Machine Learning Prediction of Consumer Travel Insurance Purchase Behavior

Conference Paper · October 2022

DOI: 10.1109/ICCCNT54827.2022.9984470

4 authors:

Maksuda Akter Rubi
Daffodil International University
16 PUBLICATIONS **71** CITATIONS
SEE PROFILE

Md. Hasan Imam Bijoy
Daffodil International University
58 PUBLICATIONS **183** CITATIONS
SEE PROFILE

Shanjida Chowdhury
Southeast University (Bangladesh)
33 PUBLICATIONS **140** CITATIONS
SEE PROFILE

Khairul Islam
Norwegian University of Science and Technology
6 PUBLICATIONS **13** CITATIONS
SEE PROFILE

Machine Learning Prediction of Consumer Travel Insurance Purchase Behavior

Maksuda Akter Rubi
Department of General Educational Development
Daffodil International University
Dhaka-1341, Bangladesh
rubi.ged@diu.edu.bd

Md. Hasan Imam Bijoy
Department of Computer Science and Engineering
Daffodil International University
Dhaka-1341, Bangladesh
hasan15-11743@diu.edu.bd

Chanjida Chowdhury
Department of General Educational Development
Daffodil International University
Dhaka-1341, Bangladesh
shan_chydiu.ged@daffodilvarsity.edu.bd

Md. Khairul Islam
Department of Computer Science and Engineering
Daffodil International University
Dhaka-1341, Bangladesh
khairul15-9838@diu.edu.bd

***Abstract*—Travel insurance covers the expenses and losses related to travelling which is helpful security for travelers domestically or abroad. The goal of the study is to see whether a consumer is interested in buying travel insurance or not based on their information of age, employment type, graduation information, annual income, family size, the existence of chronic diseases, frequent fly or not and ever travel or not which are considered as independent variables. The dataset is collected from an internet website named, Kaggle where some traveler's information is gathered through a survey of a tourist group. There were 3980 rows of data (Number of Travelers) available, with 9 columns. 10 different types of classification algorithms named logistic regression, KNearest Neighbors (KNN), Gaussian Naive Bayes (Gaussian Naive Bayes), Multinomial Naive Bayes (MNB), Decision Tree Classifier (DT), Random Forest (RF), Support Vector Clustering (SVC), eXtreme Gradient Boosting (XGBoost), Stochastic Gradient Descent (SGD) and Gradient Boosting Classifier (GBC) are performed to select a model with the best accuracy. Among all the algorithms Random Forest, Decision Tree Classifier, and Stochastic Gradient Descent provide the highest accuracy of 88% in predicting whether a consumer would decide to purchase travel insurance or not. The suggested model would be a better choice for insurance companies to make decisions on how to target their desired person and save money with the best profit.**

Keywords—Insurance Prediction, Traveler's Behavior, Travel Insurance, Purchase Insurance, Machine Learning.

I. INTRODUCTION

Insurance is a highly competitive industry. Insurance firms compete for customers by offering a comprehensive travel insurance package and favorable contract terms. Currently, insurance companies are implementing operations aimed at generating trust, building a brand, and providing value - added services, that entails paying more emphasis to the service's "packaging" than that of the service itself. Insurance services are "packaged" in a variety of ways. This comprises the employees, the design of the outlets, the range of services, the agility of customer service, innovations (Internet, call centers), and the reputation of the insurance company or insurance package customized for the consumer. Knowledge of not only marketing strategies, but also client preferences and incentives for purchasing services such as insurance and travel insurance is crucial for insurance companies.

Tourism has grown to become one of the most vital areas of the world economy [1]. Tourism is one of the largest and fastest growing sectors in the world [2], [3]. As a worldwide phenomenon, it is a vital part of the service industry and has a significant effect on the economy [4], [5], [6]. Tourism is a major driver of the country's socio-economic growth. The tourism business depends on the decision of tourists to travel to conserve or improve its current condition. Numerous travel insurance policies can influence tourists' travel decisions, boosting the country's economic growth. A comprehensive travel insurance coverage is meant to cover risks and financial losses that may occur while you are travelling, whether within the country or abroad, such as medical expenditures, lost luggage, or trip cancellation costs. Individual clients, as well as service companies and middlemen, can get travel insurance from insurance companies. Thus, it is critical to comprehend and possess sufficient knowledge regarding the factors that influence the travel insurance purchasing behavior of customers. Using Machine Learning Model this study aims to see whether a consumer is interested in buying travel insurance or not based on their demographic information.

