2024 4th Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 23 - 25, 2024

Titanic Survival Predictor Using Machine Learning Algorithms

Mohammed Ali Shaik
School of computer science & Artificial Intelligence
SR University
Warangal, Telangana State, India
niharali@gmail.com

P Sairam
School of computer science & Artificial Intelligence
SR University
Warangal, Telangana State, India
sairam12497@gmail.com

P.Rahul
School of computer science & Artificial Intelligence
SR University
Warangal, Telangana State, India
pothurahul33@gmail.com

T. Sadhvik
School of computer science & Artificial Intelligence
SR University
Warangal, Telangana State, India
sunnysadhvik94@gmail.com

P.Nithin
School of computer science & Artificial Intelligence
SR University
Warangal, Telangana State, India
panagantinithin@gmail.com

2024 4th Asian Conference on Innovation in Technology (ASIANCON) | 979-8-3503-5421-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ASIANCON62057.2024.10837935

Abstract— **Presumably, one of the most notorious and catastrophic catastrophes in history is the sinking of the RMS Titanic. Unfortunately, the Titanic sank during her inaugural voyage on April 15, 1912, early in the morning after slamming into an iceberg. Out of 2224 passengers and crew, approximately 1502 perished. Social stratification is a situation where people within a society are separated into various ranking on a hierarchical basis. These groups, also known as social classes are ranked as a social order that is made up of a person's education level, income level Next similar models based on machine learning will be chosen to come to a conclusion whether some groups of individuals are candidates to survive the long space travel. The results of application of the distinct machine learning models were illustrated down one after the other and compared through a general viewpoint of precision. Initially, to get a better feeling of how the data is distributed and what the traits are, the datasets were first examined and prepared. Following that, we ran a variety of machine learning models including Logistic regression, decision trees, random forests, naive Bayes classifier and k nearest neighbors (kNN) for prediction. While conducting our research, we benchmarked the performance of each algorithm using strict assessment techniques like evaluation of variables like the required computational effort, accuracy, precision, recall, and F1 score.**

Keywords— Feature engineering, Machine learning, Model Evaluation, Exploratory Data Analytics

I. INTRODUCTION

The Titanic Survivors Project examines and forecasts the chances of survival for those who rode on the tragic RMS Titanic [1]. We carefully analyze the Titanic dataset using machine learning approaches in order to identify the elements that affect survival and build models that reliably predict survival probability [2]. We start with preprocessing the data to handle missing values and encode variables, and then we analyze the data exploratorily to look for relationships between passenger characteristics and survival rates [3]. We create predictive models by applying various methods, such as logistic regression [1] and decision trees [5], and then fine-tuning the parameters through cross-validation to maximize the models' performance [4]. The assessment of the model offers valuable insights into the forecast accuracy and highlights significant elements that affected the survivorship of the Titanic [5]. The project concludes by summarizing the results and outlining potential directions for future study to improve prediction accuracy and expand historical comprehension [6].

II. LITERATURE REVIEW

The main task of our study, Titanic Survivors' Analysis of Theory and Previous Studies, is to scrutinize previous studies and research on disaster survival prediction, namely the sinking of Titanic in 1912 [1]. The area of research is assessed; evaluating heterogeneous approaches and strategies used in the original study such as model evaluation measures, feature selection, and machine learning algorithms [7-10]. Besides this we have also taken care in detail to check how various methods perform well in accurately predict survival outcomes of passengers having different demographic profiles on the basis of ticket class, cabin location and various other variables [11-15]. The aim is to better the procedures and definitely build upon the database of knowledge in which the preexisting two have contributed [16].

