iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** V **Month of publication:** May 2021

DOI: https://doi.org/10.22214/ijraset.2021.34416

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue V May 2021- Available at www.ijraset.com

Comparative Analysis on Stroke Prediction using various Supervised Machine Learning Techniques

Shobhandeb Paul[1], Santanu Saha[2], Suvasish Paul[3], Souradeep Kundu[4], Taniya Mitra[5] Avali Banerjee[6]
[1, 2, 3, 4, 5, 6]*Department of Electronics and Communication Engineering, Guru Nanak Institute of Technology, West Bengal, India, Pin: 700110*

Abstract: ***In a recent study by WHO, it showed that, a human gets trapped in a serious medical emergency known as stroke, that is caused by the sudden interrupt of blood supply to the brain, which is also the death cause of 11% population globally. In this research we have identified the major parameters and performed predictive analysis and compared the results using five different machine learning algorithms i.e., Random Forest, Decision Tree, Support Vector Machine, K Nearest Neighbour and Logistic Regression. After performing the analysis, Random Forest Algorithm gave the best result.***
Keywords: Stroke, Dataset, Machine Learning Techniques

I. INTRODUCTION

An average human being by its 40's or 50's is attacked with the most common diseases like diabetes, high-blood pressure, hypertension and some are even prone to the heart diseases as well. The sad truth about the recent time research by WHO (World Health Organization), shows that these are mainly caused by our unhealthy and improper lifestyle added to negligence towards our health. This paper mainly focuses on the use of modern analysis techniques such as Machine Learning to get a perfect picture of this medical emergency and how the medical facility can be improved so that it can be helpful for both, the patient and the doctors.

II. PROBLEM FORMULATION

A. Risk Factors for Stroke

Stroke is mainly caused whenever there is a sudden rupture in the blood vessel that further adds to the insufficient supply of the blood to the brain, followed by the symptoms such as sudden fatigue, numbness or weakness in the arms, legs or in any part of the body, it may be possible that a particular side of the body is affected. Sometimes the victim also faces dizziness or blurry vision while performing some basic tasks as a regular routine.

B. Datasets Characteristics

To perform this research the dataset was taken from Kaggle.com, we started analysing the dataset and tried out to find some major symptoms that are related to the person may be suffering from stroke. The dataset focuses on the parameters such as age, hypertension, heart disease, avg. glucose level, BMI as the major parameters we found is shown in the figure below:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

The dataset contains 5110 rows and 12 columns, on which some data pre-processing may be applied, so that the best suited algorithm maybe applied to get the perfect accuracy.

III. MACHINE LEARNING

The field of Machine Learning is mainly divided into three categories: Supervised Learning, Unsupervised Learning and Reinforcement Learning. Further the Supervised Learning is divided into: Regression and Classification. In this research paper, the dataset we will be dealing with is a Classification type of problem. Here, we're classifying whether a person is a victim of stroke or not, based upon some specific parameters that are present in the dataset.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue V May 2021- Available at www.ijraset.com

A. Data Visualization

Some of the known data visualizations in Machine Learning were applied to view and analyse the data the data in a better way.

1) Barplot

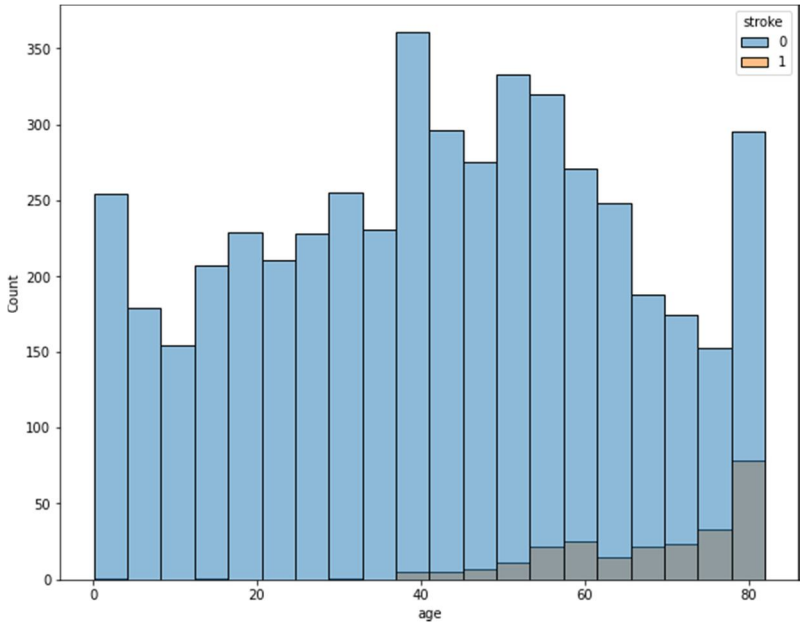


Fig:1 : Bar-plot of the number of persons who are victim of stroke and in which age group

