Table of Contents

List of Tables:

List of Figures:

2. EXPERIMENTATIONS AND RESULTS ANALYSIS

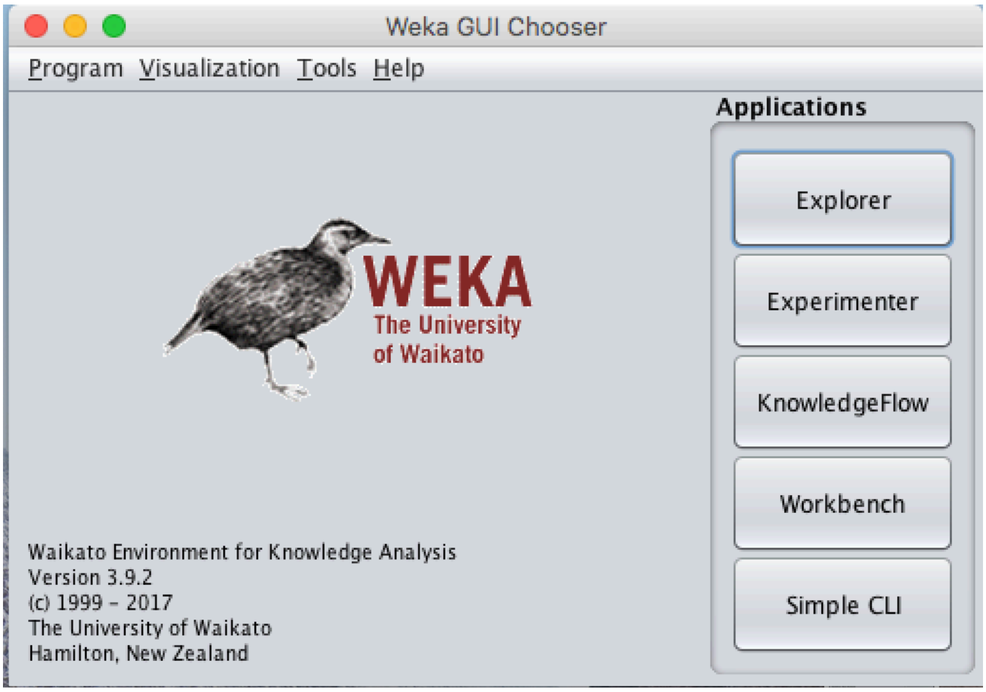
2.1 Introduction

This Section will examine and test the different ML algorithms that are based on rule induction this is since the API that is built to map back the results to the correspondent of the ASDTests app is built fundamentally on rules that generate if-then statements. All experiments were carried-out in WEKA (Figure 2.1) for the following reasons as discussed by (Brownlee, 2018), (i) equal comparison as all experiments are executed in a single platform (ii) a Graphical User Interface that sets the process of applied ML in a concise way (iii) free and open source, meaning it is immediately available for use and lastly (v) it contains cutting-edge algorithms with a comprehensive list of rule-based classifiers, as well as others. To uncover the predictive performance, a suite of evaluation metrics was engaged including predictive accuracy, precision, recall, and error rate. We have employed a recently published dataset from UCI to derive the rulesets from the rule-based algorithms.

2.2 Testing Tool

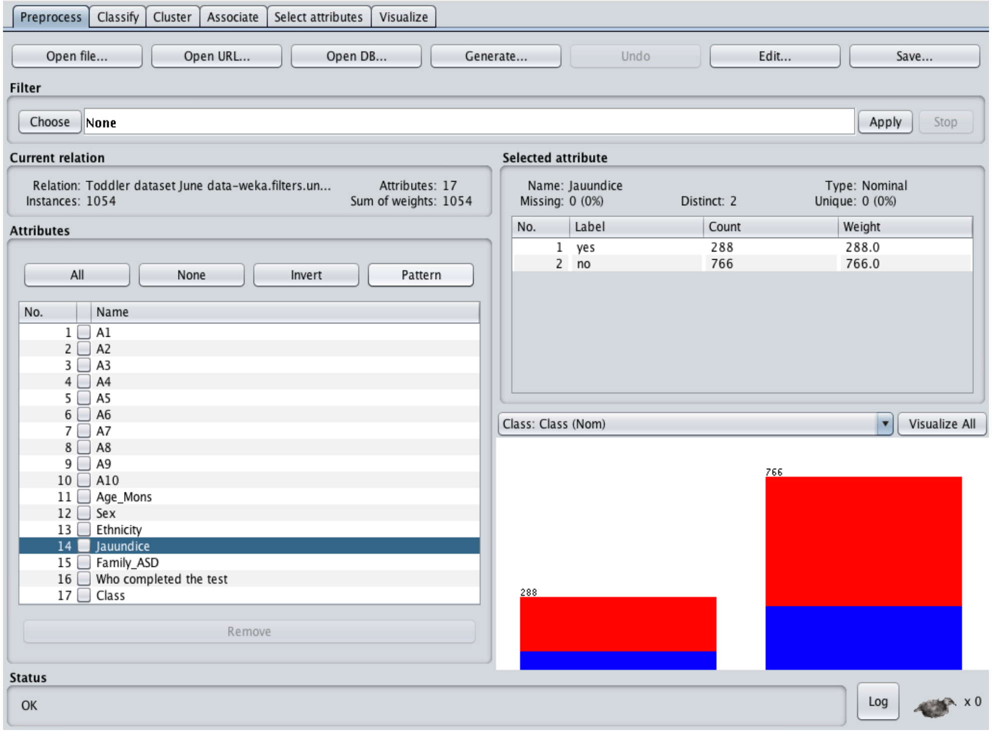
Applied ML employing all the rule-based classifiers were implemented in WEKA (Witten, Hall, & Eibe, 2016) . WEKA or Waikato Environment Knowledge Analysis was curated at University of Waikato, New Zealand and is built on java platform that is distributed under GNU (General Public License). The name is said to be pronounced to rhyme with Mecca a flightless bird only to be found in New Zealand. WEKA is tested to run in various mainstream Operating Systems such as Macintosh, Windows, and Linux. WEKA supports both supervised and unsupervised learning with different tasks such as pre-processing, classification, clustering, association, feature selection and visualisation and is primarily focused on predictive modelling.

