PREDICTIONS OF SARS-COV2 COVID-19 USING ENSEMBLE AND OTHER CLASSIFIER OF MACHINE LEARNING

[1]Syeda Sumaira Mohammadi, [2]Dr. S. Sathessh Kumar
[1]PG Scholar, M.Tech, Dept of CSE, Shadan Women's College of Engineering and Technology, Hyderabad, TS.
sumairasyed20@gmail.com
[2]Assoc Professor, Dept of CSE, Shadan Women's College of Engineering and Technology, Hyderabad, TS.

ABSTRACT

COVID-19 was considered a pandemic by the World Health Organization. Since then, world governments have coordinated information flows and issued guidelines to contain the overwhelming effects of this disease. At the same time, the scientific community is continually seeking information about transmission mechanisms, the clinical spectrum of the disease, new diagnoses, and strategies for prevention and treatment. One of the challenges is performing the tests for the diagnosis of the disease, whose technique adopted for the detection of the genetic material of COVID-19 requires equipment and specialized human resources, making it an expensive procedure. We hypothesize that machine learning techniques can be used to classify the test results for COVID-19 through the joint analysis of popular laboratory tests' clinical parameters. Machine learning techniques, such as Random Forest, Multi-Layer Perceptron, and Support Vector Machines Regression, enable the creation of disease prediction models and artificial intelligence techniques to analyze clinical parameters. Thus, we evaluated the existing correlations between laboratory parameters and the result of the COVID-19 test, and developed two classification models: the first classifies the test results for patients with suspected COVID-19, and the second classifies the hospitalization units of patients with COVID-19, both according to the laboratory parameters. The models achieved an accuracy above 96%, showing that they are promising to the classification of tests for COVID-19 and screen patients by hospitalization unit.

I. INTRODUCTION

A new type of respiratory infection was reported in China, from an outbreak of pneumonia caused by a later identified agent, SARS-Cov-2. Infection with the new coronavirus (called COVID-19) has been proving to be a challenge for global health systems. Among the challenges faced in combating the pandemic are the cost and processing time of tests to detect the disease. For diagnosis, the gold standard is the polymerase chain reaction (PCR), which analyzes the presence of viral nucleic acid. However, several factors, such as the type of sample used in the reaction, can influence the test result, causing variations in the sensitivity of the technique depending on the viral load and the material used. Due to the various difficulties inherent in the diagnostic process, of laboratory tests such as blood or urine tests. In addition, PCR is highly complex, as it requires specific professionals and specialized techniques, and a single undetected result with RT-PCR for SARS-CoV-2 does not exclude the diagnosis of infection.

These aspects make diagnosis difficult and end up increasing the possibilities of disease transmission and, consequently, the contagion curves. In Brazil, there are deficiencies, both in the lack of test kits and in the slow processing of these tests. Bioinformatics is a field of knowledge that can manipulate large volumes of clinical data through machine learning techniques. Modern tools provide better data visualization, favoring the extraction of correlations and patterns from data analysis. In addition, machine learning techniques provide the creation of models that allow predictions.

Among the applications of bioinformatics are the diagnosis of diseases and the simulation of prognoses through predictive models. Random Forest (Random Forest, RF), Multilayer Perceptron (Multi-Layer Perceptron, MLP) and Support Vector Machine (Support Vector Machine, SVM) are examples of machine learning algorithms that are able to model complex processes. Technological advancements in high-throughput cell biology have enabled researchers to examine the landscape of biomolecules (i.e., DNA, RNA, proteins, metabolites, etc.) associated with a phenotype of interest.

Next-generation sequencing technologies have revolutionized the profiling of DNA and messenger RNA (mRNA), allowing genomes and transcriptomes to be sequenced quickly and economically. Mass spectrometry allows us to efficiently identify and quantify proteins, metabolites and lipids in cells, capturing underlying cellular variations in response to physiological and pathological changes.

Consequently, large-scale studies on the genome, the transcriptome, the proteome, the metabolome, the liposome, etc. have created a plethora of data associated with these "-omens" also known as "omics" data. In this regard, machine learning (ML) algorithms have been developed to elucidate complex cellular mechanisms, identify molecular signatures, and predict clinical outcomes from large biomedical datasets. Traditionally, ML based single-omics analyses provide assorted perspectives on cellular processes with respect to a particular me. However, isolated omics studies frequently fall short when identifying the cause of multifaceted diseases such as cancer, cardiac diseases, diabetes, etc.

