International Journal of Innovation, Creativity and Change. www.ijicc.net
Volume 12, Issue 3, 2020

Classification of Compliance Tax Payments for New Private Personal Tax using the Data Mining Method

Agus Bandiyono[a], Dewa Bagaskara[b], [a]Polytechnic of State Finance STAN, Tangerang, Indonesia, [b]Polytechnic of State Finance STAN, Tangerang, Indonesia, Email: [a]agus.bandiyono@gmail.com, agusbandiyono@pknstan.ac.id, [b]bagaskaara@gmail.com,

Increasing the number of new personal taxpayers from year to year is not followed by an increase in the level of compliance with tax payments. The number of new individual taxpayers who make tax payments is still below the target set in 2015 and 2016. The low level of tax compliance can be seen in the Bulukumba Primary Tax Service Office. Classification of tax payment compliance for new individual taxpayers can be done to provide guidance and information on whether a taxpayer is a compliant taxpayer or not. This information can be used to develop appropriate plans and strategies to improve tax payment compliance for new personal taxpayers. The classification and information were obtained from using methods of *data mining.* This study used the *data mining* method with techniques, attributes, classes, algorithms, and components or variables for the selection of a predetermined model. This study used an algorithm *decision tree* or C4.5 to classify *data mining*. The making and selection of models were done to be used as a reference model for classification. The model chosen in this study shows that the age variable has the most important role to play in classifying whether a new individual taxpayer is likely to comply or not related to his tax payment.

Key words: *Tax compliance, Data mining, Tax accounting, Management information system, Tax, State finance.*

Introduction

Based on the Republic of Indonesia Law number 6 of 1983 concerning general provisions and tax procedures as amended lastly with Law of the Republic of Indonesia number 16 of 2009, taxpayers can be divided into three categories, namely tax collector/tax collector, taxpayer agency and personal taxpayer. When seen from the total amount in each category of

taxpayers, taxpayers who have the largest number are those with the category of individuals. This large amount certainly provides a large tax potential, but in practice, the realisation of tax receipts for individual taxpayers is still less than what should be achieved.

Newly registered personal taxpayers are increasing from year to year. Therefore, the supervision of the newly registered personal taxpayer must be maximised to obtain maximum tax revenue. In practice, the supervision of this new individual taxpayer has not been able to run optimally when viewed based on the number of new taxpayers who make tax payments compared to the target set. The number of new individual taxpayers who make tax payments is always below the target set in 2015 and 2016 (Alisawi Shaymaa et al., 2019).

Based on the Minister of Finance Regulation number 234 /PMK.01 /2015 concerning the Organization and Administration of the Ministry of Finance and the Minister of Finance Regulation 206.2 /PMK .01 /2014 concerning the Organization and Work Procedures of the Vertical Directorate General of Taxes, the DGT Organization can be divided into Headquarters, Regional Offices, Tax Service Offices (KPP), and Service Offices, Counseling, and Tax Consultation. The Tax Service Office is a unit that has a direct duty to carry out tax revenues. The Tax Office consists of the Large Taxpayer Tax Office, Intermediary Tax Office, and Primary Tax Office. Primary Tax Office has to carry out counselling, service, and supervision of taxpayers in the areas of income tax, value-added tax, sales tax on luxury goods, other indirect taxes, land and building tax in the area of authority based on statutory regulations. Primary Tax Office is the Tax Office that has the highest number of units compared to other Tax Offices, which is 309 units compared to Large Taxpayer Tax Office with 4, and Middle Tax Office with 28 units (Puspitasari et al., 2019).

Being the unit with the most number, KPP Pratama has an important role to play in tax revenue activities at DGT. Primary Tax Office must collect tax receipts from taxpayers who are registered in their territories and who are not registered with the Middle Tax Office or Large Taxpayer Tax Office. Primary Tax Office has a greater number of taxpayers compared to Middle Tax Office and Large Taxpayer Office. Of the 309 KPP Pratama, some KPPs can achieve their tax revenue targets, and some KPPs cannot reach the specified tax revenue targets.

One of the KPP Pratama at DGT that cannot reach its tax revenue target is the KPP Pratama Bulukumba. In 2015, the realisation of Bulukumba Tax Office tax revenue was 247.06 billion rupiahs. The realisation was only around 86.12% of the 2015 tax revenue target of 286.87 billion rupiahs. The percentage of tax revenue realisation of the Tax Office in Bulukumba declined in 2016, which was only about 67.97% of the target set. The realisation of tax revenues in 2016 Bulukumba KPP Pratama was around 250.56 billion rupiahs with a tax revenue target of 368.65 billion rupiahs. Tax receipts at the Extensions and Counseling Section at KPP Pratama Bulukumba is the section responsible for receiving new personal