TABLE I. COMPARISION TABLE OF RESEAFCH

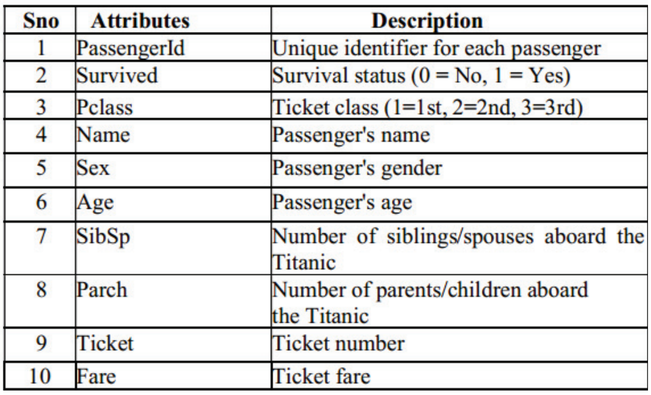
References	Dataset Utilized	Method	Performance Metrics (Accuracy)
[1]	Titanic dataset	LR DT RF	81.11% 99.29% 80.4%
[2]	Titanic Dataset	DT	85%
[3]	Training dataset	SVM KNN LR	82.82% 79.79% 79.12%
[4]	Titanic dataset	LR RF GB	72% 84% 86%
[5]	Titanic dataset	ANN	78%

We reviewed the results of the crucial studies that establish the survival rates of passengers according to their demographics and socioeconomic position as one of the methods of literature evaluation [17-18]. We also look into using up the the mark machine learning techniques like Support Vector Machines(SVM) [19], Random Forests [20], and Decision Trees [21] on dataset of Titanic that is used for survival prediction. We will as well have a look at the impact of feature engineering approaches (i.e., creating a new variable and impute missing values) on rising the model performance [22]. In addition, we importantly do research that investigates what role such passenger's attributes as age, gender and family size might take in their chance of survival [25]. We also scrutinize the disadvantages and limitations that the previous models were faced with, like the existence of overfitting and bias towards the actual data, and we learn the ways to overcome those limitations [23]. Thus, the final result we aim to reach is a precise prediction with our model by combining the viewpoints of many previous studies under the realization of the shortcomings of both the achievements and the limitations [24].

III. Proposed Methodology

A. *Dataset Used:*

TABLE II. Dataset Attributes Description



Sno	Attributes	Description
1	PassengerId	Unique identifier for each passenger
2	Survived	Survival status (0 = No, 1 = Yes)
3	Pclass	Ticket class (1=1st, 2=2nd, 3=3rd)
4	Name	Passenger's name
5	Sex	Passenger's gender
6	Age	Passenger's age
7	SibSp	Number of siblings/spouses aboard the Titanic
8	Parch	Number of parents/children aboard the Titanic
9	Ticket	Ticket number
10	Fare	Ticket fare

B. *Data Processing and Transformation:*

Critical steps of data cleaning the Titanic data set such as missing values filling, uncovering outliers, dimension reduction are crucial for ensuring the quality and accuracy of the data. Initially, data sets are handled using standard procedures like imputation and deletion to which parameters such as Age, Cabin and Embarked are applied. Anomalies like Passenger ID and Ticket are removed since they they don't provide any useful information and only make the dataset large cumbersome As an example, the statistical anomalies such as the numerical variables like Age and Fare might be flagged by the neutral design and resulting measures would not affect the research hypothesis or model estimates. Categorical variables are standardized, that is they encode one-hot encoding or label encoding, when necessary, in order to ensure all the data, have the same format. The implementation of feature engineering procedures is aimed at making the machine learning models more efficient by including newly developed features or restructuring the already existing ones. Real-valued parameters are standardized or normalized to a common scale for a rapid algorithm training of machine learning. Similarly, to prevent data redundancy and ensuring data sources conformity duplicate rows are detected and removed saving time and resources. In the final count, these data cleaning methods have significantly improved the carefully compiled Titanic dataset and make it suitable for a further machine learning, increasing the accuracy of final conclusions.

