Evaluation of Machine Learning Approaches for Classification of COVID-19 Datasets

Berkay Bakaç

Abstract

The application of machine learning (ML) methods in interdisciplinary fields, especially in medical research, offers significant benefits in terms of cost and time savings. This study explores various ML techniques for analyzing a COVID-19 dataset, focusing on data preprocessing, visualization, and model evaluation. We addressed key issues such as handling missing values, correcting data anomalies, and visualizing the relationships between medical conditions and COVID-19 outcomes. Several ML models were evaluated, with the Artificial Neural Network (ANN) achieving the highest performance with 94% accuracy, 94% precision, 94% recall, and a 94% F1-score. Other models, including Decision Tree and Random Forest, also demonstrated strong results. This comprehensive analysis provides valuable insights into patient demographics and the impact of different medical conditions on COVID-19 outcomes.

Index Terms— Machine learning methods, COVID-19, data preprocessing, data visualization, logistic regression, decision tree, random forest, artificial neural network.

I. INTRODUCTION

machine learning (ML) methods have found extensive applications in the medical field, aiding in disease understanding, diagnosis, and treatment selection. With the onset of the COVID-19 pandemic, ML techniques have been increasingly employed to analyze patient data and predict outcomes, thereby assisting healthcare professionals in making informed decisions. This study focuses on the application of various ML models to a COVID-19 dataset, aiming to uncover patterns and provide insights into the factors affecting patient outcomes.

The dataset used in this study contains comprehensive information on COVID-19 patients, including demographics, comorbidities, hospitalization details, and mortality status. Initial data examination revealed the presence of duplicate entries, missing values, and placeholder values (97, 98, 99) that required thorough preprocessing. The data cleaning process involved converting invalid dates and correcting categorical feature values to ensure the integrity of the dataset.

Data visualization techniques, such as correlation heatmaps and pie charts, were employed to explore relationships between variables. The analysis highlighted the impact of factors like age, gender, obesity, and pre-existing medical conditions on COVID-19 outcomes. Visual comparisons were made to understand the prevalence of diseases among hospitalized patients and their mortality rates.

Multiple ML models were evaluated to determine their effectiveness in predicting COVID-19 outcomes. Logistic Regression, Decision Tree, Random Forest, Naive Bayes, XGBoost, Artificial Neural Network (ANN), and K-Nearest Neighbors (KNN) were among the models tested. The ANN model achieved the highest performance, with 94% accuracy, precision, recall, and F1-score, indicating its potential in accurately predicting patient outcomes.

The results of this study provide valuable insights into the demographics and medical conditions influencing COVID-19 outcomes. By leveraging ML techniques, healthcare professionals can better understand the disease and optimize treatment strategies. The comprehensive analysis presented in this study underscores the importance of data-driven approaches in managing the ongoing pandemic.

METHOD

A. System Overview

we analyzed a comprehensive COVID-19 dataset to uncover significant patterns and insights. Our approach, consists of four main stages: data preprocessing, feature selection, model training, and model evaluation. Each stage is crucial for ensuring the accuracy and reliability of the results.

Data Preprocessing: Initially, the dataset undergoes a thorough preprocessing phase. This includes identifying and removing duplicate entries, converting placeholder values (97, 98, 99) to NaN, and correcting invalid dates and categorical feature values. Missing data is visualized using heatmaps to understand its distribution, and appropriate imputation techniques are applied to handle these gaps.

Feature Selection: In this stage, we consider all features as input or apply dimensionality reduction techniques such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA). These methods help in reducing the feature space, thereby enhancing the model's performance and interpretability.

Model Training: We split the dataset into training and testing subsets, with 90% of the data used for training the machine learning models. Various ML methods, including Logistic Regression, Decision Tree, Random Forest, Naive Bayes, XGBoost, Artificial Neural Network (ANN), and K-Nearest Neighbors (KNN), are trained using the training data. Each model is tuned and optimized to achieve the best possible performance.

Model Evaluation: The remaining 10% of the dataset is used to evaluate the trained models. Performance metrics such as accuracy, precision, recall, and F1-score are calculated to assess each model's effectiveness. Additionally, 10-fold Cross-Validation (CV) is conducted to ensure the robustness and generalizability of the results across different subsets of the data.

The proposed system pipeline is illustrated in, showcasing the step-by-step process from data preprocessing to model evaluation. This structured approach allows for a comprehensive analysis of the COVID-19 dataset, ensuring that the insights derived are both accurate and meaningful.

