ISSN No:-2456-2165

Datamining Application for the Prediction of Binary Classification Problems

Soni PM
Professor
Dept of MCA
SNGIST
Ernakulam, India

Akshara Shylajan
DDMCA, D10
Dept of MCA
SNGIST
Ernakulam, India

Athira VP
DDMCA, D10
Dept of MCA
SNGIST
Ernakulam, India

Gopika Subhash
DDMCA, D10
Dept of MCA
SNGIST
Ernakulam, India

Shilpa Sebastian
DDMCA, D10
Dept of MCA
SNGIST
Ernakulam, India

Abstract:- Nowadays this world is flooded with lots of information such as environmental data, financial data etc. It very difficult for manually classifying, and summarize these data. Data mining find a solution for this. This will acquire useful models from bulky amount of data. Different data mining techniques involve classification, aggregation, clustering etc. By using classification we can group the data based on a common factor. It also helps in identifying which set of category a new observation belongs to. It is very important in classification to achieve greatest accuracy. Feature selection select relevant features that contribute most to our expected result according to some statistical score and remove redundant features. It is a data preprocessing task that improves the classification accuracy. In this research, different datasets are chosen from weka tool and they are breast cancer, diabetes, credit card loan payment and labor. Due to the complexities of database, it is arduous for making judgment. In order to solve this problem, this paper proposes a framework that combines feature selection and classification algorithms. The result shows that our approach will not only reduce the size of data but also provide elevated classification accuracies than other methods. The application is developed as a website using python.

Keywords:- *Data Mining; Classification; Feature Selection.*

I. INTRODUCTION

Data mining is a process of extraction of useful information and patterns from huge data [1].Data mining is the process of applying these methods to data with the intention of uncovering hidden patterns. Data mining is being used in several applications like banking, insurance, and hospital and Health informatics [2] [3]. In latest years, due to the ease of use of outsized amount of data, a procedure called data mining has engrossed a lot of concentration in society and also in research area. It is a process which will find helpful information and samples from hefty amount of data that is helpful for resolution.

Main purpose of this technique is to extract new patterns and help in decision making. This method is chiefly based on machine learning algorithms, and it focuses on prediction of objects value or class affiliation based on its features. Classifications, aggregation, prediction, clustering, are the major functionalities of data mining.

Classification is a data analysis task. It took a major role in field of data mining, statistics, and neural network over years. By using classification we can group items. When a new observation arrived, by using classification we can decide to which class the new item belong to. It is very important in classification to achieve maximum accuracy. Each classification technique has its own merits and demerits. Some classification technique work well with this certain dataset and some others work well with other sets. One of such classification technique is feature selection, which will select best features from features available in dataset. Feature selection technique is used for selecting subset of relevant features from the data set to build robust learning models [4]. In this paper , a comparative study of various classifiers and feature selection methods were carried out and found out suitable classifier and feature selection method for the different data sets such as credit, labour, diabetics and breast cancer.

In section II, we discussed about various datasets and tools used in this experiment.. In sectionIII we made a discussion on the major concepts in the research that are classification, prediction, feature selection etc…Section IV suggested a proposed model for the experiment. . In section V a detailed analysis and discussion about the result obtained from the experiment is done. The project is implemented as website in python using Django Frame work and is named as “Prediction system”. Implementation is described in section VI Conclusion is given in section VII and references in section VIII.

ISSN No:-2456-2165

II. DATASET AND TOOLS

A. Dataset

For this experiment different datasets selected from weka tool are breast cancer, diabetes, credit and labor.

➢ *Diabetics*

Diabetics is a diseases that arise due to heavy blood glucose, or due to production of insulin is not sufficient, or because the body's cells are not properly react to insulin, or both. Heart problems, kidney failure, blindness, nerve injury and blood vessels harm etc. can be caused because of diabetes. So it is necessary to detect the diabetes in its early stages.