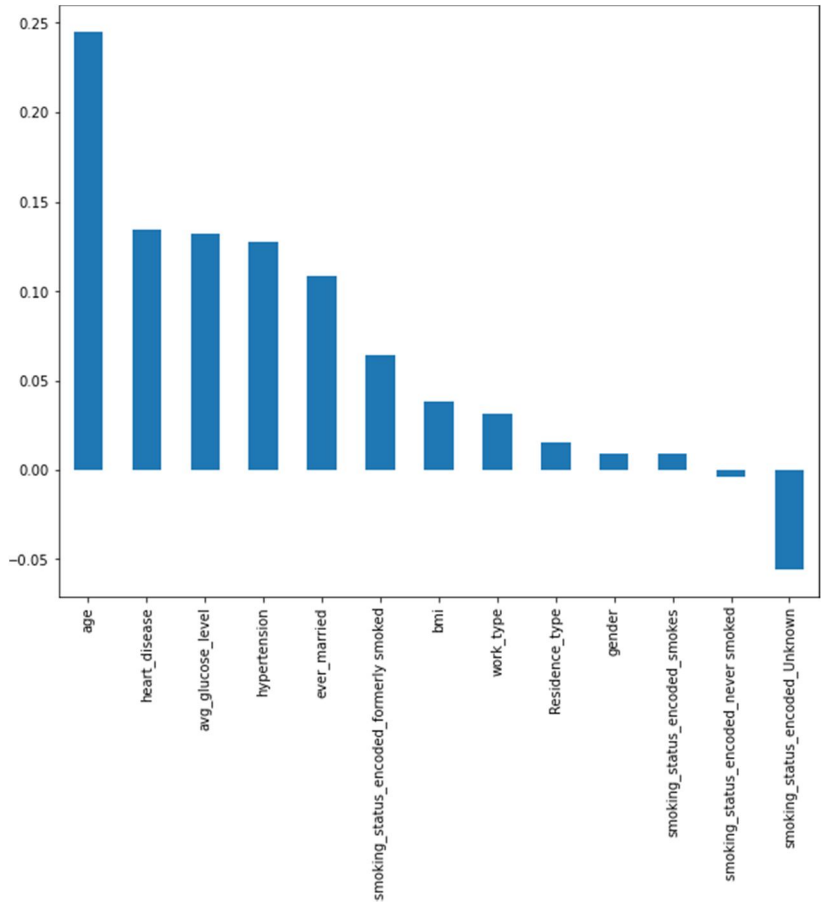


Fig.2: Bar-plot on the person suffering from stroke with specific parameters

2) Heatmap

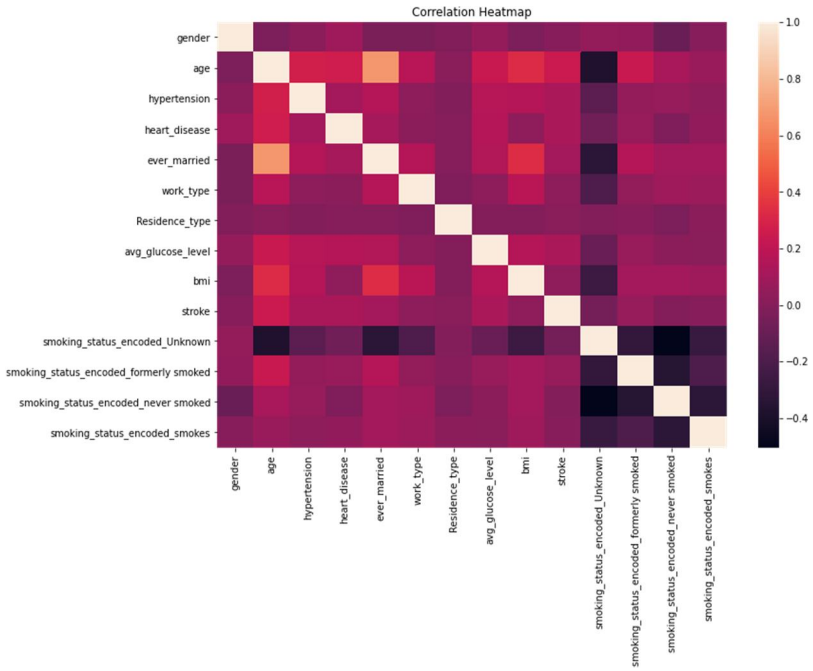


Fig.3: Heatmap showing the correlation between the parameters

B. Data Pre-processing

After visualising the dataset, we found that there are certain columns that are needed to be dropped so that the Machine Learning model would able to give perfect results without any overfitting or underfitting problems. The dataset after pre-processing is shown below:

	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
0	67.0	0	1	228.69	36.6	1
2	80.0	0	1	105.92	32.5	1
3	49.0	0	0	171.23	34.4	1
4	79.0	1	0	174.12	24.0	1
5	81.0	0	0	186.21	29.0	1

Fig.4: Dataset after performing data pre-processing

C. Random Forest

On performing Random Forest Algorithm on the given dataset, the accuracy was found to be 99%, which is found to be the best among the other applied algorithms. The classification report of the model is given below:

	precision	recall	f1-score	support
0	1.00	0.99	0.99	973
1	0.99	1.00	0.99	972
accuracy			0.99	1945
macro avg	0.99	0.99	0.99	1945
weighted avg	0.99	0.99	0.99	1945

Fig.5: Classification Report after performing Random Forest Algorithm

D. Decision Tree

On performing Decision Tree on the given dataset, the accuracy was found to be 98%, which is also best for given dataset on which the model performed very well. The classification report of the model is given below:

