IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Analysis of Data Mining Algorithms Using Weka Rules ZeroR, JRip and NNge for Hypothyroid Disease Classification

Sushilkumar Rameshpant Kalmegh

Professor
Department of Computer Science, Sant Gadge Baba Amravati University,
Amravati (M.S.) 444 602, India

Abstract: In this paper data mining approaches for Hypothyroid disease classification are presented. The Weka for disease classification using data mining algorithms is described in this paper. The algorithm evaluation using Weka is described in this paper. The data mining algorithms Rules ZeroR, JRip and NNge are proposed for disease classification using Weka. The algorithms are evaluated using Weka. These algorithms can be used to classify the disease. These algorithms are compared on the basis of accuracy.

***Index Terms* - Classification, Hypothyroid, JRip, NNge, Weka, ZeroR**

I. Introduction

As the volume of data increases, inexorably, the proportion of it that people understand decreases alarmingly. Lying hidden in all this data is information. In data mining, the data is stored electronically and the search is automated or at least augmented by computer. Even this is not particularly new. Economists, statisticians, and communication engineers have long worked with the idea that patterns in data can be sought automatically, identified, validated, and used for prediction. What is new is the staggering increase in opportunities for finding patterns in data. Data mining is a topic that involves learning in a practical, non theoretical sense. We are interested in techniques for finding and describing structural patterns in data, as a tool for helping to explain that data and make predictions from it.

As the Internet of medical Things emerge in the field of medicine, the volume of medical data is expanding rapidly and along with its variety. Disease classification involves using machine learning algorithms to analyze and categorize data related to various diseases. The goal is to develop models that can accurately predict or classify the presence or absence of a particular disease based on input features. Disease classification helps organize information in a way that makes sense. By categorizing diseases, we create a systematic way to understand and communicate about them. It provides a common language for healthcare professionals to discuss and share information about diseases. In this given research paper Hypothyroid Disease data was used. Comparative analysis of ZeroR, JRip and NNge with test mode Use Training set was done using Weka Classifier Rules. This paper is organized into Six parts. First part discusses the Introduction followed by the literature required for analysis of methods implemented. Third one is System Design followed by datasets used for analysis. Fifth is the Performance Analysis and then Conclusion.

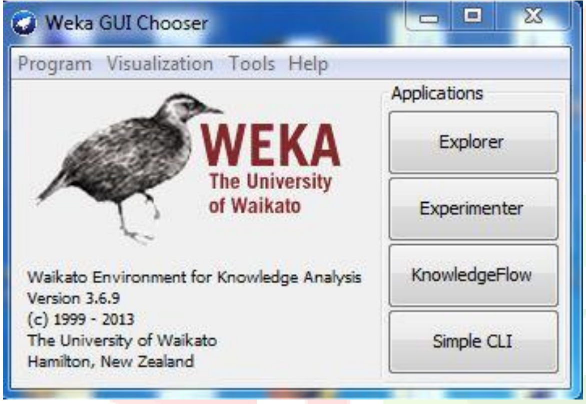
II. LITERATURE SURVEY

1.1 WEKA

Weka was developed at the University of Waikato in New Zealand; the name stands for Waikato Environment for Knowledge Analysis The system is written in Java. It runs on almost any platform and has been tested under Linux, Windows, and Macintosh operating systems and even on a personal digital assistant. The Weka workbench is a collection of state-of-the-art machine learning algorithms and data pre processing tools. It is designed so that we can quickly try out existing methods on new datasets in flexible ways. It provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning. As well as a variety of learning algorithms, it includes a wide range of pre processing tools. This diverse and comprehensive toolkit is accessed through a common interface so that its users can compare different methods and identify those that are most appropriate for the problem at hand. All algorithms take their input in the form of a single relational table in the ARFF format. The easiest way to use Weka is through a graphical user interface called Explorer as shown in figure 1. This gives access to all of its facilities using menu selection and form filling.