FEATURE ID	FEATURE NAME
F1	Pregnant
F2	Plasma glucose
F3	Diastolic Blood Pressure
F4	Triceps Skin Fold Thickness
F5	Serum-Insulin
F6	Body Mass Index
F7	Diabetes Pedigree Function
F8	Age
Class	Diabetic or Non- Diabetic

Table 1:- Dataset of Diabetic Patients

➢ *Breast cancer*

Breast cancer is the mainly frequent cancer among womans. If cancer is detected earlier that is on starting stage then it can be cured. But the problem is its prediction is very difficult. So it is important to predict the cancer by the symptoms as early. Risk factor of breast cancer can be genetic, or some life styles such as alcohol intake, can make it more likely to happen. For this we can use data mining techniques that is classification and feature selection which will help for the prediction easily

FEATURE ID	FEATURE NAME
F1	Age
F2	Menopause
F3	Tumor-size
F4	Inv-nodes
F5	Node-caps
F6	Deg-malig
F7	Breast
F8	Breast-quad
F9	Irradiat
F10	Class

Table 2:- Datasets of Breast Cancer Patients

➢ *Credit*

Credit data set is used for credit risk assessment. Many peoples and business organization will apply for different type of loan from bank. So bank should decide whether to allow loan or not based on certain conditions and factors However, due to the complexity of database, judgement making is difficult for the credit managers.

FEATURE ID	FEATURE NAME
F1	checking_status
F2	duration
F3	credit_history
F4	purpose
F5	credit_amount
F6	savings_status
F7	employment
F8	installment_commitment
F9	personal_status
F10	other_parties
F11	residence_since
F12	property_magnitude
F13	age
F14	other_payment_plans
F15	housing
F16	existing_credits
F17	job
F18	num_dependents
F19	own_telephone
F20	foreign_worker
F21	Class of loan good/bad

Table 3:- Datasets of Credit Risk Assessment

➢ *Labor*

Labor is the sum of substantial, rational, and communal stab used to produce possessions and services in the economy. It equipment the effort, knowledge, and service wanted to change raw materials into completed products and services. We will decide on whether that labor should continue in that company according to his or her performance.

ISSN No:-2456-2165

FEATURE ID	FEATURE NAME
F1	Duration
F2	Wage-increase-first-year
F3	Wage-increase-second-year
F4	Wage-increase-third-year
F5	cost-of-living-adjustment
F6	working-hours
F7	Pension
F8	standby-pay
F9	shift-differential
F10	education-allowance
F11	statutory-holidays
F12	Vacation
F13	longterm-disability-
F14	contribution-to-dental-plan
F15	bereavement-assistance
F16	contribution-to-health-plan
F17	Class

Table 4:- Datasets of Labor

B. Weka Tool

Weka is an abbreviation of Waika to Environment for knowledge Analysis. It is a machine learning software developed at the University of Waikato, and written in Java. It is free software supported in Linux, windows, OS x operating system. Licensed by GNU General public .Weka contains apparatus for classification, clustering, association, regression, data pre-processing and visualization. It is very useful for developing new machine learning schemes. In order to download the weka first of all open the disk image and drag the standalone version of Weka into your Applications folder. By double clicking on the weka.jar file start weka. The download takes 120 megabytes. In weka we can pre-process a dataset, feed it into a learning scheme, and analyse the resulting classifier and its performance, all without writing any program code at all. Getting to know data is integral part of the work, and many data visualization facilities and data pre-processing tools are provided [5].

Features of Weka tool are:

- Free availability
- Collection different techniques.
- Easy to use...
- Portability

The implementation of the project were done in python,

III. CONCEPTS USED

A. Classification

Classification is the mainly used data mining technique. By using classification, we can group features. And also can make out to which group a new observation can fit in to. Learning and classification are the two methods entail in classification process. In Learning, by using classification algorithm the samples are scrutinized. It is a Data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts[6]. In Classification, to calculate more or less the accuracy of the classification rules, test data are used.

Variety of classification model:

- Decision tree induction
- Bayesian classification
- Neural Network
- Support Vector Machines
- Classification based on association

B. Prediction

Regression technique can be personalized for prediction. To mock-up the association between one or more dependent attributes and independent attribute regression analysis is used Already acknowledged attributes are independent and what we want to forecast are response variable. Many real-world problems are not simply prediction. Due to relying on compound connections of manifold predictor variables sales quantity, stock amount, and product malfunction rates are all very difficult to envisage.

Types of regression methods

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

C. Feature Selection

It is a technique used for selecting a compartment of chief features from our data set. By eliminate most immaterial and unneeded features from the dataset classification accuracy can be improved

Feature selection consists of three steps [7].