taxpayers over the past few years that have not reached the specified target. This is in line with the realisation of total tax receipts at KPP Pratama Bulukumba that is less than the specified target. For new individual taxpayers registered in 2014 and 2015 at KPP Pratama Bulukumba, seen from the payment of taxes in 2016, 135 taxpayers routinely pay their taxes every month, and 4,015 taxpayers do not comply in carrying out their tax payment obligations. The percentage of new individual taxpayers who regularly make their tax payments is only around 3.25%. The percentage of regular tax payments to the new individual taxpayer at the KPP Pratama Bulukumba can be said to be quite small. With only 3.25% new individual taxpayers who make regular tax payments every month, the realisation of tax revenue in the Tax Extensification and Counseling Section as the person in charge of tax revenue for this new personal taxpayer will be difficult to achieve. The Tax Extensification and Counseling Section at KPP Pratama Bulukumba has made efforts to increase tax revenue from this category of new individual taxpayers. To improve the performance of tax revenue from new personal taxpayers, there are alternative methods that can be used as a tool. The method is to extract information that does not appear in a business process, which can be used to improve the performance of the business process.

Hidden information can be obtained, for example, in the research of González and Velásquez (González & Velásquez, 2013), where the research aims to find the characteristics and detect taxpayers who use fake tax invoices in Chile. The results of this research are in the form of patterns *fraud* that can be used to predict whether taxpayers use false tax invoices or not (*fraud* or not *fraud*). In line with this research, this study seeks to obtain hidden information about tax revenue from new individual taxpayers. The information that can be obtained is related to whether a new individual taxpayer is likely to be a compliant taxpayer. To explore hidden information, techniques of data mining can be used.

Data mining is a technique that can be used to find this information in a data set. As the understanding of *data mining*, according to Zaki and Meira Jr. (Zaki et al., 2014), which is a process of finding insightful, interesting, and new patterns and models that are descriptive, easy to understand, and predictive of large-scale data. The results of *data mining* can be in the form of associations, clustering, classification, or prediction. *Data mining* techniques can be used as a tool to increase tax revenue; in this case, the tax revenue of new individual taxpayers.

Bulukumba Tax Office was chosen as the object of the study because tax receipts at the Bulukumba Tax Office have not reached the target set in the past few years. Besides, the small percentage of new individual taxpayers at KPP Pratama Bulukumba who are compliant in making tax payments is also a factor in choosing the object of research. Bulukumba Primary Tax Office oversees the Taxation Service, Counseling, and Consultation Office (KP2KP) which will be used as an additional attribute examined in this study, namely office

type attributes. The existence of additional attributes can add information that will be obtained in the *decision tree* generated from the method *of data mining*.

Research Methods

The lack of tax payment compliance from new individual taxpayers is an issue that will be examined in this study to apply alternative solutions to the problem. The alternative problem solving that will be carried out in this study is expected to be useful in increasing compliance with tax payments by personal taxpayers.

The alternative problem-solving in this study is to use *data mining techniques* to classify and predict whether new personal taxpayers are compliant or not related to tax payments. Things to consider in this research method are the techniques, attributes, classes, algorithms chosen, and components or variables for the selection of models that will be used in the *data mining* method.

This study uses *data mining* techniques with the *Cross Industry Standard Process of Data Mining* (CRISP-DM) which consists of several stages, namely *Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation,* and *Deployment*.
The stages of *data mining* with the CRISP-DM method are explained as follows.

Business Understanding. The core of the stage *business understanding* is to understand the business picture of the object under study and whether the plan *data mining* to be carried out is appropriate and used as an alternative solution to the problem in the object under study. *Business understanding* consists of several stages such as understanding business objectives, understanding the situation, determining goals *data mining*, and making work plans that will be carried out. This stage also includes the determination of classes to be used in classifications using data mining.

The class that is used as a classification in the data mining process in this study is the tax compliance class for new individual taxpayers, registered in 2014 and 2015. The tax payment compliance in this study is divided into two classes, namely compliant and non-compliant. The explanation of each class is as follows.

Obedient. Compliance classification in this study is where the new personal taxpayer since registered always fulfils his tax payment obligations by paying taxes every month that are seen in the test year, namely the 2016 tax payment.

The classification does not abide by this study is when there is a month in which the new individual taxpayer does not make tax payments in the year that was tested in the research that the payment of taxes in 2016.

***Data* understanding.** *This* is the process of understanding the data that will be used in *data mining*. This process starts with collecting data and understanding and analysing the data obtained.

***Data Preparation*.** Stages of *data preparation* include processing the data that is owned to be ready for use for *data mining*. From the initial data set, which attributes will be used in the process of *data mining*. This stage also includes cleaning data such as eliminating data that is not used, double data, blank data, and invalid data. This stage also includes the transformation of data to form a desired attribute in *data mining*. The results of the stage *data preparation* are a dataset that will be performed by *data mining*.

The attributes used in *data mining* in this study are the attributes that are directly attached and /or attributes that can be extracted in the new personal taxpayer data. The attributes used are as follows.

Month of registration. The attribute of the registration month is an attribute related to the month in which the taxpayer registers. The registration month attribute contains numerical data in the form of numbers from 1 to 12.