The Weka contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. Advantages of Weka include:

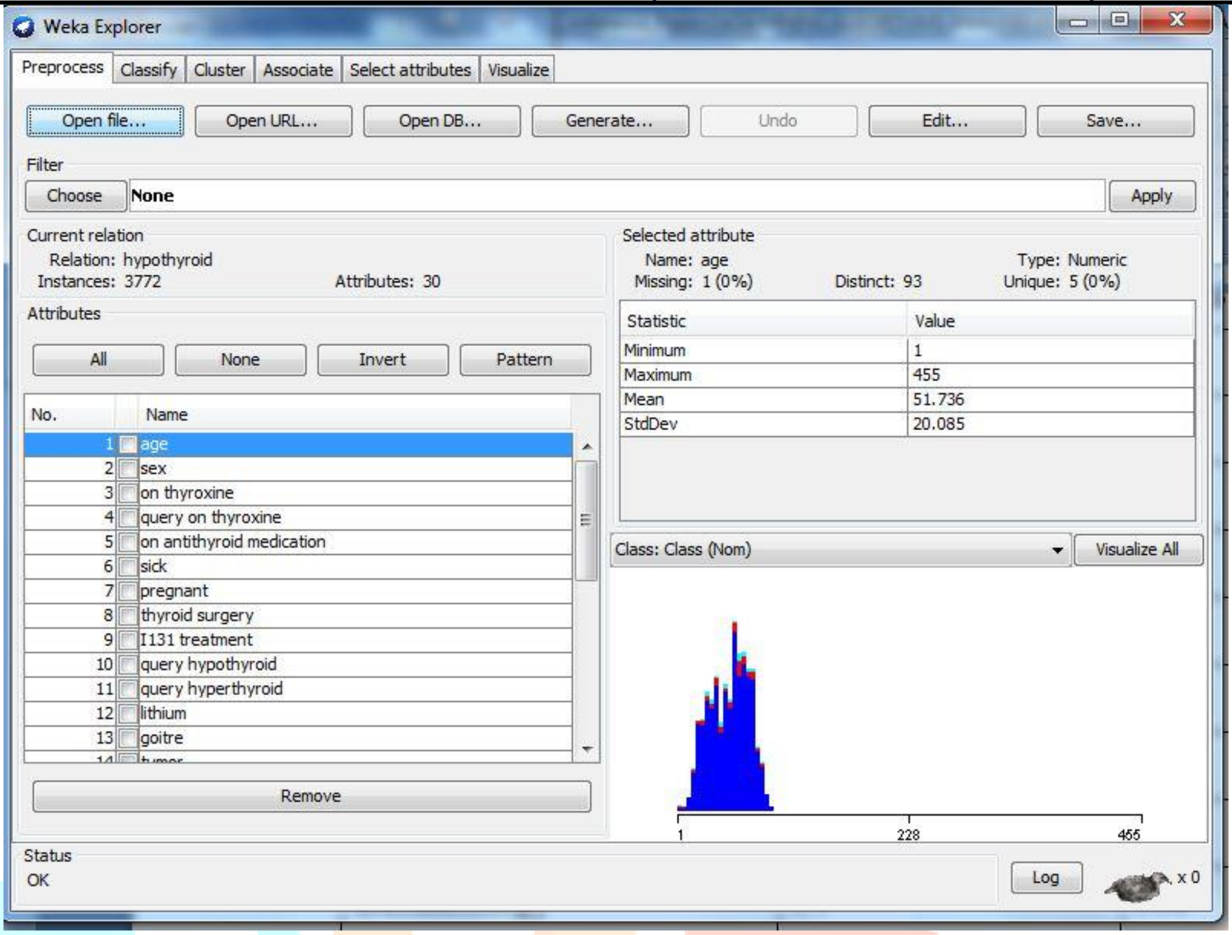
- Free availability under the GNU General Public License
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces.



Waikato Environment for Knowledge Analysis
Version 3.6.9
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand

Figure 1: WEKA GUI Explorer

Weka supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes. Weka provides access to SQL databases using Java Database Connectivity. Weka's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. The Explorer interface features several panels providing access to the main components of the workbench. Fig. 2 shows Opening of file hypothyroid.arff file by Weka Explorer and Fig. 3 shows processing of arff file for ZeroR Classifier. [1].