The rest of the paper is organized in a logical manner. The second section demonstrates a literature review. Section III introduces the work's methodology and materials. The results and discussions are shown in this section IV. Finally, Section V brings this research study to a conclusion.

II. LITERATURE REVIEW

Insurance is basically a means of spreading personal risk by giving protection against the possibility of an occurrence. Travel insurance is a sort of insurance coverage that is designed to cover many types of travelers, including those travelling for business or visiting family, as well as students, throughout their domestic and international travels [7]. External elements that customers experience (marketing and environment), according to Kotler [8], trigger consumer purchase decisions.

According to the findings of Wang et al. [9] 2019, people prioritize insurance benefits, such as insurance content and coverage value, when acquiring international travel insurance. Travel insurance is available from several insurers. Health insurance (KL) covers doctor visits and hospital stays. This insurance can cover chronic conditions. Personal liability insurance (OC) protects against claims made against the policy bearer. Sports gear or luggage insurance (BG).

Insurance for unlucky accidents (NNW), insurance for death (NWS), insurance for physical injury (NWU), insurance for medical first aid (KPPM), and insurance for rehabilitation costs (KPPM) (KRH). Insurance for trip cancellation, flight cancellation, and hotel cancellation. Insurance buyers can seek suitable support or financial reimbursement and transfer their risk during travel to the insurance business, thereby participating in the purchases of travel insurance to decrease their personal expenses and effort in these circumstances.

Predicted fraud [10] in property insurance based on nine machine learning algorithms and analyzed the strength and weakness of every machine learning models. They collected data from a major Brazilian insurance company from 2009 to 2018 with different types of data such as: income, contract time, number of policies claim etc. Algorithms are Logistic Regression, Penalized Logistic Regression, Naive Bayes, KNN, Polynomial Kernel SVM, Gaussian Kernel SVM, Deep Neural Network, Random Forest, GBM with Evaluation parameter are Accuracy, Precision, Recall, F1 Score, Kappa (Cohen's Kappa coefficient), MCC (Matthew's correlation coefficient). Overall, the Random Forest showed the best accuracy and the Naive Bayes showed the worse accuracy 84.56% and 71.16% respectively and a moderate overall performance showed by logistic regression comparison with other models. Working with a large amount of data with method of imbalance classification is the future work of this study.

Analyzed two types of insurance [11] data and predict Insurance claim and Claim status based on two case studies are health insurance data and travel insurance data. Collecting data from Kaggle with a number of 63626 and analyzed with eight machine learning supervised algorithms. Used machine learning algorithms with some feature selection techniques such as Chi-Squared Test, Recursive Feature Elimination, and Tree Based Feature Importance. After applying feature selection technique, the accuracy increased. Overall, for both types of datasets the Random Forest Classifier gave the best accuracy among the all algorithms. To find the problem of huge data imbalance using resampling of the dataset this the future scope of this study.

Before starting the auction, predicted probable-end price of online auction with some machine learning algorithms to solve the price prediction problems [12]. Collected data from eBay over the 2 months period based on different category with 1700 instances. Used different types of machine learning algorithm such as Linear Regression, Polynomial Regression, CART, Decision Tree, Neural Networks with three way of comparison Regression, Multi-Class Classification, Multiple Binary Classification tasks. Trained the models with 1300 instances and also tested with 400 instances. The best accuracy was given by Neural Networks 96%.

Wong et al. [13] study on machine learning actuarial science and also review the paper of ratemaking and reserving based on almost 120 thesis work. GLM (Generalized Linear Model), generalized additive models, random forests, gradient boosted machines, support vector machines and neural networks are the most common models in the field of actuarial science. In terms of time period from 2000 to 2020(August), the publication of machine learning increased last two years in both Pricing and Reserving. Most of the thesis work of different conference/journals is pricing prediction than Reserving. Finally, in terms of publication by models, Decision Tree gave the best result for structured problems and neural networks gave best output for unstructured problems. In addition, XGBoost model is the most famous for pricing framework and neural networks for reserving framework.

