djp

Improving Tax Administration with **Artificial Intelligence**: Bringing **Artificial Neural Networks** (ANN) into Practice

29 September 2022

Arifin Rosid

Chief of Tax Policy Impact
Directorate of Tax Potential, Revenue, and Compliance, DGT, MoF

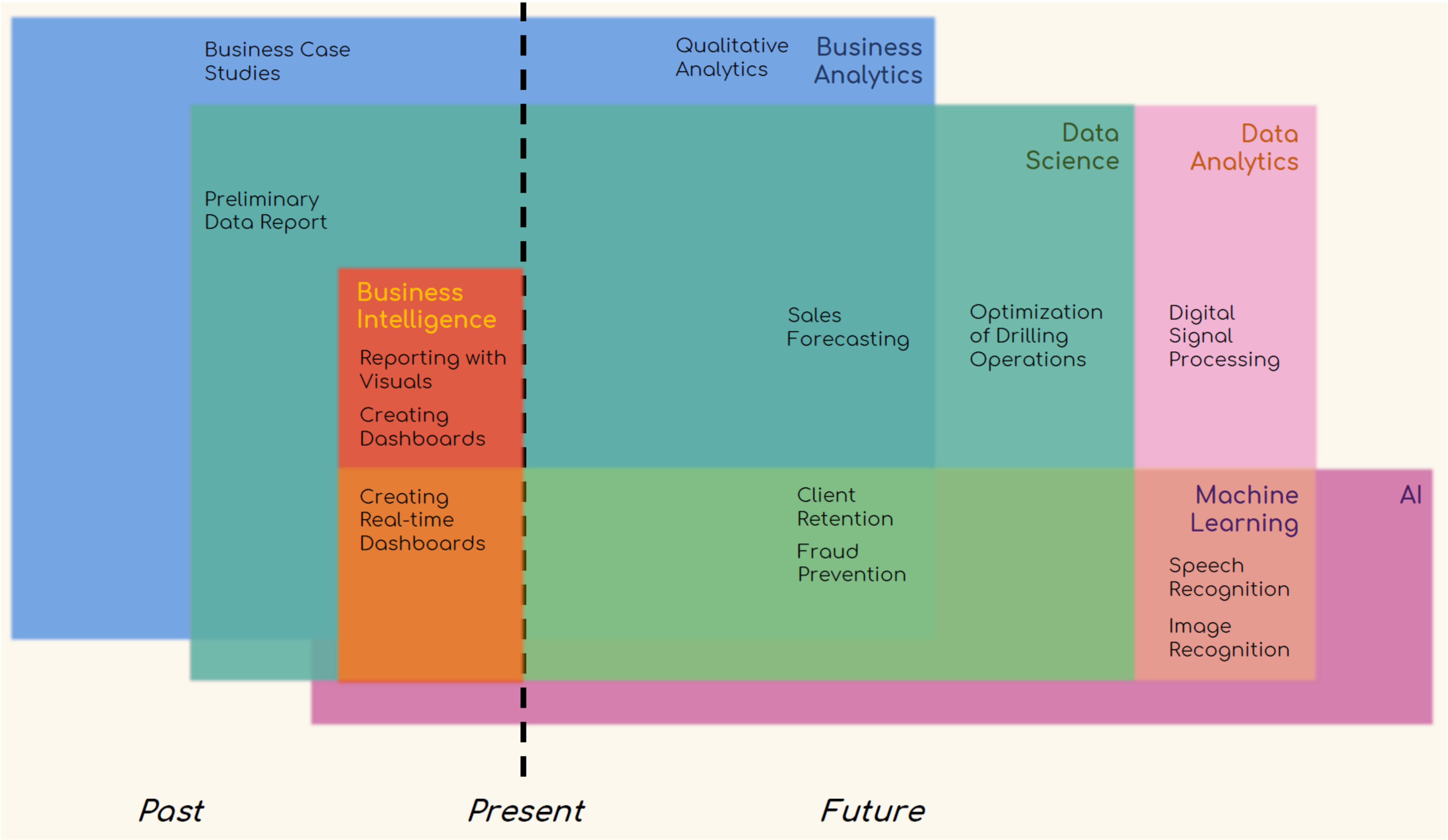
Contents (and caveats)

- Introduction and conception
- Live demo
- The real case using administrative data

What Makes AI So Popular Today?

- AI is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence.
- Artificial intelligence leverages computers and machines to mimic the problem-solving and decision-making capabilities of the human mind.
- Gaining popularity due to: (i) increased computing power and storage, (ii) availability of big data, (iii) advancement in algorithms.

Where are we?



Business Case Studies
Qualitative Analytics
Business Analytics
Data Science
Data Analytics
Preliminary Data Report
Business Intelligence
Reporting with Visuals
Creating Dashboards
Sales Forecasting
Optimization of Drilling Operations
Digital Signal Processing
Creating Real-time Dashboards
Client Retention
Fraud Prevention
Machine Learning
AI
Speech Recognition
Image Recognition
Past
Present
Future

Machine Learning

- Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that **humans learn**, gradually improving its accuracy
- Machine learning is an important component of the **growing field of data science**.
- Through the use of statistical methods, algorithms are trained to make **classifications or predictions**, and to **uncover key insights** in data mining projects.

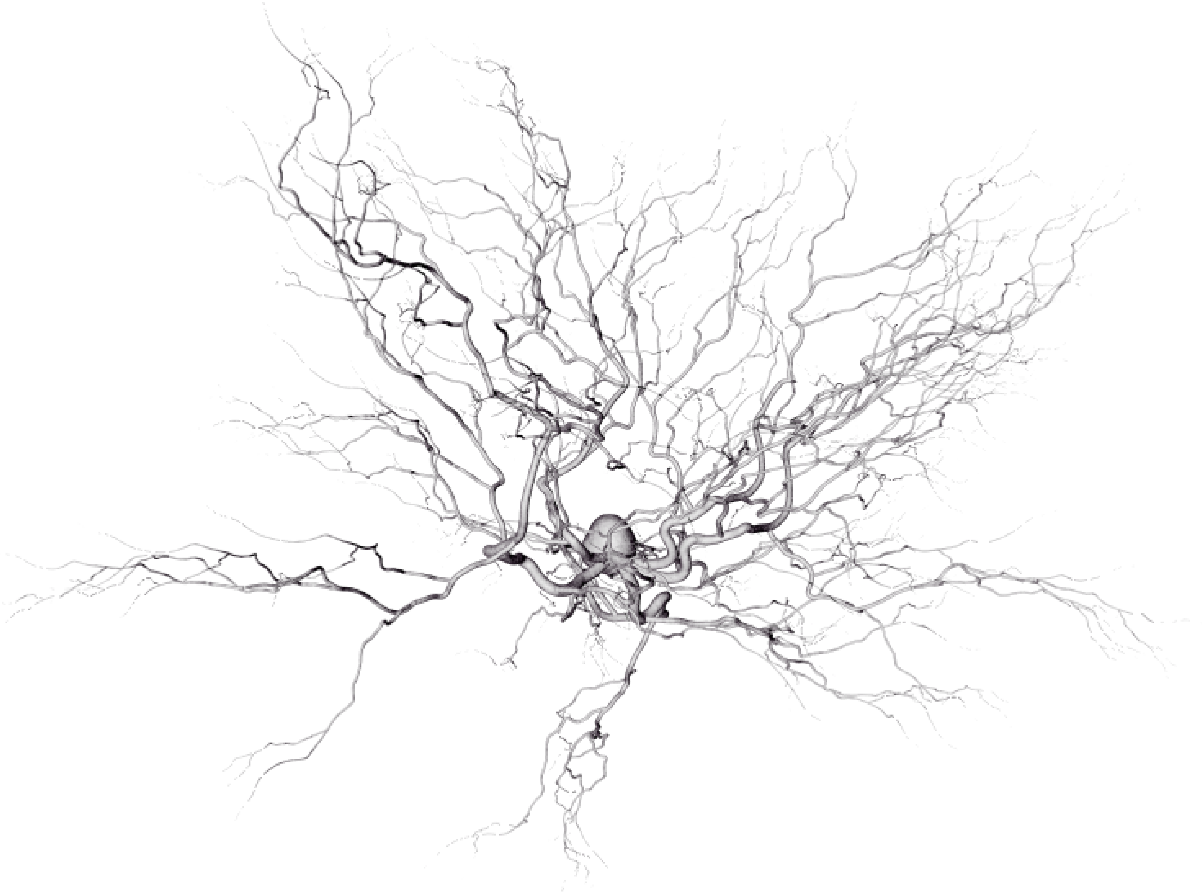
What is ANN?

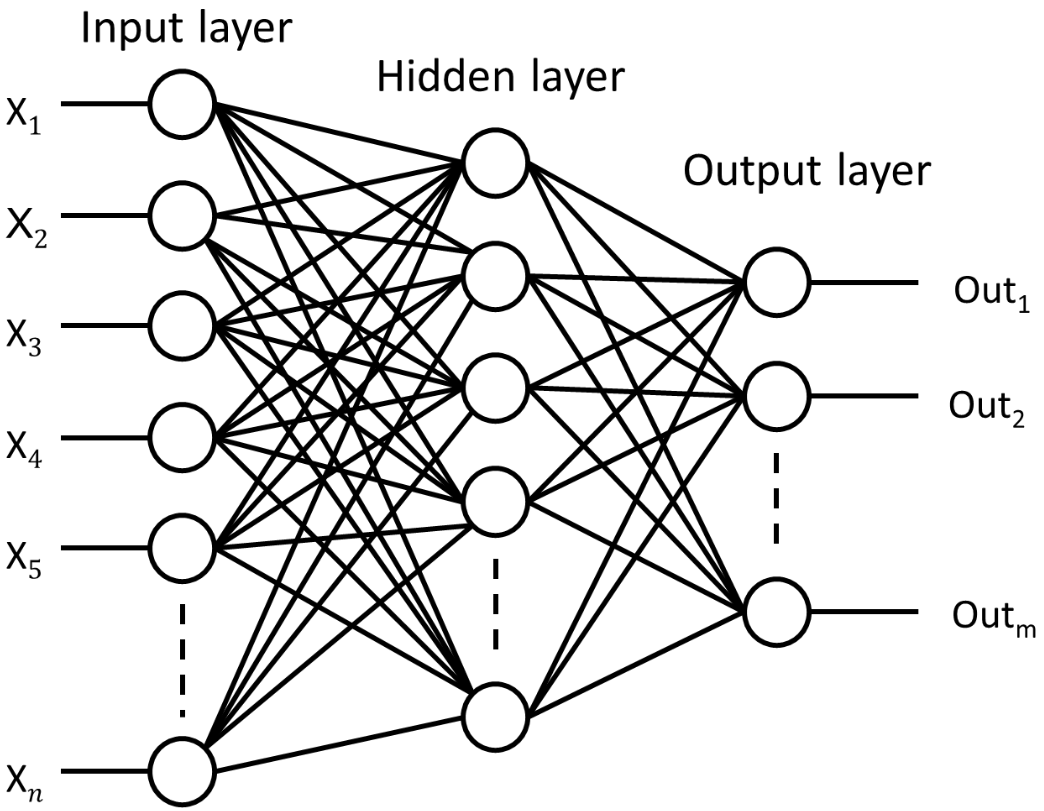
- An artificial neural network (ANN), also known as neural network (NN), is a computational structure composed of a number of interconnected data processors called neurons or nodes.
- They are inspired from the structure of biological nervous systems.
- Neural networks are powerful modeling tools because they can detect complex, nonlinear relationships between inputs and outputs.