Birth month. The month of the birth attribute is an attribute that indicates the month in which the taxpayer was born. Birth month attribute contains data about the birth month of the new personal taxpayer in the form of numbers 1 through 12, which is numerical data.

Age. The age attribute indicates the age of the taxpayer when the tax payment compliance is seen; that is, the age at the end of 2016. The age attribute in this study will be classified into four age categories, namely ages 17 to 30 years, ages 31 to 40 years, ages 41 to 50 years, and ages over 50 years. The age attribute is nominal.

Office Type. Office type attributes are attributes that indicate the type of office where the taxpayer fulfils his tax obligations. Office type attributes are divided into two, namely KPP and KP2KP.

Gender. The sex attribute indicates the sex of the Taxpayer. The gender attribute contains nominal data about the gender of the new individual taxpayer, male or female.

Type of business. The business type attribute is an attribute that contains the type of business carried out by the taxpayer. Attributes of business types of taxpayers will be categorised into five categories, namely trade, services, industry, natural products, and free work. Attributes of type of business are nominal attributes.

Distance with KPP / KP2KP. Distance attribute with KPP / KP2KP contains data about the distance of residence of the new personal taxpayer with the KPP / KP2KP where he registers and fulfils his tax obligations. Distance attributes with KPP / KP2KP are categorised into four categories, namely 0 to 10 kilometres, 11 to 20 kilometres, 21 to 30 kilometres, and over 30 kilometres. The distance attribute with KPP / KP2KP is nominal.

Modelling. Modelling is the stage in which it will be performed several times modelling for classification process. *data mining* of the several models produced, one model will be chosen as the model to be used in research. Some of the models produced will have the best model measured based on several variables such as accuracy, ROC area, *precision* and recall, size, *kappa* statistic as well as *tree size* (the size of the tree) is generated. These variables are explained as follows.

Accuracy. Accuracy is related to how a model can classify data appropriately in each class. The accuracy value can be calculated based on the classification results in the *confusion matrix*. Examples of *confusion matrices* for two classes can be seen in Table 1.

Table 1: *Confusion Matrix* Two Classes

Classification		Results of Classification	
		Yes	No
Class	Yes	TrueTrue Positive (TP)	False Negative (FN)
	No	False Positive (FP)	True Negative (TN)

Source: Gorunescu, Florin. 2011.

Based on Table 1, Gorunescu (Gorunescu, 2011) explains the calculation of the accuracy value by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

In the Weka application, classification accuracy is indicated by the value of *correctly classified instances* that are from 0% to 100%. The higher the value of *correctly classified instances*, the better the resulting model accuracy.

Kappa statistics. The function of the *statistical kappa* according to Vierra and Garret (Viera & Garrett, 2005) is that the *statistical kappa is* intended to give the reader a quantitative measure of the amount of *agreement* between observers. The interpretation of values *kappa statistic* can be seen in Table 2.

Table 2: Interpretation Value *Kappa Statistic*

Number	Value *Kappa Statistic*	Agreement(Agreement)
1.	<0	*Less than chance agreement*
2.	0:01 to 0:20	*Slight agreement*
3.	0.21-0:40	*Fair agreement*
4.	0:41–0.60	*Moderate agreement*
5.	0.61–0.80	*Substantial agreement*
6.	0.81–0.99	*Almost perfect agreement*

Source: Viera, Anthony J., and Joanne M. Garrett. 2005.

Precision and *recall*. According to Witten, Frank, and Hall (2011), *precision* is a comparison between documents taken that are relevant to documents taken. *Recall* is a comparison between relevant documents taken with the total relevant documents.

Precision can be interpreted as data "a" that is classified correctly for a category "a" (*true positive*) compared to the total data classified for a category "a" (*true positive* + *false positive*). *Recall* can be interpreted as data "a" which is classified correctly for a category (*true positive*) compared to the amount of data "a" (*true* + *false negative*). The calculation of *precision* and *recall* based on Table 1, can be seen more clearly as follows.

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Area Under ROC Curve (ROC Area). Gorunescu (Gorunescu, 2011) describes the ROC (Receiver Operating Characteristic) Curve or ROC area as the intersection between *true positive* and *false positive* where the *true positive* rate as the Y-axis and *false positive* rate as the X-axis. The higher the ROC area, the better the classification model. ROC area values can be interpreted in five categories starting from the lowest classification *failure* to the best, *excellent classification*. The interpretation of ROC area values can be seen in Table 3.

Table 3: Interpretation of ROC Area Value

No.	ROC ValueArea	Classification
1.	0.90 - 1.00	Excellent Classification
2.	0.80 - 0.90	Good Classification
3.	0.70 - 0.80	Fair Classification
4.	0.60 - 0.70	Poor Classification
5.	<0.60	Failure

Source: Gorunescu, Florin. 2011.

