MACHINE LEARNING & DATA ANALYSIS

COVID-19 INFECTION CLASSIFICATION WITH MACHINE LEARNING

Submitted By

HEMENDRA JAMPALA (00695281)

Under The Guidance of

PROF. TRAVIS MILLBURN

Overview

- Introduction
- Dataset Details
- Data Visualizations
- Data Preprocessing
- Trained Models
- Results
- Future Enhancement
- Conclusion

Introduction & Problem Statement:

- Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).
- Worlwide

  Total cases : 67 M / Recovered: 43 M / Deaths: 1.54M
- Initially prompt and accurate molecular diagnosis of COVID-19 was very challenging
- The use of modern technology with AI and ML dramatically improves the screening, prediction, contact tracing, forecasting, and drug/vaccine development with extreme reliability.

- This Model helps to predict whether the patient will be Covid Positive or Negitive

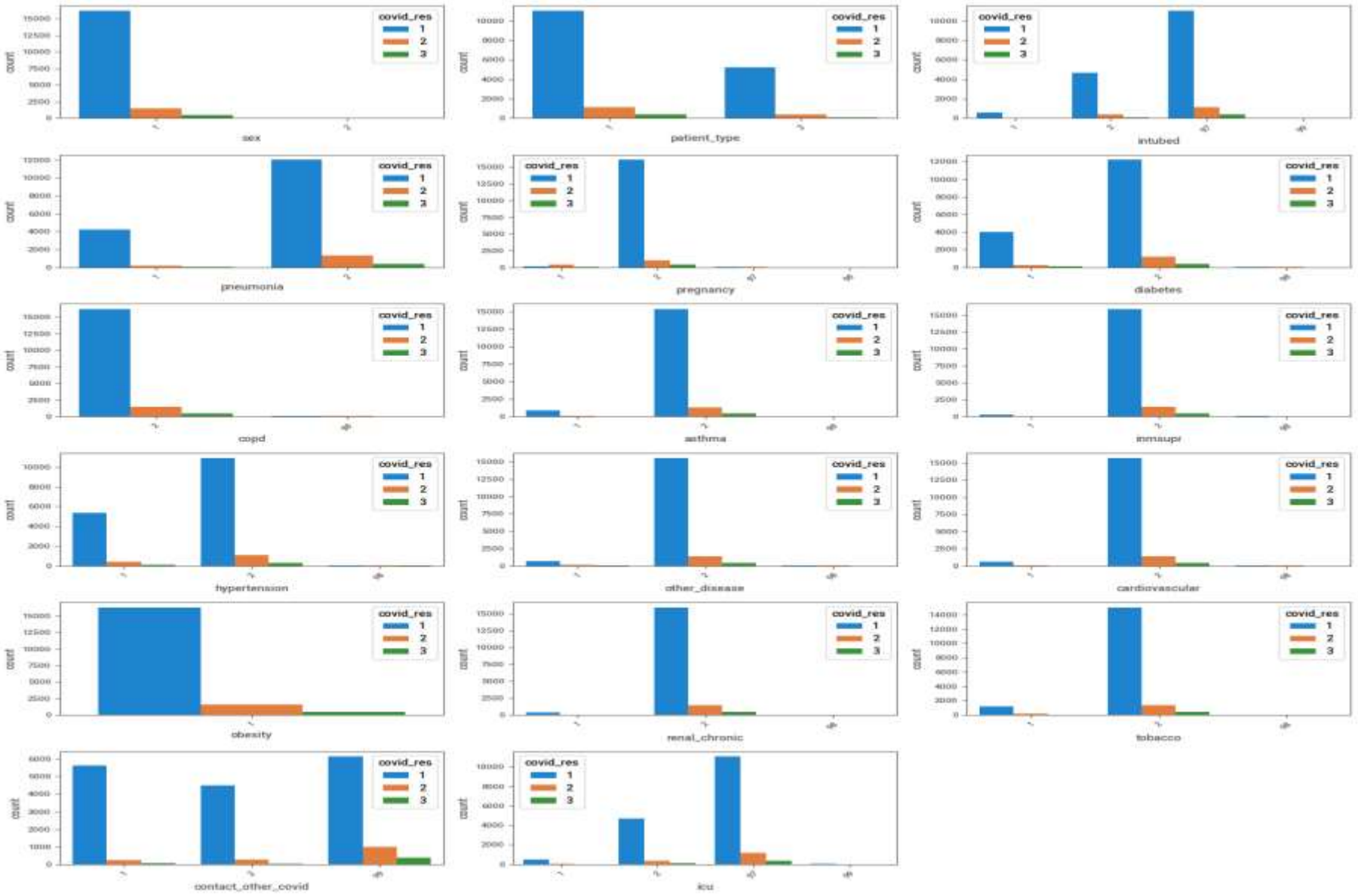
DATASET DETAILS:

- The Dataset is from Kaggle.
- Dataset was released by the Mexican government.
- It has 20K Samples and 23 Features.