This evidence suggests that an inclusive view of cellular processes, constructed by integrating information within and across -omens, is required to provide a comprehensive picture of the biological mechanisms ML-empowered integrative analysis has emerged as a key player in studies involving multiple omics data. By analyzing different omics layers together, ML-based integrative methods provide a holistic view of biological processes, offer new mechanistic insights on the phenotype of interest, and facilitate the advancements in precision medicine [26]. For example, Hoadley et al. employed ML-based integrative clustering in a comprehensive study of twelve different types of cancer which resulted in a new molecular taxonomy of diverse tumor types.

They integrated genomics, epigenetics, transcriptomic, and proteomics data utilizing cluster-of-cluster assignments (COCA) to obtain clinically relevant sub-types. In, canonical correlation analysis (CCA) with dimensionality reduction was employed for jointly analyzing microRNA (miRNA) and gene expression data. This analysis provided insight into the mechanisms of head and neck squamous cell cancer and its response to treatment via cetuximab. In another study, performed integrative analysis of somatic mutations, RNA expression, and DNA methylation data associated with chronic lymphocytic leukemia (CLL). This study identified new factors predictive of clinical outcome by employing a latent variable modeling approach.

To identify markers of body fat mass changes in obesity, proteomics and metabolomics data were integrated to create a “transonic” dataset whose individual features went through z-score transformation prior to independent component analysis (ICA). It was noted that a combined Tran’s omics dataset better discriminates lean and obese subjects as compared to single-omics data. For improving drug sensitivity in breast cancer, genomics, epigenetics, and proteomics, data were integrated using a multitier multiple kernel learning (MKL) approach.

This study showed that the predictive performance achieved by multitier learning was found to be better than that obtained by any individual view, where a ‘view’ describes a particular representation of the input data. Integrative analysis of biomedical data with ML can be performed in a variety of ways. For example, the simplest approach is to construct a large feature matrix by directly concatenating features from different datasets.

Each feature may go through z-transformation for standardization across all biological samples, followed by ML-based feature selection for molecular signature extraction and biomarker identification. Another common integrative analysis approach is to transform data from heterogeneous sources into joint latent profiles.

SCOPE OF THE PROJECT

The scope of the project is designing an application which will provide information about college for a new user who don’t know about the college. Simply user can ask the question to system according to that question the system will give the answer to user. By this application user can easily find where he needs to go inside the college.

OBJECTIVE

The COVID-19 pandemic has also had substantial impacts on air quality. As a primary method of slowing down the spread of COVID-19, initially, a lot of countries-imposed lockdown or confinement measures to enforce strict social distancing regulations. As a result, businesses and shops were closed, manufacturing activities were either stopped or shrunken while the number of vehicles in cities has declined dramatically [24]–[26]. Therefore, lockdown and confinement measures played a critical role in curtailing emissions and in improving overall air quality. Improved air quality refers to the reduction of concentration of criteria pollutants such as NO2, SO2, PM10, PM2.5, CO2 in the air. According to the International Energy Agency (IEA), the global energy demand decreased by 3.8% in the first quarter of 2020 compared to the same period of 2019 because of the sudden reduction in economic activities and mobility [24]. Many recent studies on urban air quality have estimated the impacts of lockdown and confinement measures on various criteria pollutants. These studies mainly estimate the business-as-usual concentration of the pollutants for 2020 based on climate variables using ML algorithms. Finally, the impacts of the lockdown and confinement on air quality were assessed by comparing the estimated baseline concentration with actual concentration of pollutants in 2020.

PROBLEM STATEMENT

In addition, there are ML-based frameworks that fuse data as a step toward building a model, e.g., multiple kernel learning or network modelling approaches. Notably, the accumulation of large biomedical data and the inevitable benefits of studying multiple omics together present new challenges and opportunities for developing novel computational approaches customized for integrative analysis.

