Knowledge Extraction for Business Information System using C5.0 Tree Algorithm

Anitha A, Mathivanan R, Sundarraj R, Indhulekha J

Abstract: *Usually people can predict that some products will be purchase by the males and some will be purchased by the females, but there are some hidden factors behind the data. When the data was analyzed ,Analysts comes to know those hidden factors in the dataset. In this study,C5.0 algorithm is used which is highly approachable compare to other decision tree algorithms. So that it is easy understand the data patterns and the decision that can be made by the Entrepreneur. Normally the products like beer, meat, crispy chips and so on will be purchased by the males and the products like chocolates, soft drinks will be purchased by the females, but when the data was analyzed it is predicted that which gender would buy which product that can't be predicted by the normal peoples . In this project, it is proposed to apply C5.0 algorithm for finding the target customer group. Identifying specific customer group is necessary to improve profit in sales domain. Accuracy attained with proposed model is 81.6%. For each category of product, the interested gender group is identified*

Keywords: ***Accuracy, Decision Tree, Knowledge Extraction, Prediction.***

I. INTRODUCTION

Data Science is a field which uses algorithms and scientific methods to gain knowledge from data. It is the concept to combine the statistics, data analysis, machine learning, and their methods to understand and analyze the data. Science is that the study of where the info comes from, what it represents and the way can it's became valuable resources within the business. It also employs mathematics, statistics techniques like machine learning, data mining, etc. Machine learning is the artificial intelligence tool that processes the data in data science. The main advantage of data science is to facilitate decision making. The enterprises can find when and where their products can sell better. It is important because knowledge is important. Data Analytics is the examining of data set in order to draw conclusion about information they contain. The process of evaluating data using analytics to examine every data is Data analysis. Here data are collected, analyzed(decision making) and concluded to produce new information. The aspects of data analytics are qualitative and quantitative technology. Qualitative includes the process of non-numerical data and the quantitative includes the process of numerical data. There are two types in data analysis. They are, Exploratory data analysis which finds the data pattern and relationships and the other is Confirmatory data analysis which postulates about the data is true or not using statistical technique. The advanced type of data analysis are Data mining, predictive analysis, machine learning and big data analytics and also text mining. Data mining is a technique of data analytics where large amount of pre existing databases are analyzed or examined to create new information. Here in data mining, patterns and relationships between the data sets are identified to resolve problems through data analysis. Data mining is a tool which allow us to predict future. Business intelligence contains technology utilized by enterprise for data analysis of business information. It helps different organizations by providing better decision making by using latest tools and method. It involves data analytics, data processing and massive data. It also involves various process and procedure which helps in collection of data, sharing them and reporting them to provide better decision making. With recent technologies in BI tools, users can view the output through visual so that they need not depend on an IT staff for the output. The main difference between the BI and data analysis is that the BI helps in making business decisions using past results where data analysis helps in making predictions which helps you in the future. SPSS for predictive modeler is a statistical and data mining software application from IBM to solve business and research problems. The software name stood for Statistical Package for social science. It is used to assemble predictive models and conduct other analytic tasks. It has visual interface which allows users to hold the statistical and data mining algorithms without programming. The main aim of this modeler is to produce less complexity in predictive models. The SPSS modeler has been for many applications like forecasting demand or sales, education, telecommunication, risk management, entertainment, healthcare quality improvement, etc. Decision tree is a tree based classification model. It helps to identify groups among data, discover relationship between them, and predict future events. It provides highly visual classification of data and explain analysis to non-technical audiences. The procedure in decision tree is utilized for segmentation, stratification, prediction, data reduction and variable screening, interaction identification and category merging.

Revised Manuscript Received on February 05, 2020.

* Correspondence Author

Anitha A*, Information Technology, Francis Xavier Engineering college , Tirunelveli, India. Email: draanitha@yahoo.com

Mathivanan R, Information Technology, Francis Xavier Engineering College , Tirunelveli, India. Email: mathivananr59@gmail.com

Sundarraj R, Information Technology, Francis Xavier Engineering College , Tirunelveli, India. Email: rsundarr29@gmail.com

Indhulekha J , Information Technology, Francis Xavier Engineering College , Tirunelveli, India. Email: indhulekhaj98@gmail.com

The main feature of this tree is that it identifies homogeneous groups with high or low risk. During individual cases it is easy to predict values.