Overview the classification algorithms

	Advantages	Disadvantages
Logistic regression	▪ Models are often very accurate ▪ Works well on small datasets ▪ Predicts probabilities ▪ Easy to interpret, in particular, the influence of each input variable	▪ It can only provide linear solutions ▪ Problems with high collinearity of the input variables
Linear discriminant analysis	▪ Typically, very fast building the model ▪ Works well on small datasets ▪ Optimal if data assumptions are fulfilled	▪ More restrictive assumptions than other methods (e.g., logistic regression) ▪ Usually, needs data preparation ▪ Sensitive to outliers, only applicable to linear problems
Decision trees	▪ Robust to outliers ▪ Model and decision rules are easy to understand ▪ Can handle different data types ▪ Fast in prediction and no assumptions on variable distributions needed. ▪ Can handle missing values	▪ Can be computationally expensive to train ▪ Large trees tend to overfitting ▪ Most of the time it does not find the optimal solution ▪ Prefers variables with many categories or numerical data
Neural networks	▪ Good performance on large datasets ▪ Very good at allowing nonlinear relations and can generate very complex decision boundaries ▪ Non-parametric, no distribution assumptions needed ▪ Can handle noisy data ▪ Often outperforms other methods	▪ Training can be computationally expensive ▪ Results and effects of input variables are hard to interpret (black box algorithm) ▪ Tends to overfitting and does not always find the optimal solution

Biological NN vs ANN architecture





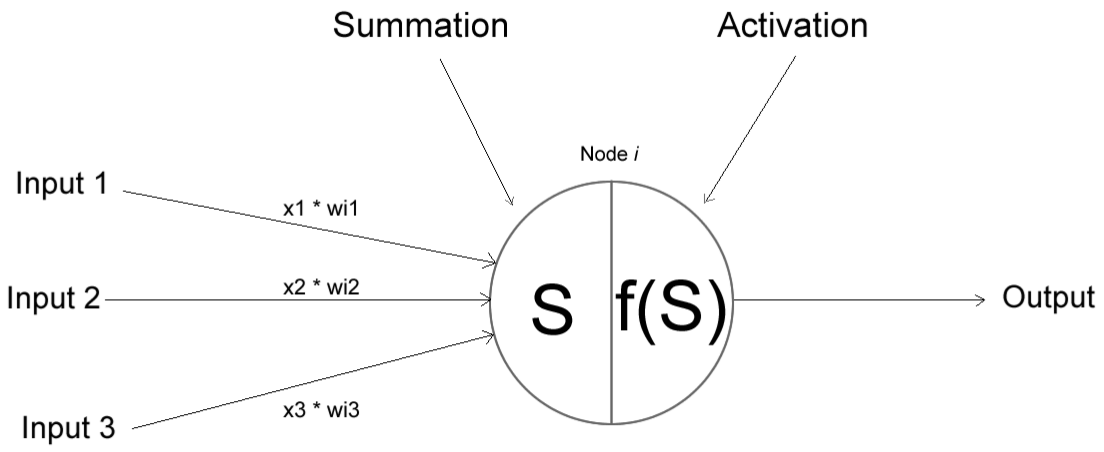
Two Classes of ANN

The neural networks could be divided into two big classes:

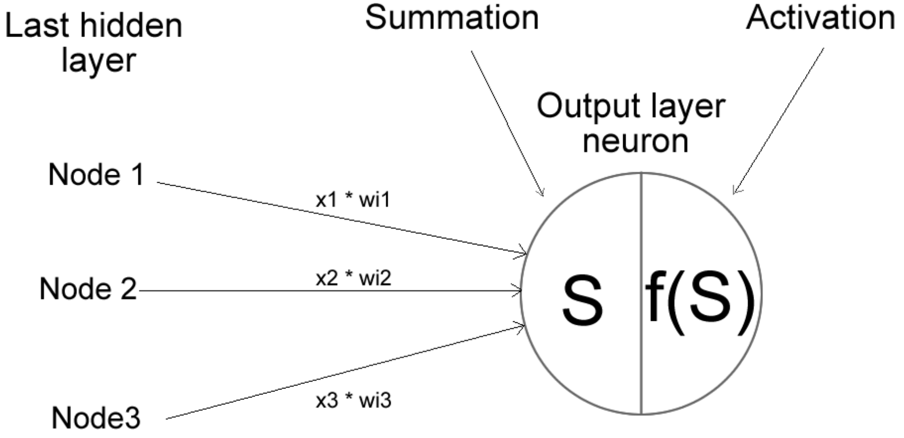
- **Feedforward networks**. In a feedforward network, the information (signals) is transmitted in one direction only: from the input to the output layer (passing through the hidden layers).
- **Recurrent networks** (aka feedback or interactive networks). In a recurrent network, the information can travel both ways (from input to output and conversely) by using loops.

What Happens Inside of a Neuron?

Operations performed in a neuron: (i) summation and (ii) activation



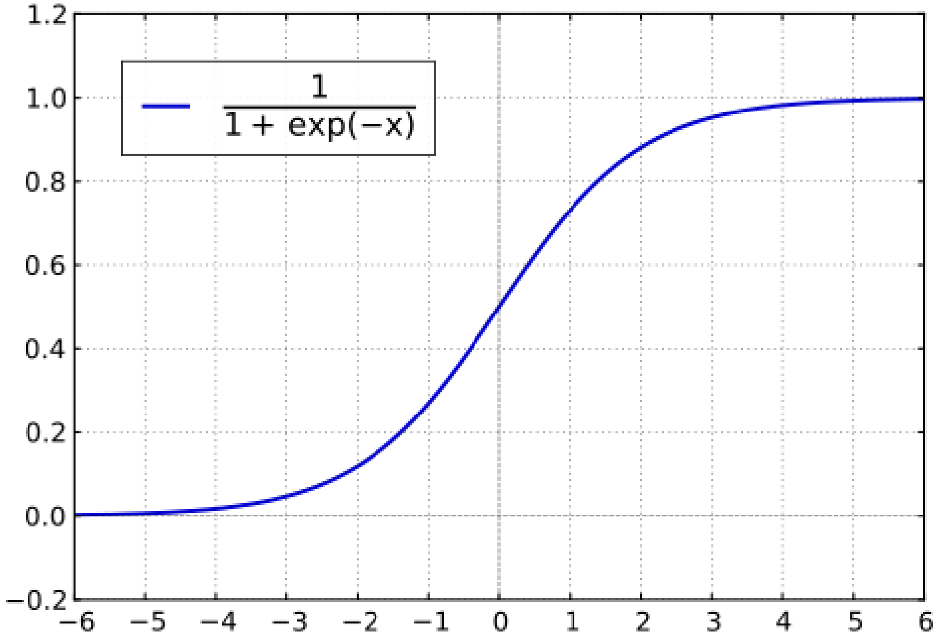
$$S = \sum_j x_j w_{ij} + b_i$$



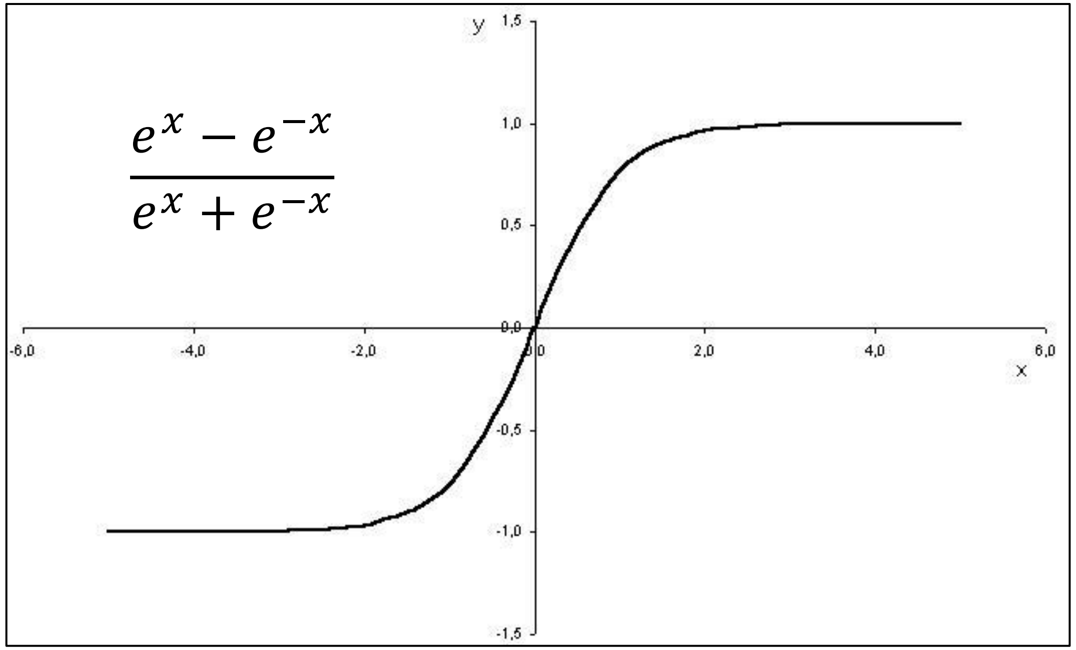
$$S = \sum_j x_j w_{ij} + b_i$$

Types of Activation Functions

Logistic function (sigmoid)

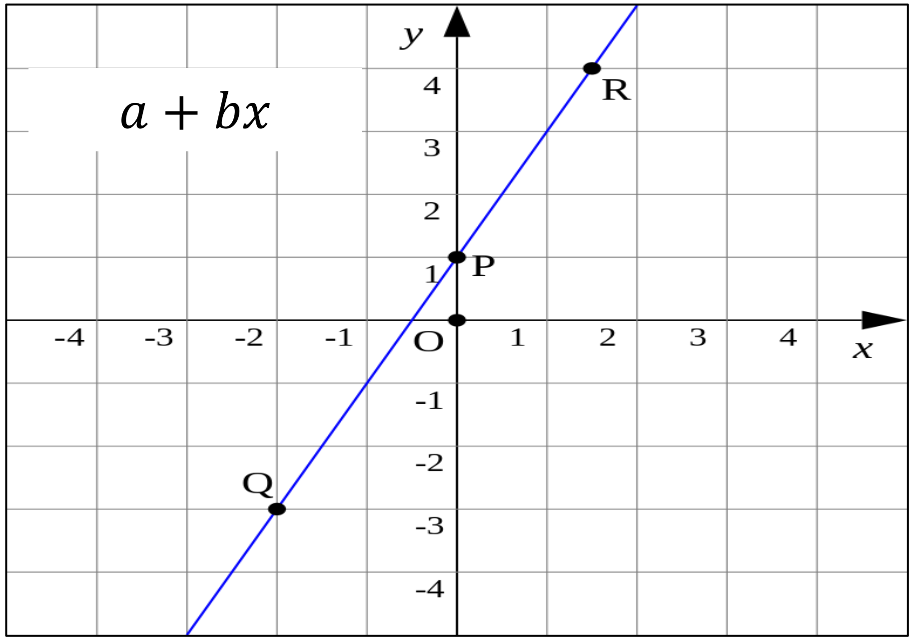


Hyperbolic tangent function (tanh)

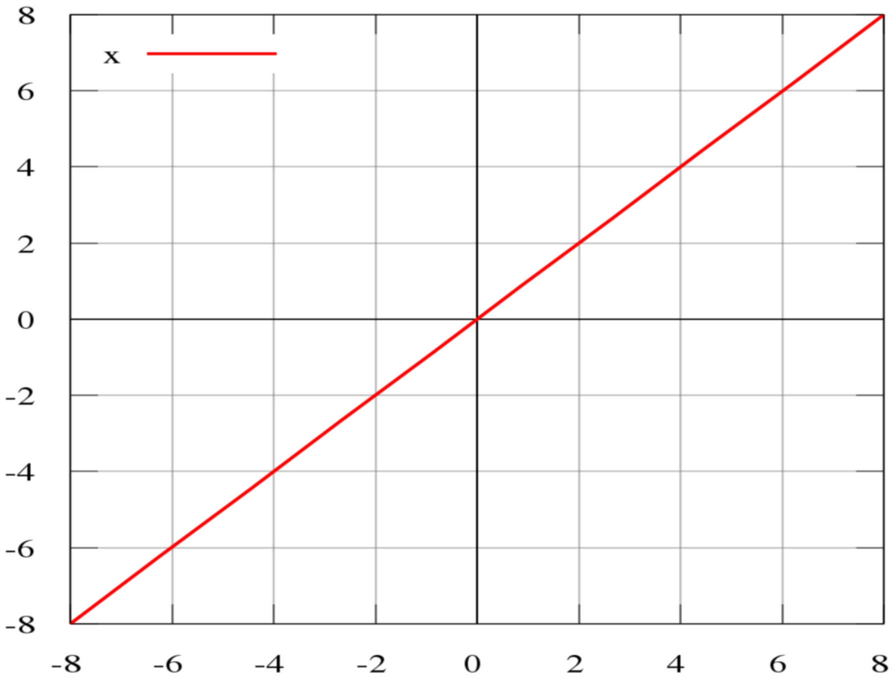


Types of Activation Functions

Linear function



Identity function



Multilayer Perceptron (MLP)

- A feedforward network with three or more layers (one input, one output and one or more hidden layers) is called **multilayer perceptron** (MLP).
- It usually employs a **sigmoid** or a **hyperbolic tangent** function as an activation function.
- A feedforward network that uses a radial basis function as an activation function is called RBF neural network (not discussed here).

Neural Networks Learning Process

The learning process of a neural network is composed of four phases:

1. Initialization
2. Feed forward
3. Error evaluation
4. Backpropagation and weights adjustment

Neural Networks Learning Process (2)

Phase 1—Initialization: At this phase, the first set of weights is generated (usually at random).

Phase 2—Feed forward: At this phase, the signal is transmitted from the input layer to the output layer, passing through the hidden layer(s). The hidden layer neurons perform the operations of summation and activation.

Neural Networks Learning Process (3)

Phase 3—Error evaluation

Here, the output of the network is compared with the actual values of the dependent variable. Usually, the sum squared error is computed, with this formula:

$$SSE = \sum (y - \hat{y})^2$$

At each step of the learning process, the algorithm checks whether the error has improved (i.e., decreased) compared to the previous step. **If the improvement is under a given threshold, the algorithm stops.** If not, it goes to the next phase.