Evaluation. This is the stage where an evaluation of the results of a classification in *data mining* using models that have been on stage *modelling.* Evaluations related to the classification algorithm *decision tree* will be performed both for the model and the resulting decision tree.

Deployment. The *deployment* stage is the stage of implementing model *data mining* that has been chosen as a model in research for the benefit of alternative solutions to the problems encountered. The resulting model can be used to determine the classification of a *data test.*
The algorithm used for classification in *data mining* is the algorithm *decision tree* or C4.5 which in the Weka application is named the J48 algorithm.

Wu *et al.* (2008) explained that the C4.5 algorithm is a development of the CLS and ID3 algorithms. C4.5 algorithm produces classification results described as *decision trees*. C4.5 can also produce classifications in the form *rulesets* of more comprehensive.

Picture of *decision tree,* according to Cios *et al.* (2007), is *a decision tree* composed of *nodes* and *branch* iceconnecting nodes. *Nodes* that are located at the bottom of the *decision tree are* called *leaves* which are a class. *The node* at the top of the *decision tree is* called *root*. All *nodes* except *leaves* and *root are* called *decision nodes*.
The object of this research is the data of new personal taxpayers at KPP Pratama Bulukumba registered in 2014 and 2015 and the tax payment data for 2016 as data *training*. The test data or *data test* in this study is the data of new taxpayers newly registered in 2017. The population of the data *training* is all-new personal taxpayers registered in 2014 and 2015. Samples of data that are processed as data are *training* obtained after cleaning the data in the DataStage *preparation* of the population. The data used for test data or *data test* are all individual taxpayers newly registered in 2017 at the KPP Pratama Bulukumba.

Results and Discussion

Data in this study indicate the symptoms of (*unbalanced data imbalance dataset*). The data in the study show that the composition of the data that are not balanced is equal to 135 data is

the compliant class, and 4,015 data is the non-compliant class data. Overcoming data that are not balanced can be done by the method *oversampling. Oversampling* can be done one of them with *Synthetic Minority Over-Sampling Technique* (SMOTE) as research by Chawla *et al.* (2002). This research will use SMOTE as an alternative to *modelling*. The type of model in this research related to the composition of the data is modelling on (*unbalanced data imbalance dataset*) and data that is *over templated* using SMOTE with a percentage of 100% and 200%.

Classification with *data mining* using the C4.5 algorithm will produce a decision tree. The decision tree consists of the number of leaves and the size of the tree. One of the techniques informing decision trees in *data mining* is *running*. The results of the decision tree concerning *pruning* can be divided into two types, namely a decision tree made by pruning leaves and a decision tree made without any pruning processes. The number of leaves and size of the tree produced in the decision tree is one component in determining which model will be selected in the study.

The technique of pruning leaves can be divided into two, namely *online running* and *post running*. *Online running* is trimming by providing a limited number of *instances* of at least *(minimum number* instances) that a leaf can be formed. *Post running* is a trimming process by changing the value of the *confidence factor* in the classification. At the stage of *modelling* in the study, the pruning process will be used so that there are three alternative modelling associated with trimming the leaves that decision trees are produced without trimming the leaves, decision trees are generated by pruning leaves *online pruning*, and decision trees are produced by pruning *post running*.

The classification models produced in this study amounted to 19 models named M1 through M19. The formation of 19 models is based on variation *oversampling* 100% and 200%(SMOTE) and technique *pruning*. Model variations on leaf pruning techniques are models without leaf pruning (*unpruned*), *online running* with variants *MinNumObj* used are 5, 10, 15 and 20, and *post pruning* with values of *confidence factor* 0.1, 0.25, 0.5, and 0.75.

Based on the models that have been made, namely 19 classification models, one model will be chosen to be used as a research model along with the results of the decision tree. To choose a model, this study considers several things such as accuracy, *kappa statistics*, ROC area, *precision* and *recall*, number of leaves and tree size. A summary of the entire model formed along with the factors selecting the model can be seen in Table 4.

Table 4: Summary of Models M1 through M19.

Model	Accuracy	*Kappa Statistic*	*Area ROC*	Total Leaf	Tree Size
M1	96.9639%	0.165	0.604	64	86
M2	94.3128%.	0.6475	0.908%	145	240
M3	93.5003	0.5887	0.882	76	123
M4	93.0264%	0.5228	0.814	35	52
M5	92.8459%	0.5029	0.832	24	35
M6	92.733%	0.4853	0.810	13	18
M7	92.733%	0.4853	0.810	13	18
M8	92.8684%	0.4988	0.808	15	22
M9	93.9291%	0.6063	0.873	70	117
M10	94.3128%	0.6475	0.908	136	226
M11	94.2487%	0.769	0.944	187	320
M12	93.1664%	0.7227	0.915	81	138
M13	92.8056%	0.7043	0.913	75	126
M14	91.6171%	0,629	0.860	41	63
M15	91.2139%	0.6083	0.837	30	45
M16	91.8081%	0.6298	0.854	33	51
M17	93.2725%	0.719	0.907	81	135
M18	93.9092%	0.7508	0.928	118	208
M19	94.2912%	0.7707	0.943	172	297

