IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 3, March 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203006 |

Integrating Data Mining Techniques for Decision Tree Model Assessment

Hareesh Edupuganti

Data Engineering Manager, DELOITTE CONSULTING, Austin, TX, USA

ABSTRACT: Decision trees are among the most widely used supervised learning models in data mining because of their interpretability and efficiency in handling structured data. However, effective evaluation of decision tree models requires careful application of data mining strategies that integrate both predictive performance and interpretability. This paper demonstrates the implementation of data mining techniques using the Weka platform to evaluate decision tree models against real-world datasets. The study highlights comparative analysis between classifiers, explores model accuracy and error rates, and emphasizes the interpretability of decision paths for decision-making. By applying practical data mining strategies, this research illustrates how decision tree evaluation can enhance knowledge discovery and improve the reliability of predictive analytics across domains.

KEYWORDS: Data Mining, Decision Tree Model, Weka, Classification, Implementation

I. APPLYING DATA MINING TECHNIQUES

Data mining has emerged as a critical discipline for extracting patterns and insights from large and complex datasets. Among the various approaches, decision tree models stand out due to their transparent structure, ease of interpretation, and ability to handle both categorical and continuous data. Decision trees are frequently applied in domains such as healthcare, finance, education, and behavioral analytics, where model interpretability is as important as predictive accuracy.

The evaluation of decision tree models, however, requires systematic application of data mining strategies. While advanced models like artificial neural networks (ANN) often achieve high predictive performance, they are often criticized for their “black-box” nature, limiting their interpretability. In contrast, decision trees provide explicit rules for classification, but their effectiveness must be validated through rigorous data mining practices such as cross-validation, attribute selection, and error-rate analysis.

This research applies data mining techniques to evaluate the performance of decision tree models using the Weka platform, an open-source data mining tool. The Weka environment supports classification, clustering, and association tasks, making it an ideal platform for implementing and comparing predictive models. A case study on youth behavioral data is presented to demonstrate the methodology, highlighting not only predictive accuracy but also the interpretability of generated decision paths.

Weka interface

Weka (Waikato Setting for Expertise Analysis) is a data mining system dispersed under public license GNU-GPL: it is actually cost-free software application that may be with ease utilized, replicated, examined, moderated and also circulated and it is actually guarded coming from appropriation efforts that will restrain these user rights.

Considering its own attributes, our experts locate that it is actually a device which, first off, has an interactive user interface which includes 4 user-machine communication methods (Fig. 1).:

- **Explorer:** is the best made use of method and the most descriptive.
- **Experimenter:** valuable mode to review the functionality of different predictive versions (practices).
- **KnowledgeFlow:** allows the aesthetic shows of modelling style with attached objective components.
- **Simple CLI (Simple Client):** supplies a console to implement the functions of the system through commands; it makes it achievable to execute any kind of procedure held through Weka straight, although it carries out ask for a detailed demand of the application.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203006 |

Figure 1: Applications of the Weka interface

Within the **Explorer** technique, it offers extensive support to the overall method of data mining (Fig. 2):

1. Accessibility to data sources, exploration and option of data as well as information handling.
   - **Preprocess**: functions targeted at importation, change (app of filters) and records extraction.
   - **Visualize**:functions targeted at the visual images of information making use of graphic approaches.

2. Anticipating as well as definitive modelling. Puts together a variety of data mining methods for the obtention of understanding styles:
   - **Classify** (classification and regression): anticipating modelling (monitored understanding).
   - **Cluster** (grouping) and **Associate** (association rules): detailed modelling (without supervision discovering).
   - **Select attributes**:option of anticipating characteristics.

Lastly, it costs highlighting the opportunity of **unit extensibility**: it permits the individual to customize Weka by integrating brand-new functionality built in Java code, using its own framework as well as item oriented functional design. This exemplifies the principal conveniences rather than various other closed code data mining systems (office programs). In Witten & Frank (2005) you can easily find a detailed description of the various methods of communication with Weka.