For example, heterogeneous data with mixed variable types, and missing values in one or more omics can substantially hinder the data integration and analysis. In addition, when integrating multiple omics data, the dimensions of the dataset can grow into hundreds or thousands of variables, while the number of observations or biological samples remains limited. This disparity is called the curse of dimensionality or the p >> n problem, where p is the number of variables and the number of samples. Moreover, the rarity or class imbalance in the data can also lead to results that are biased or less accurate.

A class imbalance problem arises when rare events are analyzed and compared against events that happen much more frequently, a common occurrence in omics datasets. Furthermore, standard integrative frameworks may not be suitable for large-scale multi-omics analysis due to computational and storage limitations.

EXISTING SYSTEM

The novel coronavirus (nCoV-2019) outbreak in Wuhan, China has spread rapidly nationwide, with some cases occurring in other parts of the world. Although most patients present with mild febrile illness with patchy pulmonary inflammation, a significant portion develop severe acute respiratory distress syndrome (ARDS), with a current case fatality of 2.3-3%. Diagnosis is based on clinical history and laboratory and chest radiographic findings, but confirmation currently relies on nucleic acid-based assays. The latter are playing an important role in facilitating patient isolation, treatment and assessment of infectious activities. However, due to their limited capacity to handle an epidemic of the current scale and insufficient supply of assay kits, only a portion of suspected cases can be tested, leading to incompleteness and inaccuracy in updating new cases, as well as delayed diagnosis. Furthermore, there has not been enough time to assess specificity and sensitivity. Conventional serological assays, such as enzyme-linked immunoassay (ELISA) for specific IgM and IgG antibodies, should offer a high-throughput alternative, which allows for uniform tests for all suspected patients, and can facilitate more complete identification of infected cases and avoidance of unnecessary cross infection among unselected patients.

II. LITERATURE SURVEY

"Professional chat application based on natural language processing.

There has been an emerging trend of a vast number of chat applications which are present in the recent years to help people to connect with each other across different mediums, like Hike, WhatsApp, Telegram, etc. The proposed network-based android chat application used for chatting purpose with remote clients or users connected to the internet, and it will not let the user send inappropriate messages. This paper proposes the mechanism of creating professional chat application that will not permit the user to send inappropriate or improper messages to the participants by incorporating base level implementation of natural language processing (NLP). Before sending the messages to the user, the typed message evaluated to find any inappropriate terms in the message that may include vulgar words, etc., using natural language processing. The user can build an own dictionary which contains vulgar or irrelevant terms. After pre-processing steps of removal of punctuations, numbers, conversion of text to lower case and NLP concepts of removing stop words, stemming, tokenization, named entity recognition and parts of speech tagging, it gives keywords from the user typed message. These derived keywords compared with the terms in the dictionary to analyze the sentiment of the message. If the context of the message is negative, then the user not permitted to send the message

Real world smart chatbot for customer care using software as service (SaaS) architecture.

It's being very important to listen to social media streams whether it's Twitter, Facebook, Messenger, LinkedIn, email or even company own application. As many customers may be using this streams to reach out to company because they need help. The company have setup social marketing team to monitor this stream. But due to huge volumes of users it's very difficult to analyses each and every social message and take a relevant action to solve users' grievances, which lead to many unsatisfied customers or may even lose a customer. This papers proposes a system architecture which will try to overcome the above shortcoming by analyzing messages of each ejabberd users to check whether it's actionable or not. If it's actionable then an automated Chatbot will initiates conversation with that user and help the user to resolve the issue by providing a human way interactions using LUIS and cognitive services. To provide a highly robust, scalable and extensible architecture, this system is implemented on AWS public cloud.[7]

An Overview of Artificial Intelligence Based Chatbots and An Example Chatbot Application.

Chatbot can be described as software that can chat with people using artificial intelligence. These software are used to perform tasks such as quickly responding to users, informing them, helping to purchase products and providing better service to customers. In this paper, we present the general working principle and the basic concepts of artificial intelligence based chatbots and related concepts as well as their applications in various sectors such as telecommunication, banking, health, customer call centers and e-commerce. Additionally, the results of an example chatbot for donation service developed for telecommunication service provider are presented using the proposed architecture.