Filter: None

Current relation
Relation: hypothyroid
Instances: 3772
Attributes: 30

Selected attribute
Name: age
Type: Numeric
Missing: 1 (0%)
Distinct: 93
Unique: 5 (0%)

Statistic	Value
Minimum	1
Maximum	455
Mean	51.736
StdDev	20.085

Class: Class (Nom)

Attributes

No.		Name
1	□	age
2	□	sex
3	□	on thyroxine
4	□	query on thyroxine
5	□	on antithyroid medication
6	□	sick
7	□	pregnant
8	□	thyroid surgery
9	□	I131 treatment
10	□	query hypothyroid
11	□	query hyperthyroid
12	□	lithium
13	□	goitre
14	□	tumor

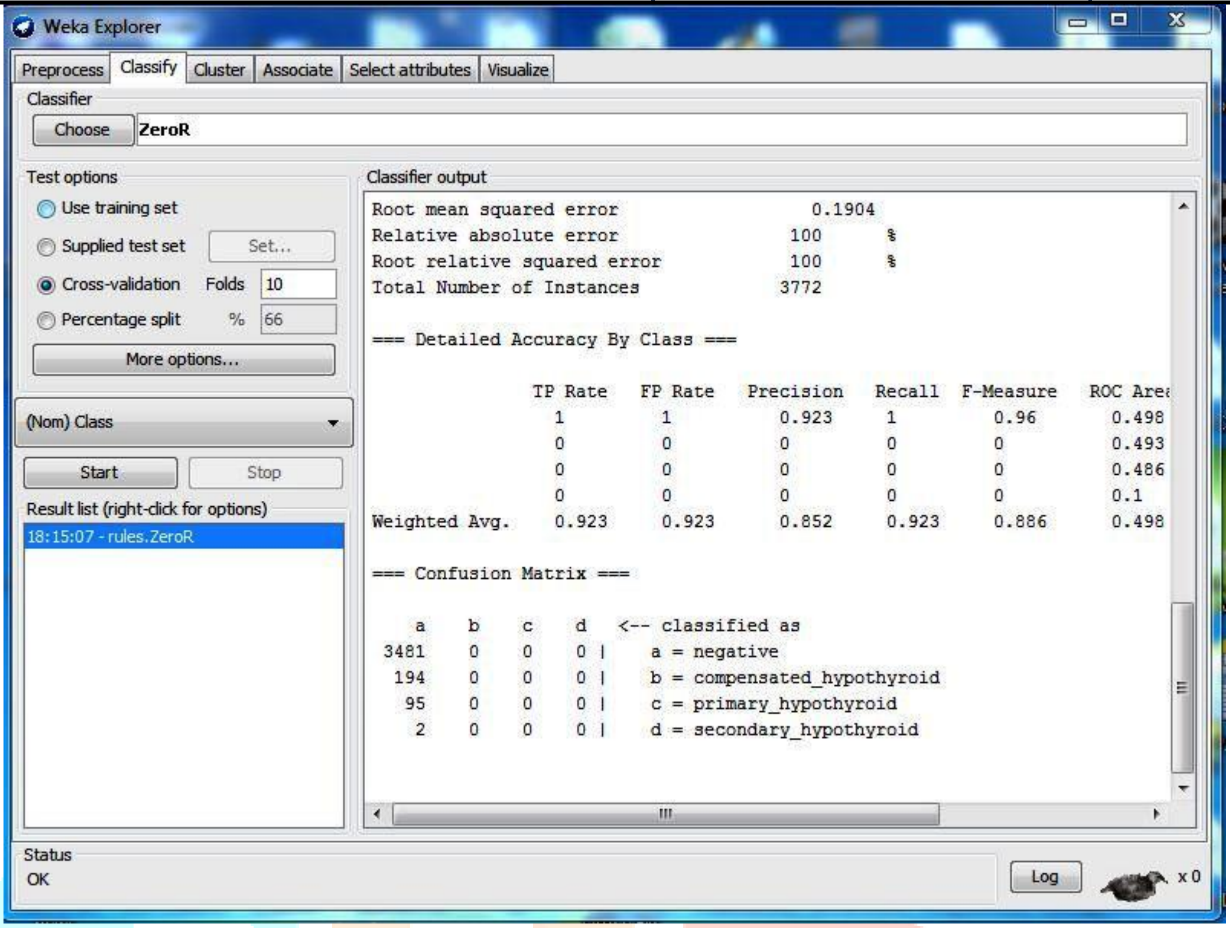
Status
OK

Figure 2: Opening of hypothyroid.arff file by Weka Explorer

1.2 Classification

Classification may refer to categorization, the process in which ideas and objects are recognized, differentiated, and understood. An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category. In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering or cluster analysis, and involves grouping data into categories based on some measure of inherent similarity. Classification is a data mining algorithm that creates a step-by-step guide for how to determine the output of a new data instance. The tree it creates is exactly that: a tree whereby each node in the tree represents a spot where a decision must be made based on the input, and to move to the next node and the next until one reach a leaf that tells the predicted output. Sounds confusing, but it's really quite straightforward.

There is also some argument over whether classification methods that do not involve a statistical model can be considered "statistical". Other fields may use different terminology: e.g. in community ecology, the term "classification" normally refers to cluster analysis, i.e. a type of unsupervised learning, rather than the supervised learning [1],[2].



Classifier: ZeroR

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) Class

Result list (right-click for options)

18:15:07 - rules.ZeroR

Classifier output

```
Root mean squared error                  0.1904
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              3772

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                   1         1         0.923     1         0.96       0.498
                   0         0         0         0         0          0.493
                   0         0         0         0         0          0.486
                   0         0         0         0         0          0.1
Weighted Avg.      0.923     0.923     0.852     0.923     0.886      0.498

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 3481    0    0    0 |    a = negative