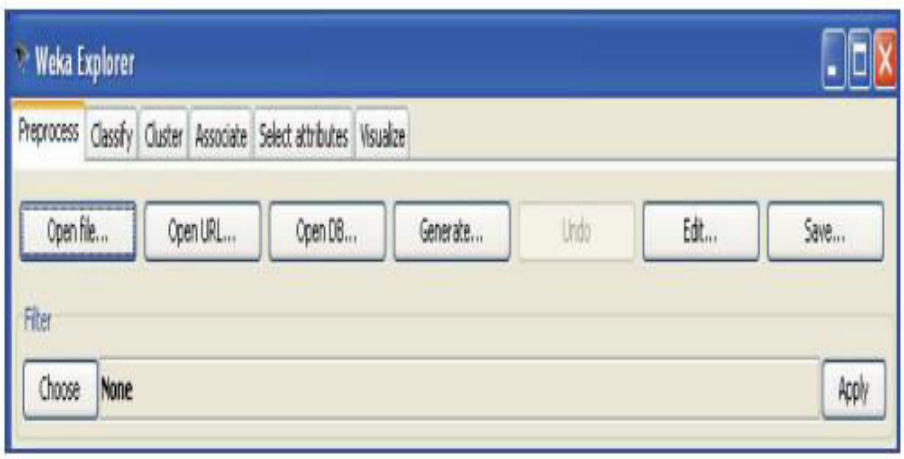

Figure 2: Weka Explorer interface

The aim of this work is to show how implementing structured data mining strategies enables a more reliable evaluation of decision tree models, ensuring that predictive insights are both accurate and actionable for decision-makers.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203006 |

II. CASE STUDY

Once our team have actually offered the methodological manner of the strategies involved in this job, within this part our experts aim to contribute comparative factors of the relevant information provided through these methods in an administered circumstance. These comparative elements describe the predictive power (reliability) of the knowledge models produced and the definitive parts that add insightful market value to the decision making procedures (categories). Hence we target to give an extra including perspective, ideally, of DM methodology, due to the fact that we provide common analysis guidelines if you want to compare the results obtained.

Regardless, coming from the evaluation of these end results we do certainly not aim to reach out to substantial conclusions associated with the situation from which the records used comes, yet instead our intent is to disclose to the readers a series of technical tools that permit our company to discover expertise patterns in a way that is actually virtually automated and also, however, to make it much easier to interpret the definitive factors connected with the evaluation of the models secured.

Coming from a first sample of 9300 youngsters aged between 14 and also 18 years, through which info concerning variables interfering in the intake of addictive compounds was actually accumulated, we decided on a sub-sample of 2526 youngsters. Our team are interested in examining the connection in between the intake/ non usage of marijuana (outcome variable) one of people checked and also the causes the subject matters surveyed eat consuming or otherwise consuming drugs (input variables). Especially, our experts collected fifteen feasible main reasons (variables) for consuming medicines and also eleven factors (variables) for not taking in; the possible response per of these variables is dichotomous (yes/no). On the other hand, if in the first example (total) the amount of individuals of marijuana is actually virtually 18%, as opposed to 78% of non customers, the sub-sample decided on series a greater equilibrium in between consumers/non buyers (44.4%/ 55.6%); this balance (or even harmonizing) is actually warranted by methodological aims, as there need to be actually a comparable amount of admittances in each of the outcome classifications (consumes/does not consume), to ensure they could be equally embodied in the modelling phase (detection of classificatory designs).

In a treatment with Traveler, first off our experts loaded (Open up documents ...) in the Preprocess area the information to become studied, whose design (data source) has been actually adjusted to the Weka layout: Arff documents. Once the records file levels, it is achievable to check out the variables they contain (Fig. 3). It is actually additionally possible to go through information in the CSV (punctuation delimited) style coming from Weka, although it is certainly not achievable to import data sources in the even more commonly used formats like Excel, Accessibility, SPSS, and so on. Nevertheless, there is actually the possibility of changing these various other even more common styles into the indigenous Arff format coming from the data mining system RapidMiner (cost-free, available code software application, which additionally allows the use of the protocols included in Weka). For instance, if the information resource remains in SPSS format, our experts can show whether our company are interested in extracting the names of the variables and/or their labels in one more format (Arff, within this case), as well as whether our company are interested in utilizing the tags of the market values (option be actually nonpayment) as opposed to the mathematical market values.