Throughout the experimental stage, the Explorer module of WEKA was utilised as it can expose the structure of the problem to the predictive modelling process. The Explorer GUI is divided into 6 tabs (Figure 2.2) such as Preprocess, Classify, Associate, Attribute Selection, and Visualise also this module provides a descriptive summary of statistics once the toddler dataset is loaded. For the scope of the experimentation and classification task at hand we will be focused on the Preprocess, Classify and Select Attributes tab.



Waikato Environment for Knowledge Analysis
Version 3.9.2
(c) 1999 - 2017
The University of Waikato
Hamilton, New Zealand

Figure 2.1: Weka Main Graphical User Interface



Filter: None

Relation: Toddler dataset June data-weka.filters.un... Attributes: 17
Instances: 1054 Sum of weights: 1054

Name: Jauundice Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	yes	288	288.0
2	no	766	766.0

No.	Name
1	A1
2	A2
3	A3
4	A4
5	A5
6	A6
7	A7
8	A8
9	A9
10	A10
11	Age_Mons
12	Sex
13	Ethnicity
14	Jauundice
15	Family_ASD
16	Who completed the test
17	Class

Class: Class (Nom)

OK

Figure 2.2: WEKA Explorer Module

2.3 Methods Used

Two types of Feature Selection (FS) methods were engaged before the model tries to learn from the training data. These are (i) Correlation-based Feature Subset Selection or implemented as CfsSubsetEval in WEKA (CFS) and (ii) ChiSquaredAttributeEval (ChiSquared). These methods were selected for the following intention below;

a) CFS is a filter that works well with small datasets and does not invoke high computational costs linked with repeatedly imploring a learning algorithm and usually outperforms the wrapper with limited datasets (Hall, 1999). The toddler dataset contains 1054 examples and 17 features which is considered a small dataset, thus discovering the correlation of a feature with respect to the class variable would be efficient for the algorithm to learn from.

b) ChiSquared, as we aim to select the most predictive features for positive or negative labelled traits characterising autism according to a quality metric that measures the dependency between the feature and the target class.

The threshold for ChiSquared remained at default set to 1.79 in WEKA so any variables reaching a score above the minimum threshold will be obtained else it will be excluded. Fundamentally, all variables acquiring a score above zero will be presented to Doctors, Clinician, Psychologist or Medical Practitioners.

For learning the predictive models, several predictive rule-based algorithms were employed to examine the binary classification outcomes of autism traits. To enumerate these algorithms are PART (Partial Decision Trees), RIPPER (Repeated Incremental Pruning to Produce Error Reduction) implemented as JRip in WEKA, Furia (Fuzzy Unordered Rule Induction), Ridor (Ripple Down Rule Induction), PRISM, and DTNB (Decision Tree Naïve Bayes).

2.4 Evaluation Measures Used

In binary classification task, as a standard point of reference there are two classes labelled these are the positive class and the negative class. The positive class matches to an unusual case that our model is trying to predict that is characteristics of autism traits and is often uncommon than the negative class that is no features of autism. A classification task can be evaluated in numerous ways to achieve specific objectives as for the context of this study which falls on a medical case it is important that in practical problem solving we differentiate certain types of errors (Novaković́, 2017). To differentiate certain types of errors we then utilise the confusion matrix to efficiently assess the true performance of the predictive models. A confusion matrix summarises the performance of a model with regards to test data. It is a two-dimensional matrix, the true class of the test data is indexed in one dimension and the assigned model class is indexed on the remainder of the dimensions (Webb & Sammut, 2017). As discussed by (Miller & Forte, 2017), the following evaluation metrics that is a derivative of the confusion matrix are

(i) Precision (P), percentage of the quantity of correctly predicted instances of the positive class to the total number of predicted instances of the positive class.