Source: processed from data processing in Weka 3.8.1

Factors of simplicity of decision trees produced became the first factor to be considered in the selection of the model chosen in this study. A simple decision tree will interpret the decision tree to determine the class of compliance for paying a new individual taxpayer more easily. Conversely, a decision tree with a large number of leaves and tree size will make interpretation difficult and more complicated when used to determine the class of new personal taxpayers. Based on this, this study will eliminate the classification model that has a large decision tree size. This study will eliminate models that have tree sizes above 100 and/or models that have several leaves above 50. Classification models that have tree sizes above 100 and/or the number of leaves above 50 are models M1, M2, M3, M9, M10, M11, M12, M13, M17, M18, and M19. After the models are eliminated in the first stage of model selection, the next stage is that the surviving models will be chosen based on their accuracy. The summary that persists after eliminating the first step along with its accuracy value is shown in Table 5.

Table 5: Model Summary in the Second Stage Model Selection

ModelModel	Accuracy
M4	93.0264%
M5	92.8459%.
M6	92.733%
M7	92.733%
M8	92.8684%
M14	91.6171%
M15	91.2139%
M16	91.8081%step

Source: processed from data processing in Weka 3.8.1

The second in the selection is to look at the model based on its classification accuracy. The accuracy is indicated by the value of *correctly classified instances*. The level of accuracy shows what percentage of a model can classify *instances* correctly according to their class. The accuracy of the eight models is not much different. All of them have a high degree of accuracy, which is above 90%. The lowest level of accuracy is model M15, that is 91.2139% and the highest level of accuracy is model M4, that is 93.0264%. In the second stage of model selection, eight existing models will be retained because all of them have a high degree of accuracy with little difference. The eight models will be continued to enter the third stage of model selection.

The third step in selecting the model is to eliminate the model based on the value *kappa statistical*. Eight classification models still survive in the third stage of model selection. A summary of the model in the third stage of the model selection along with its values *kappa statistic*s is shown in Table 6.

Table 6: Summary of the Model in the Third Stage Model Selection of the

Model	*Kappa Statistics*
M4	0.5228
M5	0.5029
M6	0.4853
M7	0.4853
M8	0, 4988
M14	0.629
M15	0.6083
M16	0.6298

Source: Compiled from data processing in Weka 3.8.1.

In 6 of the existing eight models, five models have a value of *kappa statistic* with the category *moderate* agreement. Category *moderate agreement* in the *kappa statistics* is a model with a value *statistical kappa* between 0.41 to 0.60. Models with values *kappa statistic moderate agreement,* namely models M4, M5, M6, M7, and M8. In addition to the five models with a value *kappa statistical* in the category *moderate agreement*, there are three models with value *kappa statistical* in the substantial agreement category. Category *kappa statistic* with a category of *substantial agreement* is a model with values *kappa statistic* between 0.61 up to 0.80. The model with a value *statistical kappa* in the category *moderate agreement* is M14 with a value *kappa statistical of* 0.629, model M15 with a value *statistical kappa of* 0.6083 rounded to 0.61, and model M16 with a *statistical kappa of* 0.6298. In the third stage, the selection of models in this study will eliminate models with lower values *kappa statistical*, namely for models with a category of *kappa statistics moderate agreement*. The models eliminated in the third stage are models M4, M5, M6, M7, and M8.

The fourth step in the selection of models in this study is to compare the value of the ROC area and the value of *precision* and *recall*. This stage will choose a model that will be used as a research model by comparing the ROC area values and values *precision* and *recall*. The model that will be chosen in the fourth stage is among the three surviving models, namely M14, M15, and M16. A summary of the fourth stage of the model showing the ROC area and *precision* and *recall* is shown in Table 7.

Table 7: Summary of Fourth Phase Selection Model onModel

Variable	M14	M15	M16
area ROC	0.860	0.837	0.854
Precision -Obey	0.789	0.777	0.817
Recall - Obey	0.591	0.570	0.575
Precision - Not Complying	0.932	0.929	0.930
Recall -Not Complying	0.973	0.972	0.978

Source: Compiled from data processing in Weka 3.8.1

Based on the value of the three models ROC area classification, all three have a value of ROC area with category *good classification,* i.e. ROC area ranges between 0.8 and 0.9. To determine the model to be selected based on the ROC area value, it will be selected for the model with the highest ROC area value because all three models have the same ROC area classification with a little difference between models. The model with the highest ROC area value is model M14 with a value of 0.860. Model M14, when seen in the value of *precision* and *recall*, has the highest value for the value *recall* in the compliant class and the value of *precision* for the non-compliant class. The value of *recall* model M14 for the obedient class of 0.591 is the highest compared to model M15 with 0.570 and modelM16 with 0.575. The value of *precision* in the non-compliant class for model M14 of 0.932 is the largest compared

to modelM15 with 0.929 and model M16 with a value of 0.930. Based on the results of the selection of the fourth stage by using the ROC area value as well as the value of *precision* and *recall*, the model that has been chosen to represent this research is model M14.

The model chosen was model M14 will see how the decision tree is produced. Model M14 produces a decision tree with 41 leaves and tree size of 64. About which is used as the root or *root node* in model M14 is the age attribute. The age attribute divides age into four categories, namely 17-30, 31-40, 41-50, and over 50. On the age attribute as the root of the decision tree, model M14 classifies a new individual taxpayer with ages of 17-30 and ages over 50 in non-compliant classes. At the ages of 31-40, the classification will continue with the next attribute, Business Attributes. For ages 41-50, the classification will also continue with the next attribute, Office Type attribute.

Based on the results of the decision tree of the selected model in the study, namely model M14, it can be interpreted that the attribute that has the most important role in determining whether a taxpayer is compliant or not is the age attribute. The age attribute is used as the root of the tree in the decision tree model M14. Compliant taxpayers are for those with ages of 31-40 and 41-50. Based on the decision tree produced, taxpayers aged 17-30 and over 50 will tend to be non-compliant in making tax payments.

The next interpretation is on the attributes that affect the classification of taxpayer payment compliance in the form of business type attributes. Business types with trade categories will tend to produce taxpayer compliant classifications when compared to other types of businesses.

The next attribute that affects classification is the distance attribute with the KPP /KP2KP. Taxpayers who have a distance of residence with KPP / KP2KP, namely 0-10 and over 30 kilometres, are more likely to be classified in the compliant class compared to other distance categories.

The next attribute which can be seen in the classification is the gender attribute in which taxpayers with male gender tend to be greater to be classified in the class of obedience compared to female taxpayers.

The LTO type attribute has a weight that is balanced enough to be classified in an obedient class. The KP2KP office type has two patterns of formation of an obedient class, while the KPP has three patterns of formation of an obedient class. The attributes of the month of registration and the month of birth have different patterns depending on the pattern of the previous attributes in the decision tree.

No previous studies were found that used methods of *data mining* that have the same attributes as this study. The interpretation of the decision tree in model M14will be linked to related research that can be compared. One study that can be used as a comparison is Schuetze's study (Schuetze, 2002). Schuetze (2002) investigated the profile of non-compliant entrepreneurial taxpayers in Canada in 1969-1992.

One of the results of this study is that the level of non-compliance of taxpayers decreases with age. In other words, entrepreneur taxpayers have improved tax compliance in line with increasing age. There is a match between the results of Schuetze's research (Schuetze, 2002) with the decision tree results in model M14. In the age category in model M14 that is 17-30 years of age tend to be disobedient, and in the next age, category tends to be obedient, namely at 31-40 and 41-50 years of age. What distinguishes the results of this study with Schuetze's research (Schuetze, 2002) is that the age category over 50 years shows results that tend to be disobedient while Schuetze's research (Schuetze, 2002) shows the results of entrepreneurial taxpayers who are increasingly obedient at the age of 55-64.

Another result in Schuetze's research (Schuetze, 2002) is in the category of entrepreneurial taxpayers' jobs, where the job category *sales* tend to be more compliant compared to other jobs such as *services, professional and technical,* and *construction*. This is in line with the results of the decision tree in model M14 where taxpayers with the category of a trading business are more likely to be classified in the class of compliance than other types of work.

The sex attribute in this study can also be attributed to the Schuetze study (Schuetze, 2002). The results of Schuetze's research (Schuetze, 2002) show that there is a slight difference in the level of non-compliance for the gender category of entrepreneurs taxpayers. Entrepreneur taxpayers with husband status tend to be a little more obedient compared to entrepreneurial taxpayers with wife status. This is in line with the results of this study wherein the gender attribute, taxpayers with male gender tend to be more obedient than taxpayers with the female gender.

The results of the classification using model M14 for the *data of the* Individual Taxpayer newly registered in 2017 at the KPP Pratama Bulukumba can be linked to the discussion on the classification model. In the age attribute, taxpayers have a greater likelihood to be classified in the adherent class at the age of 31-40 and 41-50. Conversely, at the age of 17-30 and over 50, the taxpayer tends to be classified in the non-compliant class.

The results on the age attribute show that of 60 new registered taxpayers who are 2017 classified as a compliant class, a total of six taxpayers have an age attribute of 31-40 years and 54 taxpayers have an age of 41-50. In 60 taxpayers with obedient classes, there are only

taxpayers with age categories 31-40 and 41-50 years, and there are no taxpayers aged 17-30 and over 50.

All age categories are found in 1,705 taxpayers whose classification results are not compliant. There are 1,705 taxpayers with non-compliant classification, consisting of 346 taxpayers aged 17-30, 590 taxpayers aged 31-40, 519 taxpayers aged 41-50, and 250 taxpayers with age over 50.