- Screening: It removes inappropriate and problematical predictors. Problematic predictors are one with a lot of missing values.
- Ranking: enduring predictors are sorted and consign ranks to features based on its magnitude.
- Selecting: It makes out the rift of features by keeping the most imperative feature and sieve all others.
- Three methods in feature selection.

- Filters
- wrappers
- Embedded.

In general, feature subset selection methods, which were derived from the evaluation function, have been classified in two broad categories as filter and wrapper methods [8]

➢ *Filter Method*

Filter methods are used as a preprocessing step. A large set of features is available for us .We can sort the features based on some values that they are obtained in several statistical tests. The disadvantage of filter approach is that the features could be correlated among themselves. The methods that use the filter approach are independent of any particular algorithm as the function that they use for evaluation relies completely on properties of the data [9].

➢ *Wrapper Method*

In wrapper method we train a model using use a subset of features. We can insert or eliminate features from the selected subset by using some assumptions we draw from the previous model. This method is usually very classy. Some paradigm for wrapper methods:

- forward feature selection
- backward feature elimination
- Recursive feature elimination.

➢ *Forward Selection*:

It is an iterative method. When we start this method, there is no feature in. after each iteration we add new feature to the set .This will repeat until certain conditions met.

➢ *Backward Elimination*:

In this method , we start with asset of features and After each iteration some features will remove. The removed features are one with low significance. This will continue until a removal will not improve the performance.

➢ *Recursive Feature Elimination*:

In this method after each iteration a best and bad model will create. After completing execution it create a model with the remaining features.

➢ *Embedded Method*

It unite the features of both filter and wrapper methods. It put into practice by algorithms that have their own built-in feature selection methods. The combination of wrapper and filter approach is known as hybrid method [10].

IV. PROPOSED MODEL

Proposed model of this research work is portraied in figure 1. The four different data sets were classified using different classification algorithms in weka. Kstar, Naive bayes, J48, random forest, and JRip were experimented and found the best algorithms that produce highest accuracy for a specific data set. Again the same data set were classified using the algorithm that produces highest accuracy after applying feature selection. Various feature selection methods were available in weka such infogain, gainratio, classifier, cfssubset, correlation attribute etc. The proposed model found the best feature selection method as well as classification method and use these combinations to predict the behavior of such binary classification problems in the corresponding application..

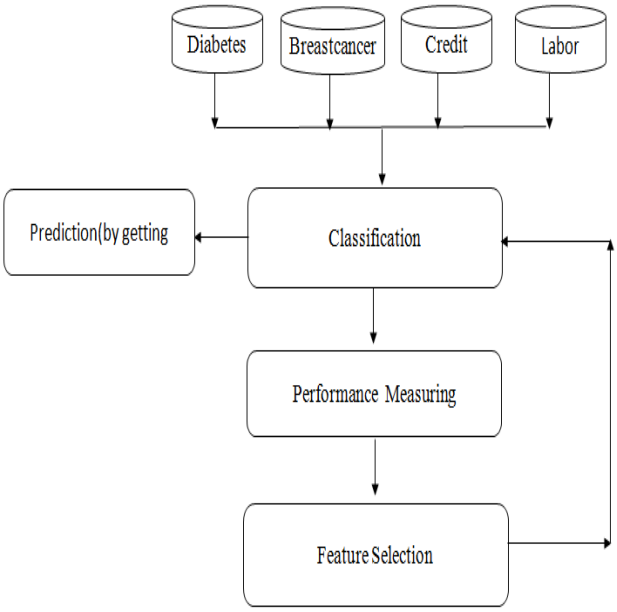

Fig 1:- Proposed Model Diagram

V. RESULTS AND DISCUSSION

Table 5 represents different accuracies obtained by different data sets using different classification algorithms such as Naïve Bayes, RandomForest, J48, Jrip,Kstar and OneR. From the table VI we can found out that it is important to apply feature selection to achieve improved accuracy in classification. Table 6 represents the improvement of classification accuracy after various feature selection methods such as InfoGain, GainRatio, Classifier, cfssubseteval and coorelation. The graphical representation of accuracies before feature selection with various classifiers are depicted in figure 2 and accuracies after feature selection with best classifier in figure 3. The application is developed using the best feature selection method and best classifier for the corresponding problems.