Predicted claim occurrence non-life (auto) insurance with various models of machine learning and compare the performance of those model [14]. In result, decrease cost, improve business and optimize business strategies in insurance business industry. Collected data with 1,488,028 instances from a Brazilian insurance company named Porto Seguro. To get the more accurate prediction result, applied various machine learning algorithms such as: Logistic Regression, XGBoost, Random Forest, Decision Tree, Naïve Bayes, and K-NN with evaluation parameter of Confusion Matrix, Kappa Statistics, Sensitivity and Specificity, Precision and Recall, F-Measure. The Random Forest gave the best accuracy 87% among the algorithms. The comparison of this dataset with ML algorithm and Deep learning algorithm is one of the future tasks of this thesis work and another is to analysis another insurance dataset whether the Random Forest will give the best accuracy or not.

III. Proposed Methodology

The objective of this study is to create a model that can accurately predict whether or not a tourist would purchase travel insurance. To achieve our aim, we must go through a number of processes, including data collection, data preparation, model implementation, and so on. Data training is then carried out with the use of algorithms. Furthermore, the test data is used to assess the system's performance. The study's workflow is depicted in Fig. 1.

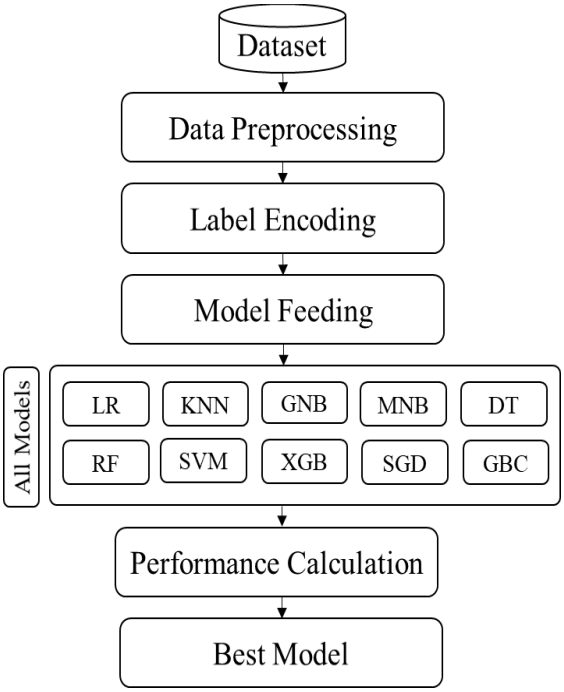

Fig. 1. Workflow Diagram.

A. Dataset

To create maximum accuracy, the expert system requires a massive amount of data. We obtained our data from Kaggle, an internet website. The dataset contains some core traveller information that was gathered through a survey of a tourist group. There were 3980 rows of data (Number of Travellers) available, with 9 columns. Table I shows a visualization of the data obtained.

TABLE I. DATASET.

Independent Variable	Dependent Variable	Class	Training Dataset	Testing Dataset
Age, EmploymentType, GraduateOrNot, AnnualIncome, FamilyMembers, ChronicDiseases, FrequentFlyer, EverTravelledAbroad	TravelInsurance	Yes (1)	3179	795
		No (0)		

The correlation matrix shows how all variables are connected to one another and measures the most frequently utilized variable in model feeding. The correlation matrix of our working dataset is shown in Fig. 2.