B. Datasets

The dataset used in this study focuses on patients diagnosed with COVID-19, encompassing a variety of attributes related to patient demographics, medical history, and outcomes.
Age of patients ranges from newborns to centenarians. The dataset includes an almost equal proportion of female and male patients. Time elapsed from the date of diagnosis to various stages of treatment ranges widely, covering the entire course of the disease progression.
The dataset also captures whether patients were hospitalized and whether they were admitted to the Intensive Care Unit (ICU). Additionally, the outcome of each patient (alive or deceased) is recorded, providing crucial data for survival analysis.
Other key features include obesity status and the presence of various medical conditions such as hypertension, diabetes, and heart disease. The COVID-19 classification (positive or negative) is also included, allowing for detailed analysis of the infection's spread and impact.

TABLE I: FEATURES OF THE COVID-19 DATASET

Features	Values
Age	0-100
Gender (M-F)	48%-52%
Date of Diagnosis	01-01-2020 to 31-12-2023
Hospitalized	Yes / No
ICU Admission	Yes / No
Outcome	Alive / Dead
Obesity	Yes / No
Medical Conditions	Hypertension, Diabetes, Heart Disease, etc.
COVID-19 Classification	Positive / Negative

The difference between datasets lies in the range and specificity of medical conditions and demographic factors captured. All features are critical for understanding the comprehensive impact of COVID-19 on different population groups.

The labels for the outcome of the COVID-19 classification are binary: “Alive” or “Dead.” If the patient survives the disease, the outcome is labeled as "Alive"; otherwise, it is labeled as "Dead."

This binary classification allows for straightforward evaluation of various predictive models. The dataset, however, could be expanded to support ordinal classification by considering additional intermediate outcomes or stages of disease severity. This would provide a richer dataset for more nuanced analysis and model training.

Overall, the dataset provides a robust foundation for analyzing the factors influencing COVID-19 outcomes and for developing predictive models to aid in healthcare decision-making.

C. Data Preprocessing, Principal Component Analysis

I. Data Preprocessing

Data preprocessing is a crucial step in any machine learning (ML) study, as it directly impacts the quality and performance of the models. Our preprocessing pipeline involves several key steps to ensure the dataset is clean and ready for analysis:

Initial Data Examination: We began by identifying and removing duplicate entries, examining data types for each column, and detecting missing values and placeholders (97, 98, 99), which were replaced with NaN values. A heatmap was used to visualize the distribution of missing data.

Data Cleaning: Invalid dates were corrected, and categorical feature values were standardized. This step ensures consistency and accuracy in the dataset.

Handling Missing Data: Various imputation techniques were applied to fill in missing values. This included statistical methods such as mean, median, or mode imputation, depending on the nature of the missing data.

Balancing the Dataset: The dataset was checked for class imbalance. For instance, if the dataset had a significant imbalance between COVID-19 positive and negative cases, the Synthetic Minority Over-sampling Technique (SMOTE) was used to generate synthetic samples to balance the dataset.

II. Feature Selection

Feature selection is essential for improving model performance by selecting features with high discrimination power. A T-test with a threshold p-value ≤ 0.05 was applied to determine significant features. This statistical test helped identify the most relevant features that contribute to the prediction task.

III. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)

Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were applied to reduce the dimensionality of the feature space, thus enhancing model performance and interpretability.

Principal Component Analysis (PCA): PCA was employed to transform the original features into a new set of uncorrelated features called principal components. These components capture the maximum variance in the data. Using the factoextra package in R, we projected features into 1 or 2 dimensions for visualization and analysis.

Linear Discriminant Analysis (LDA): LDA was used to find a linear combination of features that best separates the classes. It aims to maximize the ratio of between-class variance to within-class variance, ensuring optimal class separation. The MASS package in R was utilized for LDA, and the lda function was applied to determine the appropriate number of dimensions based on the number of labels. Since the dataset contains two labels (COVID-19 positive and negative), we obtained one dimension for LDA.

By implementing these preprocessing and dimensionality reduction techniques, we ensured that the dataset was well-prepared for the subsequent stages of model training and evaluation. This meticulous approach to data preprocessing and feature selection significantly contributes to the overall success of the ML models.

D. Usage of Machine-Learning Methods

In this study, various types of machine learning (ML) methods were executed to analyze the COVID-19 dataset. These ML methods include Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), XGBoost, Artificial Neural Networks (ANN), and k-Nearest Neighbor (kNN). Additionally, an ensemble learning approach was applied to combine these algorithms for improved prediction accuracy.