(ii) Recall (R), quantity of correct predictions pertaining to the positive class over all the members of the positive class in the toddler dataset.

(iii) F_1-measure is the Harmonic Mean between P and R.

(iv) Accuracy, percentage of the quantity of correctly classified examples according to the total number of classified examples.

Table 2.1: Confusion Matrix for Classification of Autism Traits

Actual Class		Predicted Class	
		Negatives	Positives
	Negatives	TN (True Negatives)	FP (False Positives)
	Positives	FN (False Negatives)	TP (True Positives)

$$Precision = \frac{TP}{(TP+FP)} \tag{2.1}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{2.2}$$

$$F_1 - measure = 2 * recall * \frac{precision}{recall+precision} \tag{2.3}$$

$$error\ rate = \frac{(FN+FP)}{(FN+FP+TP+TP)} \tag{2.4}$$

$$Accuracy = 1 - Error\ rate \tag{2.5}$$

Throughout the training stages of the model, 10-fold cross-validation (cv) as illustrated in Table 2.2 was utilised as it offers unbiased results during the experimentations for the predictive classification models, as it ensures that predictive models derived are not overfitted and learning is not biased (Kohavi, 1995). The reason for having an unbiased model is the training set in cv is slightly smaller than the actual dataset. A technique used is k-fold cross-validation, data is split into k roughly equal-sized parts. Training is repeated k times, each time the selection process for the data is one allocated for testing and k-1 for training which gives an advantage of testing a segment of the data that is not involved in the training (Theodoridis, 2015).

Table 2.2: 10-fold Cross-Validation

Fold1	1	2	3	4	5	6	7	8	9	**10**	
Fold2	1	2	3	4	5	6	7	8	**9**	10	**Testing**
Fold3	1	2	3	4	5	6	7	**8**	9	10	Training
Fold4	1	2	3	4	5	6	**7**	8	9	10	
Fold5	1	2	3	4	5	**6**	7	8	9	10	
Fold6	1	2	3	4	**5**	6	7	8	9	10	
Fold7	1	2	3	**4**	5	6	7	8	9	10	
Fold8	1	2	**3**	4	5	6	7	8	9	10	
Fold9	1	**2**	3	4	5	6	7	8	9	10	
Fold10	**1**	2	3	4	5	6	7	8	9	10	

2.5 Results Analysis

In this section, the focal point will be on the performance of the rule-based classification models. Performance that is based on Precision, Accuracy, Recall and Error Rate by applying these evaluation metrics the model selection process would not be biased with just observing each of the models Accuracy. Model evaluation usually involves computing the level on which the classification model recommended utilising the assigned categorical label to a class that best matches the actual classification of the case (Novakovic´, 2017). The reason for implementing the evaluation measures is to have a comparison alley amongst the rule-based models and rank them according to the best performing classifier. Also, two viewpoints of analysis will be implemented in this section to realise which rule-based classifiers will best suit the classification problem (i) Analysis without Feature Selection (FS) (ii) Analysis with Feature Selection (FS).

2.5.1 Feature Selection Analysis

The feature selection methods that were applied area ChiSquaredAttributeEval and CfsSubsetEval and the reason for the selection is stated at Section 2.4. Chi-squared attribute evaluator employed Ranker as its search method and the preferred variables were ranked according to worthiness by computing the chi-squared statistic with respect to the dependent variable. The selected variables ranked accordingly are A9, A6, A5, A7, A4, A1, A2, A8, A3 and Jaundice. The highest rank of worthiness was obtained by A9, reaching a score of 351.32. While, the lowest score is Family_ASD at 0.19, describing if there is an inherent history of ASD in the lineage o the family. The dimensionality of the dataset was reduced to 11 variables from 17.

While the succeeding method is CfsSubsetEval (CFS) utilised Ranker as its search method. This Feature Selection technique selected the variables based from highest to lowest ranking are A1, A2, A4, A5, A6, A7, A8, A9, A10, and Jaundice. All variables were calculated with a merit of 100% except for A10 and Jaundice which received 60 and 50 percent respectively. Although, Jaundice was not selected initially by the FS method of ChiSquared but, it can be observed that CFS have elected the variable, which supplements the reason stated in previous section 2.4.

The selected variables above for the Toddler dataset would be applied for the ML algorithms such as rule-based classifiers to build the predictive model and extract the rule-sets that would be used to map back to QChat 10 and produce a comprehensive report to the correspondent of the ASDTests application.

Table 2.3 displays the summary of the selected variables for the Toddler dataset which lists the variable name and number of chosen variables upon application of the FS methods. Both the FS methods have selected the top 10 variables with different criteria as for ChiSquared the purpose of variable selection is for which attributes achieves a high score of worthiness and the dependent variable, Class was also included which, indicates a positive or negative label of Autism traits. While, for the CFS method it selects the variables based from highest to lowest ranking.