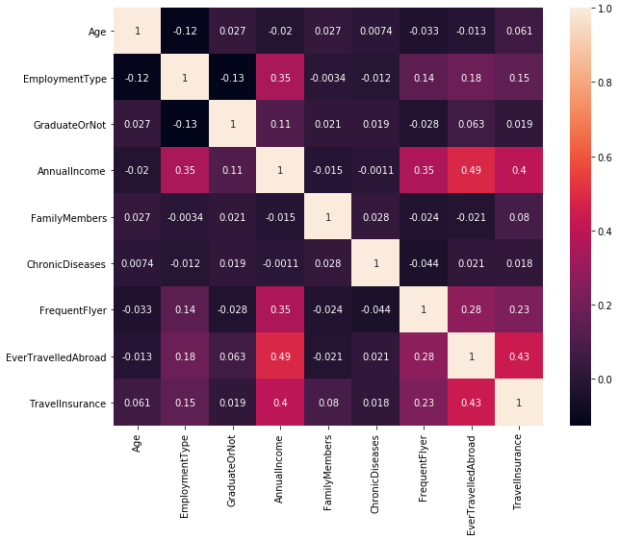

Fig. 2. Coorelation Matrix of our Dataset

B. Data Preprocessing with Label Encoder

A data preprocessing or cleaning setup is required in a machine learning project. To prepare the dataset for model fitting, we check for missing values, nan values, categorical values, and a variety of other factors. Most comment classes for categorical data and mean values for numerical data are used to fill in a missing value. After that, we use a label encoder to transform our category data into numerical values in a machine-readable format which is present in Table II and Table III.

TABLE II. SAMPLE RAW DATASET BEFORE PRE-PROCESS.

AG	ET	GN	AI	FM	CD	FF	ETA	Class
31	Govt.	Yes	400000	6	1	No	No	No
34	Private	Yes	500000	4	1	No	No	Yes
28	Private	Yes	700000	8	1	Yes	Yes	Yes
36	Govt.	Yes	750000	3	0	Yes	No	No

TABLE III. SAMPLE DATASET AFTER PRE-PROCESS.

AG	ET	GN	AI	FM	CD	FF	ETA	Class
31	0.	1	400000	6	1	0	0	0
34	1	1	500000	4	1	0	0	1
28	1	1	700000	8	1	1	1	1
36	0	1	750000	3	0	1	0	0

C. Model Feeding

There are a variety of machine learning techniques used to construct the model that predicts whether a tourist would purchase travel insurance. To model feeding, we first split our dataset into two halves, with 80% being used for training and 20% being used for testing. Then we try to cover as many classification algorithms as possible, and we're able to apply the most common 10 supervised algorithms to our dataset successfully. Those algorithms are Logistic Regression (LR) [10], KNearest Neighbors (KNN) [11], Gaussian Naive Bayes (GNB) [12], Multinomial Naive Bayes (MNB) [13], Decision Tree Classifier (DT) [14], Random Forest (RF) [15], Support Vector Machine-Clustering (SVC) [16], XGBoost (XGB) [17], Stochastic Gradient Descent (SGD) [18], Gradient Boosting Classifier (GBC) [19].

D. Performance Calculation

To measure the performance of the applied algorithm, we are considering accuracy score, confusion matrix (TP, FP, FN, TN), precision value, recall and F-1 score for selecting the best model for our proposed system and those factors presenting in "(1-4)".

$$Accuracy = \frac{True\ Purchase + True\ Not\ Purchase}{Total\ No.\ of\ Sample} \quad (1)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

IV. RESULTS AND DISCUSSIONS

The final goal of a machine learning project is to see how much better the applied model performs. The classification algorithm can certainly deliver precise results based on class and provide superior accuracy. Based on our 8 independent variables models, it seemed as if we were witnessing a computer correctly anticipate the precise outcome. We use 10 different types of classification algorithms to train our data and evaluate model performance. We got our accuracy from our model as follows: Random Forest, Decision Tree Classifier, and Stochastic Gradient Descent provide the highest accuracy of 88%, then KNearest Neighbors and Gradient Boosting Classifier provide the second highest accuracy of 83% and 82%, respectively, and the rest of the algorithm shows a satisfactory level of medium and poor accuracy. Fig. 3 shows a graphical view of ten applied model accuracy.