C. *Proposed Model:*

In order to enhance the precision of survival forecasts, we plan to implement cutting-edge machine learning techniques in the system of our proposed solution for the titanic survivor problem. We would like to use an array of techniques, for example, logistic regression, random forests, k-nearest neighbors, and maybe more complicated techniques like gradient boosting and neural networks. The integration of these approaches is aimed at obtaining computation-based regression models that better visualize the complex relationships that constitute mortality. In addition, we plan to encompassing the feature engineering method that leads to building advanced model and getting useful analysis from the dataset. Besides that, we plan on applying cross-validation methods as well. This will serve as an insurance policy to make sure that our model will remain capable of capturing the underlying patterns of the data.

To start with, our idea is expected to surpass the mere human inspection and traditional statistical analysis techniques which mostly produce imprecise and inconsistent forecasters in the Titanic survival.

D. *Classification Models:*

- Logistic Regression: Logistic groups two entities by finding the possibility of a certain event, occurrence or observation. For LR, predictors are mapped along with their likelihoods with the help of sigmoid function. A Sigmoid function maps the real value to a range from 0 to 1 using the shape of an S curve. Because of its simplicity and straightforward interpretation, logistic is essential for hepatitis C diagnostics. It helps doctors find vulnerable individuals and act quickly to improve patient care and results.

- Decision Tree: A decision tree is widely accepted and used as means of supervised non- parametric machine learning where recursive partitioning of input variables is applied to allocate data into various subsets. This is a decision tree type Hepatitis C forecasting model which will provide an insight into what factors do have the most significant influence upon the disease course and develop a hierarchical structure as a fragmentation tool for the examples. Thus, by applying the method of horizontal splitting of data set on the features that are more integral to the cause Hepatitis C one can deduce the reasons behind the spread.

- Random Forest: Random forest (RF) is advanced plural cluster learning method The model works by the construction of various decision trees during the training. Here, random samples of the training data are selected, after which a number of random subsets are employed. This technique yields flexibility in the model, thus, significantly minimizing the risk of overfitting on the surface which consequently increases the model strength. RF as a non-parametric and nonlinear machine learning technique is widely used for complex data sets in the fields of finance, healthcare and ecology among others. Besides this, this advantageous algorithm generally comes up with the important feature scoring and sometimes restrictions. Facilitates Relevant Interpretation The RF‟s strength of getting to the variables that are actually important, the high accuracy with which it picks them up and the interpretability that it has as a machine learning model makes it a very handy tool in machine learning research and applications.

- K-Nearest Neighbors: KNN, is among the simplest supervised ML technique where it finds use in both, classification and regression. Concerning Hepatitis C progression, KNN works by locating the K closest data points that corresponds with given input feature and classifying it based on the majority among neighboring K points. Taking this route implies missing an opportunity to find the patterns in the patient data by identifying correlations and simulating unknown cases with the help of the similar ones, when dealing with a non-linear or intricate decision boundaries. Even though it looks simple to implement, KNN's is able to provide better even with larger datasets. Nevertheless, it remains a valuable tool in healthcare for making personalized treatment decisions and assessing risks in the management of Hepatitis C.

- Gaussian Naive Bayes: Gaussian Naive Bayes Rooted in Bayes' theorem is a classification algorithm, especially effective for continuous data and can be widely used in various areas like information classification, medical research, spam filtering etc. Calculating probabilities of class membership based on feature distributions within each class it works. A Gaussian distribution is used to model processes. Despite its simplicity, Gaussian Naive Bayes exhibits strong performance, especially with large data sets, due to its speed and efficiency. Its ease of use and ability to handle high-dimensional data make it a popular choice for real-world segmentation projects.

- SVM: SVM is a powerful supervise algorithm used in distribution, regressive functions. Its main goal is to find the optimal hyper-plane. At its best, it separates the data point from class in high-dimensional space. At its core, the goal of SVM is to find decision boundaries that maximize the mean difference Lessons learned. This decision boundary is defined by a hyperplane in a feature space, where distance b/w the hyperplane, each class, and neighboring data points, is known support vectors. The main power of the SVM is its ability to do the high-quality data and be efficient Overloading is prevented, especially where quantity exceeds quantity of the specimens. Furthermore, it is not significantly affected by the presence of redundant features. Its focus on supporting vectors, which are critical data points that influence decision making boundaries.