  194    0    0    0 |    b = compensated_hypothyroid
   95    0    0    0 |    c = primary_hypothyroid
    2    0    0    0 |    d = secondary_hypothyroid
```

Status

OK

Figure 3 : Processing of arff file by ZeroR Classifier on Test Mode Use Training Set

1.3 ZeroR Classifiers

It is a rules-based classifier that is the simplest classification method which relies on the target and ignores all predictors. It simply predicts the majority class. Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods. It is a simple classification method that works with mode for the prediction of nominal data and mean for the prediction of numeric data. It is usually referred to as majority class method.[1], [5].

1.4 JRip Classifiers

This implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which is proposed by William W. JRip is an inference and rules-based learner (RIPPER) that tries to come up with propositional rules which can be used to classify elements. JRip implements heuristic global optimization of the rule set. Classes are examined in increasing size and an initial set of rules for a class is generated using incremental reduced-error pruning. An extra stopping condition is introduced that depends on the description length of the examples and rule set. JRip (RIPPER) proceeds by treating all the examples of a particular judgment in the training data as a class, and finding a set of rules that cover all the members of that class. Thereafter it proceeds to the next class and does the same, repeating this until all classes have been covered [1] [3] [4].

1.5 NNge Classifiers

NNge is a nearest-neighbor method for generating rules using non-nested generalized exemplars. Non-Nested Generalized Exemplars (NNGE) is an algorithm introduced by Brent, in 1995. It performs generalization by merging exemplars, forming hyper rectangles in attribute space that represent conjunctive rules with internal disjunction. The algorithm forms a generalization each time a new example is added to the database, by joining it to its nearest neighbor of the same class. The algorithm learns incrementally by first classifying, then generalizing each new example. When classifying an instance, one or more hyper rectangles may be found that the new instance is a member of, but which are of wrong class. The algorithm prunes these so that the new example is no longer a member. Once classified, the new instance is generalized by merging it with the nearest exemplar of the same class, which may be a single instance or a hyper rectangle [1] [3] [4].