Intelligent travel chatbot for predictive recommendation in echo platform

Chatbot is a computer application that interacts with users using natural language in a similar way to imitate a human travel agent. A successful implementation of a chatbot system can analyze user preferences and predict collective intelligence. In most cases, it can provide better user-centric recommendations. Hence, the chatbot is becoming an integral part of the future consumer services. This paper is an implementation of an intelligent chatbot system in travel domain on Echo platform which would gather user preferences and model collective user knowledge base and recommend using the Restricted Boltzmann Machine (RBM) with Collaborative Filtering. With this chatbot based on DNN, we can improve human to machine interaction in the travel domain

Chatbot Using a Knowledge in Database Human-to-Machine Conversation Modeling

A chatterbot or chatbot aims to make a conversation between both human and machine. The machine has been embedded knowledge to identify the sentences and making a decision itself as response to answer a question. The response principle is matching the input sentence from user. From input sentence, it will be scored to get the similarity of sentences, the higher score obtained the more similar of reference sentences. The sentence similarity calculation in this paper using bigram which divides input sentence as two letters of input sentence. The knowledge of chatbot is stored in the database. The chatbot consists of core and interface that is accessing that core in relational database management systems (RDBMS). The database has been employed as knowledge storage and interpreter has been employed as stored programs of function and procedure sets for pattern-matching requirement. The interface is standalone which has been built using programing language of Pascal and Java.

III. PROPOSED METHODOLOGY

In this chapter, various supervised machine learning approaches are used. This section provides a general description of these approaches. This ML-based AI is however different from the symbolic rule-based AI. Unlike the rule-based AI, where decisions are made based on some PR=0 -defined rules, ML-based AI learns from annotated classified datasets, examples, and experiences. In ML-based AI, a model is developed based on the information from the dataset, when it used for prediction. Also, the algorithm can learn to optimize models based on the dataset and policies for a specific task, for example, a screening process with an acceptable high false alarm policy. Initially, when there exists no relevant model in a model library, the four-layered Bayesian modeling in the preceptor extracts the statistical model of the system using decision trees. The Bayesian modeling consists of four layers for an arbitrary focus level m. Here, in this paper, the decision tree level m is considered as the focus level m for the CDS. The CDS initiates the reasoning mode when an anomaly in the user’s health is detected or a request is placed by the user.

1. Dataset
2. Pre-Processing
3. Splitting
4. Apply Algorithm
5. Visualization
6. Accuracy

➢ **Data set**

A data set is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question.

➢ **Pre-Processing**

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way.

➢ **Splitting**

Data splitting is the act of partitioning available data into. Two portions, usually for cross-validator purposes. One portion of the data is used to develop a predictive model. And the other to evaluate the model's performance.

- Training Data: Used for train the model or given as input to the to the learning model
- Testing Data: Used for test the model or given as input to the model for prediction.

➢ **Apply Algorithm**

In this we are using support vector machine algorithm to predict accuracy. It is a non-probabilistic supervised machine learning approaches used for classification and regression. It assigns a new data member to one of two possible classes. It defines a hyperplane that separates n-dimensional data into two classes.

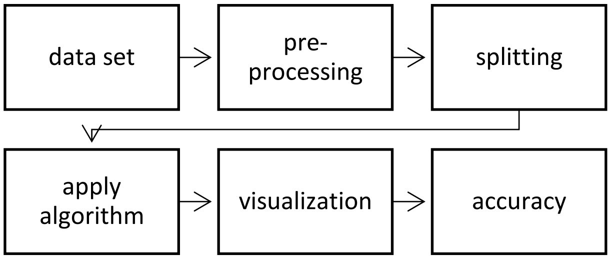
➢ **Visualization**

Visualization is a technique that uses an array of static and interactive visuals within a specific context to help people understand and make sense of large amounts of data. The data is often displayed in a story format that visualizes patterns, trends and correlations that may otherwise go unnoticed.