IV. Results

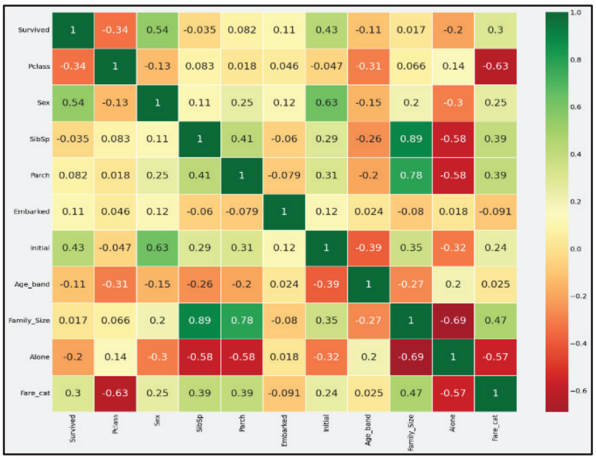


Fig. 1. Correlation Matrix Generated

The correlation heatmap displays the connections between the number of elements in the dataset and the features of the dataset. In a nutshell, it is a device to evaluate how variables are connected with each other. Through their very nature, a Pareto chart will help you get rid of the most unimportant aspects of your project. Hence, you will be able to use it to immediately spot the prevalent features and connections among the components that were not discarded. The redness of the polygon describes the strength of the correlation between them. Darker shades of blues and yellows imply strong correlations, while lighter tints show weak correlation parameters.

This heatmap enables you to look for the most influencing features which are related to close to each other. With your inputs being shaped by this know-how, you will also be able to identify the potential features, type of the model you intend to use and maybe other additional studies. Firstly, it assists in checking for possible multicollinearity issues that occur when two or more variables have a strong similar correlation and may render models less useful or unable to be understood.

A. Classification Report

Classifier metrics of a certain project are summarized in this table to simplify its understanding. The random forest, logistic regression, SVM (RBF) [4', SVM (Linear) [5], KNN [6], Naive Bayes [7], and Decision Tree [8] classification algorithms that are used to assess performance include the Random Forest [2]. The classifier exactness that is given as a measure for the right prediction percents, ranges from 0.75 to 0.83. Recall, which measures the rate of positive results among those who are actually infected, lies between 65 and 76 %. F1-Score- the mean score of recall and precision, is in the range of (0.72 - 0.76) Support that ride the waves of actual cases that exist in each class is identical—103—and that is true for all classifiers. It is fascinating, though, that accuracy, which calculates the accuracy scores overall, ranges from 80.59% to 83.58%. Essentially, making this assessment allows the tuning and selection of the model that will be further employed at the machine learning pipeline in a proportional way.

TABLE III. Classifier Metrics Results

Classifier	**Precision**	**Recall**	**F1-Score**	**Support**	**Accuracy**
Logistic Regression	0.79	0.69	0.74	103	81.34
SVM(RBF)	0.83	0.70	0.76	103	83.58
SVM(Linear)	0.78	0.72	0.75	103	81.71
Random Forest	0.82	0.67	0.74	103	81.34
KNN	0.79	0.69	0.74	103	81.34
Naive Bayes	0.75	0.76	0.75	103	81.34
Decision Tree	0.81	0.65	0.72	103	80.59