Table 2.3: Summary of Selected Features for the Toddler Dataset

DATASET	FEATURE SELECTION METHOD	# OF FEATURES	SELECTED FEATURES
TODDLER	Chisquared	10	A9, A6, A5, A7, A4, A1, A2, A8, A3, Jaundice
	CfsSubsetEval	10	A1, A2, A4, A5, A6, A7, A8, A9, A10, Jaundice

2.5.2 Error rate

Figure 2.3 displays the performance of the rule-based classification models grounded on the classifiers PART, JRip, Furia, Ridor, Prism and DTNB without Feature Selection application. The lowest Error (E) rate which is calculated in Equation 2.3 is achieved by DTNB at 2.85% and the highest E is obtained by Prism at 8.35%. Generally, all the classification models produced an error rate below 9% and has gained an average of 6.12% across the classifiers.

Table 2.4: Summary of Error Rate (E) without Feature Selection (FS)

	PART	JRIP	FURIA	RIDOR	PRISM	DTNB
CHISQUARED	6.17	6.64	6.64	5.69	5.03	7.87
CFS	4.36	7.31	6.26	8.25	4.74	6.92

Figure 2.3: Error Rate (%) without enforcing Feature Selection Methods

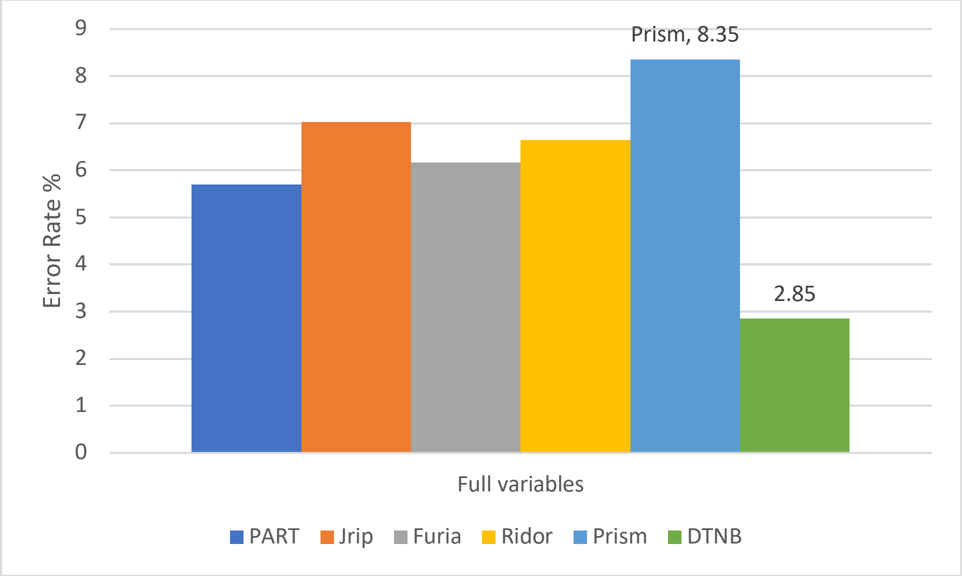


Figure 2.4 exhibits the performance of the rule-based classifiers upon application of Feature Selection methods such as Chi-Squared and CFS anchored on the classifiers PART, JRip, Furia, Ridor, Prism, and DTNB. The highest E scored is attained by Ridor classifier employing CFS FS method at 8.25%, while the lowest E was garnered by PART utilising CFS as its FS technique at 4.36%. Commonly, all the classification models generated an E rate of below 8%, which is 1% lower than without FS method applied. The highest E rate average of 7.40% was clinched by DTNB utilising both the FS methods. While the lowest E rate was scored by PART at 5.27%.

The classifiers DTNB which received the lowest E rate at 2.85% without FS method implemented and was inversely achieving a higher error rate at 7.87% with FS technique applied that has a difference of 5.02%. While Prism, which obtained a higher E without FS method scaled down its E rate when FS method was applied. Therefore, considering the context of the toddler dataset, DTNB algorithm works better with data that has greater dimensionality rather than datasets with lower dimensionality or number of features. Whilst, Prism, suits datasets that have scaled-down features from 18 to 11 that was the result of FS technique after it was implemented.