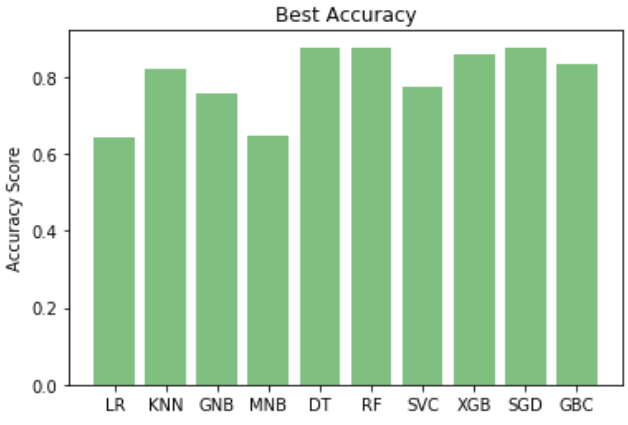

Fig. 3. Accuracy Graph of Fitted Algorithms

We can observe that each algorithm has a distinct level of accuracy, with three algorithms doing the best, two performing as expected, and the other algorithms performing moderately and poorly. Then we look at the confusion matrix,

which is a statistical categorization. We know that perhaps the confusion matrix is also known as the error matrix stated in [21], [22], [23], and it displays the four values in Table IV and Fig. 4.

TABLE IV. CONFUSION MATRIX.

Algorithms	TP	FP	FN	TN
Logistic Regression	511	0	284	0
KNearest Neigbours	460	51	92	192
Gussian Naïve Bayes	459	52	141	143
Multinomianal Naïve Bayes	345	166	115	169
Decciosn Tree Classifier	481	30	69	215
Random Forest	482	29	68	216
Support Vectore Machine	478	33	147	137
eXtreme Greadient Boost	486	25	88	196
Stochastic Greadient Descent	481	30	69	215
Greadient Boosting Classifier	500	11	122	162

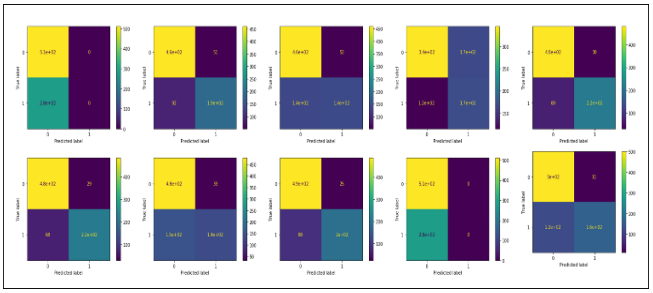

Fig. 4. Confusion Matrices for All Model.

There are some very further estimates to consider while selecting the optimal model for our research. To achieve excellent precision, we double-checked precision, recall, and F1 score and backing. We've had some fantastic results from these calculations. There were classes 0/1 representing the traveller pursuing insurance perspective as they purchase or not which means “0” is “Not Purchase Travel Insurance ” and “1” is used for “Purchase Travel Insurance”. The actual classification report for each method and dependent variable as our binary class is showing in Table V.

TABLE V. CLASSIFICATION REPORT.

Algorithms	Class	Precision	Recall	F1 Score	Accuracy
LR	0	0.64	1.00	0.78	0.64
	1	0.00	0.00	0.00	
KNN	0	0.83	0.90	0.87	0.82
	1	0.79	0.68	0.73	
GNB	0	0.77	0.90	0.83	0.76
	1	0.73	0.50	0.60	
MNB	0	0.75	0.68	0.71	0.65
	1	0.50	0.60	0.55	
DT	0	0.87	0.94	0.91	**0.88**
	1	0.88	0.76	0.81	
RF	0	0.88	0.94	0.91	**0.88**
	1	0.88	0.76	0.82	
SVC	0	0.76	0.94	0.84	0.77
	1	0.81	0.48	0.60	
XGB	0	0.85	0.95	0.90	0.86
	1	0.89	0.69	0.78	
SGD	0	0.87	0.94	0.91	**0.88**
	1	0.88	0.76	0.81	
GBC	0	0.80	0.98	0.88	0.83
	1	0.85	0.83	0.82	

Now, A Receiver Operating Characteristic curve (ROC curve) is a graph that depicts how well a classification model works at various levels of classification. This curve displays two parameters: The True Positive Rate (TPR) is a metric for determining how often something is true. The number of false positives in a given period of time. With Area Under the ROC Curve (AUC) [20], we display the ROC for our top three fitted models (RF, DT, and SGD).

