Sains Malaysiana 50(7)(2021): 2059-2077
http://doi.org/10.17576/jsm-2021-5007-20

Digital Economy Tax Compliance Model in Malaysia using Machine Learning Approach

(Model Pematuhan Cukai Ekonomi Digital di Malaysia menggunakan Pendekatan Pembelajaran Mesin)

RAJA AZHAN SYAH RAJA WAHAB* & AZURALIZA ABU BAKAR

ABSTRACT

The field of digital economy income tax compliance is still in its infancy. The limited collection of government income taxes has forced the Inland Revenue Board of Malaysia (IRBM) to develop a solution to improve the tax compliance of the digital economy sector so that its taxpayers may report voluntary income or take firm action. The ability to diagnose the taxpayer's compliance will ensure the IRBM effectively collects the income tax and gives revenues to the country. However, it gives challenges in extracting necessary knowledge from a large amount of data, leading to the need for a predictive model to detect the taxpayers' compliance level. This paper proposes the descriptive and predictive analytics models for predicting the digital economic income tax compliance in Malaysia. We conduct descriptive analytics to explore and extract a summary of data for initial understanding. Through a brief description of the descriptive model, the data distribution in a histogram shows that the information extracted can give a clear picture in influencing the results to classify digital economic tax compliance. In predictive modeling, single and ensemble approaches are employed to find the best model and important factors contributing to the incompliance of tax payment among the digital economic retailers. Based on the validation of training data with the presence of seven single classifier algorithms, three performance improvements have been established through ensemble classification, namely wrapper, boosting, and voting methods, and two techniques involving grid search and evolution parameters. The experimental results show that the ensemble method can improve the single classification model's accuracy with the highest classification accuracy of 87.94% compared to the best single classification model. The knowledge analysis phase learns meaningful features and hidden knowledge that could classify the contexts of taxpayers that could potentially influence the degree of tax compliance in the digital economy are categorized. Overall, this collection of information has the potential to help stakeholders make future decisions on the tax compliance of the digital economy.

Keywords: Accuracy; compliance; ensemble; parameter tuning; single classification; taxpayer

ABSTRAK

Bidang pematuhan cukai pendapatan ekonomi digital masih di peringkat awal. Pengumpulan cukai pendapatan kerajaan yang terhad telah memaksa Lembaga Hasil Dalam Negeri Malaysia (LHDNM) untuk mengembangkan penyelesaian untuk meningkatkan kepatuhan cukai sektor ekonomi digital sehingga pembayar cukai dapat melaporkan pendapatan secara sukarela atau tindakan tegas dapat diambil. Keupayaan untuk mendiagnosis kepatuhan pembayar cukai akan memastikan LHDNM memungut cukai pendapatan dengan berkesan dan memberi pendapatan kepada negara. Namun, ini memberikan cabaran dalam mengekstrak pengetahuan yang diperlukan dari sejumlah besar data, yang menyebabkan perlunya model ramalan untuk mengesan tahap kepatuhan pembayar cukai. Makalah ini mencadangkan model analisis deskriptif dan ramalan untuk meramalkan pematuhan cukai pendapatan ekonomi digital di Malaysia. Analisis deskriptif dijalankan untuk meneroka dan mengekstrak ringkasan data untuk pemahaman awal. Melalui penerangan ringkas model deskriptif, taburan data histogram menunjukkan bahawa maklumat yang diekstrak dapat memberikan gambaran yang jelas dalam mempengaruhi hasil untuk mengelaskan pematuhan cukai ekonomi digital. Dalam pemodelan ramalan, pendekatan tunggal dan bergabung digunakan untuk mencari model terbaik dan faktor penting yang menyumbang kepada ketidakpatuhan pembayaran cukai di kalangan peruncit ekonomi digital. Berdasarkan pengesahan data latihan dengan adanya tujuh algoritma pengelasan tunggal, tiga peningkatan prestasi telah dibuat melalui pengelasan bergabung, iaitu kaedah pembalut, pemeringkatan dan undian, dan dua teknik yang melibatkan parameter pencarian dan evolusi grid.

Hasil uji kaji menunjukkan bahawa kaedah bergabung dapat meningkatkan ketepatan model pengelasan tunggal dengan ketepatan tertinggi iaitu 87.94% berbanding dengan model pengelasan tunggal terbaik. Fasa analisis pengetahuan mempelajari ciri-ciri yang bermakna dan pengetahuan tersembunyi yang dapat mengelaskan konteks pembayar cukai yang berpotensi mempengaruhi tahap pematuhan cukai dalam ekonomi digital dikategorikan. Secara keseluruhan, pengumpulan maklumat ini berpotensi untuk membantu pihak berkepentingan membuat keputusan pada masa depan mengenai pematuhan cukai ekonomi digital.

Kata kunci: Ketepatan; model bergabung; pematuhan; pembayar cukai; pengelasan tunggal

Introduction

The trend of economic digitalization in Malaysia is followed by public information that is considered one of the main drivers of economic growth since the industrial revolution 1.0 to 4.0. Using data mining to identify hidden and potentially valuable data through an analysis of income tax compliance, big data technology has become easier and less costly (Lakshmi & Radha 2011). The task of strengthening the compliance of the digital economy with income tax in Malaysia is more critical than other developed countries (Loo et al. 2012).

In addition to the collection of income tax from the existing channel, a sudden increase in business and digital services, including web advertising, social media, e-commerce and on-line blogs, remains to be taxed accordingly. By reference to Inn (2018), when it comes to corporate income tax, companies operating in the digital economy and domiciled in Malaysia are not able to distinguish conventional economics and are often considered as income to seal business operations that make it more difficult for the IRB to collect income tax.

The Algorithm Performance Analysis using different classification techniques is carried out to ensure that the tax audit procedure for the data collection is more organized, efficient and effective. The fundamental of data mining process consists of several stages, which are data preprocessing and preparation, followed by a data mining algorithm and ended with a decision based on the resulting algorithm model. Advances in data mining applications involving classification methods have shown the need for large-scale supervised machine learning algorithms (Tretter 2003).

Processing the data for this study to find hidden transactions requires an analysis that requires quick and efficient algorithms to facilitate the interpretation of the data to help improve the understanding of the data process (Castellón González & Velásquez 2013). The best approach to classification must also be reviewed and improved in order to make decisions quickly and precisely.

The first objective of this study was to describe data features and values that affect non-taxation criteria for taxpayers through descriptive and inferential interpretation so that it can be simplified towards better understanding. Second objective was to establish predictive models such as single classification, improvement of performance through ensemble classification, and tuning of parameters to obtain the best predictive model. Ultimate goal was to analyze knowledge gained from descriptive analysis (data presentation patterns) and predictive analysis (knowledge rules) to determine the level of tax compliance of digital economy taxpayers.

RELATED WORK

Income tax compliance can be defined as the degree to which a taxpayer complies with or fails to comply with the tax laws of a country. The objective of successful tax compliance is to promote voluntary alignment with taxation through all reasonable means include an understanding from taxpayers which connected to knowledge and experience, thus impacted the level of respect for taxation and the awareness of tax compliance (Mohd Rizal et al. 2013). In research on digital economy tax compliance, taxpayers perceptions of the judicial system are seen as important factors that influence their behavior in adhering to income tax (Nellen 2015).

Machine learning is one of the main players in the field of intelligence. According to Dhrubajyoti (2017), machine learning is remarkable, as it teaches computers to process according to the standards set by the user by learning from the experience they have created. Recognizing that a single classification model algorithm still has its drawbacks, it is possible to improve the classification performance and the tuning algorithm is focused in this study by taking different classification scenarios from multiple domains for reference.

DIGITAL ECONOMY INCOME TAX SCOPE IN MALAYSIA

Any form of business transaction made through digital technology, including information delivery, distribution, advertising, promotion, marketing, supply, delivery of goods / services / transactions and all suspected payment are subject to Income Tax Act, 1967 (Risalah Ekonomi Digital LHDNM 2018).

Responsibility of every potential digital economy tax payer in Malaysia: Every businessperson in the digital economy needs to get a tax reference number; Report income/losses incurred as a result of business activities as well as digital economy technology activities; Complete the information and submit an e-B (company) form through e-filing; and Tax payments through different payment channels are provided for convenience.

CLASSIFICATION AND PREDICTION OF TAXATION

The tax authorities are obligated to recognize the non-compliance of the taxpayer promptly for further investigation. This study utilizes the classification algorithm CART 4.5, SVM (Support Vector Machine, "KNN, Naive Bayes, and MLP (Multilayer Perceptron) for the classification of the taxpayer's compliance with four functionally goals such as comply formally, required comply formally, comply materially required and not comply formally (Jupri & Sarno 2018). The results for each classification algorithm are compared and the best algorithm selected based on F-Score, accuracy and time criteria. The end results demonstrate that CART 4.5 is the best algorithm to categorize degrees of taxpayer loyalty compared to other algorithms.

Computational intelligence offers methods, techniques, and resources to automatically create specific income tax predictions based on previous observations. In this article, they proposed hybrid model classification of the CART and Naïve Bayes. This two algorithm was known to be the classification algorithm to boost tax data generalization (Madisa 2018). Human behavior, however, must be revamped in order to detect new patterns and also the knowledge base needs to be maintained.

Better option of tax audits saves time and increases tax collection efficiency (Silva et al. 2016). This study emphasizes to develop predictive models to help identify the 'fiscal bar' on a basis of the first findings appears to be a very promising one. The intention is to use multiple Bayesian Networking algorithms that establish enforcement risk levels, keep taxpayers' actions compliant or contradictory with tax authorities and can improve the precision.

It is important to study the production of evidence-based features for detecting potential taxpayers who manipulate income invoices and commit fraud information on tax payments (Castellón González & Velásquez 2013). Algorithmic models such as Decision Tree, Naïve Bayes, Self-Organizing Maps, and Neural Networks are used to identify fraud-related variables and to detect behavior patterns in cases of income tax evasion. This technique helps to generate valuable knowledge in the audit work carried out by Chilean tax administrators.

In Taiwan, the use of past corporate and individual tax data filings as a database based on decision tree algorithms and artificial neural networks was developed to enhance the efficiency of tax audits (Lin & Lin 2012). This study focuses on lowering or over-taxing as a target class and the distribution of samples. The results of the metric evaluation show that the decision tree model is more accurate in the detection of tax evasion, while the neutral network shows better performance in the corporate tax category.

The Department of Taxation and Customs Ireland has developed a predictive algorithm aimed at taxpayers, avoiding taxation and liquidating assets for tax evasion purposes (Cleary 2011). The algorithmic techniques used are logistic regression, neural networks, and decision tree processes. It predicts the probability of a case involving an audit intervention and is made accessible to taxpayers at a commitment rate of 75%. Other purposes of this model are the identification of cases with similar profiles in the audit case and the assessment of probability scores.

Lakshmi and Radha (2011) carried out taxonomy classification work by providing comparisons of several single classification algorithms under similar circumstances. The data analyzed represent the income and tax details of 365 M/s customers. MSS and Co., accountants accredited. Algorithms such as Decision Tree, Naïve Bayes, SMO, and Logistics Regression have been used to classify the data involved. Comparisons are made to help deliver high-precision results to their algorithms. Various algorithmic techniques have been studied with the aim of obtaining prior knowledge to make comparisons with this study model as summarized in Table 1.

TABLE 1. Classification algorithm and tax sector prediction algorithm

Author, year of publication	Algorithm
Jupri & Sarno (2018)	CART 4.5, SVM, KNN, NB and MLP
Madisa (2018)	CART and NB
Silva et al. (2016)	NB
Castellón González & Velásquez (2013)	CART, NB, Map Compilation and ANN
Lin & Lin (2012)	CART and ANN
Cleary (2011)	CART, ANN and LR
Lakshmi & Radha (2011)	J48, NB, SMO and LR

MATERIALS AND METHODS

In this study, we adopted Cross Industry Standard Process for Data Mining (CRISP-DM) standard data analytics research methodology introduced by Crisp (1999) that cover data science practice with five important phases namely, business understanding, data understanding, data preparation, model development, and deployment of model. Crisp-DM is the standard Data Analytic Methodology that is widely used in data analytics projects.

BUSINESS UNDERSTANDING

Business understanding phase has been explained earlier which involves defining the business goal and business question of the study. We define business goal as: to track down and predict the tax compliance and non-compliance among digital business; and to identify factors that influence the tax compliance and non-compliance of digital business. The aim is to discover how well can predictive analysis help to define target class categories once complex data from external and internal resource matching is provided, and various patterns and knowledge rules can be generated after the development of a data modeling, but are there connections between these important features and hidden information sufficient to generate valuable knowledge? Some example of digital economy businesses are taxpayers who earn income from the digital economy channel which includes business models of advertising income, affiliate income, trade, service, social media, e-commerce, blog, and online marketplaces.

DATA UNDERSTANDING

Data understanding plays a major role as data need to be collected from appropriate sources as to ensure data are relevant to answer the business question. In this study data related to digital economy are obtained through an official letter of application to the IBRM Department of Tax Operation to obtain data comprises of several external and internal sources clusters (tax assessment year 2015, 2016, and 2017). Raw data of digital economy external source were retrieved using website crawler and internal source data were obtained from inhouse database and data integration was conducted both sources of data. External data sources are filtered for company with registered name and registration number. The external data sources, are retrieved using an online web crawler software name *Kapow* to extract digital economic web pages' data and cross matched with internal data source from IRBM. This internal data includes the taxpayer's profile, tax statements, tax assessments and existing assets (including real estate, vehicles and the presence of proof of income stamp duty). There are many features involved in the classification method. The useful feature of selecting the complete attribute of this data set is that it is easier to measure only a subset of the data that is subtracted from the selected data set.

DATA PREPARATION

Data preparation involves data integration to integrate data from various sources, data exploration, to get insight of the data distribution and quality, data cleaning to handle missing, noisy, inconsistency, data reduction to get relevant attributes and reduce attribute values, and finally data transformation to prepare data format for modeling. Total of 11,706 rows of business taxpayer's data with 29 conditionals attribute and a class attribute involved in the modeling. The class attributes are the status of tax compliance namely Compliance (Comp), and Non-Compliance (Non-Comp). This indicate the abiding or non-abiding business company by tax payment. The exploration of data shows that 6,335 records are noncompliance company (Non-Comp) while 5,351 comply with income tax Act (Comp) based on the class label indicator provided by IRBM internal data sources. Figure 1 shows the distribution of compliance and non-compliance type based on data that were scrapped after integration between internal and external sources.

The attributes used in this study are Company Info (Company_Info), Tax Employer Number (Employer_No), Tax Registration Information (Tax_Register_Info), Tax Referrence File Number Found (Tax_File_Found), Employer Contact Number (Contactno), Status Assessment Yearly (Assessment_Status), Address Given in the Tax Form (Address), Tax Registration File Location (Location), Region of Tax Location (Region), Return Form Source (Source), Digital Economy Sector (Sector), Digital Economy Description Sector (Descsector), Taxpayer Bank Account Number (Bank_Acct_No), Bank Information give in Tax Return Form (Bank_Info), Tax Registration Date (Registration_Date), Tax File Active (File_Active), Submission of Tax Form Type (Submission_Type), Counting Assessment Year (Asm_Yr_Count), Year of Assessment (Assessment_Year), Property Asset Type (Asset_Type_P), Property Asset Count (Property_Count), Property Asset Amount (Asset_Value), Vehicle Asset Type (Asset_Type_V), Vehicle Asset Count (Vehicle_Count), Amount of Vehicle Loan (Loan_Amount), Vehicle Asset Amount (Asset_Value1), Stamps Asset Type (Asset_Type_S), Stamps Asset Count (Stamps_Count), Sum Amount of Stamps (Sumamountofsale), Classification Label (Compliance). Table 2 depicts the example of business tax payer data and attributes.

TABLE 2. Example Business Taxpayer Dataset

Company Info	Employer No	Tax Register Info	Tax File Found	Contact no	Assessment Status	Address	Location	Region	Source	Sector	Sector Description	Bank Acct_No	Bank Info
1	Y	C1.032023203e9	Y	Y	Non Liable	Y	Wangsa Maju	1	Yellowpages	Financial Services	Advertising	Y	Mbbb Malayan Banking Bhd
2	Y	C 1.0342838e9	Y	Y	Liable	Y	Petaling Jaya	1	Lelong	Economy Sharing	Real Estate Activities Based on Payment Policy or Contract T.T.T.L.	Y	Pubb Public Bank Bhd
3	Y	C 2.2866187e9	Y	N	Liable	N	Kl Bandar	1	Yellowpages	Manufacturing	Manufacturing of Spices and Curry Powders	Y	Mbbb Malayan Banking Bhd
4	Y	C2.209147307e9	Y	N	Liable	N	Melaka	2	Yellowpages	Retail	Wholesale of Coffee, Tea And Other Beverages	Y	Pubb Public Bank Bhd
5	Y	C1.032506602e9	Y	Y	Liable	Y	Pulau Pinang	2	Yellowpages	Manufacturing	Manufacture of Jewelry and Related Item	Y	Pubb Public Bank Bhd
6	Y	C 2.0139062e9	Y	N	Liable	N	Bukit Mertajam	2	Yellowpages	Retail	Wholesale Sales of Solids, Liquid And Gas And Related Products T.T.T.L.	Y	Citi Citibank Berhad
7	Y	C1.033288105e9	Y	Y	Liable	Y	Kuala Lumpur	1	Yellowpages	Retail	Wholesale And Retail Parts Components (Including Parts) And Motorcycle Accessories	Y	Rhbb Rhb Bank Berhad
8	Y	C1.030217404e9	Y	Y	Liable	Y	Klang	1	Yellowpages	Manufacturing	Manufacture of Other Products Made from Metal T.T.T.L.	Y	Rhbb Rhb Bank Berhad

Registration_Date	File_Active	Submission_Type	Asm_Yr_Count	Assessment_Year	ASSET_TYPE_P	PROPERTY_COUNT	ASSET_VALUE	ASSET_TYPE_V	VEHICLE_COUNT	LOAN_AMOUNT	ASSET_VALUE1	ASSET_TYPE_S	STAMPS_COUNT	SUMAMOUNTOFSALE	COMPLIANCE
Y	A	6.0	2.0	2016/2017	0	0	0	V	<=5	0	<=100K	S	<=5	<=200K	TPTH
Y	A	1.0	2.0	2016/2017	0	0	0	0	0	0	0	S	<=5	>300k	TPTH
Y	A	6.0	3.0	2015/2016/2017	0	0	0	V	<=5	<=100K	<=150K	S	<=5	>300k	TPTH
Y	A	6.0	3.0	2015/2016/2017	P	<=10	>300k	V	<=10	0	<=100K	S	<=5	>300k	TPTH
Y	A	6.0	2.0	2016/2017	P	0	0	V	<=5	0	<=100K	S	<=5	>300k	TPTH
Y	A	6.0	2.0	2016/2017	0	0	0	V	>15	0	<=100K	S	<=5	>300k	TPTH
Y	A	6.0	3.0	2015/2016/2017	0	0	0	V	>15	<=50K	<=50K	S	<=5	>300k	TPTH
Y	A	6.0	3.0	2015/2016/2017	0	0	0	V	>15	<=50K	<=100K	S	<=5	>300k	TPTH

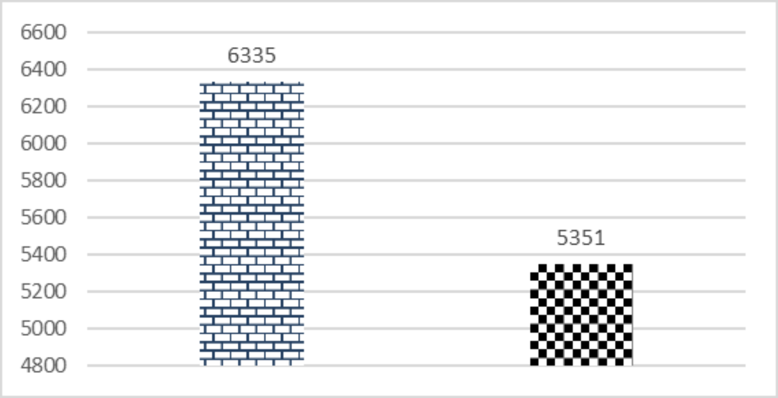

FIGURE 1. Distribution of data based on class {compliance/non-compliance} label

The histogram of Figure 2 presents the attribute {sector} which is the thirteen (13) categories of business types under the scope of taxation based on IRBM income. There are 13 categories involved with 527 digital economy taxpayers who run more than one type of business under 'others'. Total of the retail sector has the highest number with 3,964 taxpayers. Retail taxpayers or taxpayers having the branch outlets have many businesses in the digital economy. The crowd sourcing sector has a minimum number of taxpayers of 103 for making e-commerce transactions by accepting payment as income.

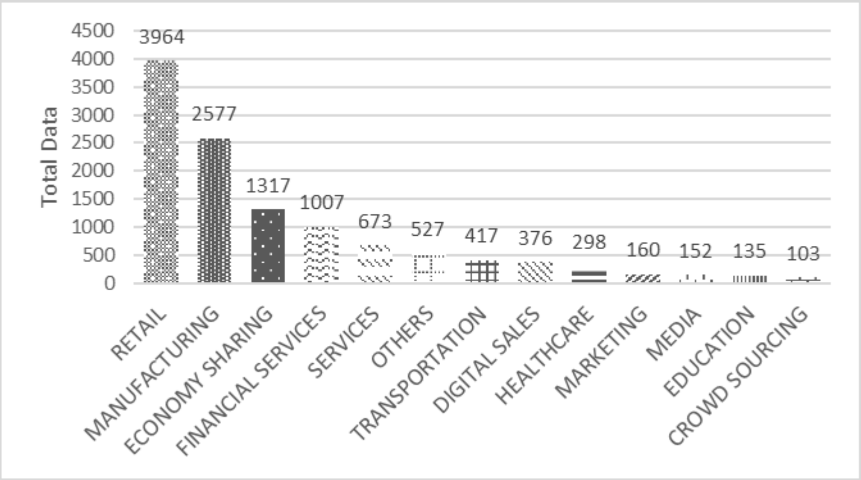

FIGURE 2. Number of data by sector

The {location} attribute is the 9 categories of business operating locations based on tax filings registered as resident in Malaysia, while the 'other' category has the highest number of data referring to the absence of a resident code or unknown location due to no registration branch collection number in IRBM internal database. Histogram in Figure 3 shows taxpayer location data. There are 2,076 of 'others' data without a trace allows the tax profile to be investigated whether or not the resident is a resident. As a result, many taxpayers have not yet submitted the legal information on which their business operations are being conducted, as external source information on the digital economy website itself is already available.

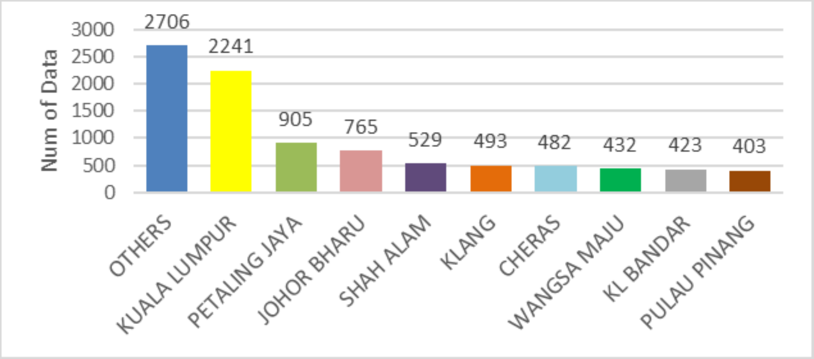

FIGURE 3. Categories of business operating locations {location}

Figure 4 describes the {source} attribute of the website used to collect orders, payment receipts, advertising, services, and ad promotion. It indicates that, the 'yellow pages' web site posted the highest number with 5,688 data points. The platform provided by the website is more user-friendly and provides a wide range of uses from the start until the business transaction completed. The search results of online 'yellow pages' also prove that the information most required on the business transaction segments of their website is the attribute {roc_no} which is the company registration number that is mandatory to match IRBM's internal source data. Online Yellow pages are very active nowadays providing essential information on potential digital economy advertisement.

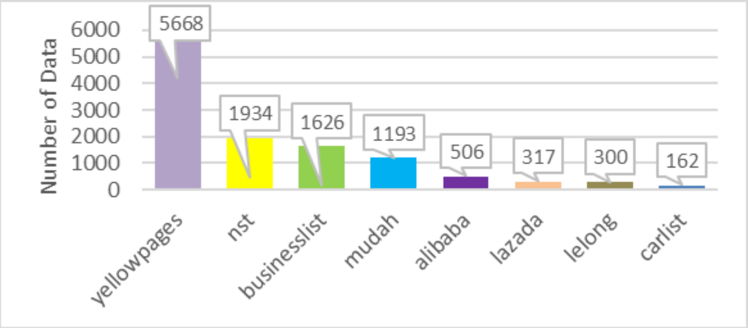

FIGURE 4. Number of data based on website Source{source}

Figure 5 depicts the tax compliance levels related to three categories i.e. property ownership, vehicle assets, and stamp duty (payments resulting from the sale of fixed assets). The presence of a record number of '<= 5' with a total of 2,408 for stamp duty, 2,833 for vehicle assets and 288 for property assets. Therefore, the factor of ownership of the assets affect the lifestyle and the economic viability of taxpayers running the digital economy making them eligible for taxable income. Record '0' has the most number of data for these 3 categories, indicates that information may be hidden from being updated in the IRBM database and may influence the results of the classification model.

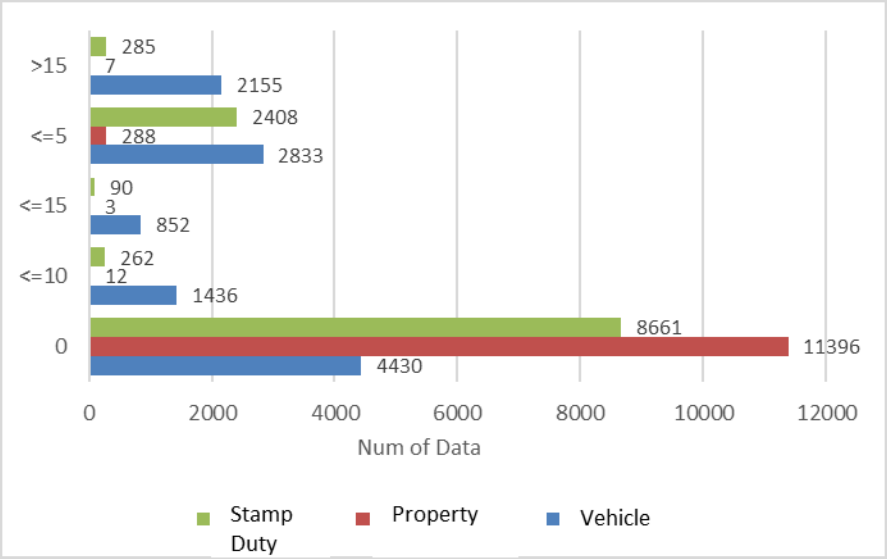

FIGURE 5. Number of data on {property_count}, {vehicle_count} and {stamps_count}

The distribution of dataset to the 'Non-Comp' and 'Comp' target classes, indicating that the 'Non-Comp' cases are taxpayers with no tax numbers, tax avoiders and fraudulent in disclosure information. The volume of 'Comp' label cases represents the low degree of tax compliance in the digital economy sector and demands for efforts to increase enforcement to taxpayers. The most important concept in 'retail' would be that of taxpayers who perform business transactions digitally. These activities include taking orders, packing, receiving payments and making deliveries involving most taxpayers as owners, agents, stockiest, and wholesalers. In the event of a non-tax compliance, the use of the company number from the retailer for future review may result in a tax audit trail. In contrast to crowd sourcing, the amount of value gained is minimal due to the lack of business activity of this type and the probability of business profit being less than operating costs.

The category of total assets under the business owner is the basis to be quoted from IRBM's internal sources to find a continuation of information on real income that is not reported, not filled in with tax forms, tax evasion and others for 'Non-Comp' tax cases. These are also supported by a type of assessment that has a tax audit code that indicates the need for an audit trail. Ownership of assets under 10 and below, is found most prominently in the histogram of real estate assets, derivative assets, and stamp duty transactions. This clearly proves that taxpayers are subject to income taxation and should not be left unattended.

DEVELOPMENT OF PREDICTIVE MODEL

Development of the model involves the use of machine learning algorithms to discover important patterns or knowledge from data. These knowledge is valuable in assisting human for decision making. There are several task in data analytics such as classification, prediction, association rules mining, deviation detection, and trend analysis. Determination of suitable data analytics task is depending on the business goal defined. In this study to achieve the tax compliance business goal, classification task is performed. Several classification algorithms are investigated and employed to find the best classification algorithms that fits the problem. The fundamental concept of classification model development can be seen in Han and Kamber (2002) (Refer to Figure 6).

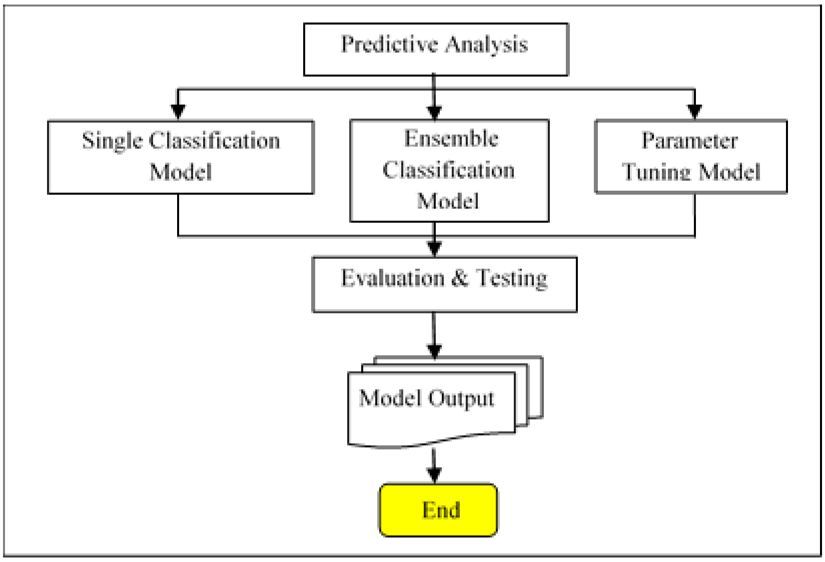

FIGURE 6. Model development methodology (Han & Kamber 2002)

Several classification algorithms employed are Classification and Regression Tree (CART), Random Forest (RF), Naive Bayes (NB), K-Nearest Neighbour, Support Vector Machine (SVM), Logistic Regression (LR), and Artificial Neural Network (ANN). Two modeling schemes are proposed in this study i.e. Single Model and Ensemble Model. Studies are also being conducted to identify significant weaknesses in a single classification model to drive the implementation of performance improvement by studying ensemble classification techniques and tuning algorithms to select the best classification accuracy. There are two disadvantages when using a single classification technique, which is that it does not provide a comprehensive solution to all types of data set studies because each of the different techniques may be appropriate for different dataset, while the other techniques may not be suitable and does not provide meaningful information during classification (Adejo & Connolly 2017).

Ensemble method uses machine learning algorithms to incorporate several single classifications to improve performance and are considered as successful techniques for solving classification problems (Pham et al. 2016). The final classification developed by ensemble method is able to incorporate the characteristics of a single classifier with the same or different factors and functions to improve performance (Mithal et al. 2017).

Wrapper technique is easy to use and is one of the first techniques of ensemble. It can often be combined with other classification algorithms such as CART, SVM, ANN, NB, and KNN. Wrapper was introduced by Brieman and the concept of sub-training on data easily obtained by random sampling was applied using a replacement method. The sub-training was carried out to train the single classifier and then the combined classification technique, which is the majority of the weighted votes, was used to combine the results of the single classifier in order to select the best classification from the best model (Breiman 1996).

Boosting also known as Adaboost is an algorithm that can be developed by improving the predictive capabilities of the classification algorithms. Introduced by Freund and Schapire in 1997, Adaboost was widely used in classifications that typically focus on difficult data values. The first weight value will be assigned to the training data set example and the weight value will be replaced during the training process on the basis of the previous basic or single classifier performance. The training process shall be stopped if the optimal weight of the training data set achieves the best classification performance (Freund & Schapire 1997).

Voting is one of the simple and popular combination of a classification technique by combining the output of several single classification algorithms, each of which is calculated in order to obtain a final classification prediction (Ali et al. 1995). The use of majority, average, minimum and maximum techniques is often used during the welding process, and best methods, such as majority, are often chosen because they produce a balanced output. For example, a data set with 2 target classes is said to have the 4 best single classifiers determined by the majority vote to see how many target classes will be selected.

EXPERIMENTAL DESIGN AND EVALUATION

The three metrics of assessment are classification accuracy, contingency table (TP, NB, FP, and FN) and classification reports which include accuracy measurement, retrieval and F-measure. Figure 7 shows the overall experimental design of this study.

Analyzes were made on the accuracy of model classification by measuring the performance of actual decisions against prediction results generated by the model (Hamsagayathri & Sampath 2017).

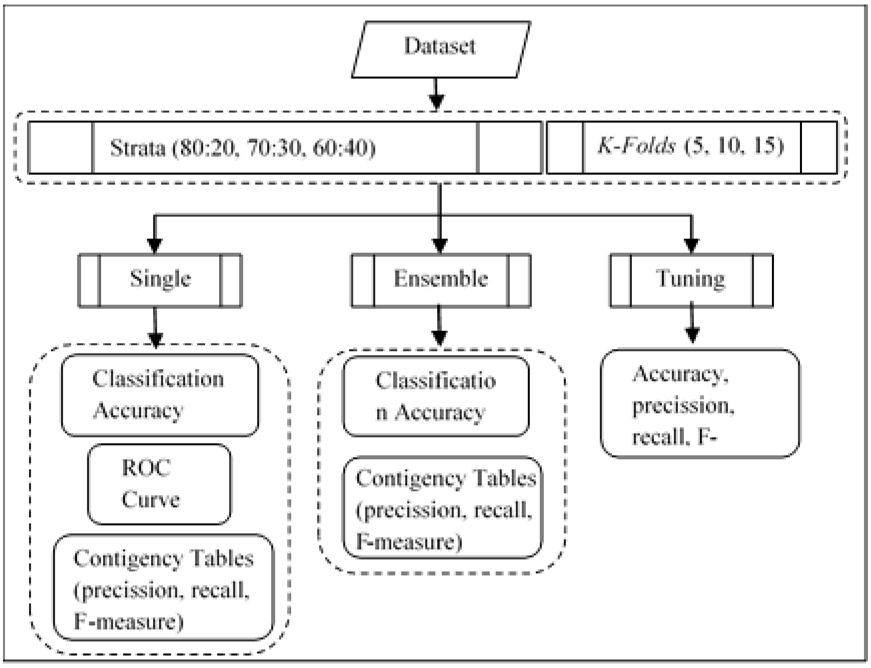
Dataset
Strata (80:20, 70:30, 60:40)
K-Folds (5, 10, 15)
Single
Ensemble
Tuning
Classification Accuracy
ROC Curve
Contigency Tables (precission, recall, F-measure)
Classificatio n Accuracy
Contigency Tables (precission, recall, F-measure)
Accuracy, precission, recall, F-

FIGURE 7. The overall experimental design of predictive analysis

Result

Single classification model

The results of each experiment on a single classification model were recorded and further analyzed. Seven (7) widely used classification models including Naive Bayes (NB), functions (SVM, LR), meta (ANN, KNN), rules and tree (CART, RF) are used and the k-fold cross-validation method (k=5,10,15) is used for percentage split of training: testing data (80:20, 70:30, 60:40). A total of 21 single classification models were developed. Each of the parameters in the model metric evaluation will show the performance of each model used.

Based on the classification accuracy, the CART model achieved the highest accuracy of 87.01% compared to the RF model of 86.98%. While in the last position is the NB model with an accuracy of 82.16%. The advantage of using CART model algorithms is the ability to classify small data sizes, with good accuracy. The overfitting reduction of the CART model was made through pre-pruning and post-pruning. Table 3 shows the results of the best model based on the k-folds sampling technique.

TABLE 3. Experimental result of accuracy on single classification models

	CART	RF	NB	KNN	SVM	LR	ANN
(k=10, train:test = 70:30)							
Time (s)	0.21	1.53	28	19.06	2.10	3.38	1.03
Accuracy (%)	85.61	85.54	81.34	82.62	81.13	81.62	83.13
(k=10, train:test = 70:30)							
Time (s)	0.29	2.19	0.35	15.12	1.54	33	1.11
Accuracy (%)	86.95	86.76	81.96	82.77	82.20	82.58	83.95
(k=15, train:test = 60:40)							
Time (s)	0.49	3.33	0.50	15.26	1.07	48	1.22
Accuracy (%)	87.01	86.98	81.27	81.38	82.45	82.91	82.34

Table 4 depicts the overall results of the contingency table. The experimental results show the matrix with TP (True Positive), FP (False Positive), TN (True Negative) and FN (False Negative). The TP parameter shows the correct classification prediction for the target class 'Non-Comp'. The FP parameter shows incorrect predictions for the target class 'Non-Comp'. The TN parameter shows the correct classification prediction for the target class 'Comp'. The contingency table analysis proved that the ANN model emerged as the best classification model for the classification of the target class 'Non-Comp' while the RF model was the best classifier in classifying the 'Comp' data. Both ANN and RF models are capable of being complete classifiers in handling small amounts of data.

TABLE 4. Contingency table of single welding model (k-folds)

	CART	RF	NB	KNN	SVM	LR	ANN
(k=5, train:test = 80:20)							
TP	578	567	579	591	554	599	602
FP	14	7	75	62	53	91	75
FN	188	196	187	182	212	167	162
TN	624	634	563	569	585	547	566
(k=10, train:test = 70:30)							
TP	892	884	885	906	847	910	925
FP	14	10	117	106	74	129	126
FN	261	269	263	257	301	238	212
TN	940	944	842	838	885	830	843
(k=15, train:test = 60:40)							
TP	1175	1165	1182	1168	1142	1214	1212
FP	19	10	183	173	94	169	194
FN	346	356	343	350	399	311	302
TN	1270	1279	1100	1118	1174	1114	1101

Beside accuracy, to determine the best model, several other important metrics are considered such as precision, recall and F-measure for both target classes should be obtained by looking at the actual performance of each model. In the experiment, the RF model of 99.15% achieved the highest accuracy with a value of 0.74% compared to the CART model of 98.41%. The RF model is capable to classify both target classes well with small classification error. The low sample size of 11,706 is among other advantages.

The metric recall determines the single classification model developed are either good or not dependent on the success of classifying 'Non_Comp' class or non-taxpayers. Experimental results show that the ANN model has the highest recall of 81.35% compared to other models. This is clear because of the highest TP and the lowest FN obtained by ANN model. The F-measure is a rate that takes into account the accuracy and recall factors of a model. It shows that the CART model with 86.55% exceeded the value of 0.13% compared to the RF model of 86.42% and lastly the SVM model with 82.25%. The overall results of the experiment are shown in Figure 8.

Based on the results of this experiment, four models of selected algorithms namely CART, RF, ANN, and LR will be used in performance improvement through ensemble classification and parameter tuning.

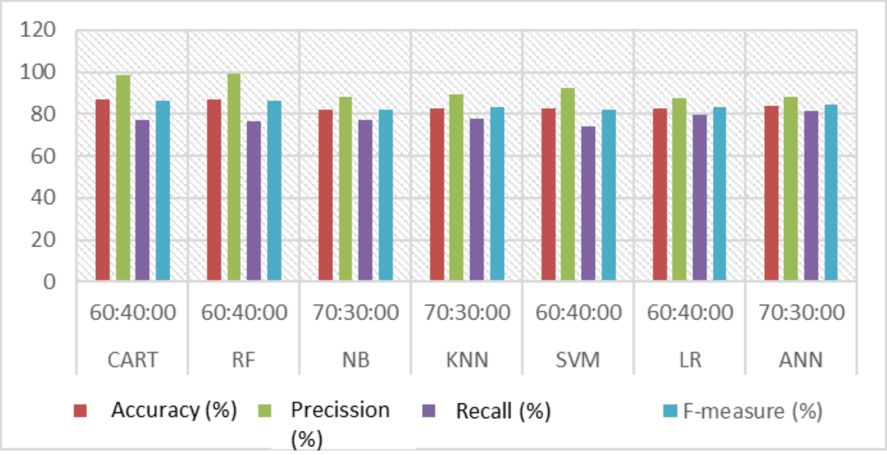

FIGURE 8. Model evaluation metrics

ENSEMBLE CLASSIFICATION MODEL

Table 5 shows the results of the ensemble classification and overall parameter tuning using classification accuracy and contingency tables (TP, TN, FP, and FN) as well as accuracy measurement, recall and F-measure. The ensemble classification approaches used are wrapper, boosting, and voting techniques, while parameter tuning approaches are grid search and evolutionary search. These techniques are supposed to improve the efficiency of a single classifier.

The wrapper technique is well known because it has been proven to be able to build a high-quality integrated ensemble model over a single model (Pham et al. 2016). This study developed k-folds cross validation method (k = 5, k = 10 and k = 15) with a percentage split of training-test validation data (80:20, 70:30, 60:40). A total of 12 classification models using wrapper techniques were developed, but only one best model of each algorithm was recorded. The results of the experiment (k = 15, validation fraction = 60:40) showed that the RF model recorded the highest accuracy value of 87.43% (0.45% difference from single classifier). However, the major difference that can be detected using this wrapper technique is the ANN model which is 1.46% difference from single classifier. It can also be seen that the CART model does not give major difference since it appears to be similar to the single model.

In boosting approach, the single Naive Bayes (NB) classifier gives the lowest accuracy compared to the other six models. A total of three classification models using Adaboost were developed and only the best ones were recorded. Adaboost are widely used in reducing classification error, bias, and high variability data (Viaene et al. 2004). It works by increasing the capability of the single model to achieve higher accuracy. Experimental results indicate that NB (Adaboost) has succeeded in improving accuracy by reducing errors in each data that are misclassified. This is evidenced by the improved accuracy performance on the weakest NB classification model by a value of 83.63% from 81.96% using the parameter (k = 10, percentage split of validation = 70:30). Majority voting is another boosting method used in this study. The results showed that the model's accuracy performance (k = 10, validation split = 70:30) recorded (less noticeable) 87.10% (0.15% difference) compared to the best single classification models listed. This is because the technique is more suited to the significant imbalance class dataset whereas in this study the class dataset is approximately balance.

Through grid search techniques, several single classifier models have been identified as among the best classification models carried out by tuning the parameters to further strengthen the model's capabilities. A total of 12 models were successfully developed, but only 4 high-precision algorithm models are listed in Table 4. Experimental results for the tuned CART model (k = 15, validation split = 60:40) recorded the highest accuracy of 87.94% compared to the tuned RF model of 87.40% and lastly the LR model with 83.65%.

Evolutionary search techniques are very useful when the range and inter-correlation coefficients are known. This technique is an improvement in order to obtain the best algorithm performance. A total of 12 models were successfully developed, but only 4 of the best models were recorded. Experimental results for the CART model (k = 15, validation fraction = 60:40) recorded the highest accuracy of 87.40% compared to the tuned RF model of 87.38% and lastly the LR model (k = 15, validation fraction = 60:40)) with 84.23%.

TABLE 5. Comparison of single classification, ensemble classification and parameters tuning

	CART	RF	ANN	LR	NB	KNN	SVM
1) Accuracy (%)							
Single	87.01 (60:40)	86.98 (60:40)	83.95 (70:30)	82.91 (60:40)	81.96 (70:30)	82.77 (70:30)	82.45 (60:40)
Wrapper	87.04 (70:30)	87.43 (60:40)	84.20 (70:30)	83.51 (60:40)	-	-	-
Boosting	-	-	-	-	83.63 (70:30)	-	-
Voting	87.10 (70:30)				-	-	-
Grid tuning	87.94 (60:40)	87.40 (60:40)	84.81 (70:30)	83.65 (60:40)	-	-	-
Evolutionary tuning	87.40 (60:40)	87.38 (60:40)	85.19 (70:30)	83.65 (60:40)	-	-	-
2) ROC							
Single	1.0	1.0	0.98	1.0	1.0	0.975	0.99
3) Contigency table_							
Single	TP=1175, FP=19, FN=346, TN=1270	TP=1165, FP=10, FN=356, TN=1279	TP=925, FP=126, FN=212, TN=843	TP=1214, FP=169, FN=311, TN=1114	TP=885, FP=117, FN=263, TN=842	TP=906, FP=106, FN=257, TN=838	TP=1142, FP=94, FN=399, TN=1174
Wrapper	TP=940, FP=60, FN=213, TN=894	TP=1244, FP=72, FN=281, TN=1211	TP=920, FP=105, FN=228, TN=854	TP=1203, FP=150, FN=322, TN=1133	-	-	-
Boosting	-	-	-	-	TP=931, FP=123, FN=222, TN=831	-	-
Accuracy (%)							
Single	98.41	99.15	88.01	87.78	88.32	89.53	92.39
Wrapper	95.09	94.53	89.76	88.91	-	-	-
Voting	93.83				-	-	-
Grid tuning	94.57	95.22	92.69	88.46	-	-	-
Evolutionary tuning	95.22	97.71	94.22	88.46	-	-	-
Recall (%)							
Single	77.25	76.59	81.35	79.61	77.09	77.9	74.11
Wrapper	80.14	81.57	80.14	78.89	-	-	-
Voting	81.79				-	-	-
Grid tuning	82.45	80.86	78.01	80.39	-	-	-
Evolutionary tuning	81.05	78.59	77.71	80.39	-	-	-
F-Measure (%)							
Single	86.55	86.42	84.55	83.5	82.32	83.31	82.25
Wrapper	86.97	87.57	84.67	83.60	-	-	-
Voting	87.40				-	-	-
Grid tuning	88.10	87.45	84.72	84.23	-	-	-
Evolutionary tuning	87.57	87.11	85.07	84.23	-	-	-

Experimental results show that single CART algorithm models have the highest classification accuracy compared to other algorithms. In contrast, for ensemble classification, RF (wrapper) models achieve higher classification accuracy than single classification. This shows that the ensemble classification technique can improve the accuracy of a single weak model. The CART model's parameter tuning has the best overall capability with the highest accuracy than the single classification model and the combined classification model. As a result, the rules generated by the tuned CART model were analyzed in order to obtain meaningful knowledge along with the wrapper model (RF) of the second-best algorithm. The experiments carried out, and each model's findings can provide an essential guide for future research in the relevant field.

KNOWLEDGE ANALYSIS

Moreover, conformity with the classification ensemble and parameter tuning in the digital economy has improved the model's classification accuracy. Along with the results of the experiments conducted for the determination of the target classes 'Non-Comp' and 'Comp', two rule base algorithms namely RF and CART. This algorithm has been successfully developed to generate rules that help to define effective target classes so that useful knowledge of IRBM can be realized. The expert evaluation and verification of classification rules found that the 'Non-Comp' class can be identified effectively.

NON-COMPLIANCE CASE

The RF (k=15, validity split = 60:40) classification rules are obtained across 100 algorithm-generated trees. Five Knowledge Analysis Feedback Forms were distributed to IRBM Domain Experts who had knowledge of taxation, statistics, and data warehouses through official e-mails and printed forms. Based on the feedback, the results are summarized in Table 6.

TABLE 6. Summary of expert evaluation (5 experts)

Num	Classification rules verification	Agree	Not sure	Not agree
1.	CART - 'Non-Comp'	4	1	-
2.	CART - 'Comp'	4	1	-
3.	RF - 'Non-Comp'	4	1	-
4.	RF - 'Comp'	4	1	-

There are many advantages in using CART algorithm, but most importantly it is simple and easy to understand because it is similar to how humans make decisions with the presence of an effective 'if-then' logic. In the case of non-compliance taxpayer, a total of 2 selected rules have been generated through the RF algorithm, while 2 rules have been selected from the CART algorithm for tax-exempt tax cases. As a result, all of these rules are evaluated by the domain expert who will prove their authenticity and see their importance in order to avoid unnecessary rules. The filtering rules are carried out when a domain expert (Azuraliza et al. 2011) supports the frequency of the data. Table 7 is referred to explain the rules of the selected 'Non-Comp' classes using the RF algorithm.

TABLE 7. Expert evaluation of non-compliance (non-comp) rules

Rules description	Selected features output	Expert decision (agree or not sure)
RF rules (2nd tree) if Non-Comp = 1487 AND Digital economy sector = 2-13 AND first tax return date = 'YES' AND ownership stamp duty return = 'YES' AND number vehicle assets = 'YES' Description of Rules There are 1487 non-compliance taxpayers correctly classified with sector two till 13, where they do submit their first business tax and stamp duty return information with a specific value of vehicle assets in possession)	1. sector; receive_date 2. asset_type_s 3. vehicle_count_0	4 agree, 1 not sure
RF rules (4th tree) if Non-Comp = 1486 AND Tax file found 'YES' AND receipt of tax form first = 'YES' AND ownership of stamp duty return assets = YES AND number of vehicle assets = 'YES' Description of Rules There are 1486 non-compliance taxpayers cases that correctly classified when they registered yearly based on the receipt of IRBM's tax return. They also have several vehicle assets in the possession and submitted their stamp duty return information without hesitation	1. tax_file_found 2. receive_date 3. asset_type_s 4. vehicle_count_0	4 agree, 1 not sure

The taxpayer group from item 1 in Table 6 is classified through features such as the digital economy sector type and the date of receipt of the first-form which is considered to be general category but effectively influences the classification of the target class 'Non-Comp'. Additionally, features such as the amount of vehicle assets owned beyond the taxable profits as well as the value of stamp duty to which the acquisition of land is re-proposed. Data of this type is most common in RF algorithms after successful classification precision output using a wrapper technique with 1,487 documents of a taxpayer. Based on RF rules, all sectors of the digital economy are facing taxpayers who are tax-exempt except for crowd sourcing.

Rule 4 tree rule found an unexpected feature in the classification of the target class 'Non-Comp' where tax file indicators were found which meant that the presence of taxpayers who had reported to the IRBM branch. This should indicate that the taxpayer was committed to performing their responsibilities. However, the voluntary aspect of tax reporting and taxpayers may seek to claim tax relief for many reasons, in order to avoid the real loss of the digital economy combined with overall taxable revenues. A total of 1,486 records were detected using these rules.

COMPLIANCE CASE

This section will analyze the compliance case rules extracted from the best model in section 4.0. Table 8 shows the selected 'Comp' class target rules using the CART algorithm method in detail. The first group of taxpayers are those who have no track record of owning a vehicle, no property assets and also no stamp duty amount which signifies no purchase of real estate assets, which further enables the taxpayer's potential not to hide revenue generated from the digital economy sector as no additional revenue is reported. This group is predominantly estimated at 1,561 taxpayers and this clearly shows taxpayers reporting income tax related to conventional business activities and digital economies in the month ending December for the last 3 accounting years of 2017, 2018, and 2019. They are considered to comply with the income tax act on the scope of the imposition of digital economy tax under the context of IRBM.

The presence of the second largest class of 1,338 taxpayer records is known to have the same features as mentioned before, but there are different features in that it shows taxpayers reporting income tax related to conventional and economic activities digital is made in the month ending December 2019. There is a presence of bank account number information proving that banking transactions can occur for online income tax repayment payments or used in making financial loans such as real estate/housing and vehicle loans. This in turn provides an overview of the availability of banking status information available to the IRBM to resolve previously taxpayer cases.

TABLE 8. Expert evaluation of compliance (comp) rules

Rules description	Selected features	Expert decision (agree or not sure)
CART rules (1st branch tree) İf Comp = 1561 AND Estimated code type = 9,102,106,903 AND stamp duty return assets 'NO' AND tax income calendar year end period between 2017 and 2019 AND property asset ownership 'NO' AND tax income calendar month end period 12. Description of Rules There are 1561 compliance taxpayer cases correctly classified with type code of 9, 102, 106, and 903 but with no information on stamp duty and property asset in possession. All this occurs in December of the tax assessment year from 2017 to 2019.	1. sector receive_date 2. asset _type_s 3. vehicle _count_0	4 agree, 1 not sure
CART rules (2nd branch tree) If Comp = 1338 AND Estimated code type = 9,102,106,903 AND stamp duty asset ownership NO AND calendar year end of year between 2017 to 2019 AND property asset ownership NO AND calendar month end of account is 1-11 AND presence YES bank account number Description of Rules There are 1338 compliance taxpayers cases correctly classified with type code of 9, 102, 106, 903, and no information on stamp duty and property asset in possession, but they have the bank account number for proof. All this occur in the tax assessment year from 2017 to 2019, in another month except December		4 agree, 1 not Sure

CONCLUSION

This study proposed machine learning algorithms for classification modeling of tax compliance and non-compliance cases. Two approaches were employed namely single and ensemble classifications. In single classification, the CART algorithm performed the best among seven other algorithms and outperformed the ensemble methods. The rules extracted from the best CART model gives a wealth of knowledge that can assist the IRBM in managing digital taxation issues. Descriptive histograms conclude and correlate each other's features through preliminary and literary studies on the income tax compliance and the scope of digital economy taxation in the context of IRBM. The predictive models select the important features contributing to the classification of digital economy practitioners' compliance and non-compliance classes. The use or massive tax data lakes can further enhance the digital economy tax compliance model, and more discovered knowledge help the IRBM in making strategic decision. It will also help the government manage the revenue and plan for development programs that benefit the nation.

This research provides a progressive mechanism in identifying the selection of features in classifying the digital economy sector's level of income tax compliance to detect potential taxpayers at an early stage. Predictive analytics intend to find hidden transactions with a fast and efficient algorithm in facilitating data understanding. Overall, this study has three important research findings to the IRBM. Firstly, it supports the initiative of the big data analytics project in the IRBM, which is still in its infancy by contributing to some extent, the results of knowledge findings in machine learning regarding classification techniques. Secondly, by using descriptive and predictive model interpretation methods aims to determine the non-compliant taxpayers' category and vice versa for future use. Finally, this study's experimental results can be used as a reference and guide for future research in improving the classification model related to determining the digital economy sector's level of tax compliance in particular and analytical data in general.

ACKNOWLEDGEMENTS

We would like to thank the Inland Revenue Board of Malaysia (IRBM) for providing platform for data retrieval. This project is supported by Universiti Kebangsaan Malaysia Grand Challenge grant project under CP-2017-15/1.

REFERENCES

Adejo, O. & Connolly, T. 2017. An integrated system framework for predicting students' academic performance in higher educational. *International Journal of Computer Science & Information Technology (IJCSIT)* 9(3): 149-157. doi:10.5121/ijcsit.2017.93013

Ali, K., Pazzani, M. & Science, C. 1995. HYDRA-MM: Learning multiple descriptions to improve classification accuracy. *International Journal on Artificial Intelligence Tools* 4: 1-22.

Breiman, L.E.O. 1996. *Bagging Predictors*. Boston: Academic Publishers. pp. 123-140.

Castellón González, P. & Velásquez, J.D. 2013. Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications* 40(5): 1427-1436.

Cleary, D. 2011. Predictive analytics in the public sector: Using data mining to assist better target selection for audit. *Proceeding of the 11th European Conference on EGovernment: ECEG*. pp. 132-140.

Dhrubajyoti, D. 2017. Machine learning. *European Journal of Multidisciplinary Studies* 2(7): 255-258.

Freund, Y. & Schapires, R.E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *AT&T Labs* 139: 119-139.

Hamsagayathri, P. & Sampath, P. 2017. Decision tree classifiers for classification of breast cancer. *International Journal of Current Pharmaceutical Research* 9(2): 31-35.

Han, B.J. & Kamber, M. 2002. *Data Mining: Concepts and Techniques*. Beijing Machinery Industry Press 84: 92-99.

Jupri, M. & Sarno, R. 2018. Taxpayer compliance classification using C4.5, SVM, KNN, Naive Bayes and MLP. *International Conference on Information and Communications Technology (ICOIACT)*. pp. 297-303.

Lakshmi, R.D. & Radha, N. 2011. Machine learning approach for taxation analysis using classification techniques. *International Journal of Computer Applications* 12(10): 1-6.

LHDNM. 2018. *Risalah Ekonomi Digital LHDNM*.

Lin, C. & Lin, I. 2012. The application of decision tree and artificial neural network to income tax audit: The examples of profit- seeking enterprise income tax and individual income tax in Taiwan. *Journal of the Chinese Institute of Engineers* 35: 37-41.

Loo, E.C., Evans, C. & McKerchar, M.A. 2012. Challenges in understanding compliance behaviour of taxpayers in Malaysia. *Asian Journal of Business and Accounting* 3(2): 145-162.

Mithal, V., Nayak, G., Khandelwal, A. & Kumar, V. 2017. RAPT: Rare Class Prediction in Absence of True Labels. *IEEE Transactions on Knowledge and Data Engineering* 4347(c): 1-14. doi:10.1109/TKDE.2017.2739739.

Mohd Rizal, P., Mohd Rusyidi, M.A. & Wan Fadillah, B.W.A. 2013. The perception of tax payers on tax knowledge and tax education with level of tax compliance: A study the influences of religiosity. *ASEAN Journal of Economics, Management and Accounting* 1(1): 118-129.

Nellen, B. 2015. Taxation and today's digital economy. *Journal of Tax Practice & Procedure* 17: 17.

Pham, B.T., Bui, D.T., Prakash, I. & Dholakia, M.B. 2016. Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan Area (India) using GIS. *Catena* 149(Part 1): 52-63. doi:10.1016/j.catena.2016.09.007.

Tretter, M.J. 2003. *Data Mining*. Encyclopedia of information systems. Executive report.

Raja Azhan Syah Raja Wahab*
Sub Section of Strategic Planning
Strategic Management and Information ICT
Department of Information Technology
Inland Revenue Board of Malaysia
63000 Cyberjaya, Selangor Darul Ehsan
Malaysia

Azuraliza Abu Bakar
Center for Artificial Intelligence Technology
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor Darul Ehsan
Malaysia

*Corresponding author; email: rajazhan@hasil.gov.my

Received: 10 June 2020
Accepted: 19 November 2020