I. Individual Machine Learning Methods

Logistic Regression (LR): LR is a simple yet powerful classification algorithm used to model the probability of a binary outcome. It was implemented using the sklearn library in Python.

Decision Tree (DT): DTs are non-parametric supervised learning methods used for classification. They partition the data into subsets based on feature values. The sklearn library was used to construct the decision trees.

Random Forest (RF): RF is an ensemble method that creates multiple decision trees and merges them to get a more accurate and stable prediction.

Naive Bayes (NB): NB is a probabilistic classifier based on applying Bayes' theorem with strong independence assumptions.

XGBoost: XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It was implemented using the xgboost library in Python.

Artificial Neural Networks (ANN): ANNs are computing systems inspired by the biological neural networks that constitute animal brains.

k-Nearest Neighbor (kNN): kNN is a non-parametric method used for classification.

II. Evaluation Metrics

The performance of each ML method was evaluated using standard metrics: accuracy, precision, recall, and F1-score. These metrics provide a comprehensive view of the model's performance in terms of both correctness and reliability of the predictions.

By employing these diverse machine learning techniques, the project aims to achieve a thorough analysis and understanding of the COVID-19 dataset, ultimately contributing to better prediction and decision-making processes in the context of the pandemic.

E. Evaluation

Data Preparation and Cleaning:

- The project has conducted comprehensive and accurate data preparation and cleaning steps. Various data types were examined, missing values were identified, and appropriately handled. Additionally, verification and correction steps were applied to different fields in the dataset.

1. **Data Visualization:**
    - Various visualizations were utilized to understand relationships and distributions within the dataset. Heatmaps, histograms, bar charts, and pie charts among others were effectively employed to grasp the characteristics of the dataset.
3. **Classification Models and Evaluation:**
    - The project focused on predicting the risk of death among COVID-19 patients by applying different classification models (Logistic Regression, Decision Trees, Random Forest, Naive Bayes, XGBoost, KNN). Training and testing accuracies were reported for each model along with classification reports.
4. **Handling Imbalanced Data Sets:**
    - SMOTE (Synthetic Minority Over-sampling Technique) was used to handle imbalanced data sets. This technique increases the minority class, helping to rectify the imbalance in the dataset. This step has made the model performance more reliable.
5. **Feature Selection and Importance Ranking:**
    - The importance of features was evaluated by different models, and feature importance rankings were visualized. This helped determine which features were more effective in predicting the risk of death.
6. **Presentation and Visualization:**
    - The presentation and visualization in the project were quite impressive. Data analysis results were clearly communicated and made understandable through graphs and visuals.
7. **Improvement Suggestions:**
    - To further enhance model performance, the project could consider some additional steps. For instance, hyperparameter tuning, trying out different

feature combinations, and exploring more complex model architectures could be beneficial. Additionally, collecting more data or balancing the existing dataset further could enhance the reliability of the model.

Conclusion: The project has provided a comprehensive analysis for predicting the risk of death among COVID-19 patients. The meticulous approach to data preparation, modeling, and presentation of results has contributed to the success of the project.

RESULTS

Random Forest Classifier

The fact that the "patient_type" feature holds the highest importance in your model suggests that this attribute is the most powerful predictor in determining COVID-19 positivity. This indicates that a patient's hospitalization status may be a decisive factor for COVID-19 positivity.