➢ **Accuracy**

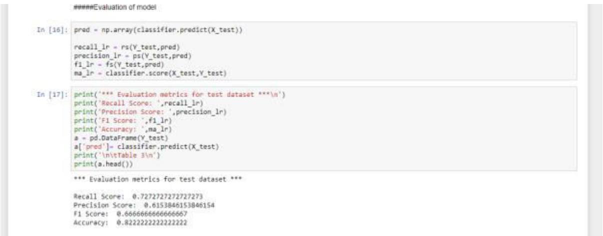
Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

IV. SYSTEM ARCHITECTURE



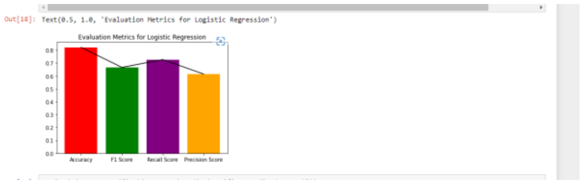
V. RESULTS

Experiments will be carried out with a larger and more balanced dataset to evaluate the generalizability of the results. We also intend to test other data imputation and machine learning techniques. We plan to analyze a dataset with less missing data and apply other machine learning techniques, both for data imputation and for training the models.



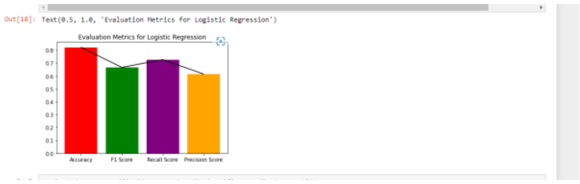
```
#####Evaluation of model

In [16]: pred = np.array(classifier.predict(X_test))

         recall_lr = rs(Y_test,pred)
         precision_lr = ps(Y_test,pred)
         f1_lr = fs(Y_test,pred)
         ma_lr = classifier.score(X_test,Y_test)

In [17]: print('*** Evaluation metrics for test dataset ***\n')
         print('Recall Score: ',recall_lr)
         print('Precision Score: ',precision_lr)
         print('F1 Score: ',f1_lr)
         print('Accuracy: ',ma_lr)
         a = pd.DataFrame(Y_test)
         a['pred']= classifier.predict(X_test)
         print('\n\tTable 3\n')
         print(a.head())

         *** Evaluation metrics for test dataset ***

         Recall Score:  0.7272727272727273
         Precision Score:  0.6153846153846154
         F1 Score:  0.6666666666666667
         Accuracy:  0.8222222222222222
```

Default Responses



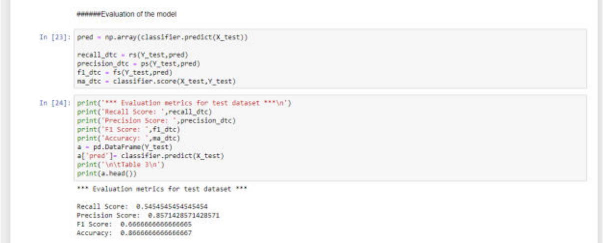
```
Out[18]: Text(0.5, 1.0, 'Evaluation Metrics for Logistic Regression')
```

Typing Tab



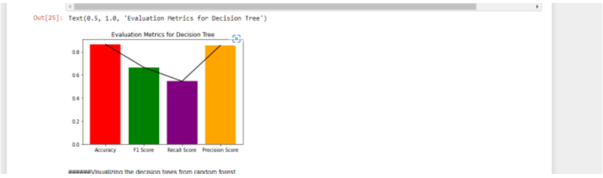
```
Out[18]: Text(0.5, 1.0, 'Evaluation Metrics for Logistic Regression')
```

Conversational Query



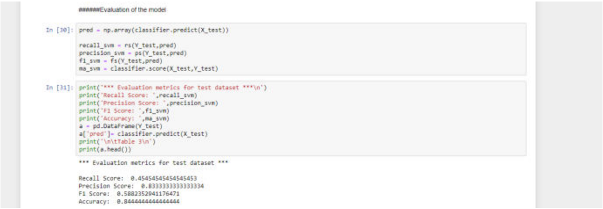
```
######Evaluation of the model

In [23]: pred = np.array(classifier.predict(X_test))

         recall_dtc = rs(Y_test,pred)
         precision_dtc = ps(Y_test,pred)
         f1_dtc = fs(Y_test,pred)
         ma_dtc = classifier.score(X_test,Y_test)

In [24]: print('*** Evaluation metrics for test dataset ***\n')
         print('Recall Score: ',recall_dtc)
         print('Precision Score: ',precision_dtc)
         print('F1 Score: ',f1_dtc)
         print('Accuracy: ',ma_dtc)
         a = pd.DataFrame(Y_test)
         a['pred']= classifier.predict(X_test)
         print('\n\tTable 3\n')
         print(a.head())

         *** Evaluation metrics for test dataset ***

         Recall Score:  0.5454545454545454
         Precision Score:  0.8571428571428571
         F1 Score:  0.6666666666666665
         Accuracy:  0.8666666666666667
```