Predictive analysis surrounds various of statistical techniques such as data mining, predictive modeling, and machine learning that that analysis the current and past data to provide predictions about future or unknown events. Simple predictive analysis means gaining information from existing data sets so as to work out patterns and predict future outcomes.

II. LITERATURE SURVEY

Researchers focused on the study of technical growth in every field[1]. Brain Tumor is one amongst the severe syndrome causes for the death of People. The Dataset of BRATS 2015 containing Brain Tumor MR Images for 20 patients. This paper discusses different Machine Learning classification algorithm. Many Machine Learning Algorithms are utilized in this study namely: Naive Bayes, Logistics Regression, Multi layer Perceptron, Support Vector Machine and Decision Tree using the Waikato Environment for Knowledge Analysis tool and MATLAB. The results obtained in this paper are 80%, 90%, 88.88%, 88.88%, and 75% for Naive Bayes, Logistics Regression, Multi-Layer Perceptron, SVM and Decision Tree respectively with the highest accuracy was given by Logistics Regression.

Predictive Modeling may be a crucial part of learning analytics[2].. The formost objectives of this study is to predict student performance based on social media traces. The attributes are performance, knowledge, score or grade. The Knowledge set is collected from a Internet Application Design Course. For grade prediction regression algorithm is employed. This paper compares the performance of classical algorithm i.e., Random Forest (RF) and K-Nearest Neighbors , called Large Margin Nearest Neighbor Regression(LMNNR). The LMNNR proved very suitable for this prediction problem, outperforming other classical regression algorithms.

According to Han Wu, more and more families are influenced by Diabetics Mellitus, most diabetics know little about their health quality or the risk factor they face prior to diagnosis [3]. This study has proposed a model based on data mining technique for predicting Diabetics Mellitus. The main objective of this study is to improve the accuracy of prediction model and to make the model adaptive to more than 1 dataset. This model is comprised of two parts: the improved K-means algorithm and regression algorithm. The dataset used in this study is Pima Indians Diabetes Dataset and toolkit utilized is Waikato Environment for Knowledge Analysis. The conclusion shows that the model get 3.04% higher accuracy of prediction.

Relevant research to the use of Predictive analysis has been discussed in a number of papers and academic articles[4]. Ahmed M.Zeki research study focused on using Urinary system dataset in UCI ML Repository , Acute inflammations. This study demonstrates the ability of Data mining to develop a prediction model for diagnosis of two familiar urinary diseases namely: Acute inflammation of Urinary bladder and Acute inflammation of Nephritis. This study evaluates the supervised machine learning decision tree algorithms Rider, oneR, J48 in terms of performance and accuracy to determine the best classification algorithm. For prediction of acute Urinary bladder, the model accuracies are Rider-100%, oneR-79.2%, J48-100% and for prediction of acute Nephritis, the model accuracies are Rider-99.2%, oneR-91.6%, J48-100%. This study proves J48(C4.5) decision tree algorithm was able to deliver 100% accuracy and 100% precision.

Muhammed Awais Shafique study focused on random forest algorithm for detecting classification of travel data with multiple sensor information[5]. The author collected data of 46 participants from three different cities in Japan, namely Niigata, Gifu, and Matsuyama. GPS and accelerometer plays a major role in travel data collection through wearable devices and that can be easily achieved by the help of Smartphone. In this study, the random forest uses 70% of data for training and remaining 30% for testing purpose. Prediction was done among train, bicycle, walk, and bike. It is concluded that the random forest algorithm produces overall prediction with high accuracy of 99.6%.

According to Rafik Khairul Amin, loans are common. Before getting a loan the applicant has to go through many step done by bank to evaluate whether the applicant is eligible or not [6]. This study proposed that decision tree tool can be used to generate a tree since it has high accuracy in decision making. Dataset used in this research is loan approval data obtained from Bank Pasar of Yogyakarta Special Region. The used data is as many as 1000 data records gathered from January till July2014. C4.5 algorithm is the successor of ID3. The result shows that the biggest precision is 78.08% reached by C4.5 algorithm with the data partition of 90%:10%.

Romana Markovis research study compares the different classification algorithm for detecting a User's interaction with Windows in office buildings[7]. This study focused on comparison using a available dataset. The algorithms like Support Vector Machines(SVM), Random Forest, and their combination with dynamic Bayesian networks(DBN) are implemented for detetcting occupants interaction with windows. Occupant Behaviour is the main source of dicrepancy between the predicted and measure energy consumption in buildings. The comparison proves that Random Forest outperformed all other algorithm for identifying the window status in office buildings. The implementation of random forest by 200 trees is more accurate than other algorithms.