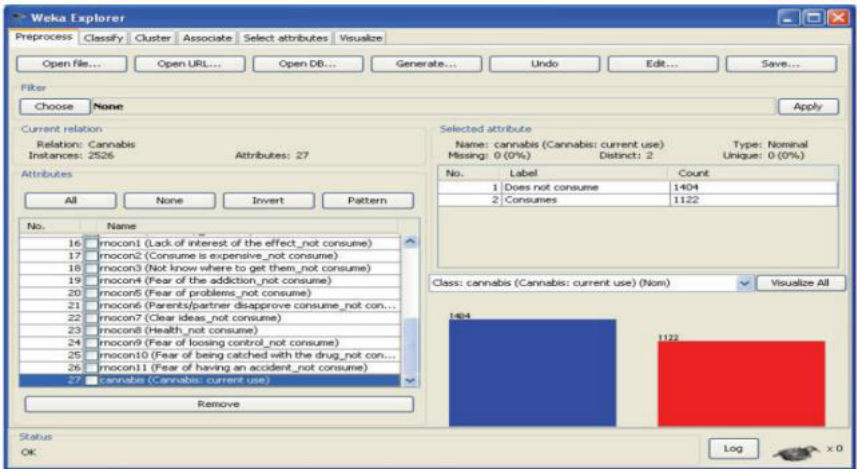

Figure 3: Weka Explorer interface: exploring variables

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203006 |

From the section Classify, our experts access the various modelling approaches incorporated in Weka (Fig. 4). For instance, our team can easily select the J48 classifier (weka.classifiers.trees.J48); this classifier uses the C4.5 protocol (Quinlan, 1993) to generate a distinction tree which resides in contract with a set of specifications determined due to the protocol (to modify all of them, click the classifier) and also other specifications determined through data mining strategy (in Exam choices). In the instance (Fig. 4) our experts have suggested that the J48 classifier utilizes 70%of the example (training data) to create the style, and the rest as test information. The result variable is actually likewise suggested (predicted variable), which through default is the last adjustable in the database. The Start switch allows us to produce the model as well as access (in Classifier result) the design's examination outcomes. It may be noted that the design has correctly categorized 599 of the 758 examination norms (79%), along with a much larger percent of favorites in the group Performs not consume (83.4%) than in the category Consumes (73.9%).

It is actually possible to access the graphic representation of the classification tree with the options of the contextual menu (Visualize plant) of the model produced (in the End result list).

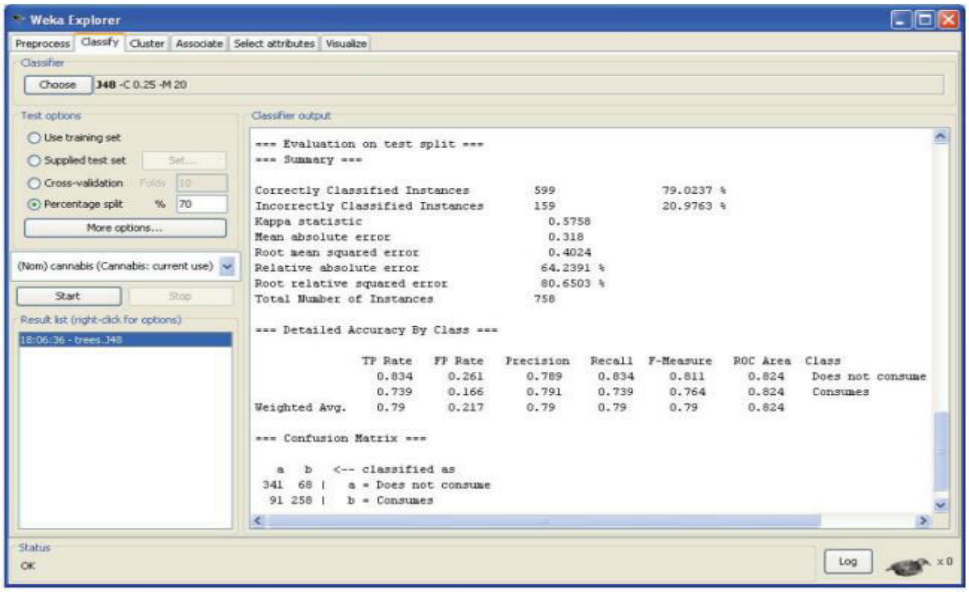

Figure 4: Evaluation of the selected decision tree model on test data

III. ARTIFICIAL NEURAL NETWORKS

Man-made Neural Networks (ANN) are records handling bodies whose framework and also functioning are actually influenced by biological neural networks. ANN were created based upon the complying with tips:

1. Data processing develops in basic factors called nerve cells.
2. The neurons transmit signals through created relationships.
3. Each correlation (communication hyperlink) has an affiliated significance.
4. Each neuron administers an account activation feature (usually non linear) to the total entry of connected nerve cells acquired (sum of items weighted depending on to the network weights), thereby securing an output worth which are going to act as the item market value which will definitely be sent to the remainder of the network.

The essential qualities of ANN are actually parallel handling, dispersed moment and versatility to the surroundings.