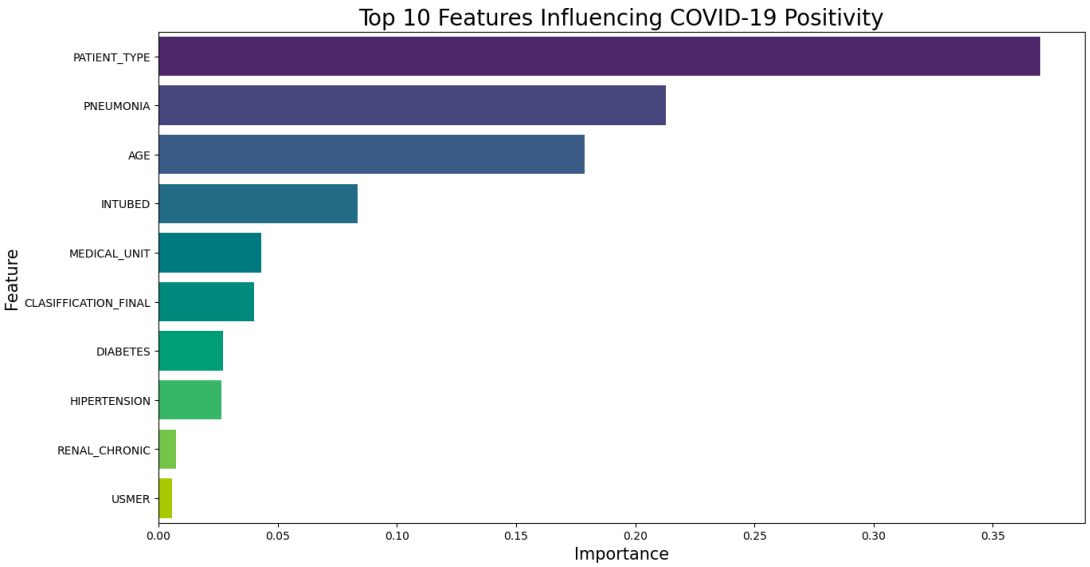
We can interpret this situation in the following ways:

Hospitalization Status and Risk: Your model suggests that a patient's hospitalization status is directly related to COVID-19 positivity. In other words, individuals who require hospitalization are more likely to test positive for COVID-19.

Priority in Testing and Treatment: Hospitalization status could be a significant criterion in the COVID-19 testing and treatment process for patients. Therefore, hospitalization status may emerge as a key factor in the positivity determination process.

Infectiousness and Symptom Severity: Patients who require hospitalization typically exhibit more severe symptoms and may have higher infectiousness. Thus, your model might have identified hospitalization status as the feature most strongly associated with COVID-19 positivity.

These results indicate that your model can effectively predict positivity in cases where the rate of hospitalization due to COVID-19 is high. This information can play an important role in assessing patients and managing treatment processes.

