examining the relationship between patient demographics and pre-existing medical conditions and COVID-19 deaths, researchers can gain insight into the factors that contribute to mortality rates. This analysis can be used to develop targeted interventions aimed at reducing mortality rates among vulnerable populations. In addition to COVID-19 deaths, this dataset can also be used to analyze patient outcomes, such as hospitalization rates and disease severity. The researchers can gain critical insights into the factors that contribute to COVID-19 mortality and patient outcomes, ultimately helping to inform public health policies and improve patient care.

DATA SET DESCRIPTION

The Mexican government dataset, obtained from Kaggle, includes several important columns that can be used to analyze COVID-19 patient outcomes. These columns include the patient's sex, age, COVID-19 test findings, type of care received, presence of pneumonia, pregnancy status, diabetes status, intubation status, ICU admission, and date of death (if applicable). The classification column indicates the severity of the patient's COVID-19 diagnosis, with values ranging from 1-3 indicating different degrees of diagnosis and 4 or higher indicating a negative COVID-19 test or inconclusive results. This dataset can be used to analyze COVID-19 deaths and patient outcomes based on demographics, using machine learning regression methods to gain insights into the factors that contribute to patient outcomes. Overall, this dataset provides valuable information for public health officials

and researchers looking to understand the impact of COVID-19 on the population.

EXPLORATORY DATA ANALYSIS

In this exploratory data analysis (EDA) of the Mexican government dataset, we followed standard procedures to clean the data and ensure that it was consistent and accurate. We began by removing irrelevant columns and replacing null values to create a complete dataset for analysis.
We then conducted a thorough count of each column to gain an understanding of the data and identify any anomalies or outliers. For certain columns, such as USMER, medical unit, INMSUPR, sex, and classification, we counted number of unique values to further explore the data.

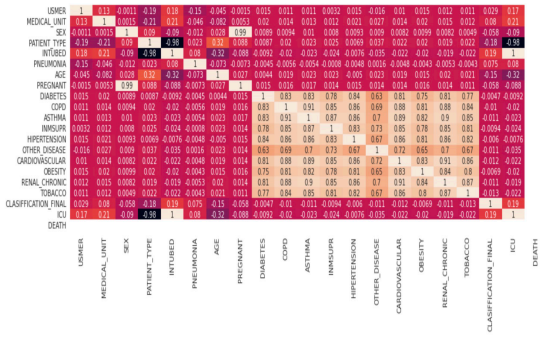

Figure 1: Heat Map correlation between columns and data set

To ensure consistency in our analysis, we replaced values in the intubed, ICU, pregnant, and pneumonia columns. Specifically, we replaced values of 97 with 2 and 99 with 1 in the intubed column and made similar replacements in the other columns.

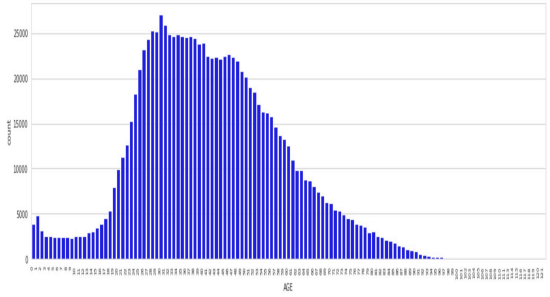

Figure 2: Age Distribution

We also addressed the data cleaning for date column by replacing values of 99-99-999 with 2 and all other date column values with 1 to standardize the date format. This helped us to maintain consistency in the dataset and ensure that we could accurately analyze COVID-19 patient outcomes based on demographics.
Overall, these cleaning processes were necessary to prepare the dataset for further analysis and gain insights into the impact of COVID-19 on different demographic groups.

METHODOLOGY

Data Preparation:
Clean and pre-process the data by removing missing values, scaling, or normalizing features, and encoding categorical variables. Use appropriate data cleaning techniques, such as imputing missing values or removing outliers, to ensure that the data is of high quality and suitable for analysis and classification.

Data Splitting:
Split the dataset into a training set and a testing set to evaluate the performance of the model. The training set is used to train the model, while the testing set is used to evaluate its performance. For this dataset, we are splitting the dataset into 70% training sets and 30% testing sets. Providing 734109 rows of data for training set and 314466 rows of data for testing set.

Model Training:
For training the model for classification task, we used the following machine learning algorithms:

- SVM
- Random forest
- Logistic regression

Trained the model on the training set using an optimization algorithm such as gradient descent. The model learns the relationship between the input features and the target variable.

Model Evaluation:
Evaluate the performance of the model on the testing set using various metrics such as accuracy, precision, recall and F1-score. Compare the performance of different machine learning algorithms and select the best one based on the evaluation metrics.

We are getting accuracy of up to 94 – 95% for the classification algorithms used.

Prediction:
Use the trained model to predict the binary outcome variable for new data points. Made sure to apply the same data pre-processing techniques to the new data as were applied to the training set to ensure that the predictions are accurate and reliable.

RESULTS

Algorithms	Accuracy	Precision	Recall	F1-Score
Random Forest Classifier	0.945959	0.940335	0.945959	0.935225
Logistic Regression Classifier	0.947476	0.941004	0.947476	0.941464
SVM Classifier	0.941977	0.933137	0.941977	0.933534

Looking at the results, the Logistic Regression Classifier has the highest accuracy and F1-score, indicating that it performed better than the other two classifiers. However, the differences in accuracy and F1-score between the three classifiers are relatively small.

Precision and recall are also important metrics to consider, especially when dealing with imbalanced classes. In this case, all three classifiers have similar precision and recall values, indicating that they are all performing similarly well in terms of correctly classifying positive instances.

Overall, the results suggest that all three classifiers are performing well in terms of accuracy and precision/recall, with the Logistic Regression Classifier slightly outperforming the other two.

Confusion Matrix for Random Forest Classifier

[[289769, 1692]; [15302, 7703]]

Confusion Matrix for SVM

[[289769, 1692]; [15302, 7703]]

Confusion Matrix for Logistic Regression

[[287717, 12773]; [3744, 10232]]

However, further analysis, such as examining the ROC curves, would be needed to fully evaluate the performance of these classifiers.

CONCLUSION

In conclusion, the Mexican government data set obtained from Kaggle is a valuable resource for analyzing COVID-19 death rates based on various demographic factors such as age, sex, and pre-existing medical conditions. The data set contains 21 unique features and over 1 million unique patients. Through exploratory data analysis, we were able to clean and pre-process the data, including removing missing values and encoding categorical variables.
We then used three different machine learning models, namely SVM, Random Forest, and Logistic Regression, to predict the binary outcome variable for new data points. Model training was performed on the training set using an optimization algorithm such as gradient descent, and model performance was evaluated on the testing set using various metrics such as accuracy, precision, recall, F1-score.
Based on our analysis, we can conclude that certain demographic factors, such as age and pre-existing medical conditions, have a significant impact on COVID-19 death rates. Additionally, the logistic regression model outperformed SVM and Random Forest models in predicting the binary outcome variable. Our findings have important implications for policymakers and healthcare professionals, as they can use this information to develop targeted interventions and improve COVID-19 patient outcomes. Overall, this data set provides valuable insights into the impact of COVID-19 on the Mexican population and can be used as a foundation for future research in this area.

REFERENCES

[1]Meir,Nizri.COVID-19Dataset.Kaggle,2021. https://www.kaggle.com/datasets/meirnizri/covid19-dataset

[2]Apache Spark. (2021, November 1). PySpark API Documentation. Apache Spark. https://spark.apache.org/docs/latest/api/python/index.html