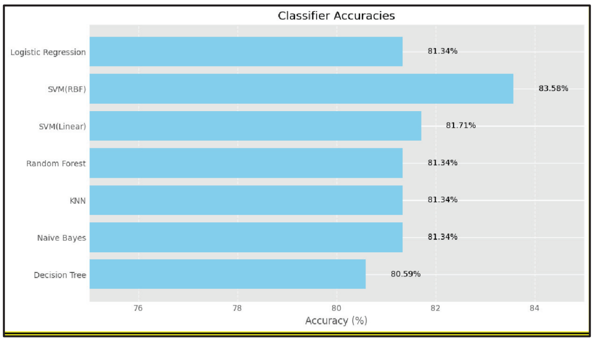


Fig. 2. Classifier Accuracies Generated

The plot's accuracy numbers show how different classifiers performed in a given categorization task. The ability of each classifier—Logistic Regression, SVM (RBF), SVM (Linear), Random Forest, KNN, Naive Bayes, and Decision Tree—to accurately classify examples into their respective classes has been the basis for evaluation.

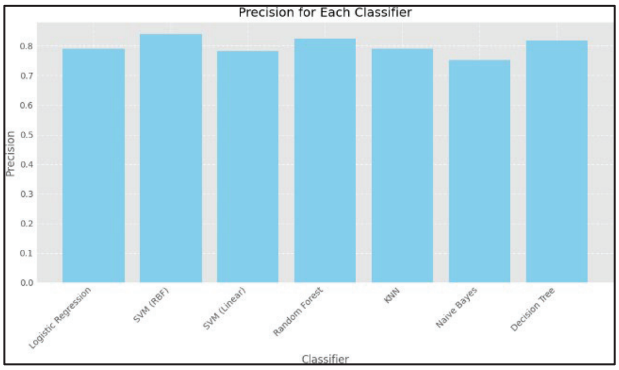


Fig. 3. Precision of each classifier Model

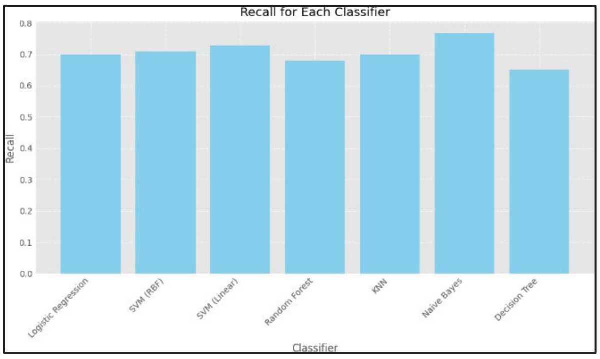


Fig. 4. Recall of each classifier Model

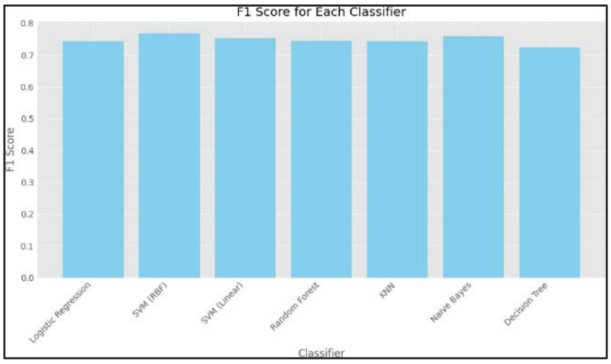


Fig. 5. F1-Score of each classifier Model

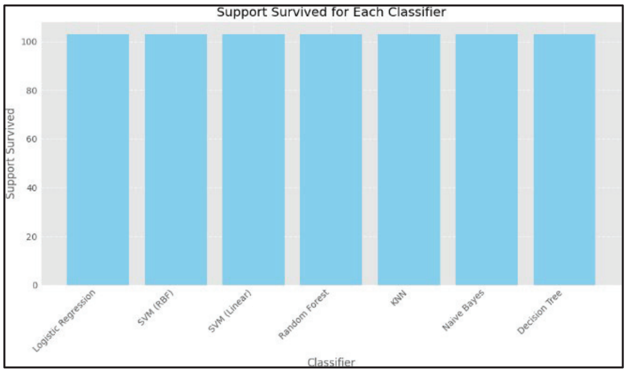


Fig. 6. Support of each classifier Model

SVM with a radial basis function (RBF) kernel outperformed rest of these classifiers in terms of accuracy, coming in second at 83.58%, and third with a linear kernel at 81.71%. The classifiers for logistic regression, random forest, KNN, and naive bayes showed comparable accuracy of 81.34%, however the decision tree fared marginally worse with an accuracy of 80.59%