The results of the classification of the *data test* when compared with Schuetze's research (Schuetze, 2002), there is a correspondence in which taxpayers with compliant classification increasingly increase with age. This is indicated by compliant taxpayers at the age of 31-40 more than the age of 17-30, which is several six taxpayers aged 31-40 compared to zero taxpayers at the age of 17-30. Taxpayers classified as compliant at 41-50 years of age are also greater than compliant taxpayers at the age of 31-40, which is 54 taxpayers compared to six taxpayers.

The difference in the classification results in this study compared to the Schuetze study (Schuetze, 2002) is seen in the absence of taxpayers who are classified in the adherent class aged over 50. In contrast, Schuetze's research (Schuetze, 2002) shows that taxpayers aged 55-64 have a level of compliance the best. The results of the classification of the age attribute are shown in Table 8.

Table 8: Results of the classification of the *Dataset* Age Attribute

Age	Compliant	Obedient
17-30	0	346
31-40	6	590
41-50	54	519
> 50	0	250

Source: Processed from the results of the classification *data test* using Microsoft Excel.

Comparison of classification results with the discussion of the model on the attributes of the type of business that is the taxpayer with the type of trading business tends to be classified into obedient class compared to other attributes. Schuetze's research (Schuetze, 2002) also shows a consistent result which entrepreneurial taxpayers in the field of *sales* show better tax compliance compared to several other fields.

The classification results show that all of the 60 taxpayers with compliant classification fall into the category of trading business types. The composition of business types at 1,705 taxpayers with non-compliant classification is 1,070 trade taxpayers, 229 service taxpayers, 277 natural product taxpayers, 128 industrial taxpayers, and one free work taxpayer. These

results are in line with Schuetze's research (Schuetze, 2002) which trade taxpayers have results of 60 compliant taxpayers compared to other types of businesses where there are no taxpayers with compliant classification. The results of the classification of the business type attributes are shown in Table 9.

Table 9: The results of the classification *data test* on the Attributes of the Type of Business

Business	Compliant	Non-Compliant
Trade	60	1,070
Services	0	229
Natural Products	0	277
Industry	0	128
Free Work	0	1

Source: Processed from the results of the classification *data test* using Microsoft Excel.

The next comparison of the results of the classification with the discussion of the model is the gender attribute. Model M14 shows the results of the decision tree on the gender attribute, which shows that male taxpayers tend to be more likely to be classified as obedient classes than female taxpayers. Schuetze's research (Schuetze, 2002) shows results with a slight difference in husband and wife taxpayer compliance. The husband's entrepreneurial taxpayer has slightly better compliance than his wife's entrepreneurial taxpayer.

The results of the classification of the sex attributes in this study for the *data test* are 60 taxpayers for the compliant class with a composition of 59 male taxpayers and one female taxpayer. For non-compliant classification, 1,113 were male taxpayers, and 592 were female taxpayers. When compared with Schuetze's study (Schuetze, 2002), the results show suitability and a mismatch. Following Schuetze's research (Schuetze, 2002) in terms of the number of compliant male taxpayers is greater than women, the discrepancy is the difference between male and female taxpayers which is quite large, which is 59 compared to 1. The classification results on gender attributes are shown in Table 10.

Table 10: Results of classification *Dataset* in Attributes of Business

Type Gender	Compliant	Obedient
Men	59	1,113
Women	1	592

Source: processed from the results of the classification *data test* using Microsoft Excel

Conclusions

The conclusions that can be drawn from the entirety of this study are, this study resulting in 19 classification models based on unbalanced data composition, data with *oversampling*

100% SMOTE, and data with *oversampling* 200% SMOTE. The components of leaf pruning informing the model consist of three types, namely without pruning, pruning *online pruning,* and pruning with a *post.* Of the 19 models in the study, one model was chosen as the model that represented the research. The model was chosen by considering accuracy, *kappa statistics*, ROC area, *precision* and *recall*, and the number of leaves and tree size. The chosen model is model M14 which has an accuracy value of 91.6171%, a value *kappa statistical of* 0.629, and a ROC area of 0.860. The values *precision* and *recall* for the obedient class in model M14 are 0.789 and 0.591, and the non-compliant class is 0.932 and 0.973. Model M14 has 41 leaves and tree size of 64. The decision tree produced by model M14 defines the age category attribute as the *root node*. Age categories 17-30 and over 50 are immediately classified in non-compliant classes. The classification for the age categories 31-44 and 41-50 continues in the following attributes.

Based on the decision tree in the M14 classification model chosen in this study, the age attribute is the attribute that has the most significant role to play in classifying whether a new individual taxpayer is likely to be compliant or not compliant concerning his tax payment.