Neural Networks Learning Process (4)

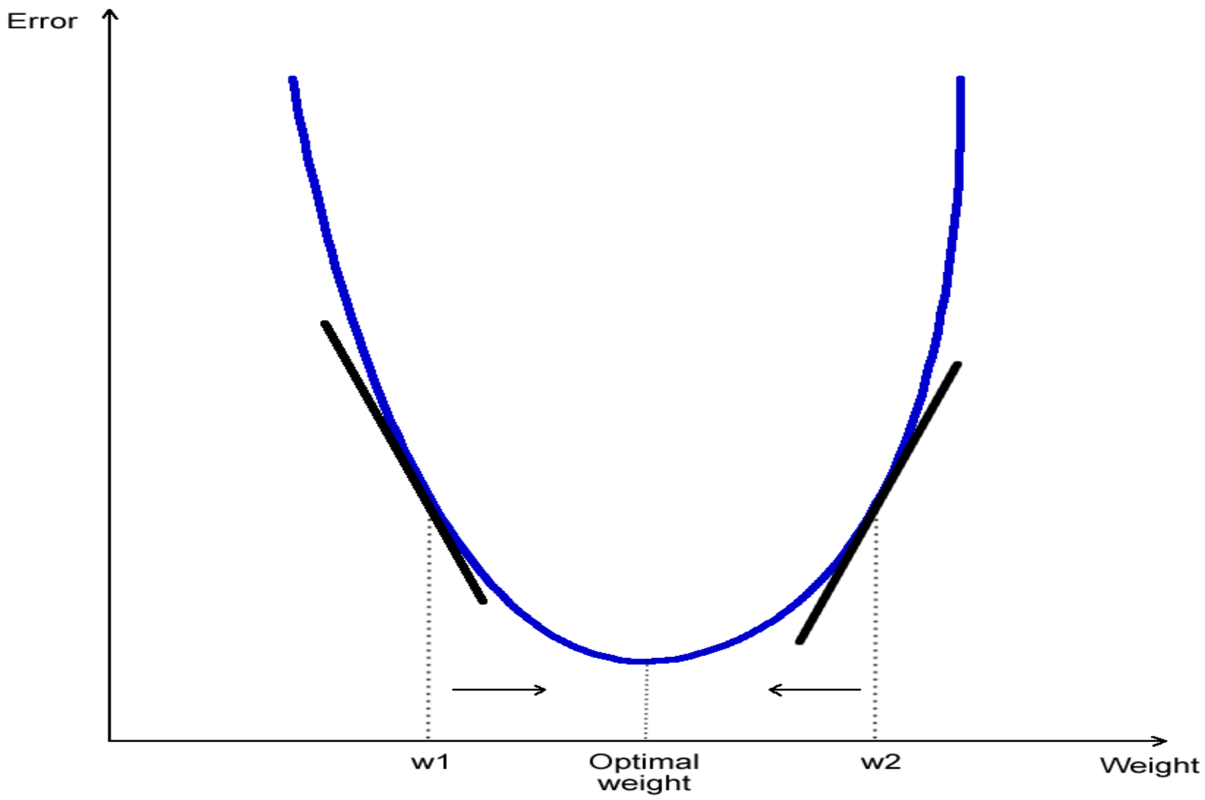
Phase 4—Backpropagation and weights adjustment

The error is propagated backwards through the network and the weights are adjusted with the goal of reducing the error. The learning process starts over. This process continues until the error decrease falls under the specified threshold.

To adjust the weights (and the biases) in the hidden layer neurons, the gradient descent technique is used.

Gradient Descent

- Gradient descent is an optimization algorithm which is commonly-used to train machine learning models and neural networks.
- Training data helps these models learn over time, and the cost function within gradient descent specifically acts as a barometer, gauging its accuracy with each iteration of parameter updates.



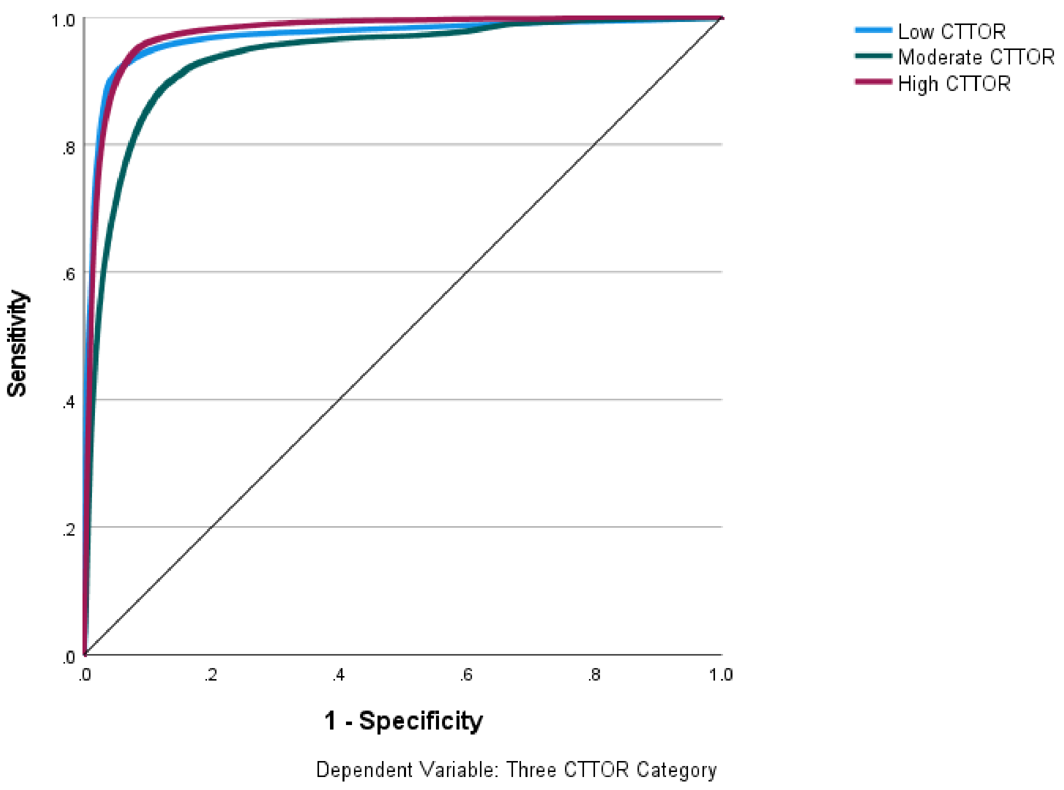
ROC Curve

- The ROC curve summarizes the performance of a neural network.
- ROC stands for **R**eceiver **O**perating **C**haracteristics.
- It is built based on two indicators:
  - *sensitivity* – the ability of the model to predict that an event will happen when it actually happens (in other words, the ability to predict the ***true positives***)
  - *specificity* – the ability of the model to predict that an event will not happen when it actually does not happen (in other words, the ability to predict the ***true negatives***).

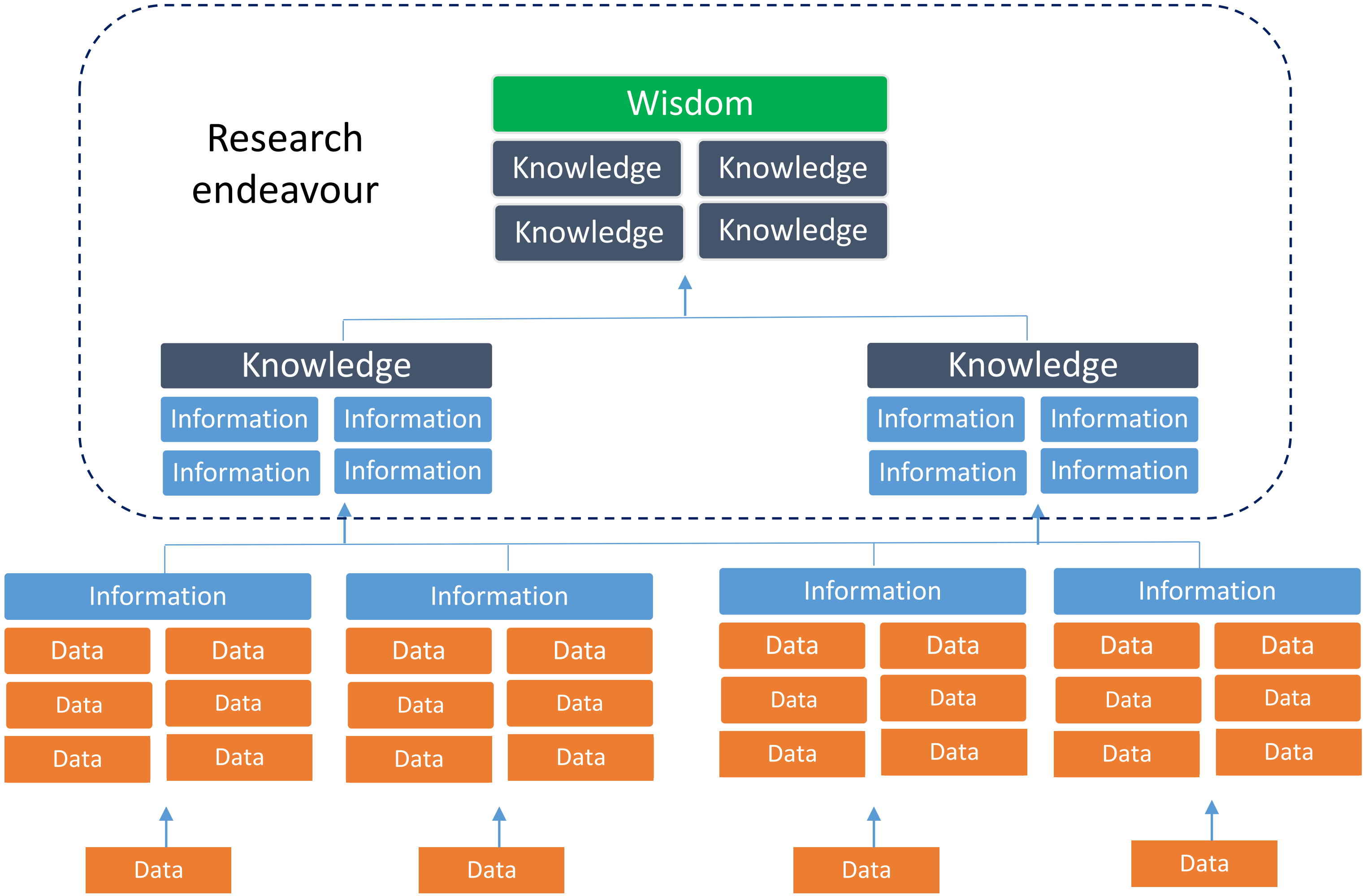
ROC Curve (2)

The accuracy of the model is given by the area under the curve (AUC). This area is comprised between 0.50 and 1.