The Table III comprises of various performance indicators are displayed horizontally in a bar plot across several classifiers. Recall emphasises the capacity to successfully collect positive examples, whereas precision shows the accuracy of positive predictions. The F1- score provides a thorough evaluation metric by striking a balance between recall and precision. Support shows how real events are distributed between classes. The best model for the classification problem can be chosen with the help of this visual representation, which enables a comparison of classifier performance. Decision-making is guided by the interpretation of each measure, guaranteeing alignment with project needs and objectives.

B. *Error Rates*

The table underneath lists the error limits of the classes tested on the dataset. These error rates are essential metrics for penetrating into the effectiveness and accuracy of machine learning algorithms.

Consequently, Mean Absolute Error (MAE) is employed to represent the mean absolute difference between forecasted and the real observations. The smaller MAE value represents the actual and projected plot lines more faithfully indicating that the model is more reliable.

Furthermore, the Non-linear Epsilon-Insensitive Support Vector Machine (SVR) loss function is also considered taking into consideration the output, which is the difference between the expected and actual values. There are no small errors and no permanent mistakes in life. Little errors will be severely punished while big mistakes will be forgiven. Thus, a better model is found by decreasing MSE, which means that the mean squared error is low.

TABLE IV. ERROR RATE GENERATED

Classifier	MAE	MSE	RMSE
Logistic Regression	0.18	0.18	0.43
SVM(RBF)	0.16	0.16	0.40
SVM(Linear)	0.18	0.18	0.42
Random Forest	0.18	0.18	0.43
KNN	0.18	0.18	0.43
Naive Bayes	0.18	0.18	0.43
Decision Tree	0.20	0.20	0.44

Fig. 7. Mean Absolute Error for different Classifier

Finally, the RMSE will be used, where it is the square root of MSE. It is more tangible for the MSE, hence being measured in the same units as the target variable. Lower (RMSE) values are showing as better prediction accuracy because it because of less difference /variances expected and obtained values.

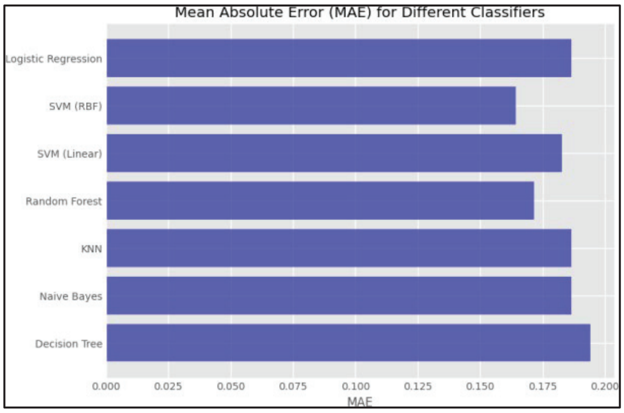


Fig. 8. Mean Absolute Error for different Classifier

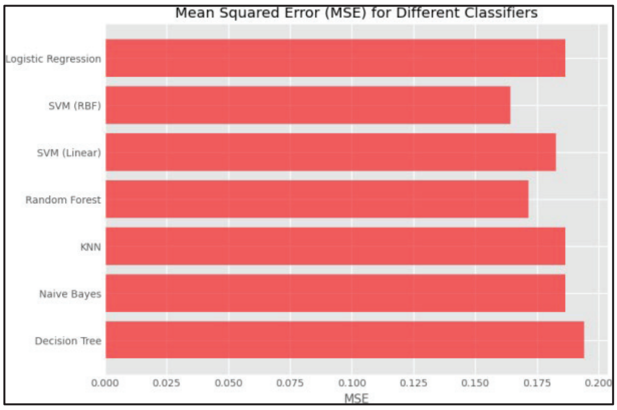


Fig. 9. Root Mean Squared Error for different Classifier

By and large, these error rate statistics may consequently characterize the extent to which the both classifiers have excelled in predicting outcomes of the data tested. Reducing prediction gates across the board are a strong aspect in a model of good quality and desired results.