	precision	recall	f1-score	support
0	1.00	0.95	0.98	973
1	0.96	1.00	0.98	972
accuracy			0.98	1945
macro avg	0.98	0.98	0.98	1945
weighted avg	0.98	0.98	0.98	1945

Fig.6: Classification Report after performing Decision Tree Algorithm

E. Logistic Regression

On performing Logistic Regression on the given dataset, the accuracy was found to be 94%, that means the model performed very well on the dataset. The classification report of the model is given below:

	precision	recall	f1-score	support
0	0.94	1.00	0.97	1591
1	0.00	0.00	0.00	96
accuracy			0.94	1687
macro avg	0.47	0.50	0.49	1687
weighted avg	0.89	0.94	0.92	1687

Fig.7: Classification Report after performing Logistic Regression Algorithm

F. K- Nearest Neighbour

On performing Logistic Regression on the given dataset, the accuracy was found to be 94%, which is same as that of Logistic Regression, the model performed well on the dataset. The classification report of the model is given below:

	precision	recall	f1-score	support
0	1.00	0.88	0.94	973
1	0.90	1.00	0.95	972
accuracy			0.94	1945
macro avg	0.95	0.94	0.94	1945
weighted avg	0.95	0.94	0.94	1945

Fig.8: Classification Report after performing K-Nearest Neighbour Algorithm

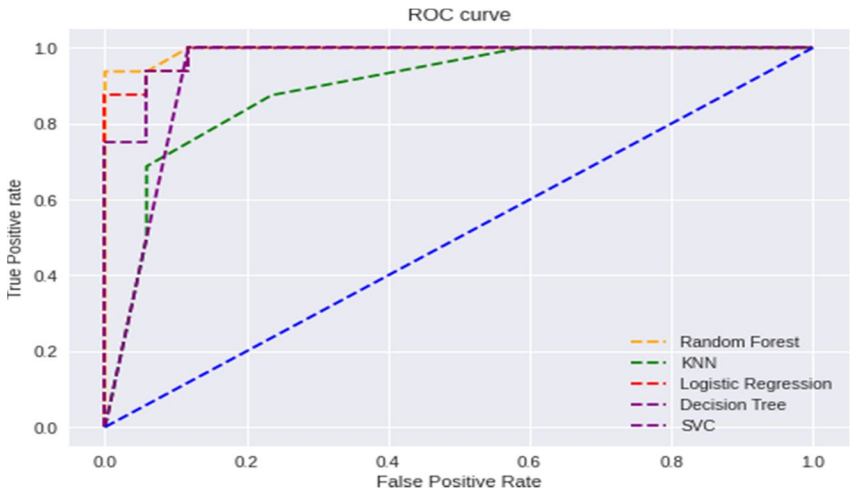
G. Support Vector Machine

On performing Support Vector Machine on the given dataset, the accuracy was found to be 76%, which is also a better score, though low than rest of algorithms. The classification report of the model is given below:

	precision	recall	f1-score	support
0	0.78	0.73	0.76	973
1	0.75	0.80	0.77	972
accuracy			0.76	1945
macro avg	0.77	0.76	0.76	1945
weighted avg	0.77	0.76	0.76	1945

Fig.9: Classification Report after performing Support Vector Machine Algorithm

H. AUC Curve



Random Forest: 0.9944

K-Nearest Neighbour:0.8952 Logistic Regression: 0.9889
Decision Tree: 0.9411

SVC: 0.9816

Fig.10: AUC Curve comparing the performance of the applied algorithms

IV. CONCLUSION

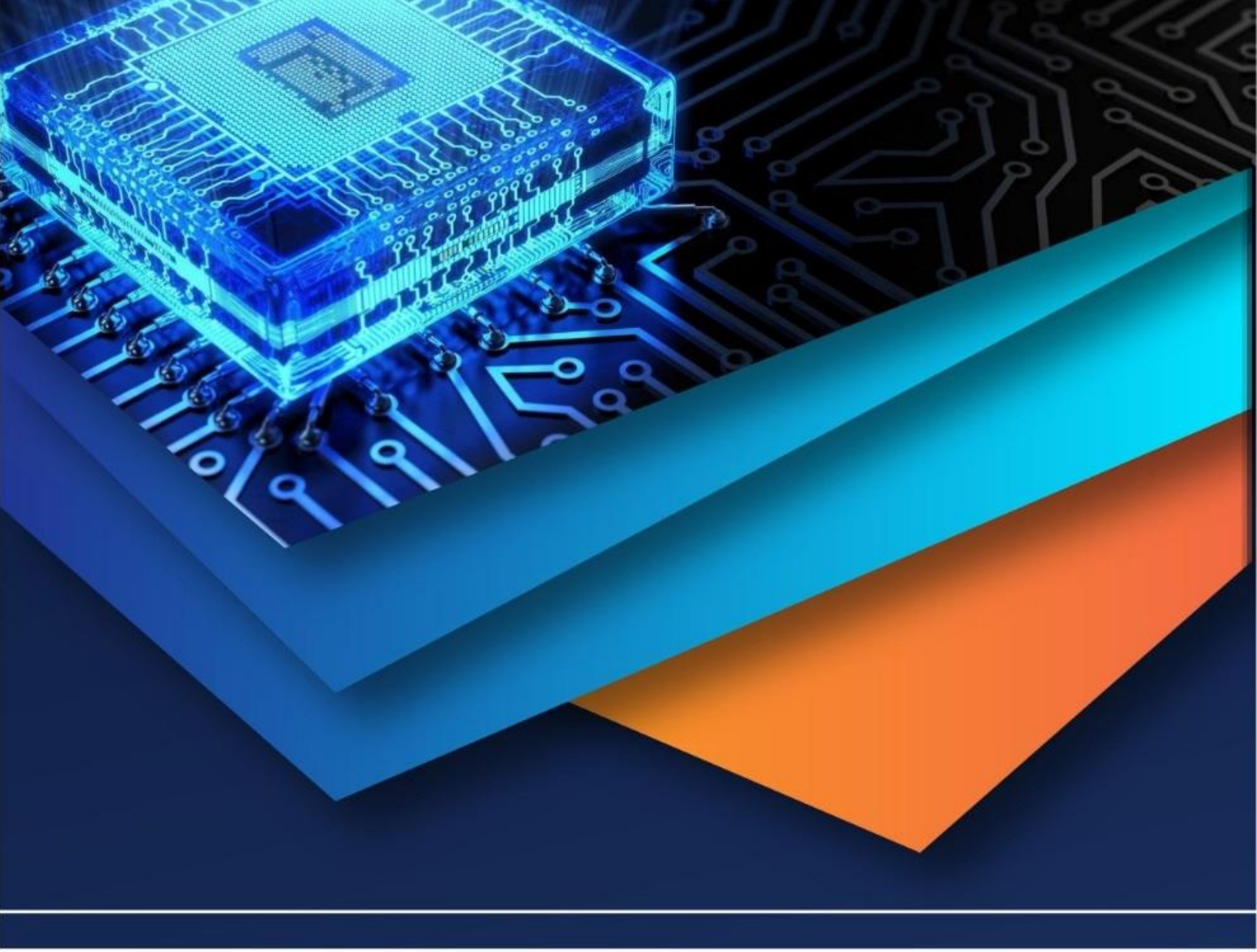
The main idea behind this research paper is to implement the new technologies, mainly Machine Learning to deal with the medical problems that would be very helpful for both the end user i.e., the patients and for the doctors and the physicians, so that a proper solution can be found out for proper treatment that too quickly and efficient steps can be taken for precautions so that this kind of medical ailments can be solved.