The processing unit is the synthetic nerve cell, which receives the entries from the neighbouring neurons and calculates an output worth, which is delivered to all the continuing to be nerve cells.

As far as the portrayal of input and result info is actually involved, our experts can easily discover connect with continuous input and outcome records, networks with discrete or even binary input and outcome information and also connect with continual input information as well as separate output records.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203006 |

An ANN is actually made up of the consecutive order of three general types of nodes or even coatings: input nodes, output nodes and also more advanced nodes (concealed layer) (Figure 5). The input nodes supervise of acquiring the preliminary market values of the data apiece situation so as to transfer them to the network. The output nodules get input and also compute the result market value.

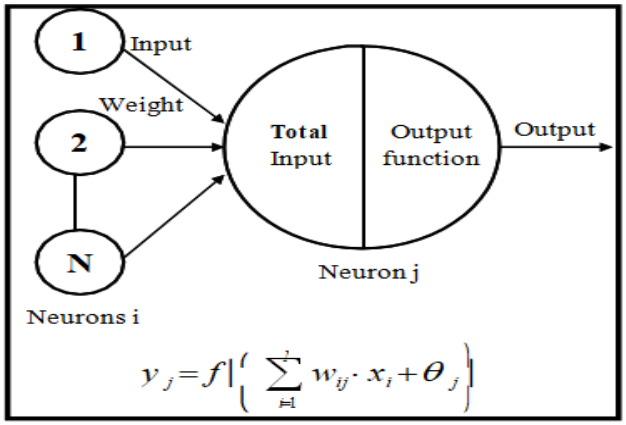

Figure 5: Generic working of an artificial neuron and its output mathematical representation

This collection of nodes utilized due to the ANN, together with the account activation functionality, makes it feasible for the ANN to simply embody non-linear relationships, which are actually the absolute most difficult as for multivariate techniques are actually regarded.
The best secondhand account activation features are: the measure functionality, identification feature, sigmoid or even logistic feature and the hyperbolic tangent. There is actually a big choice of ANN designs. A combo of the geography (amount of nerve cells as well as surprise coatings, and just how they are actually connected), the finding out standard as well as the understanding formula define an ANN version (Bigus, 1996).

It can be said that an ANN has 3 benefits that makes it incredibly appealing in information managing: flexible learning through instances, toughness in handling unnecessary and also incorrect relevant information as well as substantial parallelism.

The best secondhand strategy in the sensible uses of ANN is actually the multilayer perceptron, which was actually made prominent by Rumelhart et al. (1986).

A multilayer perceptron sort of ANN starts with an input level through which each nodule or nerve cell relates a predictor variable. These input nerve cells get in touch with each of the nerve cells making up the concealed coating. The nodes in the hidden layer in turn connect with the neurons in an additional concealed coating. The result level is composed of one (binary forecast) or additional output neurons. In this particular sort of style, the info is consistently sent from the input layer in the direction of the outcome coating.

The recognition of the multilayer perceptron is actually primarily due to the simple fact that it can working as a global feature approximator. Even more especially, a "backpropagation" network which consists of at the very least one concealed layer with enough non-linear devices can easily know any sort of kind of feature or ongoing partnership between a group of input variables (discrete and/or continuous) as well as an output variable (discrete or ongoing). This building produces multilayer perceptron systems general, pliable and also non-linear resources. A full description of the mathematical structures associated with the training stage and also the performing phase of the backpropagation formula in multilayer perceptron style may be discovered in Rumelhart et cetera (1986).

When the system is actually used to identify typically the result coating has as many nodes as the amount of classes and the nodule of the output layer with the highest possible value uses the price quote of the class which the network creates a particular input. In the grandfather clause of 2 classes it prevails to possess a node in the result layer, and the category between the two training class is actually performed through administering a trimmed lead to the nodule value.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203006 |

If among the merits of ANN is that they enable modelling any sort of operational relationship (direct or even non linear) in between variables and also, therefore, function as universal function approximators, yet another of the exceptional conveniences of this particular technique, matched up to classic modelling strategies, is that it performs not enforce any type of regulation relative to the starting data (sort of operational partnership in between variables), not either does it normally start from specific beliefs (like the kind of distribution the records comply with). An additional virtue of the approach hinges on its capacity to approximate excellent models also despite the life of sound in the details evaluated, as takes place when there is actually a presence of left out worths or outlier worths in the circulation of the variables. As a result, it is actually a sturdy method when managing concerns of sound in the information provided; however, this does certainly not imply that the cleaning requirements of the information matrix need to be relaxed.