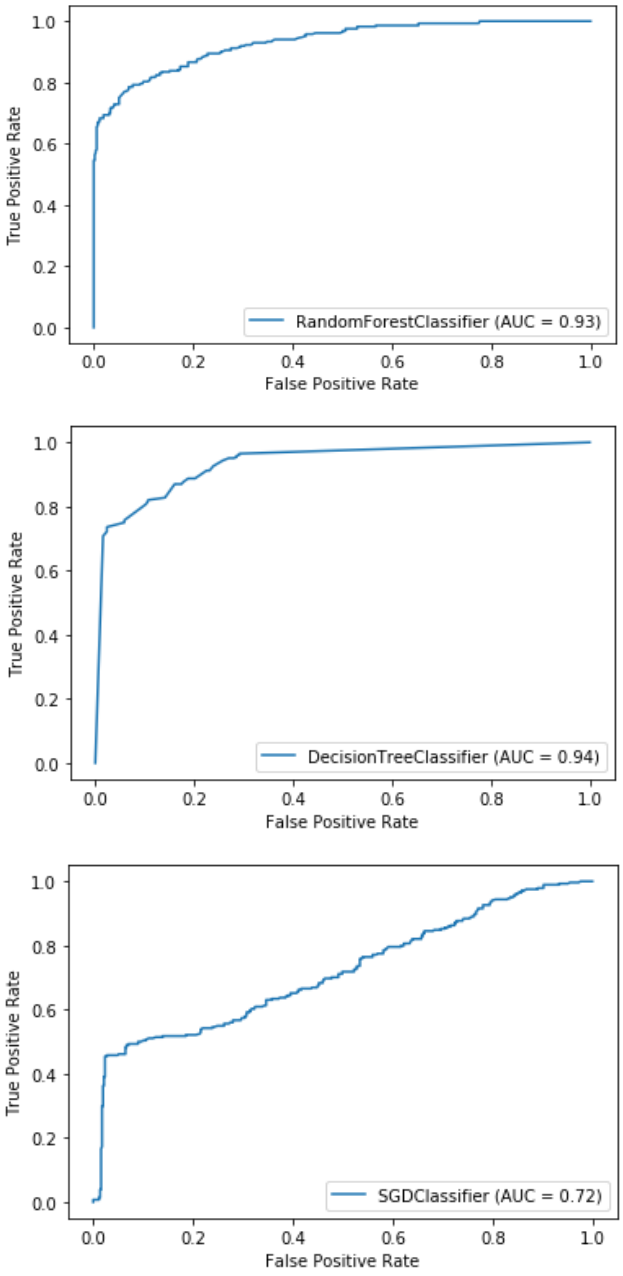

Fig. 5. ROC for Best Fitted Model (RF, DT, SGD)

V. CONCLUSION

The results of the experimental part show that the used algorithm performs here with the best accuracy. The aim was to see whether a consumer is interested in buying travel insurance or not based on their information of age, employment type, graduation information, annual income, family size, the existence of chronic diseases, frequent fly or not and ever travel or not. In this case, it would be pretty useful to build a model that will classify in the best possible way the people who are interested in buying travel insurance and those

who don't. One possible advantage would be the targeted advertisement of insurance to people who are actually interested. A targeted advertisement would save money from an advertisement campaign in order to invest in something else. Travel insurance data wherein some customers opted for the insurance while some others did not. So, this implemented model is business-oriented for insurance companies to make decisions on how to target their desired person and save money with the best profit.

REFERENCES

[1] Lopes, Sérgio Dominique Ferreira, Antonio Rial Boubeta, and Jesús Varela Mallou. "Post hoc tourist segmentation with conjoint and cluster analysis." *PASOS. Revista de Turismo y Patrimonio Cultural* 7, no. 3 (2009): 491-501.

[2] Ninemeier, Jack D., and Joe Perdue. *Hospitality operations: careers in the world's greatest industry*. Recording for the Blind & Dyslexic, 2006.

[3] Cooper, Chris, and Colin Michael Hall. *Contemporary tourism: An international approach*. Routledge, 2008.