Data sets	Naïve Bayes	Random Forest	J48	Jrip	Kstar	OneR
Breast cancer	71.68%	69.58%	75.52%	70.98%	73.43%	70.28%
Credit	75.40%	76.40%	70.50%	71.70%	69.40%	66.10%
Diabetes	76.30%	75.78%	78.83%	76.04%	70.88%	71.48%
Labor	89.47%	89.47%	73.68%	77.12%	89.47%	71.92%

Table 5:- Accuracy before Feature Selection

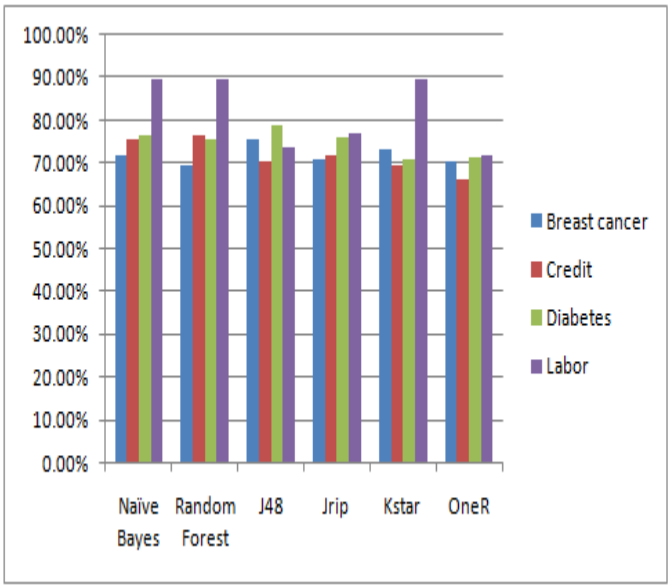

Fig 2:- Graphical Representation of Accuracy before Feature Selection

Feature selection	ACCURACY			
	Breast cancer	credit	Diabetes	Labor
InfoGainAttributeEval	80.07%	80%	82.61%	91.22%
GainRatioAttributeEval	75.17%	78.52%	79.31%	91.22%
ClassifierAttributeEval	75.87%	77.94%	80.07%	91%
CfsSubsetEval	75.87%	50.90%	79.69%	85.96%
CorrelationAttributeEval	77.27%	76.70%	81.60%	92.98%

Table 6:- Accuracy after feature selection

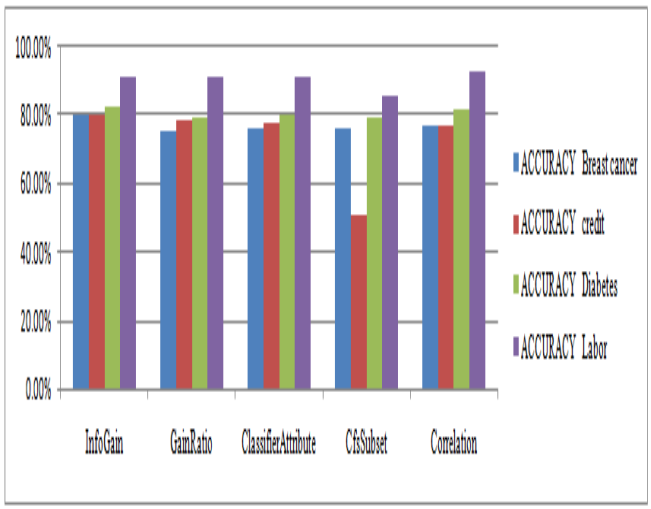

Fig 3:- Graphical Representation of Accuracy after Feature Selection

Finally we can conclude with best classifier and feature selection method for each dataset. The Table 7 displays each dataset with their suitable feature selection and classification methods.

Dataset	Feature selection	Classifier
Breast cancer	InfoGainAttributeEval	J48
Credit	InfoGainAttributeEval	Random Forest
Diabetes	InfoGainAttributeEval	Naïve Bayes
Labor	CorrelationAttributeEval	Kstar

Table 7:- List of Suitable Feature Selection and Classifier

VI. IMPLEMENTATION

The proposed model for finding the best classifier and feature selection method to predict the behavior of binary classification problems had been implemented as a website. The website was created using Python. Python is a programming language. Django, the most popular web framework is used to make the site. Python is very easy to use and it support multiple programming paradigms such as object oriented, function oriented programs. The website is named as “Prediction System”.