1.6 Hypothyroid

Hypothyroid also called as thyroid disorder, the main grounds of thyroid is iodine deficiency. Almost one-third of the world's population lives in areas of iodine deficiency. Hypothyroidism is a clinical disorder commonly encountered by the primary care physician. Untreated hypothyroidism can contribute to hypertension, dyslipidaemia, infertility, cognitive impairment, and neuromuscular dysfunction. The prevalence increases with age, and is higher in females than in males. Hypothyroidism may occur as a result of primary gland failure or insufficient thyroid gland stimulation by the hypothalamus or pituitary gland. Thyroid diseases are the commonest endocrine disorders worldwide. Thyroid dysfunction is 10 times more common in women than in men. In India, it has been estimated that about 42 million people suffer from thyroid diseases. Universal salt iodization program in India changed the thyroid status of India. Thyroid hormone is necessary for healthy growth and brain development, A condition famous as hypothyroidism occurs when the thyroid gland is not capable to produce adequate thyroid hormone. The grade of primary hypothyroidism is characterized by elevated serum levels of thyroid-stimulating hormone (TSH) and normal levels of free thyroxine (T4) and triiodothyronine (T3). [6], [7], [8], [9].

III. SYSTEM DESIGN

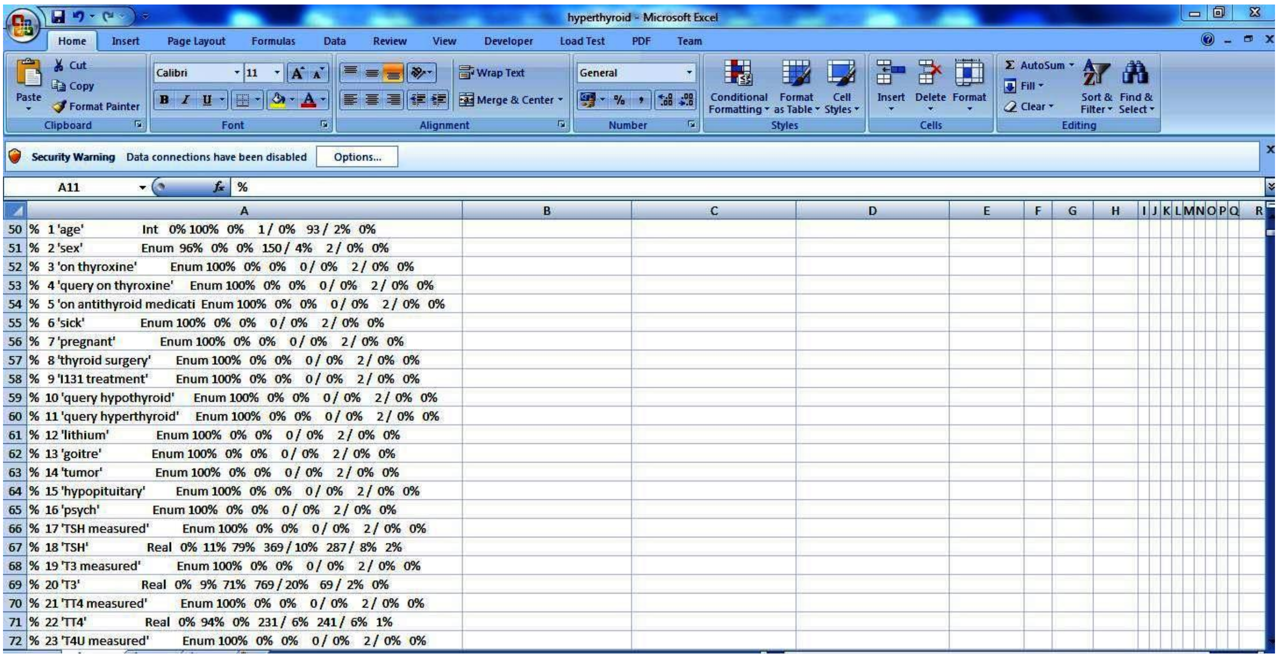
In order to co-relate hypothyroid with the categories, a model based on the machine learning was designed. As an input to the model, various number of hypothyroid features are considered which are available online. Around 3772 hypothyroid samples were collected on above repository using internet. The hypothyroid samples then separated into 4 categories Negative, Compensated Hypothyroid, Primary Hypothyroid, Secondary Hypothyroid. Finally classification is processed using WEKA Explorer. Due to classification in above 4 categories we are also able to find the Negative, Compensated Hypothyroid, Primary Hypothyroid, Secondary Hypothyroid on every data set which help to develop models that can accurately predict or classify the presence or absence of a hypothyroid based on input features.