Yurong Zhong research study focused on the study of basic principle of data mining and basic algorithms [8]. There are different types of Classification, in which the decision tree is very easy way to understand and to implement. The main objective of this paper is to combine the principle of Taylor formulae with the attribute selection of the ID3 selection, to reduce the algorithm complexity and to increase the running efficiency. This paper result show that improved simplified entropy algorithm reduce the complexity degree, and to improve the accuracy of the algorithm.

III. PROPOSED SYSTEM

A. Component Diagram

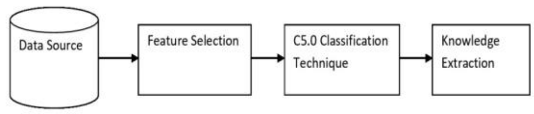


Fig. 1.Component Diagram

From the above Fig. 1, the data source node is the representation of the data sets which contains the attributes. Higher the data patterns higher the accuracy of prediction for the classification of data sets into a decision tree. The feature selection node is used to select the input and target fields (attributes) for the result. C5.0 algorithm is used to build a decision tree for the given input fields and helps to improve the turnover of any business and also this analysis assist the industrialist to take decisions on business strategies. Ultimately, the knowledge that is extracted from the decision tree can be very much useful to predict the future inferences of business strategies.

B. Algorithm

C5.0 algorithm is utilized to build either a decision tree or a rule set. This works by splitting the entire sample Dataset based on the attribute that has highest value of information gain. C5.0 node splits the subsample again and again, until it cannot be split further. The lowest-level splits are finally re-examined and those which do not contribute much are removed or pruned. It works for both categorical and continuous data sets. It offers the powerful boosting method to increase accuracy of classification. One of the most useful type of classification is Decision tree. It is a type of supervised learning algorithm (having a pre-defined target variable) that can be useful in classification. Exactly one prediction is possible for any data record in a Decision tree. C5.0 models are quite robust.Usually decision tree follows Sum of Product(SOP) representation, which is known as the disjunctive normal form. The major challenge while implementing decision tree algorithm is to identify which attribute should be consider as a root node and each level, this situation can be handled by the attribute selection. Commonly, there are two different methods for attribute selection measures they are,

1).Information Gain

2).Gini Index.

- **Information gain***:* When the information gain is used as a criterion ,it tries to estimate the information contained by each attribute. It can be measured by calculating the randomness or uncertainty of the field(X) called the Entropy. The entropy of the field is measured using the formula :

$$H(X) = \sum_{x \in X} p(x) \log_2 p(x) \qquad (1)$$

$$\text{Information Gain(IG)} = E(\text{Target}) - E(\text{Target, Attribute})$$

where, H(X)= E(X) → entropy value of the attribute X

p(x) → probability of the individual classified attribute x

- *Gini Index:* It is a metric to measure how often a random element is identified.(i.e.) An attribute with lower gini index is preferred. The general formula for calculating the gini index is :

$$\text{Gini Index} = 1 - \sum_j p_j^2 \qquad (3)$$

where, p(x) → probability of the individual classified attribute x

By using these two popular attribute selection, the attribute which is considered as the root node can be identified.

IV. EXPERIMENTAL RESULTS

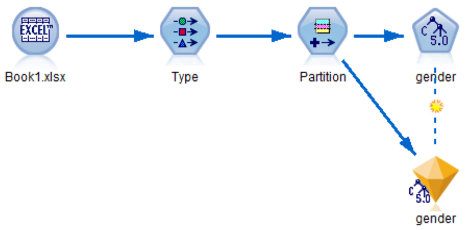
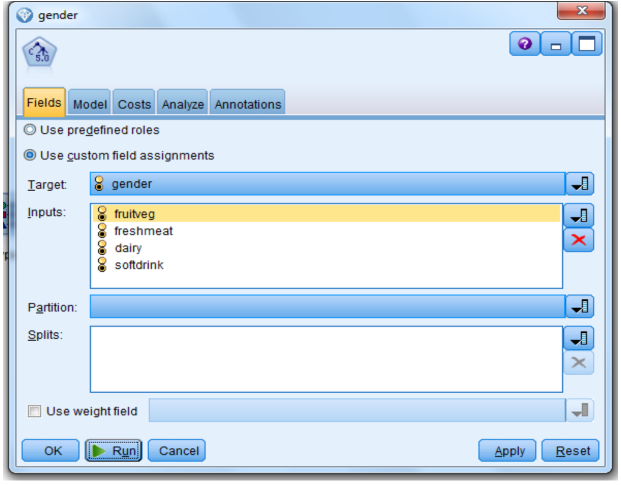