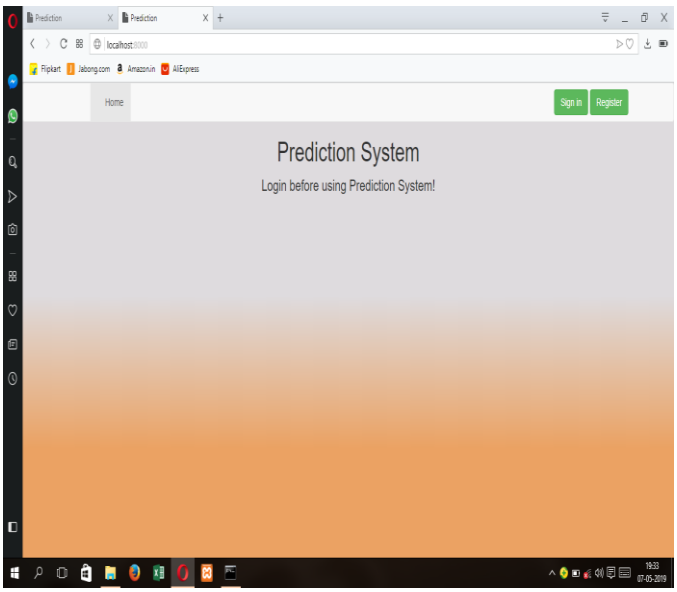

Fig 4:- Home Page

The website is created for predicting the behaviour on credit,labor,breast cancer and diabetic datasets. Fig 5 representing the main page , from there we can select the desired dataset.After selecting the intended data set ,navigate to a set of questions about choosen dataset. Here we explain the application using credit data set.. Fig 6

ISSN No:-2456-2165

representing the input screen for predicting the loan repayment capability behaviour of a customer . It consists of multiple choice question on credit data that helps to predict the customer as safe or risky.