IV. DATA COLLECTION

Hence, it was proposed to generate hypothyroid data. Consequently the national and international resources were used for the research purpose. Data for the purpose of research has been collected from the various online resources using internet. Thyroid disorder records were supplied by the Garavan Institute and J. Ross New South Wales Institute, Syndney, Australia. There are 3772 hypothyroid samples in total with 30 Attributes. The details are as shown in following table 1 & Fig. 4.

Table 1: Collection of hypothyroid Dataset

Name of Disease	Number of Features	Number of Samples
Hypothyroid	**30**	**3772**



Security Warning Data connections have been disabled Options...

A11 %

```
50 %  1 'age'              Int   0% 100%  0%    1 / 0%   93 / 2%  0%
51 %  2 'sex'              Enum  96%  0%  0%  150 / 4%    2 / 0%  0%
52 %  3 'on thyroxine'     Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
53 %  4 'query on thyroxine'  Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
54 %  5 'on antithyroid medicati  Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
55 %  6 'sick'             Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
56 %  7 'pregnant'         Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
57 %  8 'thyroid surgery'  Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
58 %  9 'I131 treatment'   Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
59 % 10 'query hypothyroid'   Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
60 % 11 'query hyperthyroid'  Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
61 % 12 'lithium'          Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
62 % 13 'goitre'           Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
63 % 14 'tumor'            Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
64 % 15 'hypopituitary'    Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
65 % 16 'psych'            Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
66 % 17 'TSH measured'     Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
67 % 18 'TSH'              Real   0% 11% 79%  369 / 10%  287 / 8%  2%
68 % 19 'T3 measured'      Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
69 % 20 'T3'               Real   0%  9% 71%  769 / 20%   69 / 2%  0%
70 % 21 'TT4 measured'     Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
71 % 22 'TT4'              Real   0% 94%  0%  231 / 6%   241 / 6%  1%
72 % 23 'T4U measured'     Enum 100%  0%  0%    0 / 0%    2 / 0%  0%
```

Figure 4 : Figure showing Attributes of hypothyroid Dataset

V. Performance Analysis

The Data so collected need a processing. Hence as given in the system design phase, all the 3772 data were processed on Negative, Compensated Hypothyroid, Primary Hypothyroid, Secondary Hypothyroid into 4 categories. The test mode "Use Training set" used for ZeroR, JRip and NNge. For processing WEKA APIs were used. The following tables shows the Confusion Matrix and True positive (TP) and False Positive (FP) rate of ZeroR, JRip and NNge.

The following tables 2, 4 and 6 shows the result for the Confusion Matrix and the Tables 3, 5 and 7 shows True Positive and False Positive rate of ZeroR, JRip and NNge for test mode "Use Training set".

Table 2: Confusion Matrix for ZeroR for Test Mode: Use Training Set

Classified as →	Negative	Compensated Hypothyroid	Primary Hypothyroid	Secondary Hypothyroid
Negative	3481	0	0	0
Compensated Hypothyroid	194	0	0	0
Primary Hypothyroid	95	0	0	0
Secondary Hypothyroi	2	0	0	0

Table 3: TP and FP Rate of ZeroR for Test Mode: Use Training Set

Class ↓	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Negative	1	1	0.923	1	0.96	0.5
Compensated Hypothyroid	0	0	0	0	0	0.5
Primary Hypothyroid	0	0	0	0	0	0.5
Secondary Hypothyroid	0	0	0	0	0	0.5
Weighted Avg	0.923	0.923	0.852	0.923	0.886	0.5

Table 4: Confusion Matrix for JRip for Test Mode: Use Training Set

Classified as →	Negative	Compensated Hypothyroid	Primary Hypothyroid	Secondary Hypothyroid
Negative	3474	1	6	0
Compensated Hypothyroid	1	190	3	0
Primary Hypothyroid	0	0	95	0
Secondary Hypothyroid	2	0	0	0

Table 5: TP and FP Rate of JRip for Test Mode: Use Training Set

Class ↓	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Negative	0.998	0.01	0.999	0.998	0.999	0.994
Compensated Hypothyroid	0.979	0	0.995	0.979	0.987	0.997
Primary Hypothyroid	1	0.002	0.913	1	0.955	0.999
Secondary Hypothyroid	0	0	0	0	0	0.539
Weighted Avg →	0.997	0.01	0.996	0.997	0.996	0.994