Conversational Query Continuation

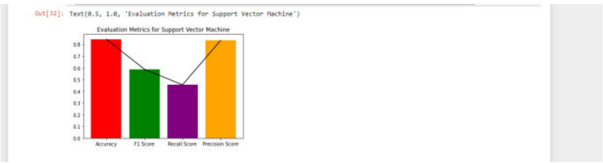


```
Out[25]: Text(0.5, 1.0, 'Evaluation Metrics for Decision Tree')
```

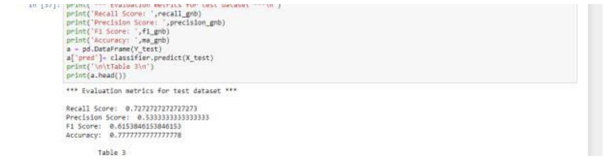
```
######Visualizing the decision trees from random forest
```



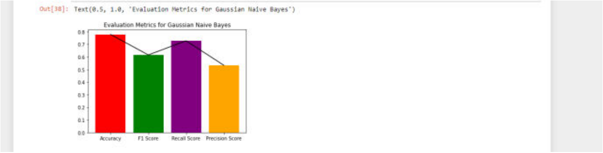
```
######Evaluation of the model

In [30]: pred = np.array(classifier.predict(X_test))

         recall_svm = rs(Y_test,pred)
         precision_svm = ps(Y_test,pred)
         f1_svm = fs(Y_test,pred)
         ma_svm = classifier.score(X_test,Y_test)

In [31]: print('*** Evaluation metrics for test dataset ***\n')
         print('Recall Score: ',recall_svm)
         print('Precision Score: ',precision_svm)
         print('F1 Score: ',f1_svm)
         print('Accuracy: ',ma_svm)
         a = pd.DataFrame(Y_test)
         a['pred']= classifier.predict(X_test)
         print('\n\tTable 3\n')
         print(a.head())

         *** Evaluation metrics for test dataset ***

         Recall Score:  0.45454545454545453
         Precision Score:  0.8333333333333334
         F1 Score:  0.5882352941176471
         Accuracy:  0.8444444444444444
```



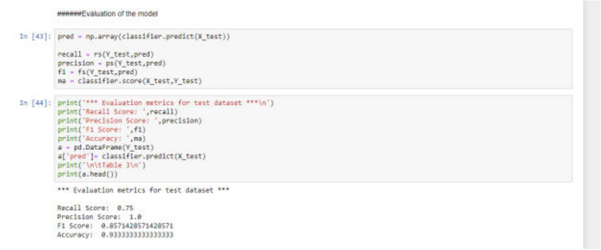
```
Out[32]: Text(0.5, 1.0, 'Evaluation Metrics for Support Vector Machine')
```



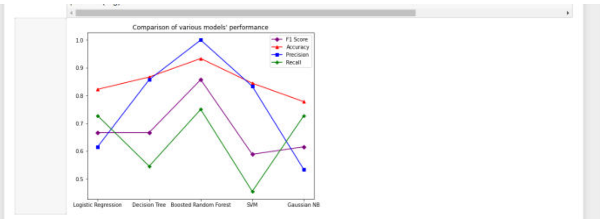
```
In [37]: print('*** Evaluation metrics for test dataset ***\n')
         print('Recall Score: ',recall_gnb)
         print('Precision Score: ',precision_gnb)
         print('F1 Score: ',f1_gnb)
         print('Accuracy: ',ma_gnb)
         a = pd.DataFrame(Y_test)
         a['pred']= classifier.predict(X_test)
         print('\n\tTable 3\n')
         print(a.head())

         *** Evaluation metrics for test dataset ***

         Recall Score:  0.7272727272727273
         Precision Score:  0.5333333333333333
         F1 Score:  0.6153846153846153
         Accuracy:  0.7777777777777778

                 Table 3
```



```
Out[38]: Text(0.5, 1.0, 'Evaluation Metrics for Gaussian Naive Bayes')
```



```
######Evaluation of the model

In [43]: pred = np.array(classifier.predict(X_test))

         recall = rs(Y_test,pred)
         precision = ps(Y_test,pred)
         f1 = fs(Y_test,pred)
         ma = classifier.score(X_test,Y_test)

In [44]: print('*** Evaluation metrics for test dataset ***\n')
         print('Recall Score: ',recall)
         print('Precision Score: ',precision)
         print('F1 Score: ',f1)
         print('Accuracy: ',ma)
         a = pd.DataFrame(Y_test)
         a['pred']= classifier.predict(X_test)
         print('\n\tTable 3\n')
         print(a.head())

         *** Evaluation metrics for test dataset ***

         Recall Score:  0.75
         Precision Score:  1.0
         F1 Score:  0.8571428571428571
         Accuracy:  0.9333333333333333
```