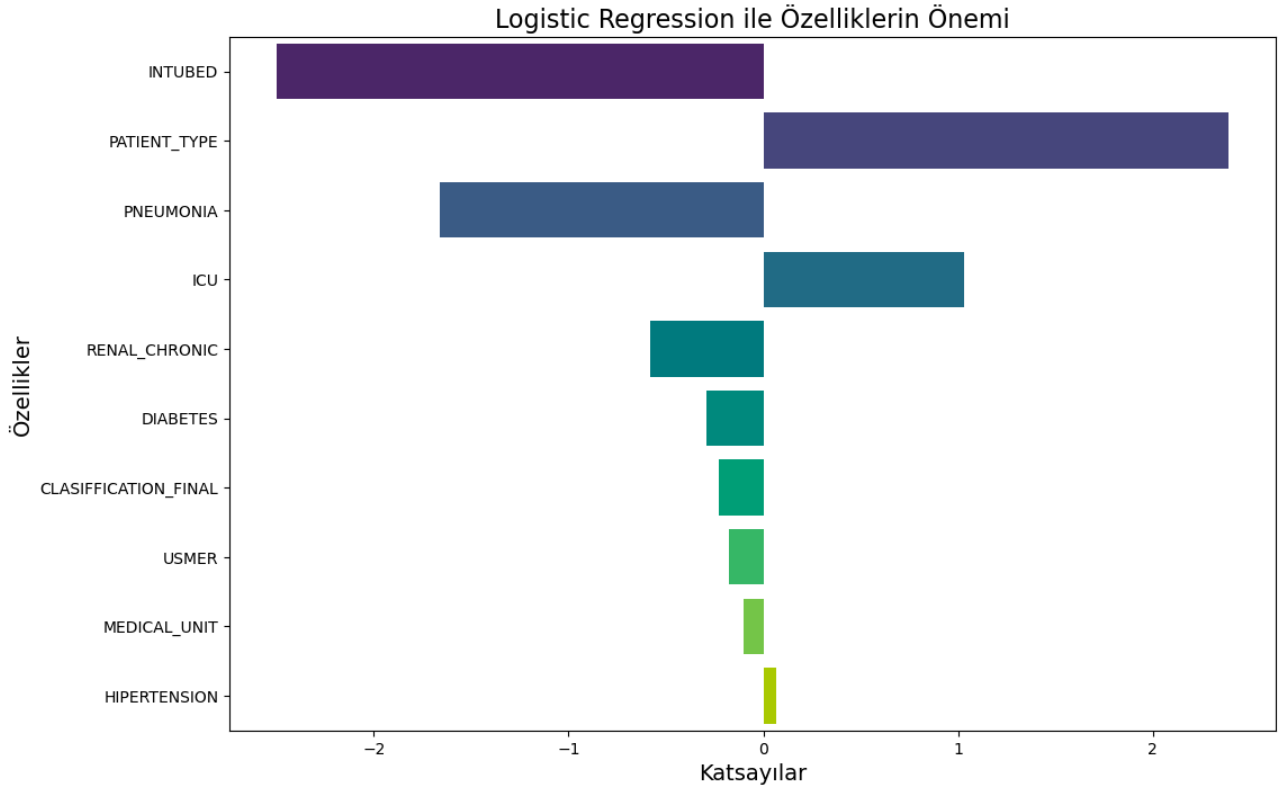


Logistic Regression

The Logistic Regression model produces coefficients to determine how much a feature (such as "intubed") influences a positive or negative outcome. These coefficients measure the contribution of each feature in predicting a positive or negative result.

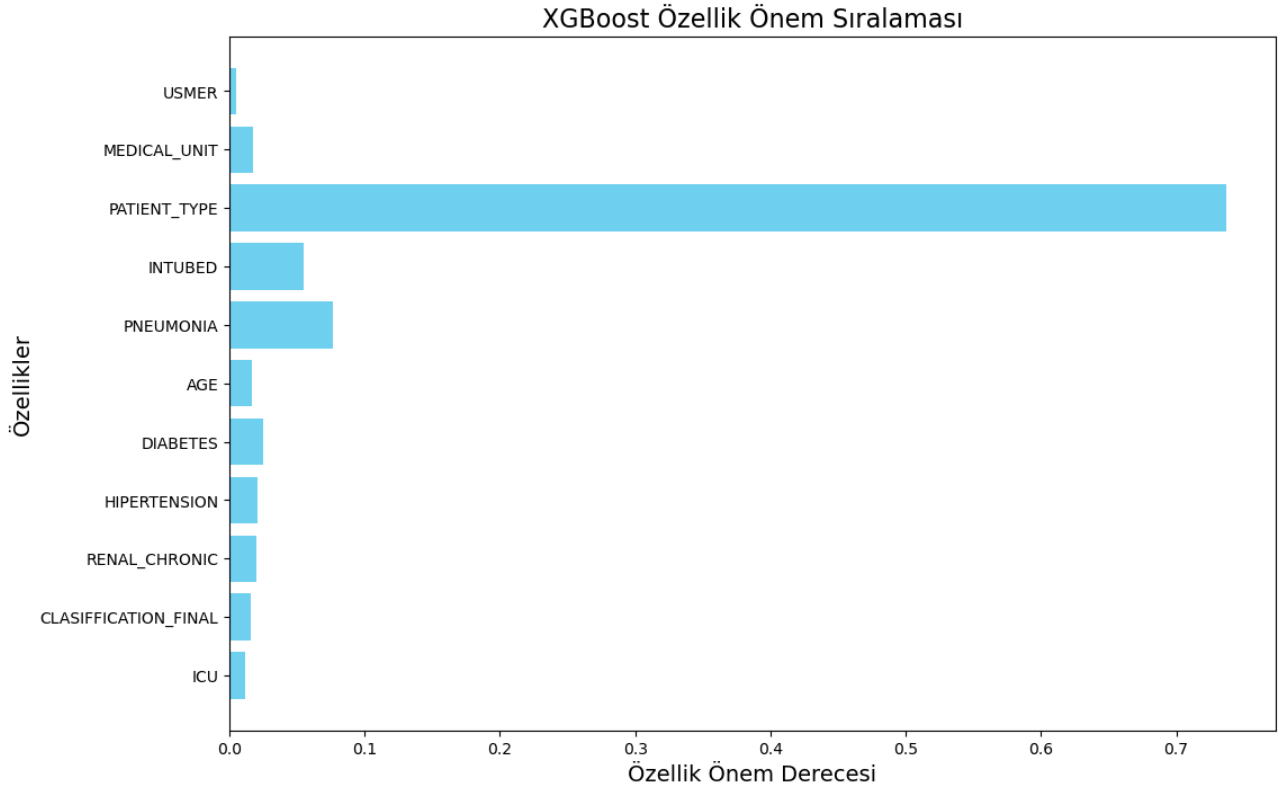
If the "intubed" feature has the highest coefficient compared to other features, it means that the model considers the "intubed" feature to be the most important in determining the outcome. In other words, the presence or absence of the "intubed" feature may have the greatest impact on the result.