Fig. 2.Proposed System

The above Fig. 2, represents the analytics of the data using C5.0 algorithm using IBM SPSS Modeler. The IBM SPSS Modeler is a tool which is used to analyze the data sets and gives the output to predict the data.


gender

Fields | Model | Costs | Analyze | Annotations

○ Use predefined roles

● Use custom field assignments

Target: gender

Inputs: fruitveg, freshmeat, dairy, softdrink

Partition:

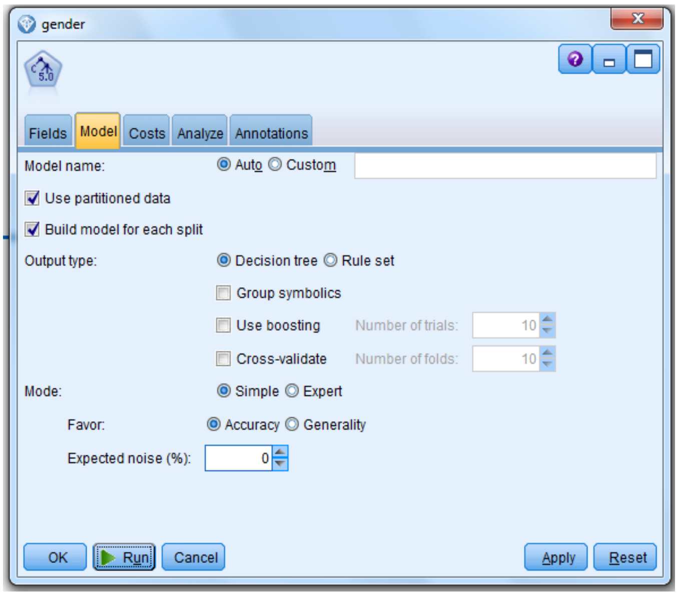
Splits:

□ Use weight field

Fig. 3. Feature Selection

The above Fig. 2, represents that the C5.0 algorithm requires the input data sets and the target fields, for classifying the datasets under some category and gives the output with the percentile of the fields that are related to the targeted field in a tree structure called a decision tree .

Here the products like fruits and vegetables, fresh meat , dairy , soft drink are set as the inputs and the gender field is taken as the output. It produces the decision tree in which the data is classified under the splitting criteria called the information gain.



gender

Fields | Model | Costs | Analyze | Annotations

Model name: ● Auto ○ Custom

☑ Use partitioned data

☑ Build model for each split

Output type: ● Decision tree ○ Rule set

□ Group symbolics

□ Use boosting Number of trials: 10

□ Cross-validate Number of folds: 10

Mode: ● Simple ○ Expert

Favor: ● Accuracy ○ Generality

Expected noise (%): 0

Fig. 4.Decision Tree Node

The above Fig. 4, shows that the C5.0 algorithm can build decision tree as well as the rule set to decide what to do with the predicted data. The decision tree contains the n numbers of node according to the data set, in which the node which has the higher information gain will taken as the root node. Hence it is proposed that the data sets are classified under the category of products like soft drinks, fruit and vegetables, fresh meat, and dairy products.

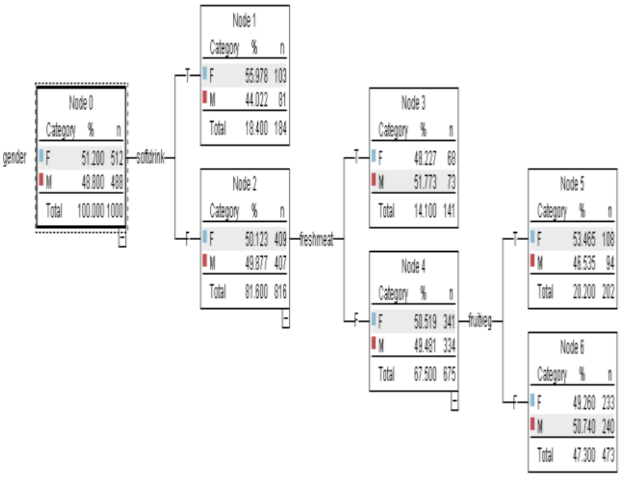


Fig. 5. Data Set Classification

The above Fig. 5, represents the classification of the products according to the target field gender. From the tree, it is known that out of 100% about 51.2% of females and 48.8% of males, purchase the products and out of which 55.978% of females and 44.022% of males (i.e.)out of 100% about only 18.4% of people buys soft drinks, and 81.6% of peoples doesn't buy soft drinks. So the C5.0 algorithm is used for the owner of the super market to make further decisions in order to improve their business strategy. From the above details it is known that the product named soft drinks are not mostly bought by the peoples. Likewise, it is predicted for each and every product, which is further used for implementing their business plans.

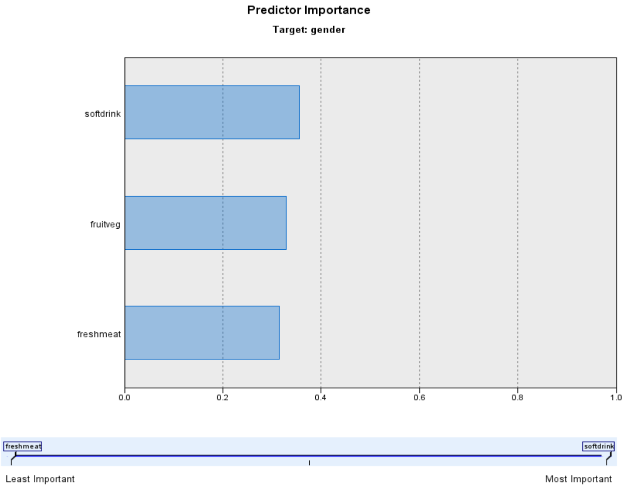
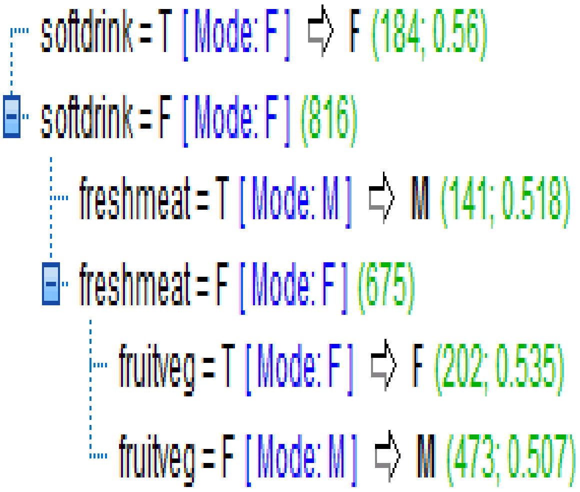


Fig. 6. Predictor Impotance

The above Fig. 6, shows that in accordance with the calculation of the information gain or the gini index the predictor importance of each and every fields are graphed above. By the above analysis it was said that the soft drink product gives the more information gain for finding the target customers of a super market data sets, and next comes the fruits and vegetables.



softdrink = T [Mode: F] ⇨ F (184; 0.56)

softdrink = F [Mode: F] (816)

freshmeat = T [Mode: M] ⇨ M (141; 0.518)

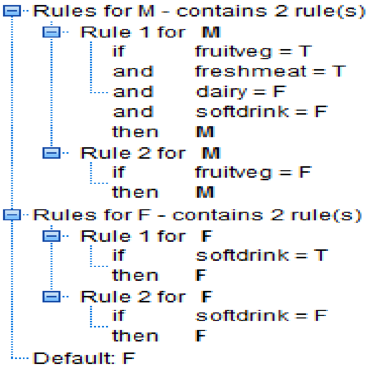
freshmeat = F [Mode: F] (675)

fruitveg = T [Mode: F] ⇨ F (202; 0.535)

fruitveg = F [Mode: M] ⇨ M (473; 0.507)