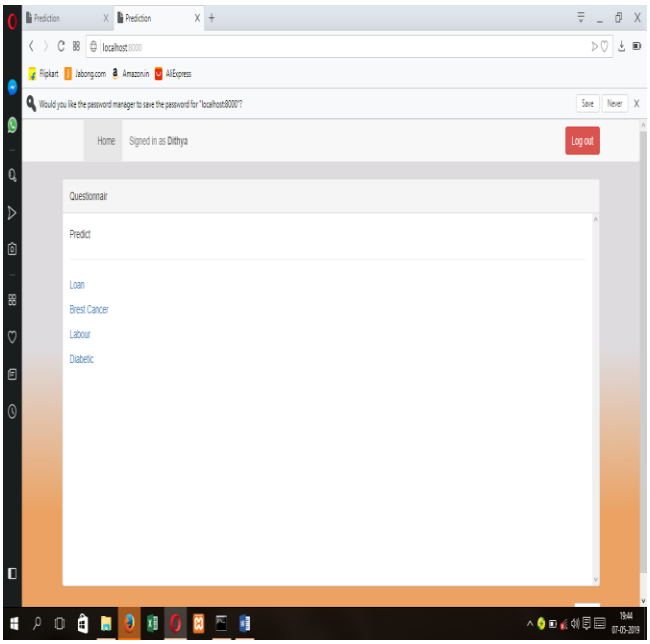

Fig 5:- Main Page

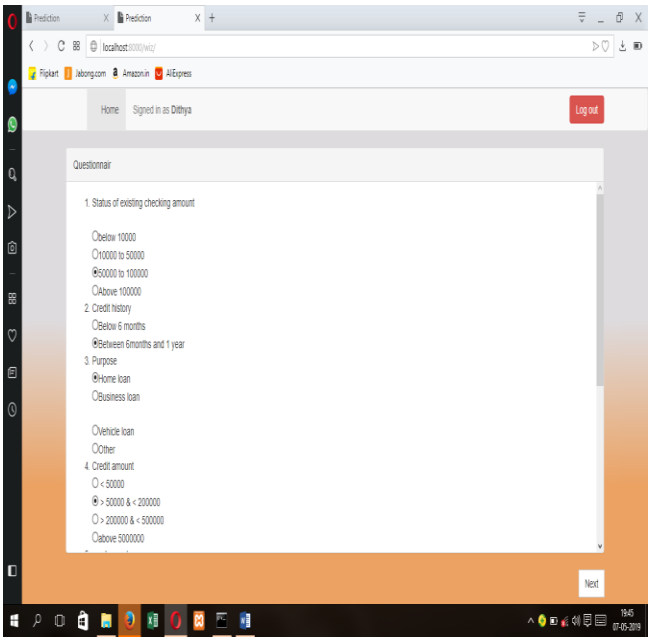

Fig 6:- Credit Data Entry Form

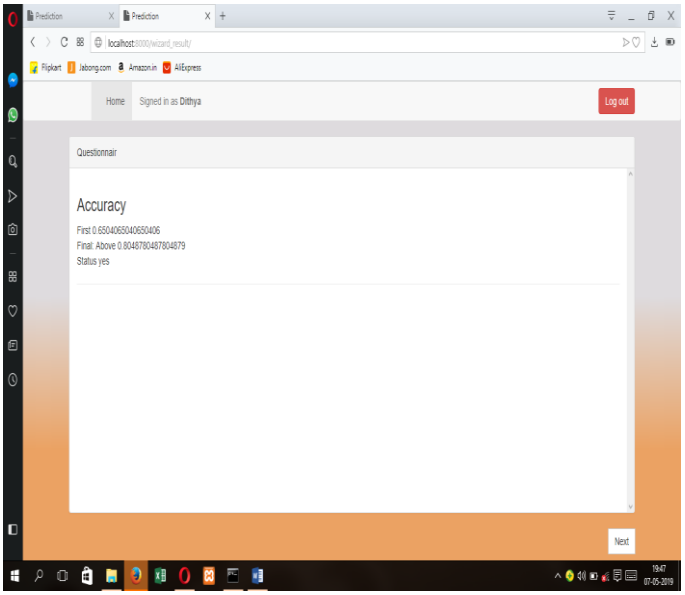

Fig 7:- Result Form

Based on how we answer the question, the accuracy will predict and suggest a solution for the problem. The accuracy for the credibility prediction and status is displayed in figure 7. The application can be used to solve multiple binary class problems such as breast cancer prediction, prediction of employment status and prediction of diabetics

VII. CONCLUSION

In this paper we introduced a comparative study about different feature selection methods and classifiers. Also created an application using the best classifier and best feature selection method for the specific dataset. This article outlines problem of multivariate data, for which to find patterns not seen by people, data mining techniques are used. The problem posed by the massive amount of data which must be analyzed by both people and data mining techniques. People face problems for determining patterns even if the dataset is not very big. The important feature selection method is one of the methods used to condense the amount of information analyzed via data mining. For conducting the work four data set such as credit, labor, breast cancer and diabetics were used. The classification accuracy will increase after making use of feature selection method. From the experiment it is found that Infogain and correlation are the feature selection techniques that produce improved classification accuracy.

REFERENCES

[1]. [DATA MINING TECHNIQUES AND APPLICATIONS Bharati M. Ramageri / Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305]
[2]. Z. Haiyang, "A Short Introduction to Data Mining and Its Applications", IEEE, 2011
[3]. Kritika Yadav and Mahesh Parmar: Analysis of Mahatma Gandhi National Rural Employment Guarantee Act Using Data Mining Technique.

International Journal of Computational Intelligence Research (IJCIR). 2017, Volume 13, Number 9 (2017), pp. 2221-2235, ISSN 0973-1873

[4]. [Improving Classification Accuracy by Using Feature Selection and Ensemble Model International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-2, May 2012]

[5]. Shilpa Dhanjibhai Serasiya and Neeraj Chaudhary," Simulation of Various Classifications Results using WEKA" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-1, Issue-3, August 2012

[6]. https://www.geeksforgeeks.org/basic-concept-classification-data-mining/

[7]. [Enhancing Classification Accuracy Through Feature Selection Methods Everton R. Reis1 , Paulo A. L. de Castro2 , Jaime S. Sichman1∗]

[8]. George John, Ron Kohavi and Karl Pfleger. Irrelevant Features and the subset Selection problem. Proceedings of the Eleventh International Conference on Machine Learning. pp. 121-129, New Brunswick, Morgan Kaufmann, 1994.

[9]. Brown G., Pocock A., Zhao M.J. and Lujan M. (2012). "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection", Journal of Machine Learning Research (JMLR).

[10]. Saeys, Y, Inza, I & Larrañaga, P 2007, „A review of feature selection techniques in bioinformatics. Bioinformatics", vol. 23, no. 19, pp.2507-2517