Registered Details for Queries

APPLICATION

- Retail and e-commerce.
- Travel and hospitality.
- Banking, finance, and fintech.
- Healthcare.
- Media and entertainment.
- Education.

VI. CONCLUSION

In this article, machine learning algorithms were successfully used to train models for two classification tasks to help combat COVID-19. The first task was to classify the positive and negative test results for COVID-19 and the second task was to classify the inpatient unit of the patient with COVID-19, being possible the units: ward, semi-intensive unit or intensive care unit. It was used a dataset containing laboratory parameters of clinical exam results of patients with suspected COVID-19.

VII. FUTURE ENHANCEMENT

In our future work, experiments will be carried out with a larger and more balanced dataset to evaluate the generalizability of the results. We also intend to test other data imputation and machine learning techniques. We plan to analyze a dataset with less missing data and apply other machine learning techniques, both for data imputation and for training the models.

REFERENCES

[1] S.-Y. Xiao, Y. Wu, and H. Liu, “Evolving status of the 2019 novel coronavirus infection: Proposal of conventional serologic assays for disease diagnosis and infection monitoring,” Journal of medical virology, vol. 92, no. 5, pp. 464–467, 2020.

[2] W. Wang, Y. Xu, R. Gao, R. Lu, K. Han, G. Wu, and W. Tan, “Detection of sars-cov-2 in different types of clinical specimens,” Jama, vol. 323, no. 18, pp. 1843–1844, 2020.

[3] G. Lippi and M. Plebani, “Laboratory abnormalities in patients with covid-2019 infection,” Clinical Chemistry and Laboratory Medicine (CCLM), vol. 58, no. 7, pp. 1131–1134, 2020.

[4] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong et al., “Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China,” Jama, vol. 323, no. 11, pp. 1061–1069, 2020.

[5] C. C. M. Davi, D. S. Silveira, and F. B. L. Neto, “A framework using computational intelligence techniques for decision support systems in medicine,” IEEE Latin America Transactions, vol. 12, no. 2, pp. 205– 211, 2014. [6] F. E. S. Alencar, D. C. Lopes, and F. M. Mendes Neto, “Development of a system classification of images dermoscopic for mobile devices,” IEEE Latin America Transactions, vol. 14, no. 1, pp. 325–330, 2016.

[7] A. Urrutia, E. Chavez, R. Motz, and R. Gajardo, “An ontology to assess data quality domains. a case study applied to a health care entity,” IEEE Latin America Transactions, vol. 15, no. 8, pp. 1506–1512, 2017.

[8] J.-j. Zhang, X. Dong, Y.-y. Cao, Y.-d. Yuan, Y.-b. Yang, Y.-q. Yan, C. A. Akdis, and Y.-d. Gao, “Clinical characteristics of 140 patients infected with sars-cov-2 in wuhan, china,” Allergy, 2020.

[9] T. Ma and A. Zhang, “Multi-view factorization autoencoder with network constraints for multi-omic integrative analysis,” in IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2018, pp. 702–707.

[10] D. Sun, M. Wang, and A. Li, “A multimodal deep neural network for human breast cancer prognosis prediction by integrating multidimensional data,” IEEE/ACM transactions on computational biology and bioinformatics, vol. 16, no. 3, pp. 841–850, 2018.