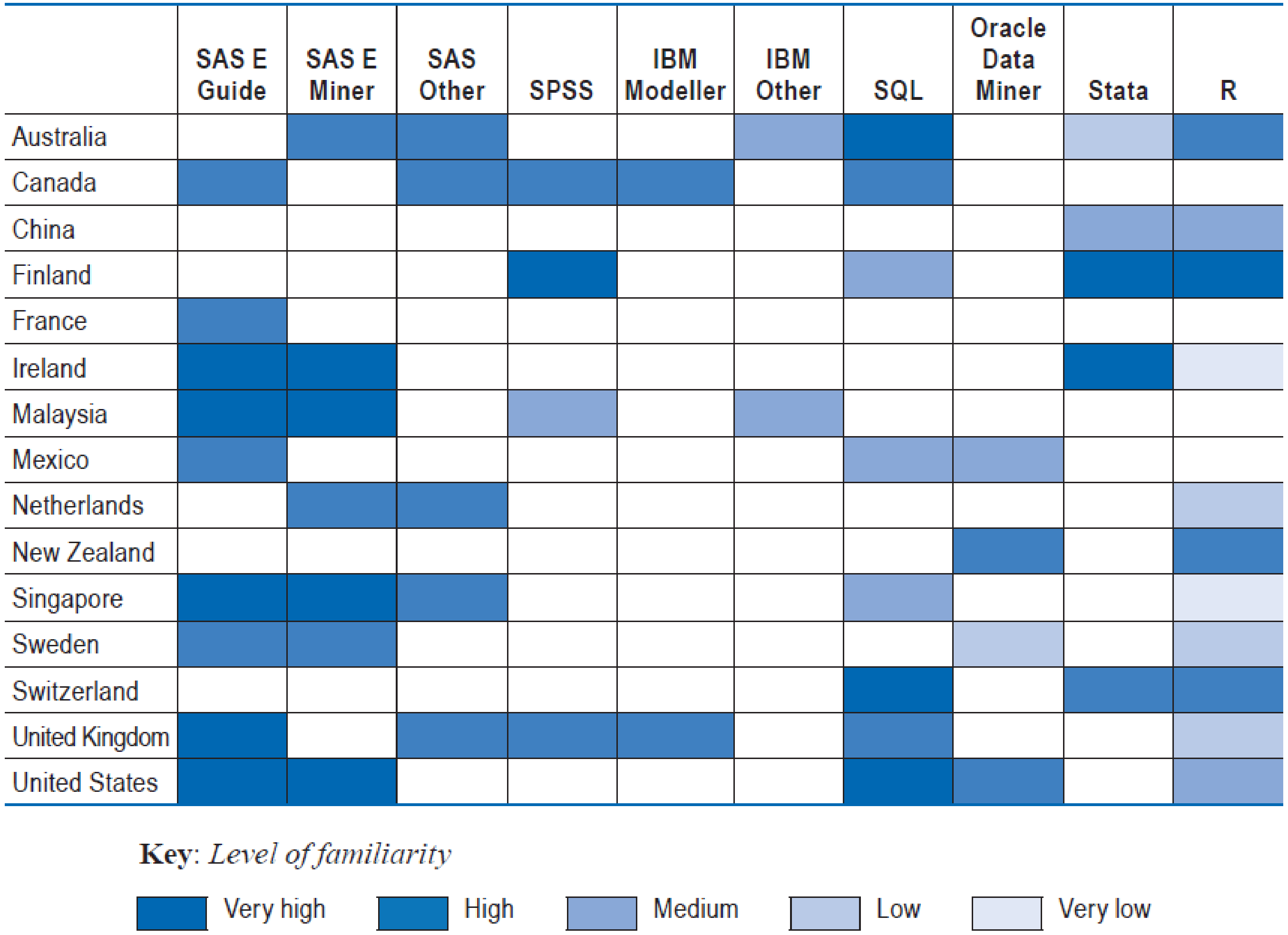
- if AUC is close to 0.50, the model is useless for prediction
- the closer AUC is to 1, the better the model



DIKW Model—a research perspective



Usage of and Familiarity with Analytics Software



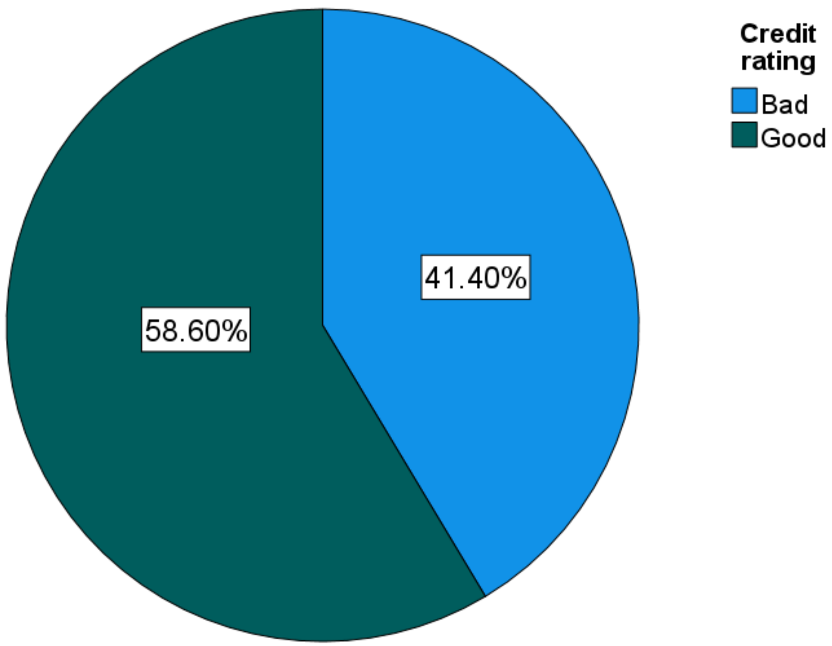
Source: Forum on Tax Administrations, Advanced Analytics survey (2015)

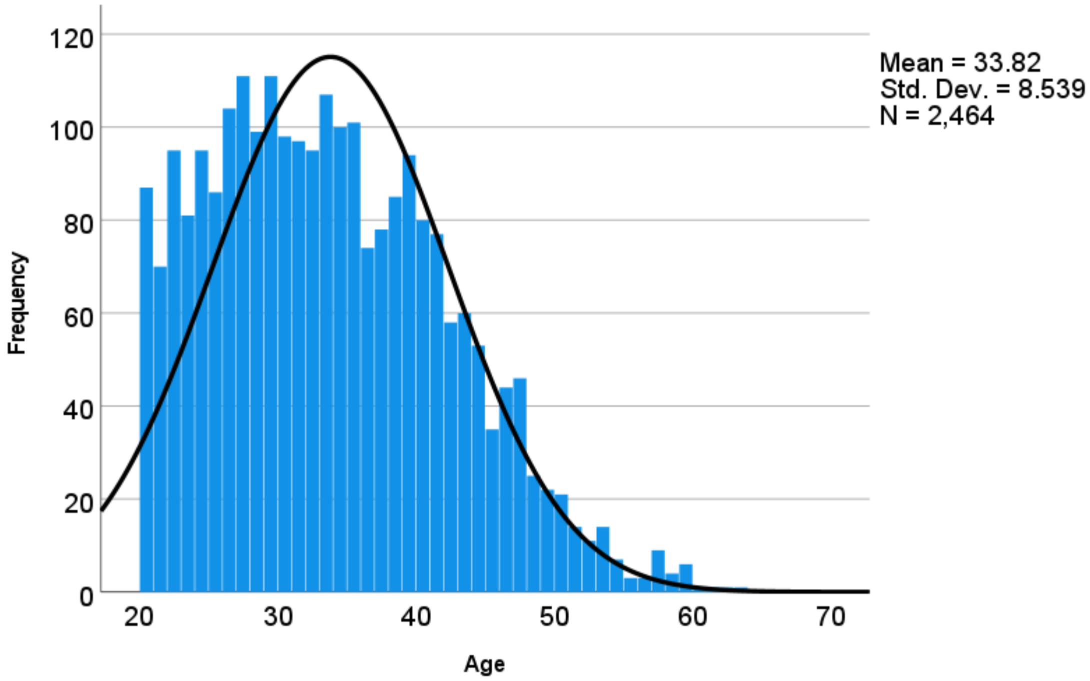
So, Let's Build Our ANN Model (based on historical data)

The data set contains information about the credit ratings of 2.464 subjects. We will use the multilayer perceptron **to predict** the credit rating category using five variables in the data set as predictors.

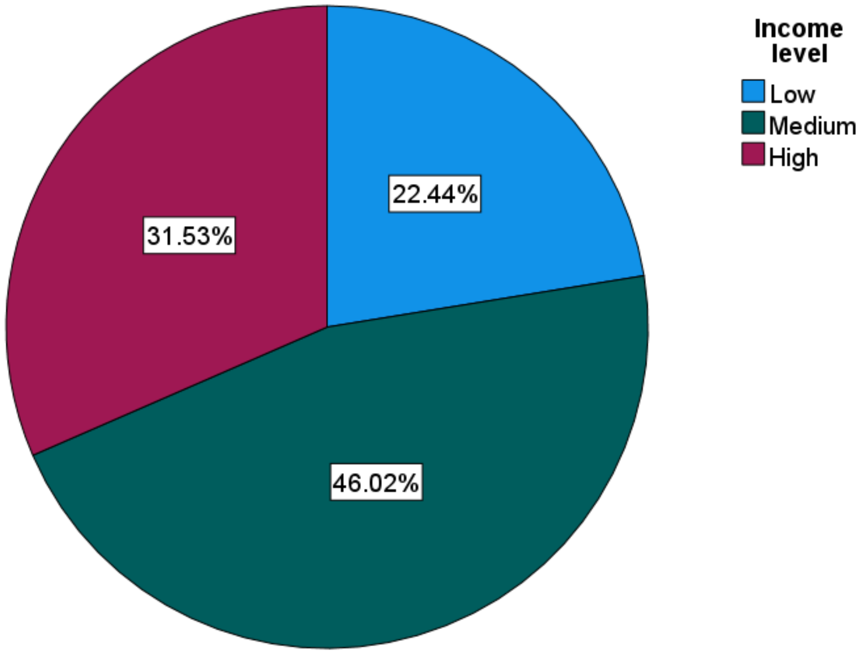
Variables	Measurement scale	Labels
Credit rating	Categorical	0=bad, 1=good, 9=no data
Age	Ratio	-
Income level	Ordinal	1=low, 2=medium, 3=high
Number of credit cards	Ordinal	1=<5, 2=5 or more
Education	Ordinal	1=high school, 2=college
Car loans	Ordinal	1= none or 1, 2=>2