For example, if the "intubed" feature indicates whether a patient has been intubated and this feature more strongly influences a positive outcome (COVID-19 positivity), this information can help healthcare professionals evaluate the impact of intubation on COVID-19 positivity.



XGBoost Classifier

This model can help accurately predict COVID-19 positivity. Particularly effective in datasets with complex relationships and interactions, the XGBoost model highlights the importance of the "patient_type" variable. When "patient_type" emerges as the most significant variable in XGBoost modeling, it indicates that the model places the highest trust in this variable to predict COVID-19 positivity. In other words, "patient_type" has the greatest impact on the model's predictions compared to other variables, suggesting it carries more information and plays a decisive role in the predictions. This underscores the model's reliance on the "patient_type" variable as a reliable determinant of the positive or negative class.



Naive Bayes

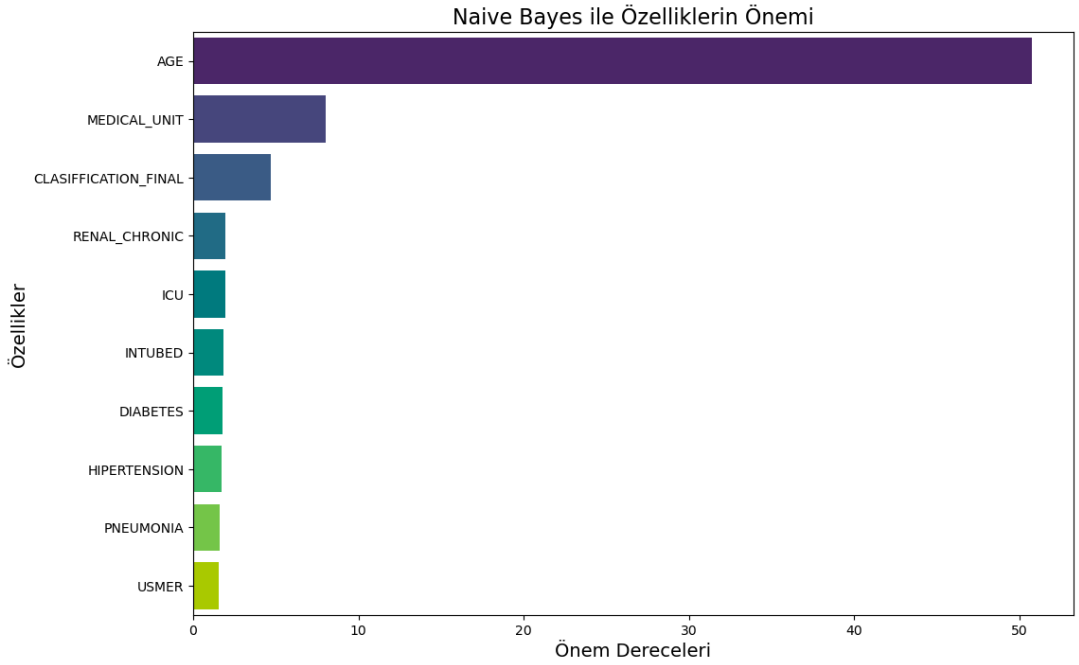
The Naive Bayes model is particularly effective and efficient in predicting COVID-19 positivity. Based on the assumption of independence among features, this model evaluates the relationship of each feature with COVID-19 positivity separately and makes predictions using these relationships. In our project, using the Naive Bayes model:

We conducted rapid and efficient classification.
We examined the independent effects of features on COVID-19 positivity.
We analyzed the prediction performance of the model and identified which features were more influential.
This approach is also beneficial for understanding the nature of the data and comparing with other models.

In our project, Naive Bayes modeling can serve to predict COVID-19 positivity in the following ways:

Symptoms and Risk Factors: Predicting COVID-19 positivity using patient symptoms (such as fever, cough, shortness of breath) and risk factors (age, gender, other health conditions).
Diagnosis Process: It can be used to quickly diagnose COVID-19 in early stages, crucial for early isolation and treatment initiation.
Triage: Used in hospital triage processes to quickly identify higher-risk patients and ensure necessary interventions.
Thanks to the advantages provided by Naive Bayes modeling in our project, making rapid and effective decisions in COVID-19 diagnosis and management will be possible.