The error rate graphs can be plotted and studied for refining findings obtained through, e.g., accuracy. These techniques assist in determining whether to chose the most appropriate and accurate algorithms. The SVM technique as numerous rates for all algorithms, therefore its decision boundary to classify data better is easy to predict.

V. Conclusion

Overall, diverse machine learning techniques, including the Naive Bayes, CNN, tree, K-NN, logistic regression and random forest, were implemented for the success of predicting the survivors among the Titanic passengers. The performance rates were chosen as protocols of the method such as accuracy, precision, recall, and F1-score.

The results demonstrated that SVM(RBF) performed better than with other prospects as for modeling survival results, getting 83.58 degrees of accuracy. The error gauge now goes up in this case on account of the point that each algorithm efficiency could quite be influenced by the frame of the dataset and the preprocessing methods applied.

References

[1] Adi Nugroho and Bistok Hasiholan Simanjuntak, "ARMA (Autoregressive Moving Average) Model for Prediction of Rainfall in Regency of Semarang-Central Java-Republic of Indonesia", *International Journal of Computer Science Issues (IJCSI)*, vol. 11, no. 3, 2014.

[2] S. Meenakshi Sundaram and M. Lakshmi, "Rainfall Prediction using Seasonal Auto Regressive Integrated Moving Average model", Indian Journal Of Research, 2014.

[3] K. Pu, X. Liu, X. Sun and S. Li, "Error Analysis of Rainfall Inversion Based on Commercial Microwave Links With A–R Relationship Considering the Rainfall Features", IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-12, 2023.

[4] M. A. Shaik, Y. Sahithi, M. Nishitha, R. Reethika, K. Sumanth Teja and P. Reddy, "Comparative Analysis of Emotion Classification using TF-IDF Vector," 2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), Erode, India, 2023, pp. 442-447, doi: 10.1109/ICSSAS57918.2023.10331897.

[5] Wei, Y. Liu, H. Song and Z. Lu, "A Method of Rainfall Detection From X-Band Marine Radar Image Based on the Principal Component Feature Extracted", IEEE Geoscience and Remote Sensing Letters, vol. 20, pp. 1-5, 2023.

[6] Mohammed Ali Shaik and Dhanraj Verma, (2022), "Prediction of Heart Disease using Swarm Intelligence based Machine Learning Algorithms", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020025-1–020025-9; https://doi.org/10.1063/5.0081719, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020025-1 to 020025-9

[7] C. Guo et al., "Correction of Sea Surface Wind Speed Based on SAR Rainfall Grade Classification Using Convolutional Neural Network", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 321-328, 2023.

[8] M. A. Shaik, R. Sreeja, S. Zainab, P. S. Sowmya, T. Akshay and S. Sindhu, "Improving Accuracy of Heart Disease Prediction through Machine Learning Algorithms", 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 41-46, doi: 10.1109/ICIDCA56705.2023.10100244.

[9] G. L. Vara Prasad, B. Ravi Teja, S. Govathoti and S. Dhanikonda, "Leveraging ARMA and ARMAX Time-Series Forecasting Models for Rainfall Prediction", ICACCS, pp. 353-357, 2023.

[10] Mohammed Ali Shaik, M. Varshith, S. SriVyshnavi, N. Sanjana and R. Sujith, "Laptop Price Prediction using Machine Learning Algorithms", 2022 International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS), Nagpur, India, 2022, pp. 226-231, doi: 10.1109/ICETEMS56252.2022.10093357.