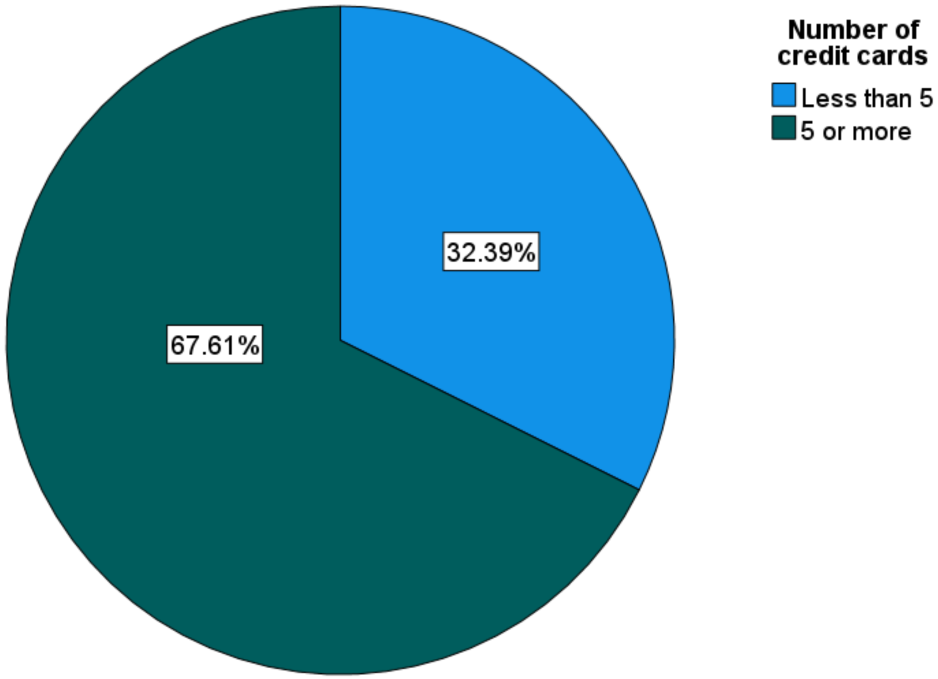
Data Description



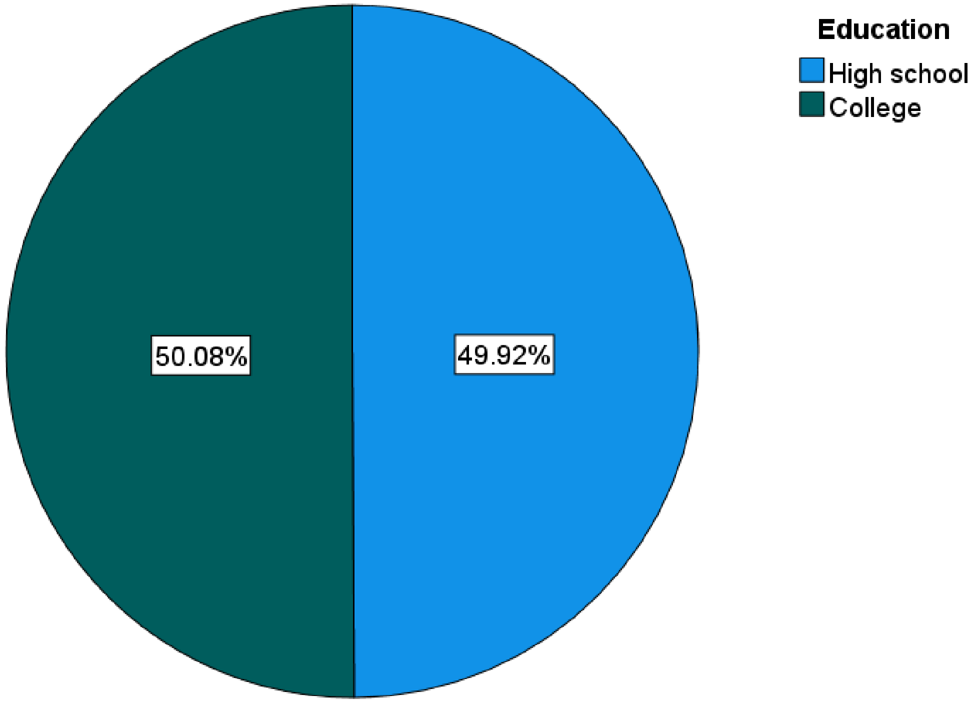


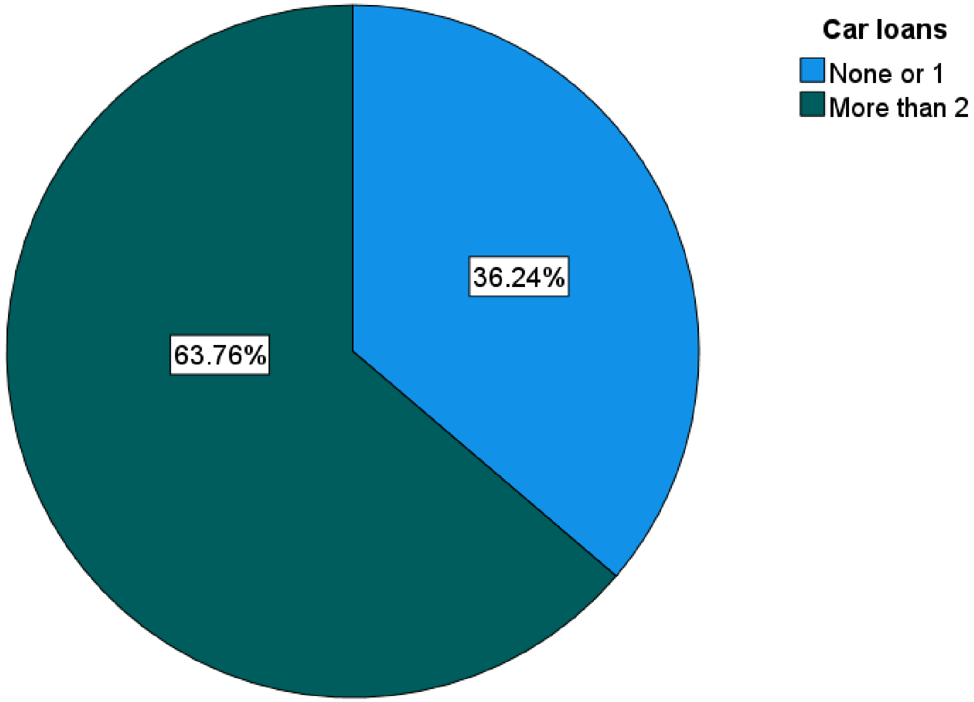
Data Description (2)





Data Description (3)





File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

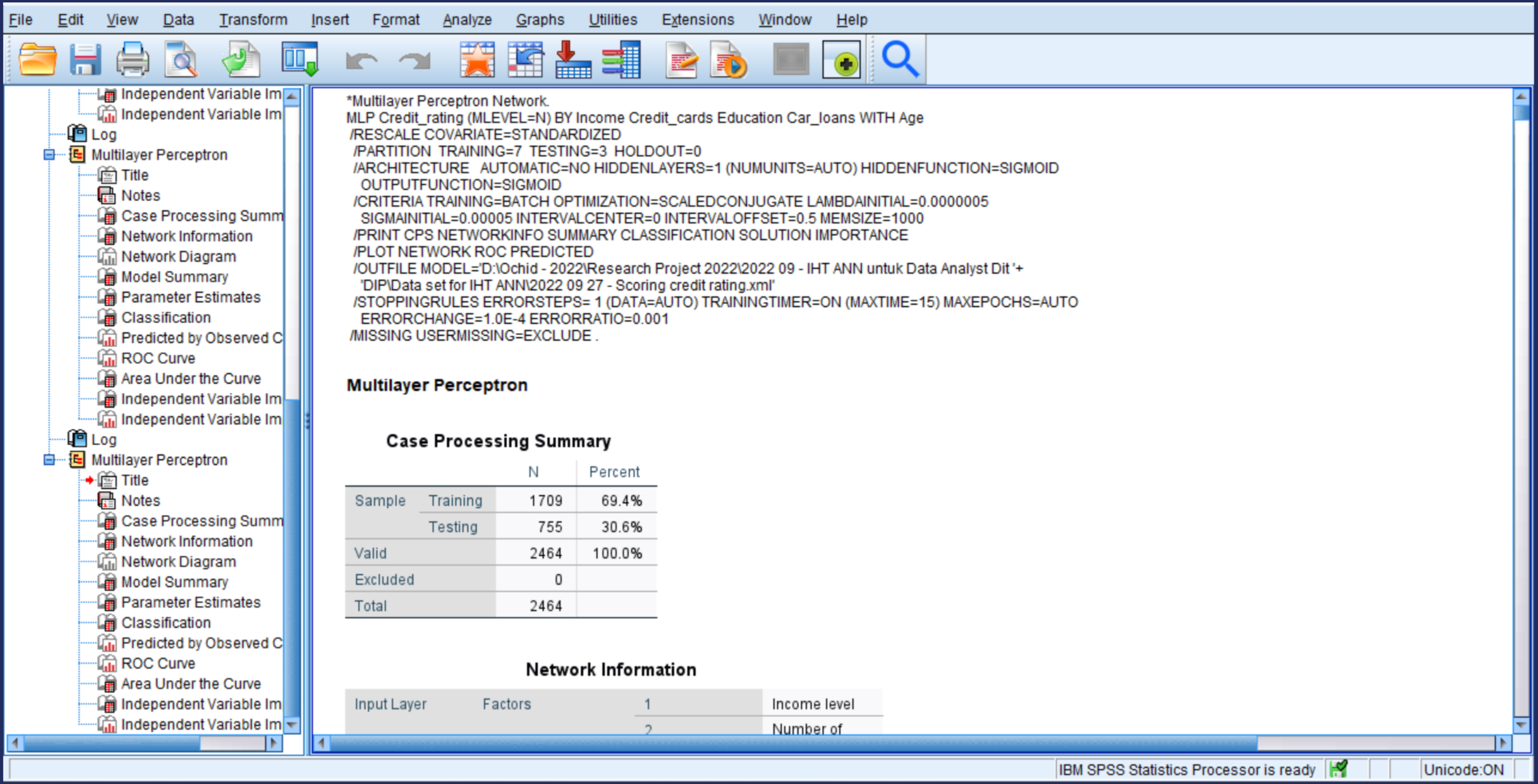
1 :

Visible: 6 of 6 Variables

	Credit_rating	Age	Income	Credit_cards	Education	Car_loans	var	var	var
1	Good	57	High	Less than 5	High school	None or 1			
2	Good	51	Medium	5 or more	High school	More than 2			
3	Good	27	High	Less than 5	High school	None or 1			
4	Good	28	Low	Less than 5	High school	None or 1			
5	Good	30	High	Less than 5	High school	More than 2			
6	Good	28	High	5 or more	High school	More than 2			
7	Good	42	High	5 or more	High school	None or 1			
8	Good	35	Medium	5 or more	High school	None or 1			
9	Good	24	Medium	Less than 5	High school	None or 1			
10	Good	36	High	Less than 5	High school	None or 1			
11	Good	30	Medium	Less than 5	College	None or 1			
12	Good	41	Medium	Less than 5	College	None or 1			
13	Good	42	Medium	5 or more	College	More than 2			
14	Good	44	Medium	5 or more	College	More than 2			
15	Good	42	High	5 or more	College	More than 2			
16	Good	39	High	5 or more	College	None or 1			
17	Good	54	High	5 or more	College	More than 2			
18	Good	37	Medium	Less than 5	High school	None or 1			
19	Good	40	Low	5 or more	High school	More than 2			
20	Good	33	Medium	Less than 5	High school	None or 1			
21	Good	31	Medium	5 or more	High school	More than 2			

Data View | Variable View

IBM SPSS Statistics Processor is ready | Unicode:ON


File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help
Independent Variable Im
Independent Variable Im
Log
Multilayer Perceptron
Title
Notes
Case Processing Summ
Network Information
Network Diagram
Model Summary
Parameter Estimates
Classification
Predicted by Observed C
ROC Curve
Area Under the Curve
Independent Variable Im
Independent Variable Im
Log
Multilayer Perceptron
Title
Notes
Case Processing Summ
Network Information
Network Diagram
Model Summary
Parameter Estimates
Classification
Predicted by Observed C
ROC Curve
Area Under the Curve
Independent Variable Im
Independent Variable Im
*Multilayer Perceptron Network.
MLP Credit_rating (MLEVEL=N) BY Income Credit_cards Education Car_loans WITH Age
/RESCALE COVARIATE=STANDARDIZED
/PARTITION TRAINING=7 TESTING=3 HOLDOUT=0
/ARCHITECTURE AUTOMATIC=NO HIDDENLAYERS=1 (NUMUNITS=AUTO) HIDDENFUNCTION=SIGMOID
OUTPUTFUNCTION=SIGMOID
/CRITERIA TRAINING=BATCH OPTIMIZATION=SCALEDCONJUGATE LAMBDAINITIAL=0.0000005
SIGMAINITIAL=0.00005 INTERVALCENTER=0 INTERVALOFFSET=0.5 MEMSIZE=1000
/PRINT CPS NETWORKINFO SUMMARY CLASSIFICATION SOLUTION IMPORTANCE
/PLOT NETWORK ROC PREDICTED
/OUTFILE MODEL='D:\Ochid - 2022\Research Project 2022\2022 09 - IHT ANN untuk Data Analyst Dit '+
'DIP\Data set for IHT ANN\2022 09 27 - Scoring credit rating.xml'
/STOPPINGRULES ERRORSTEPS= 1 (DATA=AUTO) TRAININGTIMER=ON (MAXTIME=15) MAXEPOCHS=AUTO
ERRORCHANGE=1.0E-4 ERRORRATIO=0.001
/MISSING USERMISSING=EXCLUDE .
Multilayer Perceptron
Case Processing Summary
N Percent
Sample Training 1709 69.4%
Testing 755 30.6%
Valid 2464 100.0%
Excluded 0
Total 2464
Network Information
Input Layer Factors 1 Income level
2 Number of
IBM SPSS Statistics Processor is ready
Unicode:ON

Let’s go to a live demo!

How do we predict the credit rating of future client?

The data set contains information about 100 subjects **without credit rating categories**. We will use the multilayer perceptron **to predict** the credit rating category using five variables in the data set as predictors.

Variables	Measurement scale	Labels
Credit rating (?????)	Categorical	0=bad, 1=good, 9=no data
Age	Ratio	-
Income level	Ordinal	1=low, 2=medium, 3=high
Number of credit cards	Ordinal	1=<5, 2=5 or more
Education	Ordinal	1=high school, 2=college
Car loans	Ordinal	1= none or 1, 2=>2

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

1 : Visible: 5 of 5 Variables

	Age	Income	Credit_cards	Education	Car_loans	var	var
1	37	Medium	Less than 5	College	None or 1		
2	35	High	5 or more	High school	More than 2		
3	39	Medium	5 or more	College	More than 2		
4	29	Medium	5 or more	College	More than 2		
5	22	Medium	5 or more	College	More than 2		
6	25	Low	5 or more	College	More than 2		
7	36	Medium	5 or more	High school	More than 2		
8	34	Low	5 or more	College	More than 2		
9	20	Medium	5 or more	High school	More than 2		
10	25	Low	Less than 5	High school	None or 1		
11	23	Medium	Less than 5	High school	None or 1		
12	23	Low	5 or more	High school	More than 2		
13	32	Low	5 or more	High school	More than 2		
14	41	Medium	5 or more	High school	More than 2		
15	26	Medium	5 or more	College	More than 2		
16	20	Low	5 or more	College	More than 2		
17	21	Low	5 or more	High school	More than 2		
18	22	Medium	5 or more	High school	More than 2		
19	23	Medium	5 or more	College	More than 2		
20	26	Medium	5 or more	College	More than 2		
21	26	Medium	Less than 5	College	None or 1		