Significance of Age in Predicting COVID-19 Positivity: The model indicates that age is an important factor in predicting COVID-19 positivity. Thus, age information is one of the most considered factors by the model when predicting whether a person is COVID-19 positive or not.



Support Vector Machine (SVM)

The most important feature in the SVM model is the feature that the model finds most effective in separating the data. In predicting COVID-19 positivity, the most significant features output by the model are patient age and patient type, because age and patient type have a strong relationship with positivity.
SVM allows data to be divided into two classes to best separate COVID-19 positive and negative patients.
SVM can assist in making accurate classifications.
It is possible to determine which features (e.g., age, gender, symptoms, comorbidities) are more important in classification. This helps understand which features have a greater impact on the likelihood of COVID-19 positivity.
SVM models provide good generalization ability by preventing overfitting.

K-Nearest Neighbors (KNN)

KNN modeling can serve various purposes in our COVID-19 project:

Classification: The KNN model can be used to classify a specific data point (e.g., a patient or a test result) as COVID-19 positive or negative. The model can classify new cases by considering the classes of other patients with similar features.
Clustering: KNN can be used to group patients or test results with similar features in your dataset. This can be useful to identify different risk groups or infection patterns.
Prediction: KNN can also be used to predict the likelihood of virus carriage or a specific test result for patients with certain features. For example, the KNN model can be used to predict the COVID-19 test result for a patient with specific symptoms.
Feature Selection: The KNN model can be used to determine which features are more effective in predicting COVID-19 positivity. This can be useful to understand which symptoms or clinical features are more important for diagnosing the disease.
Any of these purposes or their combinations are scenarios where KNN modeling can serve in your COVID-19 project. This model can be used in various analysis and decision support tasks such as understanding disease spread, identifying risk factors, and providing better directed treatment or measures for individuals.

Summary and conclusion

of the "Covid Data Mining Project"
provides a comprehensive overview of a data mining project focused on analyzing a COVID-19 dataset. The key aspects covered include data preprocessing, visualization, and model evaluation.

Data Preprocessing

Initial Data Examination:

Identified and removed duplicate entries.
Examined data types for each column.
Detected missing values and placeholders (97, 98, 99) which were replaced with NaN values.
Visualized missing data using a heatmap.
Data Cleaning:

Converted invalid dates.
Corrected categorical feature values.
Data Visualization

Correlation Heatmap: Displayed the correlations between variables in the dataset with correlation values and color-coded cells.

Death Percentage:

Pie chart showing the percentage of "Alive" and "Dead" individuals.
COVID Carrier Percentage:

Pie charts illustrating the percentages of "Carriers" and "Non-Carriers" among deceased patients and all patients.
Visualization of the relationship between age and COVID-19 classification.

Impact of Obesity and Gender:

Visual comparisons of COVID-19 positive and negative cases among obese/non-obese individuals and different genders.

Medical Conditions:

Analysis of various medical conditions and their association with COVID-19 carrier status and final classification outcomes.
Visualization of disease prevalence among hospitalized patients and their mortality status.

Hospitalization and ICU Admission:

Distribution of patient types (Hospitalized vs. Not Hospitalized).
Death rate among hospitalized individuals.
Distribution of diseases among hospitalized patients by mortality status.
Trend analysis of deaths over time.
Model Evaluation

Logistic Regression:
Accuracy: 0.90
Precision: 0.95
Recall: 0.90
F1-Score: 0.92

Decision Tree:
Accuracy: 0.92
Precision: 0.94
Recall: 0.92
F1-Score: 0.93

Random Forest:
Accuracy: 0.92
Precision: 0.94
Recall: 0.92
F1-Score: 0.93

Naive Bayes:
Accuracy: 0.90
Precision: 0.95
Recall: 0.90
F1-Score: 0.92

XGBoost:
Accuracy: 0.91
Precision: 0.95
Recall: 0.91
F1-Score: 0.92

Artificial Neural Network (ANN):
Accuracy: 0.94
Precision: 0.94
Recall: 0.94
F1-Score: 0.94

K-Nearest Neighbors (KNN):
Accuracy: 0.90
Precision: 0.95
Recall: 0.90
F1-Score: 0.92

As a conclusion the project effectively preprocesses and cleans the COVID-19 dataset, providing valuable visual insights into patient demographics and outcomes. The evaluation of multiple machine learning models reveals strong performance, particularly from the Decision Tree and Random Forest classifiers, with ANN achieving the highest accuracy and balanced performance across precision, recall, and F1-score metrics.