Table 2.5: Summary of Error Rate (E) with Feature Selection (FS) Applied

FILTERING METHOD	DATASET	PART	JRIP	FURIA	RIDOR	PRISM	DTNB
FULL VARIABLES	Toddler	5.69	7.02	6.17	6.64	8.35	2.85

Figure 2.4: Error Rate (%) enforcing Feature Selection Methods

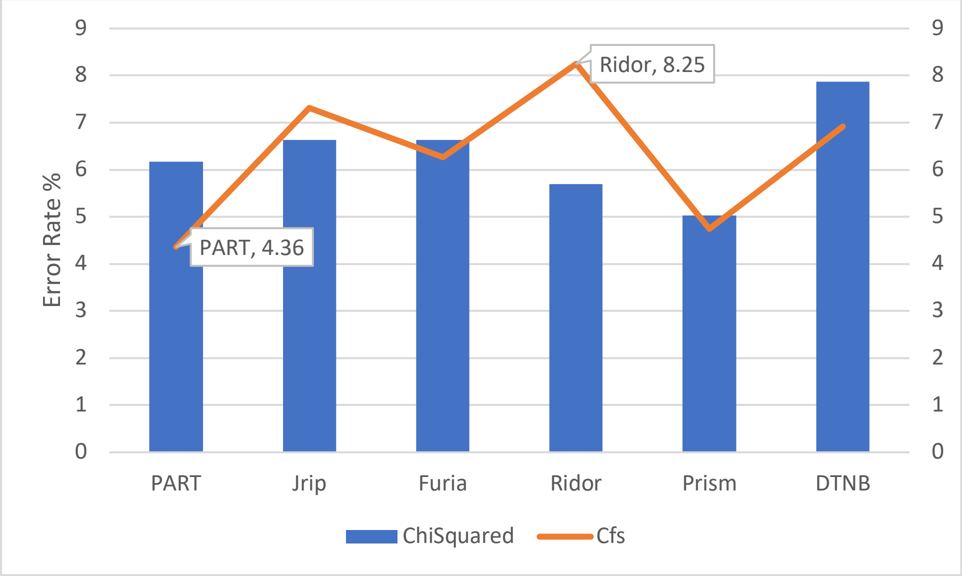


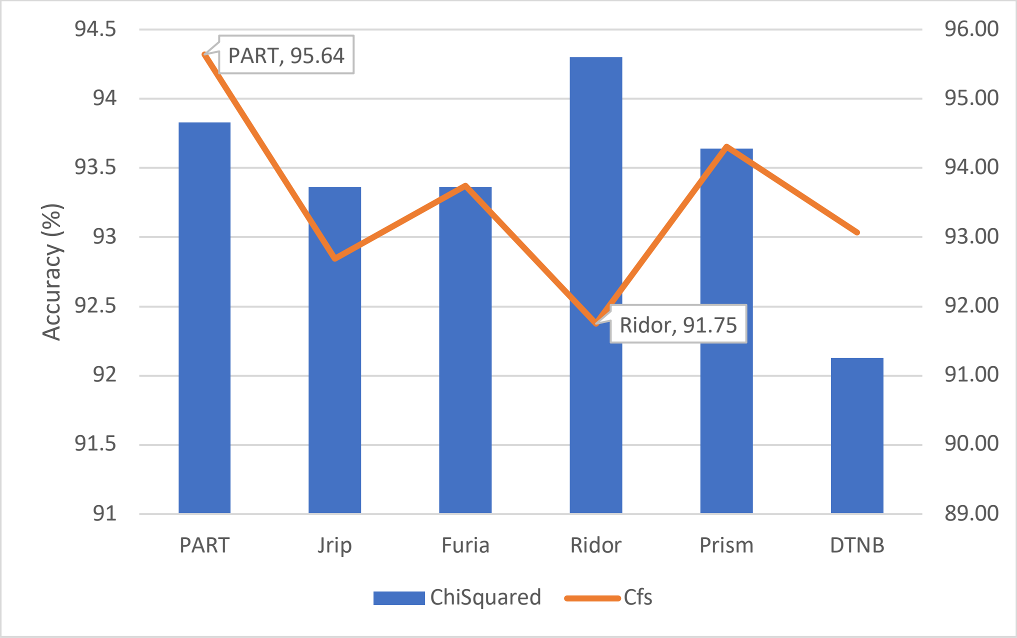
Table 2.6 shows the summary of the Classification Accuracy of the rule-based classifiers for the predictive model upon applying the FS methods. The highest accuracy registered was for PART at 95.64%, whereas the lowest accuracy from the group was obtained by Ridor at 91.75% employing ChiSquared as its FS method. Also, the highest average for predictive accuracy across the classification models was achieved by PART at 94.74% and the lowest are obtained by JRip and Ridor. Therefore, the classification correctness of the rule-based classifier PART depicts that there is a 95.64% accuracy with respect to P, R and F_1-measure that a toddler has a potential to be labelled as acquiring Autism traits.

Table 2.6: Summary of Predictive Accuracy of Rule-Based Classifiers with FS

FILTERS	PART	JRIP	FURIA	RIDOR	PRISM	DTNB
CHISQUARED	93.83	93.36	93.36	94.3	93.64	92.13
CFS	95.64	92.69	93.74	91.75	94.31	93.07

Figure 2.5 below shows that all Predictive models achieved an Accuracy of above 90%. Ridor obtained the lowest accuracy from the group at 91.75% while PART scored the highest at 95.64% outperforming the rest of the rule-based classifiers.

Figure 2.5: Classification Accuracy (%) of the rule-based classifiers after applying FS

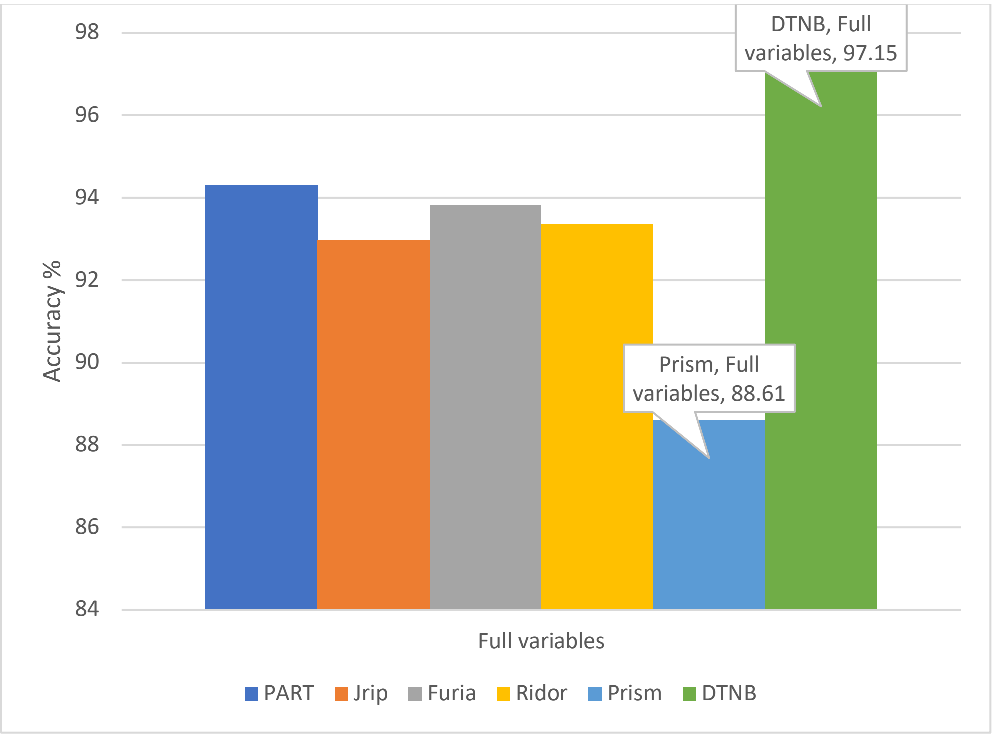


In table 2.7, the best Prediction Accuracy was achieved by DTNB at 97.15%, while the least performing model is Prism at 88.61% without employing FS technique. Also shown in Figure 2.6 that DTNB outperformed the rest of the rule-based classifiers by an average of 4.5%. The low E rates for DTNB is directly proportional with its high predictive Accuracy as seen in Figure 2.6, therefore this classifier suits the classification problem at hand by classifying the labels efficiently as for which toddlers will represent patterns of Autism with respect to the class variable that is known.

Table 2.7: Summary of Predictive Accuracy of Rule-based Classifiers without FS

DATASET	FILTERS	PART	JRIP	FURIA	RIDOR	PRISM	DTNB
FULL VARIABLES	None	94.31	92.98	93.83	93.36	88.61	97.15

Figure 2.6: Rule-based Classification Models Accuracy (%) w/out Feature Selection



2.5.3 Precision, Recall, F_1-measure

In section 2.5.2, Error rate alone is an insufficient metric itself, if the simplest classifier is designed to have 0.01 percent by generating a prediction that every toddler will not possess any Autism traits then such classifier would be useless (Miller & Forte, 2017). A more efficient metrics is the confusion matrix where Precision (P), Recall (R), and F_1-measure is derived. In this section, the ML predictive models will be evaluated in two cases (i) without Feature Selection methods applied upon model implementation (ii) selected features as summarised in Table 1. These case evaluation scenarios would be quantified against P, R and F_1-measure or Harmonic mean.

Table 2.8 summarises the evaluation performance of the rule-based classification models – PART, JRip, Furia, Ridor, Prism, and DTNB. In the Toddler dataset, the highest ratio for Precision (P) as calculated in Equation 2.1, which exemplifies the True Positives (TP) in the confusion matrix was achieved by DTNB at 97% while the lowest was obtained by Ridor at 89%. Whilst, the highest ratio for recall calculated in Equation 2.2 (R) or the number of correct predictions for the positive class over all the members of the positive class in the toddler dataset were achieved by DTNB at 97%, and the lowest was attained by Ridor at 89%. Both the Harmonic Mean (F1-Measure) of DTNB and Ridor are at 97% and 89% respectively. Therefore, high ratio for P translates to low false positive rates as DTNB achieved 0.97 that equates to 97% correctly labelled predictions. Also, the R is 0.97 that is above the 0.5 threshold. In figure 5, the outcome of the evaluation metrics depicts that the hybrid algorithm represented by DTNB outperformed the conventional rule-based classification models without the implementation of Feature Selection. Moreover, the difference between Predictive Accuracy and Recall for DTNB without FS is 0.15, while JRip and Prism both registered at negative 0.02 and 2.39. While the rest of the conventional rule-based algorithms like PART, Furia, Ridor registered a differentiation of 0.31, 0.83, and 4.36 accordingly.

Table 2.8: Evaluation Metrics (weighted avg.) for Classification Models w/out FS

Dataset	Classifier	P	R	F1-measure
Toddler	PART	0.94	0.94	0.94
	JRip	0.93	0.93	0.93
	Furia	0.93	0.93	0.93
	Ridor	0.89	0.89	0.89
	Prism	0.92	0.91	0.92
	DTNB	0.97	0.97	0.97

Figure 2.7: Evaluation Metrics (weighted avg.) for the Predictive Models without Feature Selection

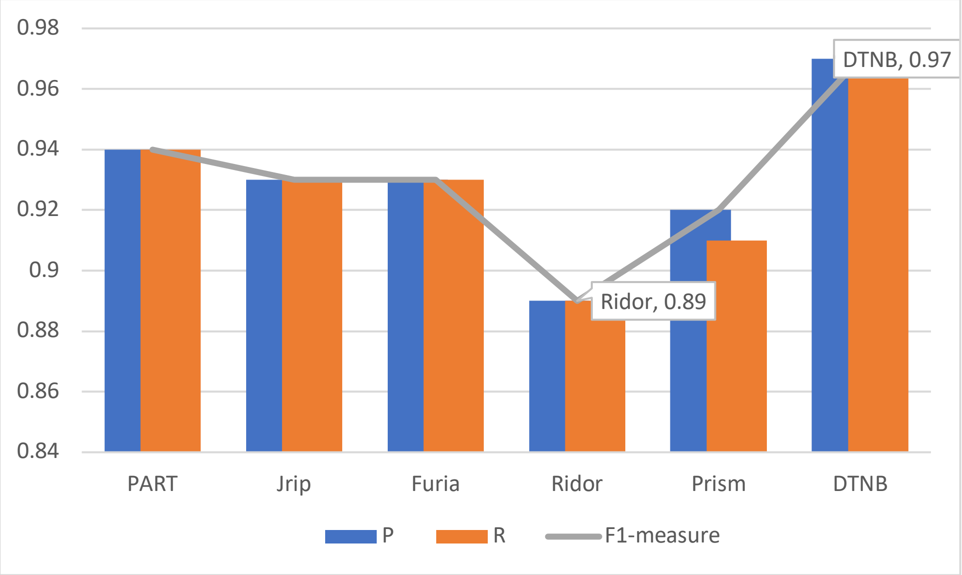


Table 2.9 covers the evaluation metrics for the selected rule-based classifiers after implementing ChiSquared as its Feature Selection technique. The highest P ratio is achieved by Prism at 0.95 that translates to roughly 95% of correctly labelled positive instances of the positive class for Autism. Whilst, the highest ratio for R was a three-way tie for PART Ridor, and Prisma at 0.94 that equates to 94% which is above the 0.05 threshold that is correctly labelled the instances that shows Autism traits. Therefore, the conventional rule-based classifier Prism outperformed the Hybrid classifier, DTNB. The reason or this is Prism works computationally fast and effective for low dimensionality dataset after drawing out the features chosen by the ChiSquared technique.

Moreover, after implementing FS techniques several critical features were removed that was significant for building an efficient model inferred from DTNB.

Table 2.9: Evaluation Metrics (weighted avg.) for Classification Models after applying FS - ChiSquared

Dataset	Classifier	P	R	F1
Toddler	PART	0.94	0.94	0.94
	JRip	0.93	0.93	0.93
	Furia	0.93	0.93	0.93
	Ridor	0.94	0.94	0.94
	Prism	0.95	0.94	0.95
	DTNB	0.93	0.92	0.92

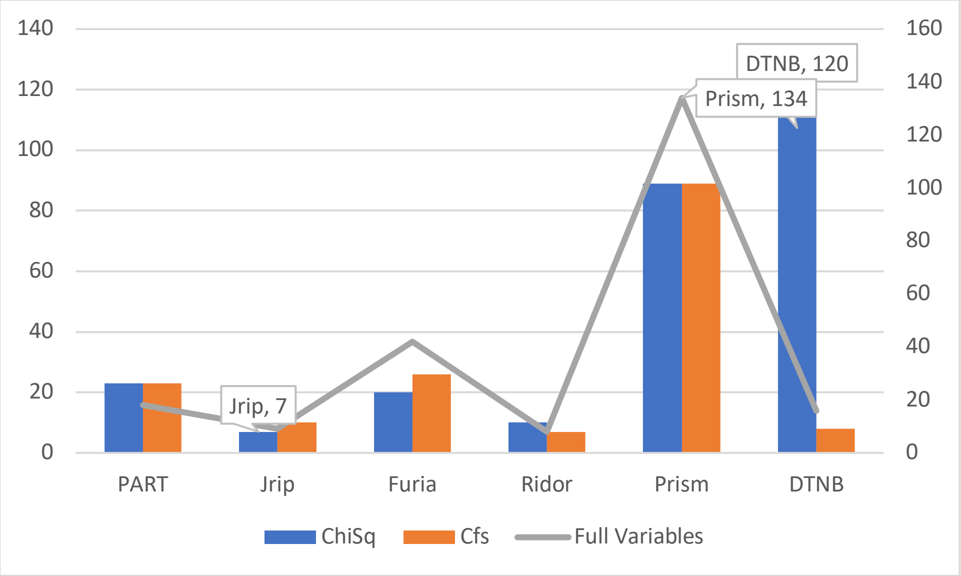
2.6 Rules derivatives

Table 2.10, displays a summary of the rules derived from the rule-based classifiers that generates if-then models which can be generally understood by novice end-users (Thabtah & Kamalov, 2017). The table highlights the algorithm which produced the highest number of rules that is Prism and the least rules was generated by Ridor. When the quantity of rules is reduced by the rule-based algorithms it allows us to efficiently understand the core basic model used to build the predictions. Therefore, the number of rules generated is not highly correlated to data dimensionality as FS method utilising ChiSquared produced 120 rules and DTNB classifier without FS produced 16 rules. Moreover, FS technique is appropriate to discard the strong assumptions of algorithms with the input or training data when raw datasets are used. Also, other features that are correlated don't provide any additional information. Hence, it's significant to have a sharp view of the dataset and determine the most common classifiers leveraged to reduce the dimensionality or only select the best variables (Bonaccorso, 2017).

Table 2.10: Summary of Rules Generated from the Rule-Based Classification Models with and without FS

Dataset	Filters	PART	JRip	Furia	Ridor	Prism	DTNB
Toddler	ChiSq	23	7	20	10	89	120
	CFS	23	10	26	7	89	8
	Full Variables	18	9	42	8	134	16

Figure 2.8: Rules Derivative with and without FS



2.7 Model Selection

Although, DTNB surfaces as the ideal predictive model to select with regards to high predictive accuracy and favourable evaluation metrics result, however the rules it produced without feature and after discretization is difficult to interpret. Since these rules will be the medium to map back a descriptive report to the ASDTest app correspondent therefore, to easily interpret the generated rules from the models we have selected the PART rules as its result is binary either 0= No and 1=Yes (Appendix 1a.) that links to a Q-Chat-10 question and upon completion of the questionnaire each question would be scored according to the delegated grades from Quantitative Checklist for Autism in Toddlers.

3. Summary

Several experiments were carried-out in WEKA workbench using the toddler dataset from UCI repository to build, evaluate and select the rule-based classification model that will produce the least efficient rules which would then be used to map back to QChat questions to generate a comprehensive report for the correspondents of the ASDTests app. The generated rules are then evaluated against their corresponding error rate, precision, recall and harmonic mean amongst others. Analytical results confess that the hybrid algorithm DTNB without FS outperforms the rest of the rule-based algorithms even with FS applied. Furthermore, DTNB produces the least rules and high predictive accuracy with respect to its precision, recall, F1-measure, and error rate with respect to the target class label compared to the conventional rule-based classifiers. Also, DTNB classifier reveals that it performs well with low data dimensionality alongside limited examples.

REFERENCES

1. Bonaccorso, G. (2017). *Machine Learning Algorithms*. Birmingham, UK: Packt.
2. Brownlee, J. (2018). *Machine Learning Mastery with Weka* (1.5).
3. Hall, M. (1999). *Correlation-based Feature Selection for Machine Learning*. University of Waikato, New Zealand. Retrieved from https://www.cs.waikato.ac.nz/~mhall/thesis.pdf
4. Kamalov, F., & Thabtah, F. (2017). A Feature Selection Method Based on Ranked Vector Scores of Features for Classification. *Annals of Data Science*, *4*(4), 483–502. https://doi.org/10.1007/s40745-017-0116-1
5. Kohavi, R. (1995). The Power of Decision Tables. Presented at the European Conference on ML. Retrieved from https://www.researchgate.net/publication/2255536_The_Power_of_Decision_Tables
6. Miller, J., & Forte, R. M. (2017). *Mastering Predictive Analytics with R* (2nd ed.). Birmingham, UK: Packt.
7. Novaković, J. (2017). Evaluation of Classification Models in Machine Learning. Presented at the Theory and Applications of Mathematics & Computer Science.
8. Quinlan, J. R. (n.d.). Generating Production Rules from Decision Trees, 4.
9. Raschka, S. (2018). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, 33.
10. Thabtah, F., & Kamalov, F. (2017). A Feature Selection Method Based on Ranked Vector Scores of Features for Classification. https://doi.org/DOI 10.1007/s40745-017-0116-1
11. Theodoridis, S. (2015). *Machine Learning a Bayesian and Optimization Perspective*. Elsevier.
12. Ting, K. M. (2017). Encyclopaedia of Machine Learning and Data Mining (pp. 260–260). Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7687-1_50
13. Weka features selection (InfoGainAttributeEval, ChiSquaredAttributeEval). (2015). Retrieved 5 September 2018, from https://stackoverflow.com/questions/21132527/weka-features-selection-infogainattributeeval-chisquaredattributeeval
14. Witten, I. H., Frank, E., Hall, M. A., & Palestro, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). San Francisco, UNITED STATES: Elsevier Science & Technology. Retrieved from http://ebookcentral.proquest.com/lib/manukau/detail.action?docID=4708912