V. SCOPE OF THE FUTURE

The best Machine Learning model can be used and integrated with web applications and mobile applications for the end users to self-assess themselves close they are prone to whether getting attacked by stroke or not, or if they are, what are possible ways they can take precautions to avoid this kind of medical ailment.

REFERENCES

[1] K. Akazawa, T. Nakamura, S. Moriguchi, M. Shimada, and Y. Nose. Simulation program for estimating statistical power of Cox's proportional hazards model assuming no specific distribution for the survival time. Computer Methods and Programs in Biomedicine, 35(3):203–12, 1991.
[2] American Heart Association. Heart Disease and Stroke Statistics 2009 Update. American Heart Association, Dallas, Texas, 2009.
[3] R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate Cox proportional hazards models. Statistics in Medicine, 24(11):1713–1723, 2005.
[4] L. E. Chambless, G. Heiss, E. Shahar, M. J. Earp, and J. Toole. Prediction of ischemic stroke risk in the atherosclerosis risk in communities' study. American Journal of Epidemiology, 160(3):259–269, 2004.
[5] C. Cortes and M. Mohri. AUC optimization vs. error rate minimization. In Advances in Neural Information Processing Systems 16. MIT Press, 2004.
[6] T. R. Dawber, G. F. Meadors, and F. E. Moore. Epidemiological approaches to heart disease: The Framingham study. American Journal of Public Health and the Nation's Health, 41:279–286, March 1951.
[7] J. M. Engels and P. Diehr. Imputation of missing longitudinal data: a comparison of methods. Journal of Clinical Epidemiology, 56(10):968–976, 2003.
[8] L. P. Fried, N. O. Borhani, P. Enright, C. D. Furberg, J. M. Gardin, R. A. Kronmal, L. H. Kuller, T. A. Manolio, M. B. Mittelmark, A. Newman, D. H. O'Leary, B. Psaty, P. Rautaharju, R. P. Tracy, and P. G. Weiler. The Cardiovascular Health Study: design and rationale. Annals of Epidemiology, 1(3):263–276, February 1991.
[9] J. Goeman. l1 penalized estimation in the Cox proportional hazards model. Biometrical Journal, 52(1):70–84, 2009.
[10] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, Recent Advances in Learning and Control, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008.
[13] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. Journal of Machine Learning Research, 3:1157–1182, 2003.
[14] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1):29–36, 1982.
[15] F. E. Harrell. Regression Modeling Strategies, With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer, 2001.
[16] K. Ikeda, H. Kumada, S. Saitoh, Y. Arase, and K. Chayama. Effect of repeated transcatheter arterial embolization on the survival time in patients with hepatocellular carcinoma. Cancer, 68(10):2150–4, 2001.
[17] T. Joachims. A support vector method for multivariate performance measures. In Proceedings of the International Conference on Machine Learning, pages 377–384, 2005.
[18] A. Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In Proceedings of the International Conference on Machine Learning, 2004.
[19] A. Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In Proceedings of the International Conference on Machine Learning, 2004.
[20] M.-Y. Park and T. Hastie. An l1 regularization-path algorithm for generalized linear models. Journal of the Royal Statistical Society: Series B, 69(4):659–677, 2007.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY