# Analyzing the Health of Engineering Student's Using, Feature Selection Algorithms - A Comparative Study

**G. Ramya[1] S. Kavitha[2]**
[1,2]Research Scholar

*Abstract*— In this paper, data mining study about the Health of Undergraduate Engineering Students. This dissertation use the feature selection algorithm based on clustering methods to analyze the importance of health and well-being in students is exemplified by the large number of studies on this topic. Past research has focused on using surveys to identify factors that affect the health, but applying machine learning tools to such data has not received much attention. Moreover this dissertation presents the comparative study between the Correlation based feature selection algorithm and the Minimum Redundancy and Maximum Relevance feature selection (mRmR), these two algorithms can be applied to the engineering students based data, for analysis the health.

***Key words:*** collection, extraction, warehousing, analysis, statistics, interestingness metrics, complexity considerations

## I. INTRODUCTION

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amount of data, not the extraction of data itself. It also is a buzzword, and is frequently also applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence, machine learning, and business intelligence. The popular book "Data mining: Practical machine learning tools and techniques with Java" (which covers mostly machine learning material) was originally to be named just "Practical machine learning", and the term "data mining" was only added for marketing reasons. Often the more general terms "(large scale) data analysis", or "analytics" – or when referring to actual methods, artificial intelligence and machine learning – are more appropriate.

### *A. Process of Data mining*

Data mining is the process of analyzing the data from different perspectives and summarizing it into useful information. It allows users to analyze data from many different dimensions and categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining has popularly treated as a synonym of knowledge discovery in databases. In general, a knowledge discovery Process consists of data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, knowledge presentation.

Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data. It is important to know that classification rules are different than rules generated from association. Association rules are characteristic rules, but classification rules are prediction rules.

## II. METHODOLOGY

### *A. Background*

In this paper describe a comparative study in data mining using some classification techniques. It includes feature selection algorithms for analysis.

### *B. Feature Selection*

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many *redundant* or *irrelevant* features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points). The archetypal case is the use of feature selection in analyzing DNA microarrays, where there are many thousands of features, and a few tens to hundreds of samples.

### *C. Subset selection*

Subset selection evaluates a subset of features as a group for suitability. Subset selection algorithms can be broken up into Wrappers, Filters and Embedded. Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset. Wrappers can be computationally expensive and have a risk of over fitting to the model. Filters are similar to Wrappers in the search approach, but instead of evaluating against a model, a simpler filter is evaluated. Embedded techniques are embedded in and specific to a model.

Many popular search approaches use greedy hill climbing, which iteratively evaluates a candidate subset of

features, then modifies the subset and evaluates if the new subset is an improvement over the old. Evaluation of the subsets requires a scoring metric that grades a subset of features. Exhaustive search is generally impractical, so at some implementers (or operator) defined stopping point, the subset of features with the highest score discovered up to that point is selected as the satisfactory feature subset. The stopping criterion varies by algorithm; possible criteria include: a subset score exceeds a threshold; a program's maximum allowed run time has been surpassed, etc.

Two popular filter metrics for classification problems are **correlation** and **mutual information,** although neither are true metrics or 'distance measures' in the mathematical sense, since they fail to obey the triangle inequality and thus do not compute any actual 'distance' – they should rather be regarded as 'scores'. These scores are computed between a candidate feature (or set of features) and the desired output category. There are, however, true metrics that are a simple function of the mutual information.

*D. Correlation based feature selection Algorithm*

The Correlation Feature Selection (CFS) measure evaluates subsets of features on the basis of the following hypothesis: "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other". The following equation gives the merit of a feature subset *S* consisting of *k* features:

$$\text{Merit } s_k = \frac{K\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}}$$

Here, K is a number of attributes.

$\overline{r_{cf}}$ is the average value of all feature-classification correlations.

$\overline{r_{ff}}$ is the average value of all feature-feature correlations.

*E. Algorithm of CFS*

input: *S*(*F*1,*F*2, ...,*FN*,*C*) // a training data set
// a predefined threshold

output: *Sbest* // a selected subset

(1) **begin**
(2) for *i* = 1 to *N* do begin
(3) calculate *SUi*,*c* for *Fi*;
(4) if (*SUi*,*c* > □)
(5) append *Fi* to *S*0 *list* ;
(6) end;
(7) order *S*0*list* in descending *SUi*,*c* value;
(8) *Fj* = *getFirstElement*(*S*0*list*);
(9) do begin
(10) *Fi* = *getNextElement*(*S* 0*list* ,*Fj*);
(11) if (*Fi* <> NULL)
(12) do begin
(13) if (*SUi*, *j* _ *SUi*,*c*)
(14) remove *Fi* from *S*0*list* ;
(15) *Fi* = *getNextElement*(*S*0*list* ,*Fi*);
(16) end until (*Fi* == NULL);
(17) *Fj* = *getNextElement*(*S*0*list* ,*Fj*);
(18) end until (*Fj* == NULL);
(19) *Sbest* = *S*0*list* ;
(20) end;

*F. Minimum Redundancy-Maximum Relevance Feature Selection*

A feature selection method that can use mutual information, correlation, or distance/similarity scores to select features. The aim is to penalize a feature's relevancy by it's redundancy in the presence of the other selected features. The relevance of a feature set *S* for the class *c* is defined by the average value of all mutual information values between the individual feature $f_i$ and the class *c* as follows:

$$\min_{S\subset\Omega} \frac{1}{|S|^2}\sum_{j\in s} I(f_i, f_j)$$

Where $I(f_i,f_j)$ is mutual information between $f_i$ and $f_j$ and |s| is the number of features S. For classes $c=(c_i,....c_k)$ the maximum relevance condition is to maximize the total relevance of all features in S

$$\max_{S\subset\Omega} \frac{1}{|s|}\sum_{i\in S} I(c, f_i)$$

Here **C** is defined by the average value of all mutual information values between the individual features feature of (fi).We can obtain the mRMR feature set by optimizing these two conditions simultaneously, either in quatient form

$$\max_{S\in\Omega}\{\sum_i I(c, f_i)/[\frac{1}{|s|i,}\sum_{j\in s} I(f_i, f_j)]\}$$

Suppose that there are *n* full-set features. Let $x_i$ be the set membership indicator function for feature $f_i$, so that $x_i$=1 indicates presence and $x_i$=0 indicates absence of the feature $f_i$ in the globally optimal feature set. Let $c_i=I(f_i;c)$ and $a_{ij}=I(f_i;f_j)$.

## III. Problem Description

In this paper uses the two popular algorithms which mean correlation based feature selection (CFS). Minimum redundancy and maximum relevance feature selection algorithms are used to analysis the mental health of engineering students. These two algorithms select the Feature from the dataset based on the requirement.

In the correlation based feature selection algorithm can calculate the correlation value for all the attributes. And produce the accurate prediction for the analysis. Compare to the MRMR it just give a better and accurate result. It easily varies the accuracy differences between attribute. So that, the prediction can done easily. If the correlation value is either -1 or 1 that the subset may be consider as a required feature. Otherwise if the correlation value is 0 that the subset may not be consider which means not a relevant feature.

"Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other".

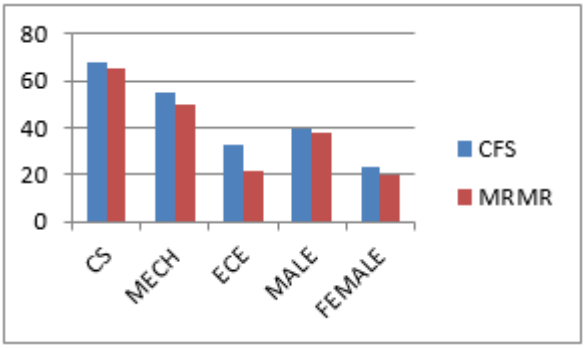
In the minimum redundancy and maximum relevance feature selection algorithm measure based feature selection algorithm can calculate the mutual information value between the features. Using this type of value it can predict the solution. Compare to correlation based feature selection algorithm, MRMR feature selection can give less accuracy.

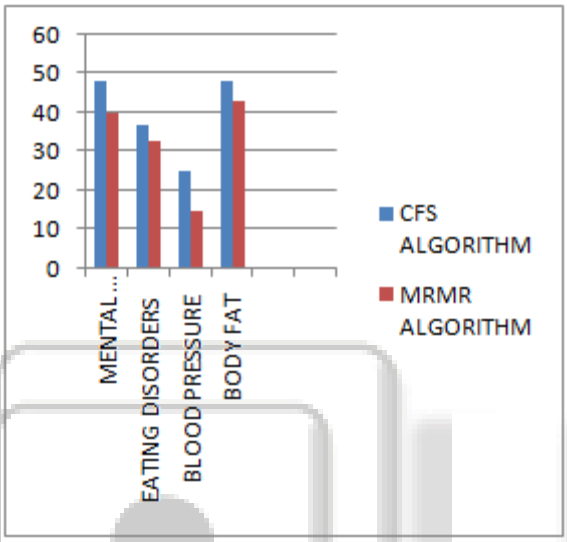
Here the paper use several data set about engineering students details like year of study, name, name of the department, blood pressure level, body fat level, no of

hours they sleep, no of hours their home works, about eating disorders.

*A. Comparative result*




## IV. Conclusion

In this paper it is expressed that the health of engineering student's data has been analyzed by using two algorithms viz. (Correlation based feature selection and minimum redundancy maximum relevance feature selection) and find out the remedy measures and had done a comparative study. Out of which it is proved the Correlation based feature selection gives a better and accurate result.