[11] B. Mirza, W. Wang, J. Wang, H. Choi, N. C. Chung, and P. Ping, “Machine learning and integrative analysis of biomedical big data,” Genes, vol. 10, no. 2, p. 87, 2019.

[12] L. Breiman, Machine Learning. Dordrecht, 1997, vol. 45, no. 15–32.

[13] F. Murtagh, “Multilayer perceptrons for classification and regression,” Neurocomputing, vol. 2, no. 5-6, pp. 183–197, 1991.

[14] V. Vapnik, The nature of statistical learning theory. Springer science & business media, 2013.

[15] A. Alimadadi, S. Aryal, I. Manandhar, P. Munroe, B. Joe, and X. Cheng, “Artificial intelligence and machine learning to fight covid-19.” Physiological genomics, vol. 52, no. 4, pp. 200–202, 2020.

[16] G. S. Randhawa, M. P. Soltysiak, H. El Roz, C. P. de Souza, K. A. Hill, and L. Kari, "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: Covid-19 case study," PLoS One, vol. 15, no. 4, pp. 1–24, 2020.

[17] H. C. Metsky, C. A. Freije, T.-S. F. Kosoko-Thoroddsen, P. C. Sabeti, and C. Myhrvold, "Crispr-based covid-19 surveillance using a genomicallycomprehensive machine learning approach," bioRxiv, 2020.

[18] L. Yan, H.-T. Zhang, Y. Xiao, M. Wang, C. Sun, J. Liang, S. Li, M. Zhang, Y. Guo, Y. Xiao et al., "Prediction of survival for severe covid-19 patients with three clinical features: development of a machine learning-based prognostic model with clinical data in wuhan," medRxiv, 2020.

[19] Y. Ge, T. Tian, S. Huang, F. Wan, J. Li, S. Li, H. Yang, L. Hong, N. Wu, E. Yuan et al., "A data-driven drug repositioning framework discovered a potential therapeutic agent targeting covid-19," bioRxiv, 2020.

[20] C. Butt, J. Gill, D. Chun, and B. A. Babu, "Deep learning system to screen coronavirus disease 2019 pneumonia," Applied Intelligence, vol. 1, no. 1, pp. 1–7, 2020.

[21] B. Ghoshal and A. Tucker, "Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection," arXiv, pp. arXiv–2003, 2020. [22] C. Feng, Z. Huang, L. Wang, X. Chen, Y. Zhai, F. Zhu, H. Chen, Y. Wang, X. Su, S. Huang et al., "A novel triage tool of artificial intelligence assisted diagnosis aid system for suspected covid-19 pneumonia in fever clinics," medRxiv, 2020.

[23] X. Jiang, M. Coffee, A. Bari, J. Wang, X. Jiang, J. Shi, J. Dai, J. Cai, T. Zhang, Z. Wu et al., "Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity," Computers, Materials & Continua, vol. 63, no. 1, pp. 537–551, 2020. [24] "Kaggle," kaggle.com/pedrojbl/diagnosis-of-covid-19-a-careful-analysis-v10-0, Acessado em: 2020-05-28.

[25] "scikit," scikit-learn.org, Acessado em: 2020-05-28. [26] A. Jedlitschka and D. Pfahl, "Reporting guidelines for controlled experiments in software engineering," in Int. Sym. Empirical Soft. Engineering (ESEM), 2005, pp. 95–104.

[27] C. Wohlin, P. Runeson, M. Host, M. C. Ohlsson, B. Regnell, and ¨ A. Wesslen, ´ Experimentation in software engineering. Springer Science & Business Media, 2012. [28] V. R. Basili, D. Rombach, K. S. B. Kitchenham, D. Selby, and R. W. Pfahl, Empirical Software Engineering Issues. Springer Berlin/Heidelberg, 2007.

[29] G. N. Wilkinson and C. E. Rogers, "Symbolic description of factorial models for analysis of variance," Journal of the Royal Statistical Society Series C, vol. 22, no. 3, pp. 392–399, 1973.

[30] J. Tukey, "Multiple comparisons," Journal of the American Statistical Association, vol. 48, no. 264, pp. 624–25, 1953.

[31] R. Feldt and A. Magazinius, "Validity threats in empirical software engineering research-an initial survey." in Int. Conf. on Software Engineering and Knowledge Engineering, 2010, pp. 374–379