Suggestions related to this research include, future studies can use different classification algorithm models and/or more data. Model M14 can be used by KPP Pratama Bulukumba and other KPPs that have similar characteristics to KPP Pratuk Bulukumba to classify newly registered personal taxpayers. Model M14 can be used to predict whether the taxpayer is likely to be compliant with his tax payment obligations. The classification results based on model M14 for newly registered individual taxpayers are expected to be used for Bulukumba Primary Tax Offices and other KPPs that have similar characteristics to the Bulukumba Primary Tax Office to implement appropriate strategies to improve tax compliance for new personal taxpayers. KPP can focus on monitoring the compliance of tax payments for new individual taxpayers with characteristics of ages 17 to 30and over 50, with types of business other than trafficking and female sex, in which these characteristics have a greater likelihood of classified in the class that does not comply with the payment of taxes.

REFERENCES

Alisawi Shaymaa, S. A. A., Hani Al-Zameli, A. A., & Kadhim Sendw, A. (2019). Tax harmonisation of tax accounting procedures and income of industrial companies. Edición Especial, 35(23).

Aprianang, D. (2016). Pengaruh Persepsi Tentang Kontraprestasi Pajak Per Sektor Terhadap Kepatuhan Perpajakan. Tangerang Selatan: Skripsi Program Studi Akuntansi Politeknik Keuangan Negara STAN.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Chawla, N. V., et al. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

Cios, K. J., et al. (2007). Data Mining: A Knowledge Discovery Approach. Springer Science+Business Media, LLC.

Cross Industry Standard Process for Data Mining (CRISP-DM). http://crisp-dm.eu/.

da Silva, L. S., Rigitano, H., Carvalho, R. N., & Souza, J. C. F. (2016, June). Bayesian Networks on Income Tax Audit Selection-A Case Study of Brazilian Tax Administration. In BMA@ UAI (pp. 14-20).

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

Gorunescu, F. (2011). Data Mining: Concepts, models and techniques (Vol. 12). Springer Science & Business Media.

Han, J., & Kamber, M. (2012). Data Mining: Concept and Techniques. Edisi ke-3. Waltham: Morgan Kauffman Publishers.

Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. International Journal of Advanced Computer Science and Applications, 4(2), 0-0.

James, S., & Alley, C. (2002). Tax compliance, self-assessment and tax administration. Journal of Finance and Management in Public Services, 2(2), 27-42

Jantan, H., Hamdan, A. R., & Othman, Z. A. (2009). Data Mining Classification Techniques for Human Talent Forecasting. InTech.

Laporan Kinerja Direktorat Jenderal Pajak tahun 2016.

Laporan Tahunan Direktorat Jenderal Pajak tahun 2015.

Larose, D. T. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. New Jersey: Jhon Wiley & Sons, Inc.

Palil, M. R. (2010). Tax Knowledge And Tax Compliance Determinants in Self Assessment System In Malaysia. University of Birmingham Research Archive.

Peraturan Menteri Keuangan nomor 206.2/PMK.01/2014 tentang Organisasi dan Tata Kerja Instansi Vertikal Direktorat Jenderal Pajak.

Peraturan Menteri Keuangan nomor 234/PMK.01/2015 Tentang Organisasi dan Tata Kerja Kementerian Keuangan.

Peraturan Pemerintah Republik Indonesia Nomor 46 Tahun 2013 tentang Pajak Penghasilan atas Penghasilan Dari Usaha yang Diterima atau Diperoleh Wajib Pajak yang Memiliki Peredaran Bruto Tertentu.

Puspitasari, L., In'am, A., & Syaifuddin, M. (2019). Analysis of Students' Creative Thinking in Solving Arithmetic Problems. *International Electronic Journal of Mathematics Education, 14*(1), 49-60. https://doi.org/10.12973/iejme/3962

Rajaraman, A., & Ullman, J. D. (2014). Mining of Massive Datasets. Cambridge University Press.

Schuetze, H. J. (2002). Profiles of Tax Non Compliance Among the Self-Employed in Canada: 1969-1992. Univesity of Victoria.

Schuetze, H. J. (2002). Profiles of tax non-compliance among the self-employed in Canada: 1969 to 1992. Canadian Public Policy/Analyse de Politiques, 219-238.

Surat Edaran Direktur Jenderal Pajak Nomor SE-37/PJ/2015 tentang Pengawasan Wajib Pajak Baru.

Undang-Undang Nomor 6 tahun 1983 tentang Ketentuan Umum dan Tata Cara Perpajakan sebagaimana telah diubah terakhir dengan Undang-Undang Nomor 16 tahun 2009.

Undang-Undang Nomor 7 Tahun 1983 tentang Pajak Penghasilan sebagaimana telah diubah terakhir dengan Undang-Undang Nomor 36 Tahun 2008.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. Fam med, 37(5), 360-363.

Widiastuti, D. (2008). Analisa Perbandingan Algoritma SVM, Naïve Bayes, dan Decision Tree dalam Mengklasifikasikan Serangan (Attacks) pada Sistem Pendeteksi Intrusi. Jakarta: Universitas Gunadarma.

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Edisi ke-3. San Fransisco: Morgan Kauffman Publishers.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

Zaki, M. J., Meira Jr, W., & Meira, W. (2014). Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press.