## V. Future Work

It is being experienced that most of the engineering students are so relaxed in their regular study while comparing their school level and not adhering routines. Moreover they are becoming an adults and spending their time and money as their desire and cultivating the irregularities.

Hence it is suggested imparting regular yoga education right from the beginning of academic session, conducting of various seminars on their personal health, hygiene & involve them in social, communal & spiritual activities. These activities may be reduced their irregularities, improvement in health, social responsibilities and creates the best citizen of United States of India.

## References

[1] forati, Alireza Moayekdikia, Andisheh Keikha A Novel Approach for Feature Selection based on the Bee Colony Optimization proceeding of International journals of computer applications(0975-8887) volume 43-No.8, April 2012.

[2] Monalisa Mandal, Anirban Mukhopadhyay An Improved Minimum Redundancy Maximum Relevance Approach for Feature Selection in Gene Expression Data proceeding of International conference on computational Intelligence (CIMTA) at procedia Technology 10(2013) 20-27

[3] Chia Huey Ooi*, Madhu Chetty and Shyh Wei Teng "Differential prioritization between relevance and redundancy in correlation-based feature selection techniques for multiclass gene expression data" published by BMC Bioinformatics 23 June 2006, 7:320 doi:10.1186/1471-2105-7-320.

[4] Megha Aggarwal, Amrita "Performance Analysis Of Different Feature Selection Methods In Intrusion Detection", proceeding of International Journal of Scientific & Technology Research volume 2, issue 6, june2013.

[5] Piyushkumar A. Mundra and Jagath C. Rajapakse "SVM-RFE With MRMR Filter for Gene Selection" From IEEE Transactions on Nano bioscience, volume 9, No.1, March 2010.

[6] Rajdev Tiwari, Manu Pratap Singh "Correlation-based Attribute Selection using Genetic Algorithm" International Journal Of Computer Applications(0975-8887) volume 4 – No.8, August 2010.

[7] Mital Doshi , Dr.Setu K Chaturvedi, Ph.D "Correlation Based Feature selection (CFS) Technique to Predict Student Performance". International Journal of Computer Networks & Communications (IJCNC) Vol.6, No.3, May 2014.

[8] Binita Kumari "Feature Subset Selection in Large Dimensionality using Correlation based GA-SVM". International journal of computer applications (0975-8887) volume 45- No.6, May 2012.

[9] Part pramokchon and Punpiti Piamsa-nga "An Unsupervised, Fast Correlation-Based Filter for Feature Selection for Data Clustering" Springer link Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)Lecture Notes in Electrical Engineering Volume 285, 2014, pp 87-94

[10] Habibu Rabiu*, M. Iqbal Saripan, Syamsiah Mashohor and Mohd Hamiruce Marhaban "3D facial expression recognition using maximum relevance minimum redundancy geometrical features" Rabiu et al. EURASIP Journal on Advances in Signal Processing 2012, 2012:213

[11] Lei Yu, Huan Liu "Efficient Feature Selection via Analysis of Relevance and Redundancy" Journal of Machine Learning Research 5 (2004) 1205–1224.

[12] Amira Sayed A. Aziz, Ahmad Taher Azar, "Genetic Algorithm with Different Feature Selection Tech"niques for Anomaly Detectors Generation" Proceedings of the 2013 Federated Conference on Computer Science and Information Systems pp. 769–774

[13] 1Pablo A. Estévez , Michel Tesmer , Claudio A. Perez , Jacek M. Zurada, "Normalized Mutual

Information Feature Selection" IEEE transactions on neural networks, vol. 20, no. 2, february 2009.
[14] Isabelle Guyon, Andr´e Elisseeff "An Introduction to Variable and Feature Selection" Journal of Machine Learning Research 3 (2003) 1157-1182.
[15] JianzhongWang, Shuang Zhou, Yugen Yi, and Jun Kong "An Improved Feature Selection Based on Effective Range for Classification" Hindawi Publishing Corporation Scientific World Journal Volume 2014, Article ID 972125.
[16] Pabitra Mitra, C.A Murthy, Sankar K. Pal " Unsupervised Feature Selection using Feature Similarity" IEEE Transactions on pattern analysis and machine intelligence, Vol.24. No.3 March 2002.
[17] Hanchuan Peng, Fuhui Long, Chris Ding "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy" IEEE Transactions on pattern analysis and machine intelligence, vol. 27, no. 8, august 2005.
[18] Ms.Barkha Malay Joshi, G.B. Jethava, Hetal B.Bhavsar Ms.Barkha Malay Joshi et al. / International Journal of Engineering Science and Technology (IJEST) ISSN : 0975-5462 Vol. 4 No.05 May 2012 2022