Table 6: Confusion Matrix for NNge for Test Mode: Use Training Set

Classified as →	Negative	Compensated Hypothyroid	Primary Hypothyroid	Secondary Hypothyroid
Negative	3481	0	0	0
Compensated Hypothyroid	0	194	0	0
Primary Hypothyroid	0	0	95	0
Secondary Hypothyroid	0	0	0	2

Table 7: TP and FP Rate of NNge for Test Mode: Use Training Set

Class ↓	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Negative	1	0	1	1	1	1
Compensated Hypothyroid	1	0	1	1	1	1
Primary Hypothyroid	1	0	1	1	1	1
Secondary Hypothyroid	1	0	1	1	1	1
Weighted Avg. →	1	0	1	1	1	1

VI. CONCLISION

The following table 8 shows the summary of Classification.

Table 8: Summary of Classification

Classifier	Rules.ZeroR	Rules.JRip	Rules.NNge
Test Mode	Use Training Set	Use Training Set	Use Training Set
Correctly Classified Instances	3481 (92.3%)	3759 (99.66%)	3772 (100%)
Incorrectly Classified Instances	291 (7.7%)	13 (0.34%)	00 (0%)

In this paper as per the previous performance analysis, Table 8 Summary of Classification shows that the Classifier NNge has the accuracy for test mode evaluate on training data is 100%, the Classifier JRip has accuracy for test mode evaluate on training data is 99.66% and the Classifier ZeroR has accuracy for test mode evaluate on training data is 92.3%.. This 100% accuracy for test mode evaluate on training data for the Classifier NNge is achieved due to the NNge algorithm forms a generalization each time a new example is added to the database, by joining it to its nearest neighbor of the same class. Overall Performance of JRip algorithm is acceptable, except some of instances from every category are classified into other category. This is because JRip implements heuristic global optimization of the rule set. Classes are examined in increasing size and an initial set of rules for a class is generated using incremental reduced-error pruning. The accuracy Classifier ZeroR for test mode evaluate on training data is worst as there is no predictability power in ZeroR. So it is concluded that Classifier NNge is the best classifier for classifying hypothyroid data.

From all the above result in the Table 2 to Table 8, it is observed that performance of Classifier NNge is Excellent as compared to Classifier ZeroR and JRip.

REFERENCES

[1] Ian H. Witten, Eibe Frank & Mark A. Hall. 2016, Data Mining Practical Machine Learning Tools and Techniques, (Third Edition, Morgan Kaufmann Publishers is an imprint of Elsevier).

[2] http://en.wikipedia.org/wiki/Classification.

[3] Sushilkumar R. Kalmegh, 2014, Successful Assessment of Categorization of Indian News Using JRip and NNge Algorithm, International Journal of Emerging Technology and Advanced Engineering (IJETAE), 4(12): 395-402

[4] Prof. Sushilkumar Rameshpant Kalmegh, 2022, COMPARATIVE ANALYSIS OF CATEGORIZATION OF CAR REVIEWS DATASET BY USING WEKA CLASSIFICATION ALGORITHM RULES NNGE AND JRIP, International Journal of Current Science (IJCSPUB), 12(1): 875-881

[5] Er. Daman Preet Kaur, Er. Parminder Singh, 2019, A Comparative Research of Rule based Classification on Dataset using WEKA TOOL, International Research Journal of Engineering and Technology (IRJET), 6(9): 2098-2102

[6] https://www.tandfonline.com/doi/abs/10.1080/20786204.2012.10874256

[7] Selvi Kumar, Pushpa Kotur, 2020, Effects of hypothyroidism in Indian women of reproductive age group – A review article, Indian Journal of Obstetrics and Gynecology Research, 7(1):1-6

[8] Gudisa Bereda, 2023, Definition, Causes, Pathophysiology, and Management of Hypothyroidism, Mathews Journal of Pharmaceutical Science, 7(1):1-5

[9] Mark P. J., Vanderpump, 2011, The epidemiology of thyroid disease, British Medical Bulletin, 99: 39-51