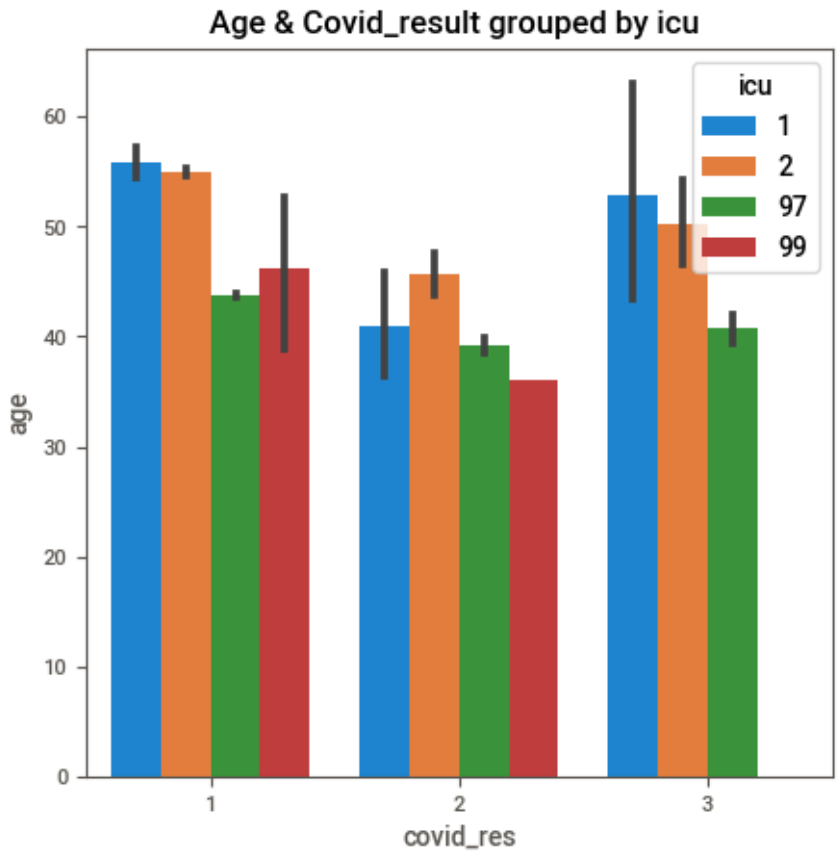
- Some of the Features in the dataset are

'age', 'patient_type', 'contact_other_covid', 'intubed', 'pneumonia', 'pregnancy', 'diabetes', 'copd', 'asthma', 'inmsupr', 'hypertension', 'other_disease', ‘gender', 'cardiovascular', 'obesity', 'renal_chronic', 'tobacco', 'covid_res', 'icu‘

DATA VISUALIZATIONS:

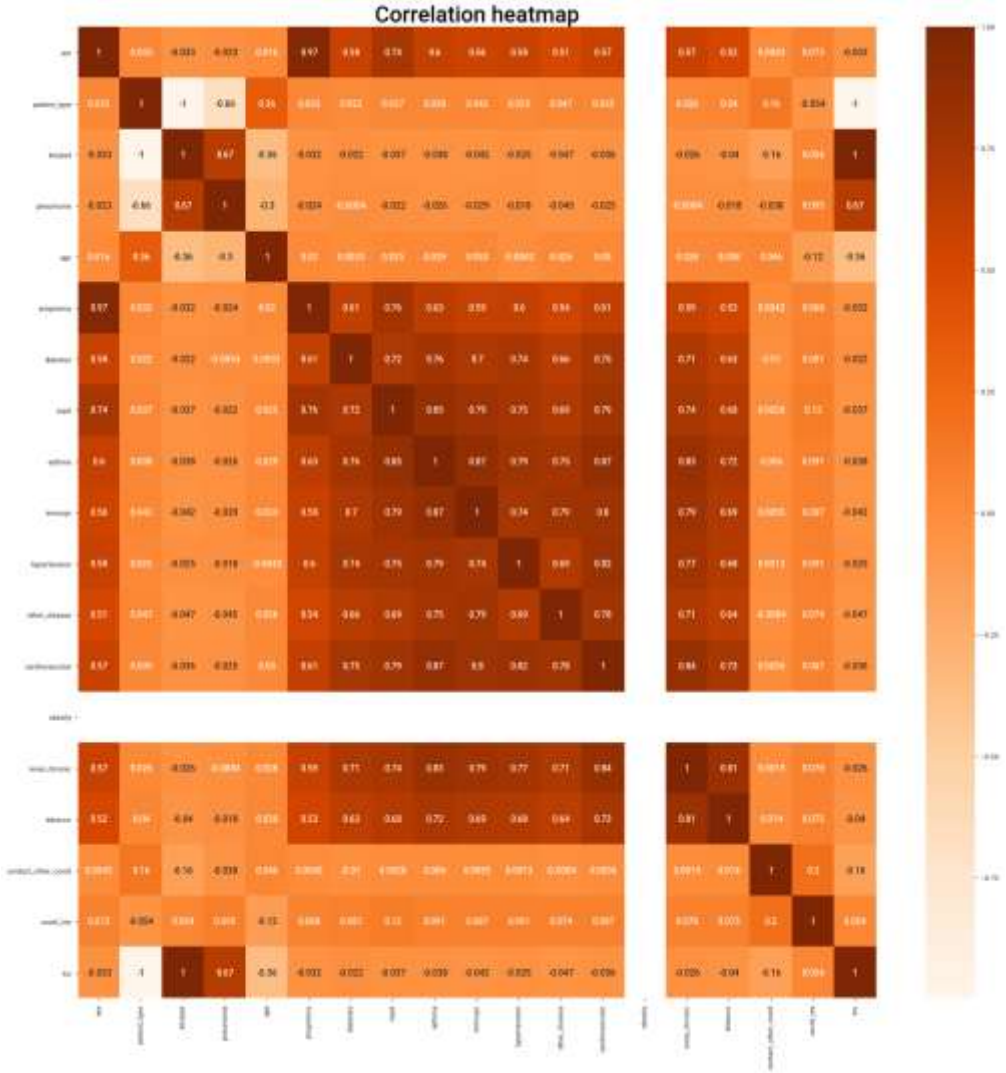


DATA VISUALIZATIONS:



Python Libs:- Pandas, Seaborn, Matplotlib

Correlation-Heatmap :



Data Preprocessing:

Applied below preprocessing techniques to the data before feeding it into our model.

- Null value treatment – Fill NA with Mean/Median.
- Label Encoding -Handling Categorical Variables
- Feature Scaling – Standard Scalar

Trained Models:

- KNeighborsClassifier
- LogisticClassifier
- DecisionTreeClassifier
- RandomForestClassifier

Model Results:

KNeighborsClassifier - Accuracy is: 0.8935528120713306
RandomForestClassifier -Accuracy is:0.9001371742112483

```
In [9]: best_model = compare_models()
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
0	Gradient Boosting Classifier	0.9072	0.0000	0.4265	0.8753	0.8825	0.3235	0.3787	6.1654
1	Ridge Classifier	0.9070	0.0000	0.4245	0.8723	0.8821	0.3212	0.3774	0.0499
2	Light Gradient Boosting Machine	0.9070	0.0000	0.4319	0.8876	0.8826	0.3181	0.3754	1.0242
3	Extreme Gradient Boosting	0.9065	0.0000	0.4332	0.8839	0.8824	0.3167	0.3721	3.0109
4	Linear Discriminant Analysis	0.9063	0.0000	0.4279	0.8750	0.8828	0.3300	0.3784	0.1574
5	CatBoost Classifier	0.9062	0.0000	0.4296	0.8866	0.8813	0.3080	0.3664	41.3097
6	Logistic Regression	0.9058	0.0000	0.4169	0.8703	0.8795	0.3012	0.3617	0.8812
7	Ada Boost Classifier	0.9045	0.0000	0.4098	0.8711	0.8769	0.2796	0.3453	0.6636
8	SVM - Linear Kernel	0.9006	0.0000	0.4157	0.8660	0.8744	0.2759	0.3267	0.6523
9	K Neighbors Classifier	0.8956	0.0000	0.3873	0.8534	0.8651	0.1996	0.2531	0.2981
10	Random Forest Classifier	0.8921	0.0000	0.4367	0.8640	0.8736	0.2859	0.3069	0.1436
11	Extra Trees Classifier	0.8914	0.0000	0.4378	0.8646	0.8728	0.2797	0.3008	0.6152
12	Decision Tree Classifier	0.8877	0.0000	0.4345	0.8596	0.8698	0.2684	0.2854	0.0859
13	Naive Bayes	0.8405	0.0000	0.4152	0.8540	0.8406	0.2008	0.2097	0.0358
14	Quadratic Discriminant Analysis	0.0300	0.0000	0.3411	0.5899	0.0121	0.0023	0.0326	0.2148

Future Enhancement:

- We can create a Web application, where ask user to provide responses for all of our input features.
- Based on user input we predict target feature like Covid-19 Positive or Negitive.
- We will display the predicted value to the user.
- We need to Deploy the ML model.

THANK YOU