Regardless, its harsh flexibility hinges on the need to have sufficient training data and also it needs additional opportunity for its own implementation than various other procedures (Shmueli et al., 2007). It deserves revealing that in ANN, in addition to the collection of instruction records to build the design and also the collection of independent records (examination information) so as to determine its own induction capability, a 3rd set of private records (verification specified) is used to stay clear of overfitting the design (during the knowing process) which may cause a too much lot of parameters or even weights relating to the issue (Hastie et cetera, 2001, p. 356).

Regardless of this, this viewpoint of ANN as a complex "black box" is actually certainly not totally real. In this sense, various tries at translating the weights of the neuronal system have emerged, of which the most extensively made use of is the alleged sensitiveness analysis (Montaño & Palmer, 2003), executed in ANN programmes as lately offered through Palmer et al. (2001), under the name of Level of sensitivity Semantic network 1.0.

IV. CONCLUSION

The usefulness of the multilayer perceptron, depends on its own ability to know basically any kind of partnership in between a set of input as well as outcome variables. On the other hand, if our team use approaches stemmed from timeless studies like straight discriminant evaluation, this does not have the ability of determining non-linear features and, for that reason, will present a lesser efficiency compared to the multilayer perceptron in distinction duties that include sophisticated non-linear connections. This paper applied data mining strategies in the evaluation of decision tree model.

REFERENCES

[1] Alam et al. 2012, Md. Hijbul Alam, JongWoo Ha, SangKeun Lee, Special strategies to creeping necessary web pages early, Expertise and likewise Details Equipment, December 2012, Amount 33, Concern 3, pp 707-734
[2] Aral S. and Passerby D. 2012, Identifying popular in addition to prone participants of social networking sites, Scientific research, vol.337, pp.337-341.
[3] Machanavajjhala in addition to Reiter 2012, Ashwin Machanavajjhala, Jerome P. Reiter: Big personal privacy: securing personal privacy in big data. ACM Crossroads, 19(1): 20-23, 2012.
[4] Pushpa Mannava, "Big Data Analytics in Intra-Data Center Networks and Components Of Data Mining", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 1 Issue 3, pp. 82-89, November-December 2016. Available at doi : https://doi.org/10.32628/CSEIT206272
[5] Pushpa Mannava, "An Overview of Cloud Computing and Deployment of Big Data Analytics in the Cloud", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 1 Issue 1, pp. 209-215, 2014. Available at doi : https://doi.org/10.32628/IJSRSET207278
[6] Pushpa Mannava, "Role of Big Data Analytics in Cellular Network Design", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 1 Issue 1, pp. 110-116, March-April 2015. Available at doi : https://doi.org/10.32628/IJSRST207254
[7] Pushpa Mannava, "A Comprehensive Study on The Usage of Big Data Analytics for Wireless and Wired Networks", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 4 Issue 8, pp. 724-732, May-June 2018. Available at doi : https://doi.org/10.32628/IJSRST207256

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203006 |

[8] Pushpa Mannava, "A Big Data Processing Framework for Complex and Evolving Relationships", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, ISSN: 2278 – 8875, Vol. 1, Issue 3, September 2012

[9] Pushpa Mannava, "A Study on the Challenges and Types of Big Data", "International Journal of Innovative Research in Science, Engineering and Technology", ISSN(Online) : 2319-8753, Vol. 2, Issue 8, August 2013

[10] Pushpa Mannava, "Data Mining Challenges with Bigdata for Global pulse development", International Journal of Innovative Research in Computer and Communication Engineering, ISSN(Online): 2320-9801, vol 5, issue 6, june 2017

[11] Sriramoju Ajay Babu, Namavaram Vijay and Ramesh Gadde, "An Overview of Big Data Challenges, Tools and Techniques"in "International Journal of Research and Applications", Oct - Dec, 2017 Transactions 4(16): 596-601

[12] Ramesh Gadde, Namavaram Vijay, "A SURVEY ON EVOLUTION OF BIG DATA WITH HADOOP" in "International Journal of Research In Science & Engineering", Volume: 3 Issue: 6 Nov-Dec 2017.

[13] Ajay Babu Sriramoju, Namavaram Vijay, Ramesh Gadde, "SKETCHING-BASED HIGH-PERFORMANCE BIG DATA PROCESSING ACCELERATOR" in "International Journal of Research In Science & Engineering", Volume: 3 Issue: 6 Nov-Dec 2017.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details