Fig. 7. Rule Generation

Eventually, a well known decisions are found with the help of the above summary (Fig.7.). The soft drinks were bought by many females rather than males, and the products like the fresh meat were bought by the females. Using these kind of decisions, owners of super markets can be aware of business strategy and the gender target, So that the owners will target specific gender for the specific product sales.



```
Rules for M - contains 2 rule(s)
    Rule 1 for  M
        if      fruitveg = T
        and     freshmeat = T
        and     dairy = F
        and     softdrink = F
        then    M
    Rule 2 for  M
        if      fruitveg = F
        then    M
Rules for F - contains 2 rule(s)
    Rule 1 for  F
        if      softdrink = T
        then    F
    Rule 2 for  F
        if      softdrink = F
        then    F
Default: F
```

Fig. 8. Rule Set

The above Fig. 8, represents the rule set that is concluded from the decision tree. (i.e.) If a person buys the products like fruits and vegetables, and fresh meat then that person might be the male. Likewise the C5.0 algorithm will also helps to predict the rule set using the business sales data.

V. CONCLUSION

This study has shown that the entrepreneur will able to identify their business tactics effectively. It is concluded that about 81.6% of accuracy has been obtained. This algorithm improves analytics strategy and it offers the powerful boosting method for improving the accuracy of classification. In this project, it is proved that C5.0 decision tree algorithm is useful in finding a target group of customers. From the purchase of items, it is possible to recognize the target group of customers to promote marketing strategy for new item. This kind of data analysis very much useful in improving sales.

REFERENCES

1. Ashwini S., D.V. Veena .:" Relative investigation of machine learning algorithm for performance analysis on brain mr images". Procedia Computer Science Vol.143,pp. 552-562(2018).
2. Elvira P., Florin L.," Predicting academic performance based on learner traces in social learning environment". IEEE Access Vol..6,pp. 72774-72785(2018).
3. Shengqi Yang H W.,.Zhangqin F ., Xiaoyi Wang J H., "Type 2 diabetes mellitus prediction model based on data mining. Informatics in medicine unlocked", Vol.10,pp. 100-107(2018).
4. Mahmood H K ., Ahmed Zeki, M.," Prediction of urinary system disease diagnosis: a comparative study of three decision tree algorithms". In: INTERNATIONAL CONFERENCE ON COMPUTER ASSISTED SYSTEM IN HEALTH 2014, CPS, pp. 58-61. IEEE.
5. Muhammad A S., Eiji Hato.:" Classification of travel data with multiple sensor information using random forest.Transportation Research", Procedia vol.22,pp. 144-153(2017).
6. Rafik Khairul A., Indwiarti., S.Yuliant.:" Implementation of decision tree using c4.5 algorithm in decision making of loan application by debtor". In: 3rd INTERNATIONAL CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGY(ICoICT) 2015, pp. 75-80. IEEE.
7. Romana Markovic., Sebastian Wolf., Jun Cao., Eric Spinnraker., Daniel Wolki., Jerome Frisch., .T.K.Christoph .:" Comparison of different classification algorithm for the detection of user interaction with windows in office buildings. Energy Procedia", Vol.122, 337-342(2017).
8. Yurong Zhong.: "The analysis of cases based on decision tree". In: 7th INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING AND SERVICE SCIENCE(ICSESS) 2016, pp.142-147. IEEE.

AUTHORS PROFILE

Anitha received her B.E(CSE) degree in 2001 from National Engineering College. She pursued her M.Tech(Computer Engg) in 2005.She has completed her doctorate degree (Ph.D-Computer Engg) in 2012 with University Grants Commission – Senior Research Fellowship. She has received Rs.45 lakhs fund from Department of Science and Technology for FIST Projects. She is presently working as Professor in Francis Xavier Engg College. She is the coordinator for IBM Centre of Excellence. Here research area is Data Mining. She has made several SCI journal publications and many international conference publications.

Mathivanan R pursuing B.Tech(IT) degree in Francis Xavier Engineering college,Tirunelveli.He has participated in various hackathons and project contests conducted by various organizations.He is awarded INSPIRE(Innovation in Science Pursuit for Inspired Research)award for the year 2013-2014. He won various prizes in coding contests.He has attended various guesttalk and Conferences in various colleges.

Sundarraj R pursuing B.Tech(IT) degree in Francis Xavier Engineering College,Tirunelveli.He has participated in various hackathons and project contests conducted by various organizations.He had co-ordinated various events for junior students.

Indhulekha J pursuing B.Tech(IT) degree in Francis Xavier Engineering College,Tirunelveli.She won 2nd prize in Poster Presentation conducted by Francis Xavier Engineering College.She got First Class with Distinction in Typewriting(English).