Data View | Variable View

IBM SPSS Statistics Processor is ready | Unicode:ON

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

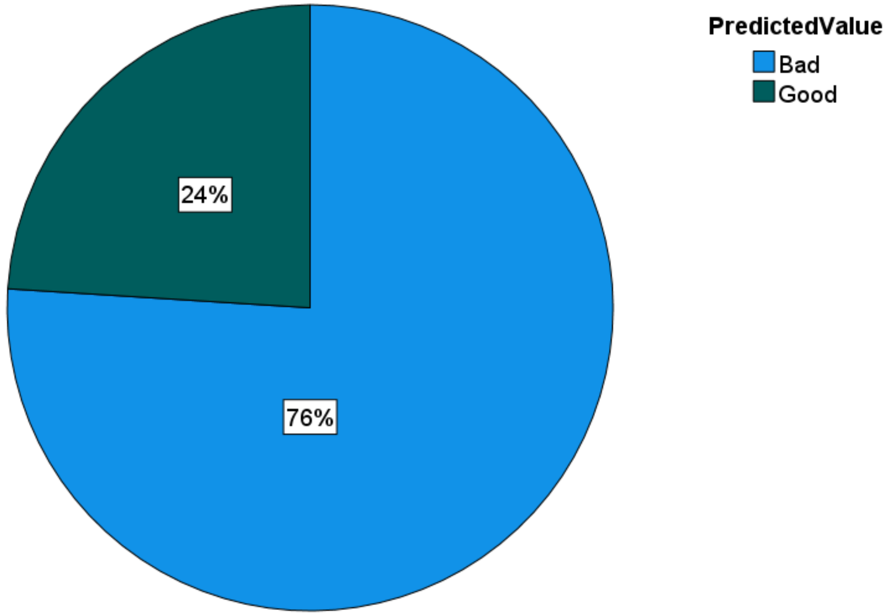
Visible: 7 of 7 Variables

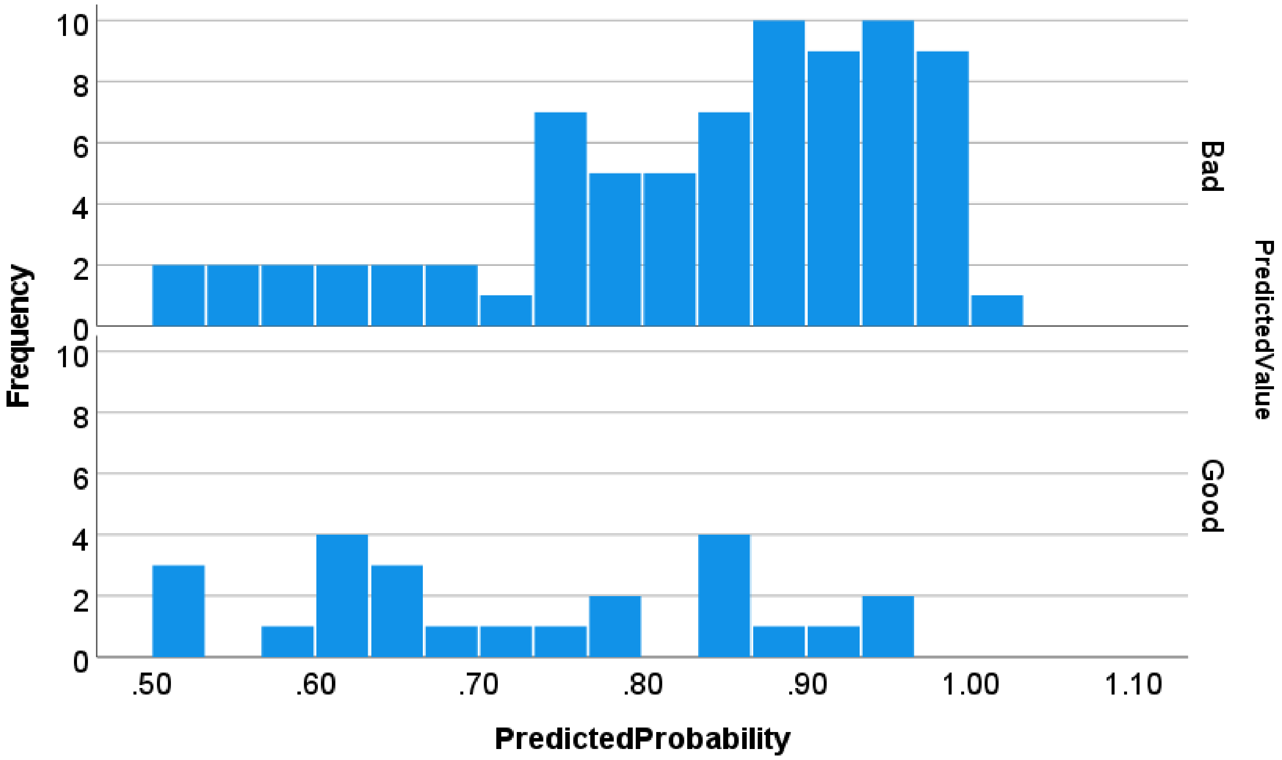
	Age	Income	Credit_cards	Education	Car_loans	PredictedValue	PredictedProbability	var	var
1	37	Medium	Less than 5	College	None or 1	Good	.94		
2	35	High	5 or more	High school	More than 2	Good	.84		
3	39	Medium	5 or more	College	More than 2	Good	.61		
4	29	Medium	5 or more	College	More than 2	Bad	.69		
5	22	Medium	5 or more	College	More than 2	Bad	.86		
6	25	Low	5 or more	College	More than 2	Bad	.97		
7	36	Medium	5 or more	High school	More than 2	Good	.53		
8	34	Low	5 or more	College	More than 2	Bad	.89		
9	20	Medium	5 or more	High school	More than 2	Bad	.90		
10	25	Low	Less than 5	High school	None or 1	Bad	.54		
11	23	Medium	Less than 5	High school	None or 1	Good	.79		
12	23	Low	5 or more	High school	More than 2	Bad	1.00		
13	32	Low	5 or more	High school	More than 2	Bad	.94		
14	41	Medium	5 or more	High school	More than 2	Good	.66		
15	26	Medium	5 or more	College	More than 2	Bad	.78		
16	20	Low	5 or more	College	More than 2	Bad	.98		
17	21	Low	5 or more	High school	More than 2	Bad	1.00		
18	22	Medium	5 or more	High school	More than 2	Bad	.87		
19	23	Medium	5 or more	College	More than 2	Bad	.83		
20	26	Medium	5 or more	College	More than 2	Bad	.77		
21	26	Medium	Less than 5	College	None or 1	Good	.85		

Data View Variable View

IBM SPSS Statistics Processor is ready Unicode:ON

Results: Predicted Value and Predicted Probability



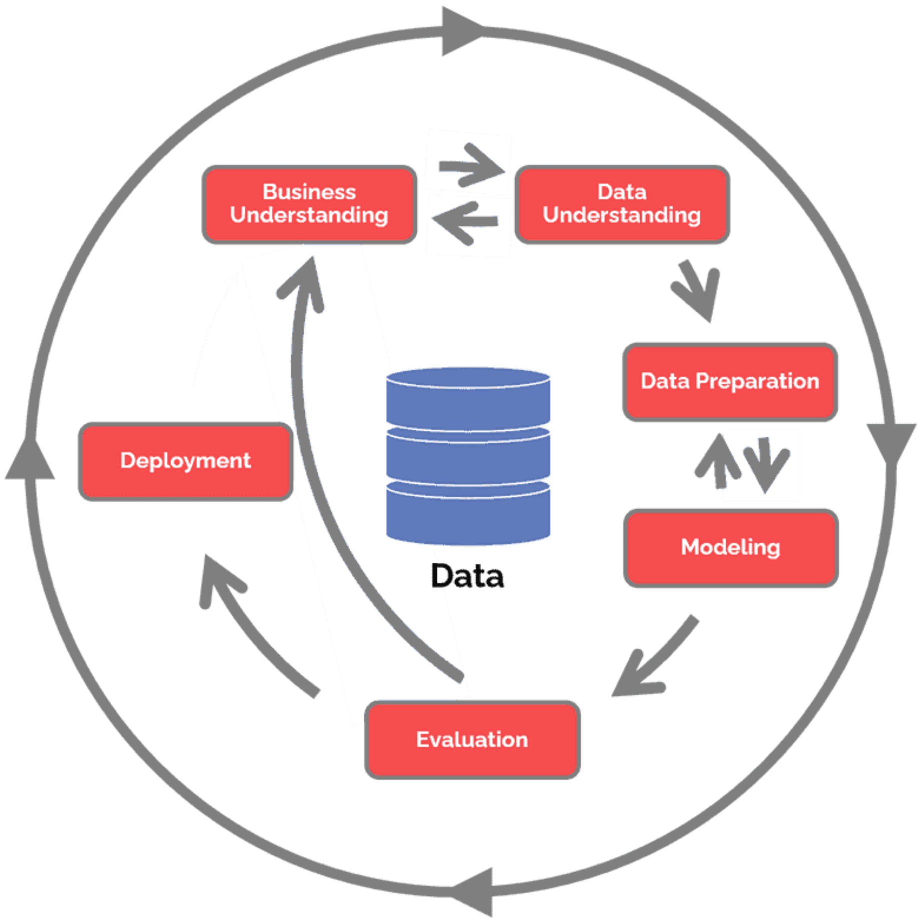


Let’s go to a live demo!

Bringing ANN into Practice: A CRISP-DM Perspective

The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata Mining (CRISP-DM) is a process model that serves as the base for a **data science** process.

1. Business understanding – **What does the business need?**
2. Data understanding – **What data do we have / need? Is it clean?**
3. Data preparation – **How do we organize the data for modeling?**
4. Modeling – **What modeling techniques should we apply?**
5. Evaluation – **Which model best meets the business objectives?**
6. Deployment – **How do stakeholders access the results?**

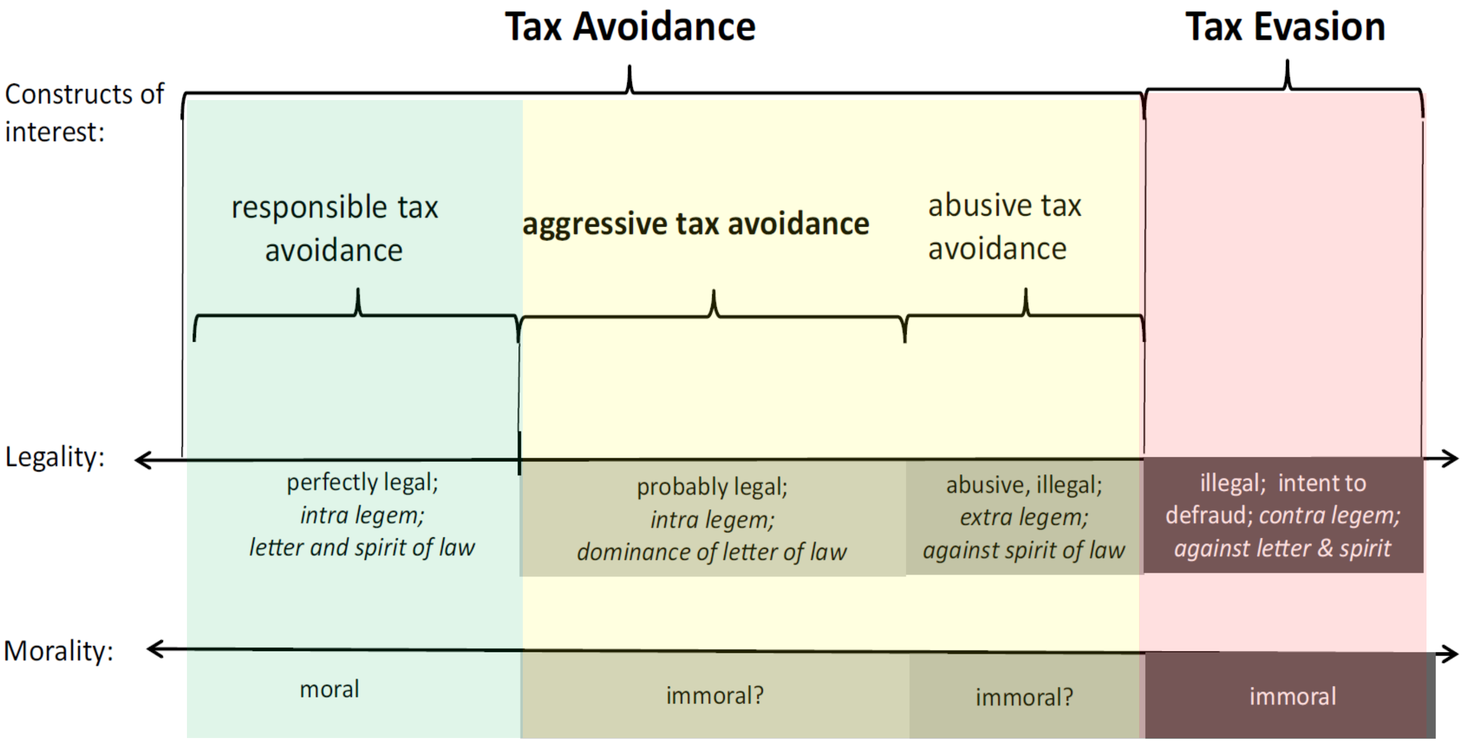


Predicting Firms' Taxpaying Behaviour Using Artificial Neural Networks: The Case Of Indonesia

The full working paper is available here:

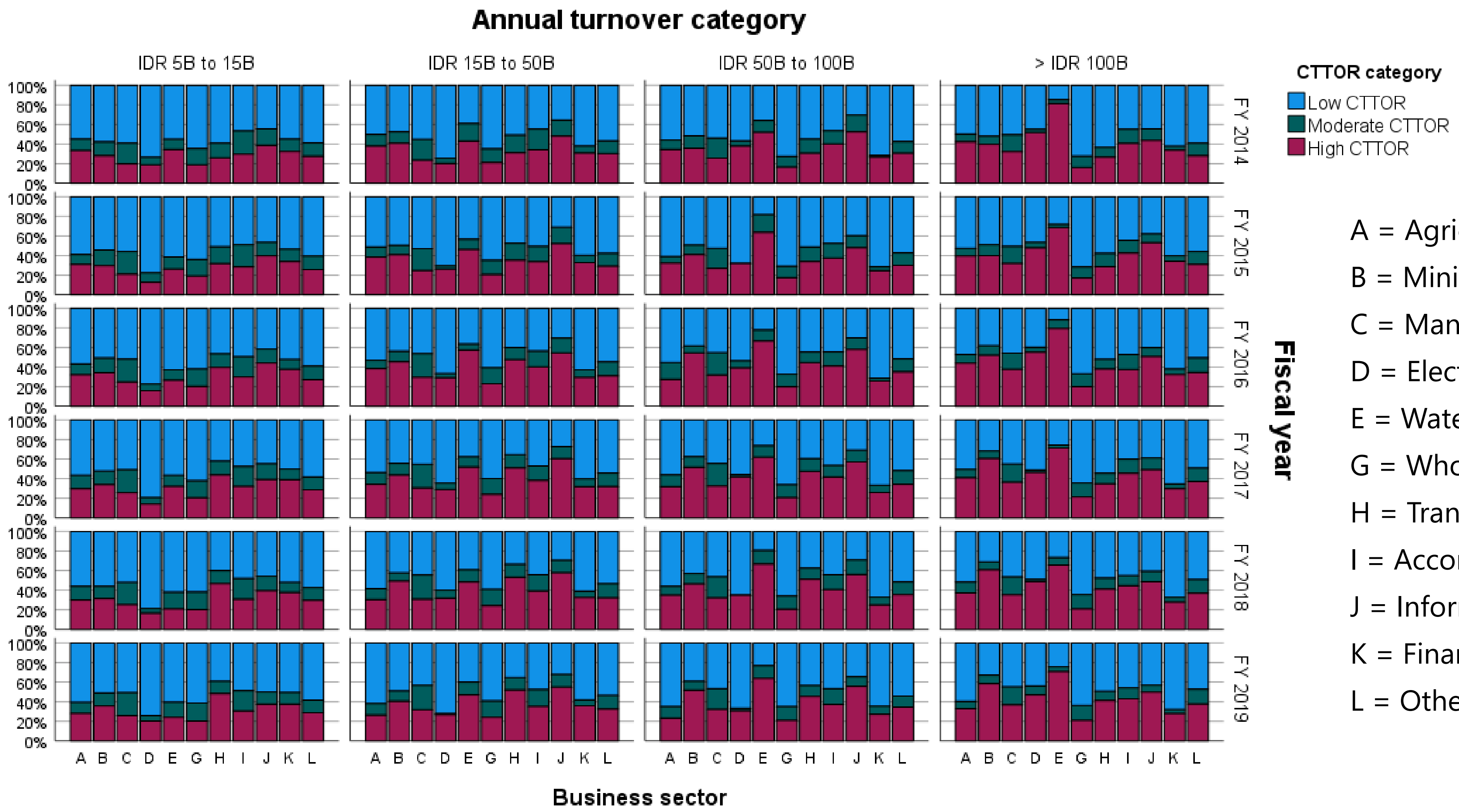
https://ssrn.com/abstract=4185966 or
http://dx.doi.org/10.2139/ssrn.4185966

Tax Compliance Behaviour: A Closer Look



- ❑ Can we accurately **predict** taxpaying behaviour?
- ❑ What are the most influential **predictors**?

Empirical Setting



A = Agriculture
B = Mining and quarrying
C = Manufacturing
D = Electricity and gas
E = Water supply, sewerage, etc.
G = Wholesale and retail trade
H = Transportation and storage
I = Accommodation and food service
J = Information and communication
K = Financial and insurance
L = Others

Empirical Data

Descriptive statistics (n = 538,254)

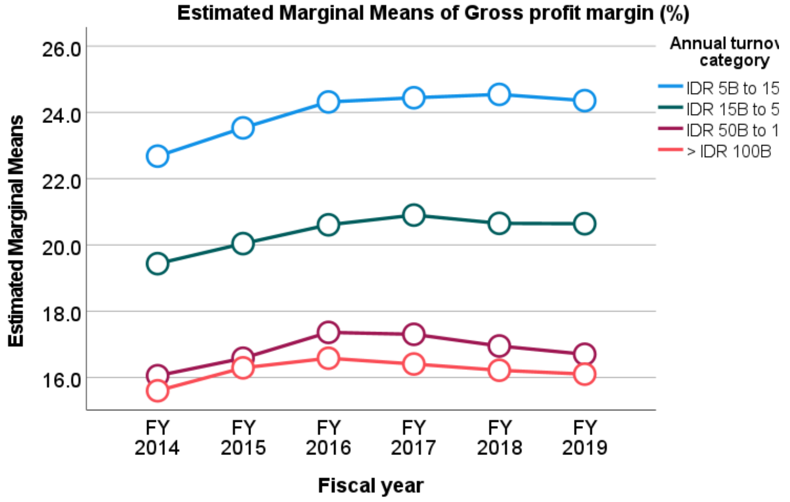
	Mean	Median	SD	Variance	Min.	Max.
Gross profit margin (%)	20.83	15.17	20.04	401.50	-98.60	100.00
Operating profit margin (%)	4.49	3.02	11.57	133.86	-175.64	100.00
Other income ratio (%)	1.20	0.04	8.62	74.38	-92.59	4,308.28
Other expense ratio (%)	1.59	0.01	9.53	90.87	-70.42	4,332.31
Positive fiscal adjustment ratio (%)	7.24	0.22	21.78	474.50	-104.07	100.00
Negative fiscal adjustment ratio (%)	0.60	0.00	4.51	20.33	-189.45	100.00
CTTOR (%)	0.89	0.42	1.35	1.82	0.00	23.37

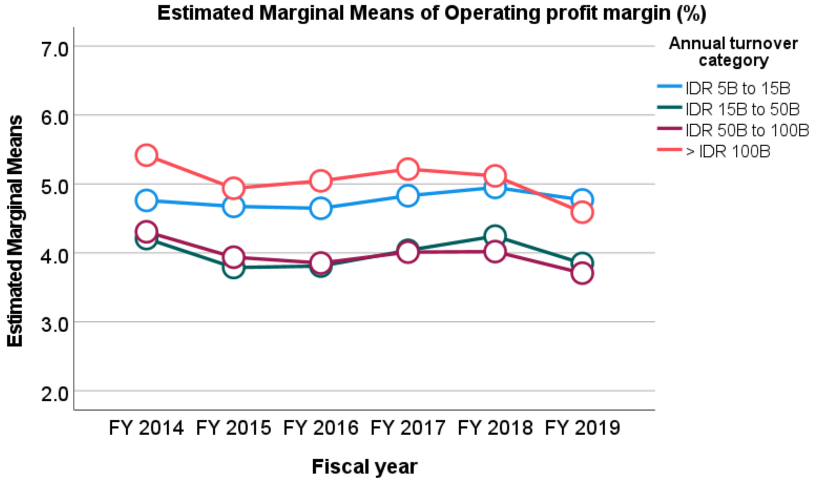
Relationship between variables under study (n = 538,254)

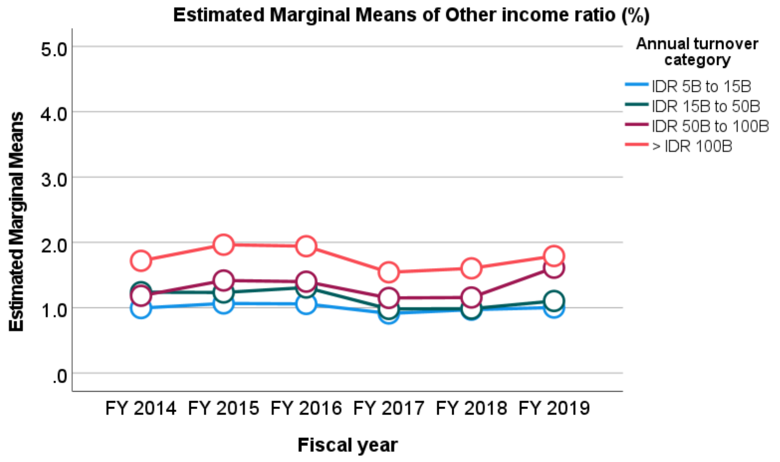
	GPM	OPM	OIR	OER	PFAR	NFAR	CTTOR
GPM	1	.476**	.023**	.080**	.065**	.065**	.349**
OPM	.476**	1	-.095**	.018**	.072**	.045**	.410**
OIR	.023**	-.095**	1	.698**	.020**	.108**	.074**
OER	.080**	.018**	.698**	1	.038**	.073**	.004**
PFAR	.065**	.072**	.020**	.038**	1	.098**	-.120**
NFAR	.065**	.045**	.108**	.073**	.098**	1	.011**
CTTOR	.349**	.410**	.074**	.004**	-.120**	.011**	1

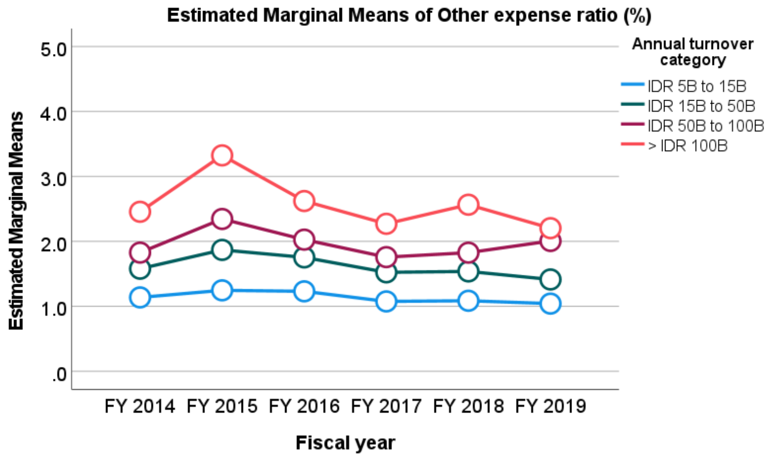
**. Correlation is significant at the 0.01 level (2-tailed).