[11] Y. J. N. Kumar, K. Shirisha, N. Niveditha, M. Swapna, P. Sagar and I. Prashanth, "Utilizing Machine Learning Algorithms for Rainfall Analysis", ICSMDI, pp. 357-362, 2023.

[12] Mohammed Ali Shaik, Praveen Pappula, T Sampath Kumar, "Predicting Hypothyroid Disease using Ensemble Models through Machine Learning Approach", European Journal of Molecular & Clinical Medicine, 2022, Volume 9, Issue 7, Pages 6738-6745. https://ejmcm.com/article_21010.html

[13] S. Majumdar, S. K. Biswas, B. Purkayastha and S. Sanyal, "Rainfall Forecasting for Silchar City using Stacked- LSTM", IEMECON, pp. 1-5, 2023.

[14] M. A. Shaik, S. k. Koppula, M. Rafiuddin and B. S. Preethi, (2022), "COVID-19 Detector Using Deep Learning", International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2022, pp. 443-449, doi: 10.1109/ICAAIC53929.2022.9792694.

[15] L. Wang, H. Chen, R. Cifelli and Z. Li, "Improving Surface Rainfall Mapping in Complex Terrain Regions Through Lowering the Minimum Scan Elevation Angle of Operational Weather Radar", IEEE Geoscience and Remote Sensing Letters, vol. 20, pp. 1-5, 2023.

[16] Mohammed Ali Shaik and Dhanraj Verma, (2022), "Predicting Present Day Mobile Phone Sales using Time Series based Hybrid Prediction Model", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020073-1–020073-9; https://doi.org/10.1063/5.0081722, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020073-1 to 020073-9

[17] D. V. Rayudu and J. F. Roseline, "Accurate Weather Forecasting for Rainfall Prediction Using Artificial Neural Network Compared with Deep Learning Neural Network", ICECONF, pp. 1-6, 2023.

[18] Mohammed Ali Shaik, Geetha Manoharan, B Prashanth, NuneAkhil, Anumandla Akash and Thudi Raja Shekhar Reddy, (2022), "Prediction of Crop Yield using Machine Learning", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020072-1–020072-8; https://doi.org/10.1063/5.0081726, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020072-1 to 020072-8

[19] J. Byun, C. Jun, J. Kim, J. Cha and R. Narimani, "Deep Learning-Based Rainfall Prediction Using Cloud Image Analysis", IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-11, 2023.

[20] Mohammed Ali Shaik and Dhanraj Verma, (2020), Deep learning time series to forecast COVID-19 active cases in INDIA: A comparative study, 2020 IOP Conf. Ser.:Mater.Sci.Eng. 981 022041, doi.org/10.1088/1757-899X/981/2/022041

[21] M. Djibo et al., "Commercial microwave link networks for rainfall monitoring in Burkina Faso: First results from a dense network in Ouagadougou", MNE3SD, pp. 1-7, 2023.

[22] Mohammed Ali Shaik, “Time Series Forecasting using Vector quantization”, International Journal of Advanced Science and Technology (IJAST), ISSN:2005-4238,Volume-29,Issue-4 (2020), Pp.169-175.

[23] A. K. Bitto, M. A. Rubi, M. H. I. Bijoy, S. D. Shuvo, A. Das and A. Chowdhury, "KGR-Rainfall: Temperature-Based Rainfall Prediction in Bangladesh with Novel KGR Stacking Ensemble", 2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1), pp. 1-6, 2023.

[24] Mohammed Ali Shaik, “A Survey on Text Classification methods through Machine Learning Methods”, International Journal of Control and Automation (IJCA), ISSN:2005-4297,Volume-12,Issue-6 (2019), Pp.390-396.

[25] J. Zombori, J. Lukács and R. Horváth, "Determination of Rainfall Probability Using Response Surface Method", SACI, pp. 000359-000362, 2023.