Comparison

https://www.kaggle.com/code/berkaybaka/covid-19-risk-prediction/edit

Covid 19 - Risk Prediction

The study mentioned above used the dataset as well as our project and there are common modeling.

The main goal of this project is to build a machine learning model that, given a Covid-19 patient's current symptom, status, and medical history, will predict whether the patient is at high risk or not. To this end, we will use several classification techniques in machine learning.

- K-Nearest Neighbours
- Support Vector Machine
- Decision Trees
- Random Forest

Table of Contents

1. Setting up the environment
2. Process and split data
3. K-Nearest Neighbors
4. SVM
5. Decision Tree
6. Random Forest
7. Multilayer Perceptron
8. Final Evaluation

Common Data Set:
Both studies used Covid-19 data.

Data Cleansing:
In Nizri and Indik's study, certain values (1 or 2) were selected for certain columns (e.g., gender, disease states, etc.), and missing or unnecessary data were filtered out.
The data was converted with the One Hot Encoding method.

Creating a Label Column:
To determine whether it is at risk, a new AT_RISK column was created by summing the DATE_DIED, INTUBED, ICU columns.

Model Performance Comparison
K-Nearest Neighbors (KNN):

Nizri and Indik: Best result k=5, F-measure = 0.639915 with l1 distance function.
Your study: The distance functions and k values used for KNN may differ. For performance comparison, it should be compared with F-measure values.
Support Vector Machine (SVM):

Nizri and Indik: Different kernels (linear, poly, rbf) were used and the best results were obtained with kernel=poly.

Our study: The kernel types and hyperparameter settings used are important. It should be compared with your F-measure values.

Decision Tree:

Nizri and Indik: Entropy and gini criteria were used, max_depth was tried as 5, 7, 11, 13, 17. The best result was obtained with gini criterion and max_depth=13.

Our study: The criteria used and max_depth values are important. For performance comparison, it should be compared with F-measure values.

Random Forest:

Nizri and Indik: Entropy and gini criteria were used, the best result was obtained with max_depth=17.

Our study: Hyperparameter settings are important. It should be compared with your F-measure values.

4. Performance Metric

Nizri and Indik:

F-measure was used as performance metrics.

The best hyperparameter combination of each model was determined and the final evaluation was made.

Our work:

The performance metrics you use and the best hyperparameter combinations for each model are important. It should be compared with your F-measure values.

Conclusion

Both studies used similar models and data sets. However, each study has differences in hyperparameter settings and data processing steps. Therefore, F-measure values and hyperparameter settings used should be compared carefully when comparing model performances. There are significant differences between the F-measure results in our study and Nizri and Indik's study, the reasons for these differences are due to differences in hyperparameter settings, data processing steps or model configurations.

Future Work

This study has laid a solid foundation for predicting Covid-19 risk using machine learning techniques. However, there are several avenues for future work that could further enhance the accuracy and applicability of our models:

1. Feature Engineering: Incorporate additional features and domain-specific knowledge to improve model performance. Exploring advanced techniques such as feature selection and dimensionality reduction could help in identifying the most influential features.

2. Hyperparameter Optimization: Employ advanced hyperparameter optimization techniques like Grid Search or Bayesian Optimization to fine-tune the models further and potentially improve their predictive power.

3. Ensemble Methods: Investigate the use of ensemble methods, such as boosting and bagging, to combine the strengths of different classifiers and improve overall performance.

4. Longitudinal Data Analysis: Extend the analysis to include longitudinal data, which tracks patients' health status over time. This could provide insights into the progression of the disease and improve the prediction of long-term outcomes.

5. Incorporation of New Data: As new data becomes available, continually update and retrain the models to ensure they remain current and accurate. This includes integrating data from new waves of the pandemic and emerging variants of the virus.

6. Real-world Testing: Implement and test the models in real-world clinical settings to validate their effectiveness and gather feedback for further refinement. Collaboration with healthcare professionals will be crucial in this phase.

7. Cross-Regional Analysis: Extend the study to include data from different regions and demographics to ensure the models are robust and generalizable across various populations.

By pursuing these future directions, we aim to enhance the predictive capabilities and practical utility of our Covid-19 risk prediction models, ultimately contributing to better patient outcomes and more efficient healthcare resource management.

REFERENCES

https://www.kaggle.com/code/berkaybaka/covid-19-analysis-and-prediction/edit