[4] Ninemeier, Jack D., and Joe Perdue. *Discovering hospitality and tourism: The world's greatest industry*. Pearson Prentice Hall, 2008.

[5] KAY, H. K. (2003), Selling Tourism, New York, Delmar Learning.

[6] Koc, Erdogan. "The role of family members in the family holiday purchase decision-making process." *International journal of hospitality & tourism administration* 5, no. 2 (2004): 85-102.

[7] Gee, Chuck Y, Dexter J. L Choy, and James C Makens. 1996. *The Travel Industry*. 3rd ed. London: Van Nostrand Reinhold.

[8] Kotler, Philip, and Ronald E. Turner. *Marketing management: Analysis, planning, implementation, and control*. Vol. 9. Upper Saddle River, NJ: Prentice hall, 1997.

[9] Wang, Shao-Ping, Li-Chun Chen, Miao-Sheng Chen, and Mou-Jian Li. "Purchasing Factors for Travel Insurance by Asian Consumers." *International Journal of Human Resource Studies* 9, no. 1 (2019): 311-329.

[10] Kleinbaum, David G., and Mitchel Klein. "Introduction to logistic regression." In *Logistic regression*, pp. 1-39. Springer, New York, NY, 2010.

[11] Zhang, Zhongheng. "Introduction to machine learning: k-nearest neighbors." *Annals of translational medicine* 4, no. 11 (2016).

[12] Ontivero-Ortega, Marlis, Agustin Lage-Castellanos, Giancarlo Valente, Rainer Goebel, and Mitchell Valdes-Sosa. "Fast Gaussian Naïve Bayes for searchlight classification analysis." *Neuroimage* 163 (2017): 471-479.

[13] Raschka, Sebastian. "Naive bayes and text classification i-introduction and theory." *arXiv preprint arXiv:1410.5329* (2014).

[14] Myles, Anthony J., Robert N. Feudale, Yang Liu, Nathaniel A. Woody, and Steven D. Brown. "An introduction to decision tree modeling." *Journal of Chemometrics: A Journal of the Chemometrics Society* 18, no. 6 (2004): 275-285.

[15] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." *R news* 2, no. 3 (2002): 18-22.

[16] Boswell, Dustin. "Introduction to support vector machines." *Departement of Computer Science and Engineering University of California San Diego* (2002).

[17] Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, and Hyunsu Cho. "Xgboost: extreme gradient boosting." *R package version 0.4-2* 1, no. 4 (2015): 1-4.

[18] Bottou, Léon. "Large-scale machine learning with stochastic gradient descent." In *Proceedings of COMPSTAT'2010*, pp. 177-186. Physica-Verlag HD, 2010.

[19] Natekin, Alexey, and Alois Knoll. "Gradient boosting machines, a tutorial." *Frontiers in neurorobotics* 7 (2013): 21.

[20] Fawcett, Tom. "An introduction to ROC analysis." *Pattern recognition letters* 27, no. 8 (2006): 861-874.

[21] Md. Hasan Imam Bijoy, Masud Rabbani, Md. Ishrak Islam Zarif, Md. Mahbubur Rahman, Md. Rakibul Hasan, and Tridip Bhowmik , "A stupendous statistics on the pandemic impact on all sectors of Bangladesh", AIP Conference Proceedings 2451, 020006 (2022) https://doi.org/10.1063/5.0095349

[22] Bijoy, Md Hasan Imam, Sumiya Alam Akhi, Md Ali Ashraf Nayeem, Md Mahbubur Rahman, and Md Jueal Mia. "Prediction of internet user satisfaction levels in Bangladesh using data mining and analysis of influential factors." *Bulletin of Electrical Engineering and Informatics* 11, no. 2 (2022): 926-935.

[23] Bijoy, Md Hasan Imam, Maksuda Akter Rubi, Shanjida Chowdury, and Siddiqur Rahman. "Growth and Trustworthiness of Online Shopping & Business during the Pandemic: A Case Study on Bangladesh." *Journal of Sales, Service and Marketing Research (e-ISSN: 2582-7804)* (2022): 6-16.