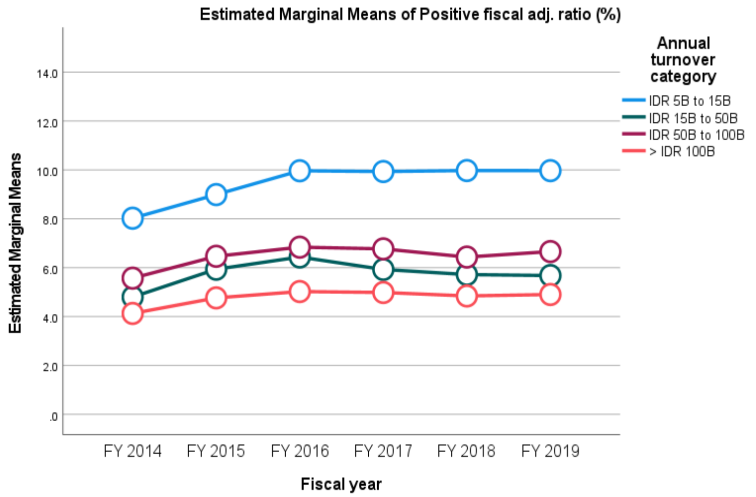
Graphical Evidence

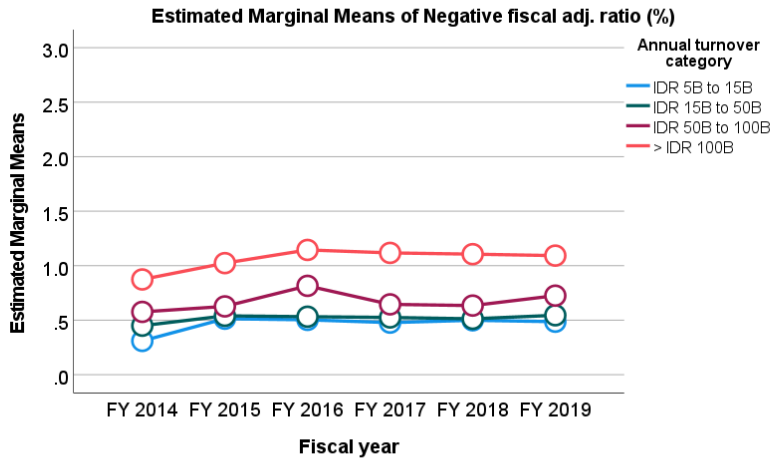










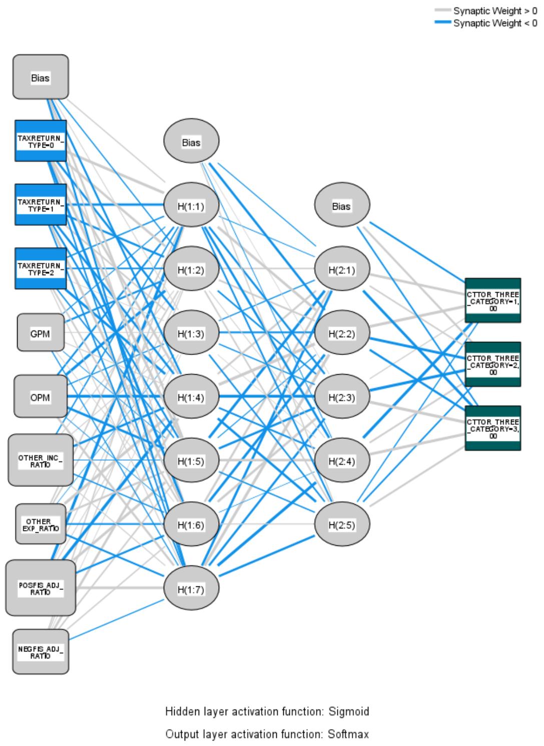


ANN Datasets

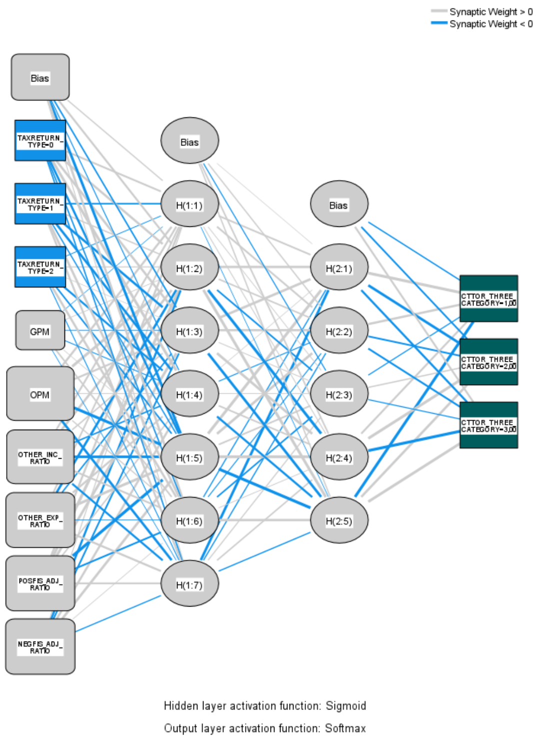
	Small firms (IDR 5 billion to 15 billion)		Medium firms (IDR 15 billion to 50 billion)		Medium-large firms (IDR 50 billion to 100 billion)		Large firms (More than IDR 100 billion)	
Sample	N	Percent	N	Percent	N	Percent	N	Percent
Training	135,140	60.3%	99,779	60.0%	36,816	59.8%	51,659	59.8%
Testing	44,634	19.9%	33,167	20.0%	12,236	19.9%	17,260	20.0%
Holdout	44,393	19.8%	33,215	20.0%	12,517	20.3%	17,409	20.2%
Valid	224,167	100.0%	166,161	100.0%	61,569	100.0%	86,328	100.0%
Excluded	14		10		3		2	
Total	224,181		166,171		61,572		86,330	

Structure of Neural Networks for the Model Prediction

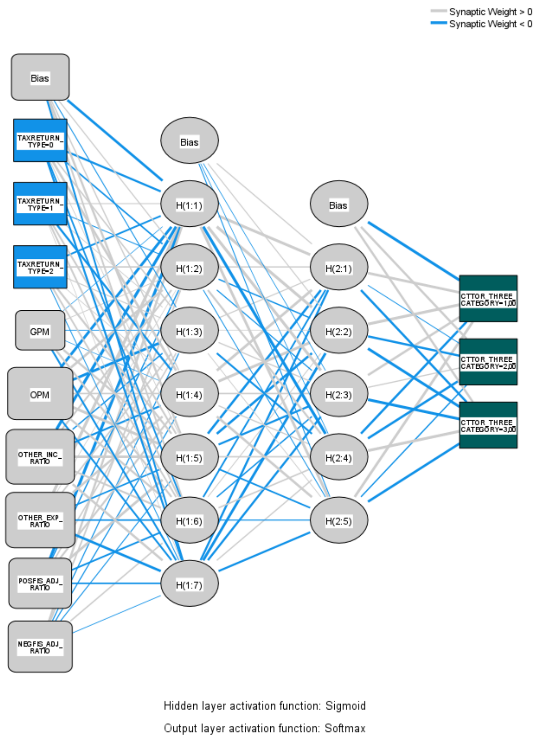
Small firms
(n = 224,181)



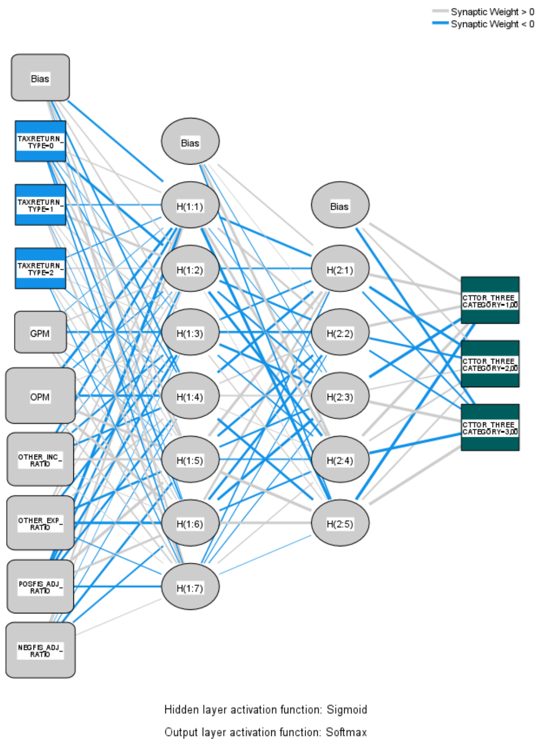
Medium firms
(n = 166,171)



Medium-large firms
(n = 61,572)



Large firms
(n = 86,330)



ANN Model Summary

		Small firms	Medium firms	Medium-large firms	Large firms
Training	Cross Entropy Error	42,037.16	22,814.93	9,211.94	10,952.95
	Percent Incorrect Predictions	10.9%	6.4%	7.7%	6.8%
	Stopping Rule Used	Max. number of epochs (100) exceeded	Max. number of epochs (100) exceeded	Max. number of epochs (100) exceeded	Max. number of epochs (100) exceeded
	Training Time	0:00:10,35	0:00:07,97	0:00:03,24	0:00:04,56
Testing	Cross Entropy Error	13,972.8	7,444.7	3,078.6	3,750.5
	Percent Incorrect Predictions	10.8%	6.4%	7.6%	7.0%
Holdout	Percent Incorrect Predictions	10.9%	6.4%	6.9%	7.1%

Accuracy of Classification

		Predicted for small firms (IDR 5 billion to 15 billion)				Predicted for medium firms (IDR 15 billion to 50 billion)			
Sample	Observed	Low	Moderate	High	% Correct	Low	Moderate	High	% Correct
Training	Low	75,006	3,103	1,847	93.8%	53,201	1,211	1,398	95.3%
	Moderate	3,472	16,862	2,876	72.6%	1,300	13,690	1,319	83.9%
	High	947	2,461	28,566	89.3%	474	703	26,483	95.7%
	Percent	58.8%	16.6%	24.6%	89.1%	55.1%	15.6%	29.3%	93.6%
Testing	Low	24,762	1,024	604	93.8%	17,716	384	431	95.6%
	Moderate	1,139	5,586	954	72.7%	418	4,520	467	83.6%
	High	333	760	9,472	89.7%	183	230	8,818	95.5%
	Percent	58.8%	16.5%	24.7%	89.2%	55.2%	15.5%	29.3%	93.6%
Holdout	Low	24,675	1,027	599	93.8%	17,804	396	430	95.6%
	Moderate	1,174	5,546	961	72.2%	455	4,602	433	83.8%
	High	270	796	9,345	89.8%	156	258	8,681	95.4%
	Percent	58.8%	16.6%	24.6%	89.1%	55.4%	15.8%	28.7%	93.6%

Accuracy of Classification (2)

		Predicted for medium-large firms (IDR 50 billion to 100 billion)				Predicted for large firms (More than IDR 100 billion)			
Sample	Observed	Low	Moderate	High	% Correct	Low	Moderate	High	% Correct
Training	Low	21,191	335	472	96.3%	27,356	763	1,030	93.8%
	Moderate	947	3,819	499	72.5%	634	6,064	616	82.9%
	High	315	252	8,986	94.1%	275	182	14,739	97.0%
	Percent	61.0%	12.0%	27.0%	92.3%	54.7%	13.6%	31.7%	93.2%
Testing	Low	7,121	114	169	96.2%	9,162	299	316	93.7%
	Moderate	308	1,270	159	73.1%	206	2,000	232	82.0%
	High	104	70	2,921	94.4%	84	63	4,898	97.1%
	Percent	61.6%	11.9%	26.6%	92.4%	54.8%	13.7%	31.6%	93.0%
Holdout	Low	7,182	106	136	96.7%	9,177	296	352	93.4%
	Moderate	299	1,331	139	75.2%	240	1,995	206	81.7%
	High	104	81	3,139	94.4%	75	72	4,996	97.1%
	Percent	60.6%	12.1%	27.3%	93.1%	54.5%	13.6%	31.9%	92.9%

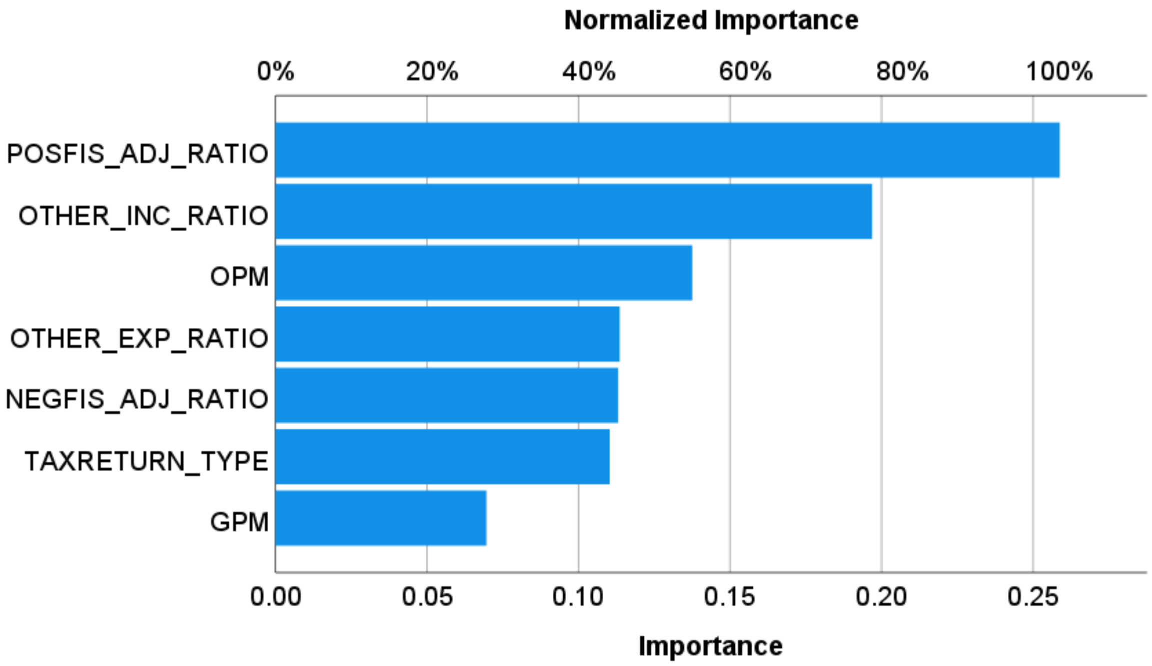
Independent Variable Importance

	Small firms (IDR 5 billion to 15 billion)		Medium firms (IDR 15 billion to 50 billion)	
	Importance	Normalized Importance	Importance	Normalized Importance
Annual tax return type	0.110	42.6%	0.098	54.5%
Gross profit margin (%)	0.070	26.9%	0.031	17.1%
Operating profit margin (%)	0.138	53.2%	0.172	95.8%
Other income ratio (%)	0.197	76.1%	0.178	99.3%
Other expense ratio (%)	0.114	43.9%	0.165	92.2%
Positive fiscal adj. ratio (%)	0.259	100.0%	0.178	99.4%
Negative fiscal adj. ratio (%)	0.113	43.7%	0.179	100.0%

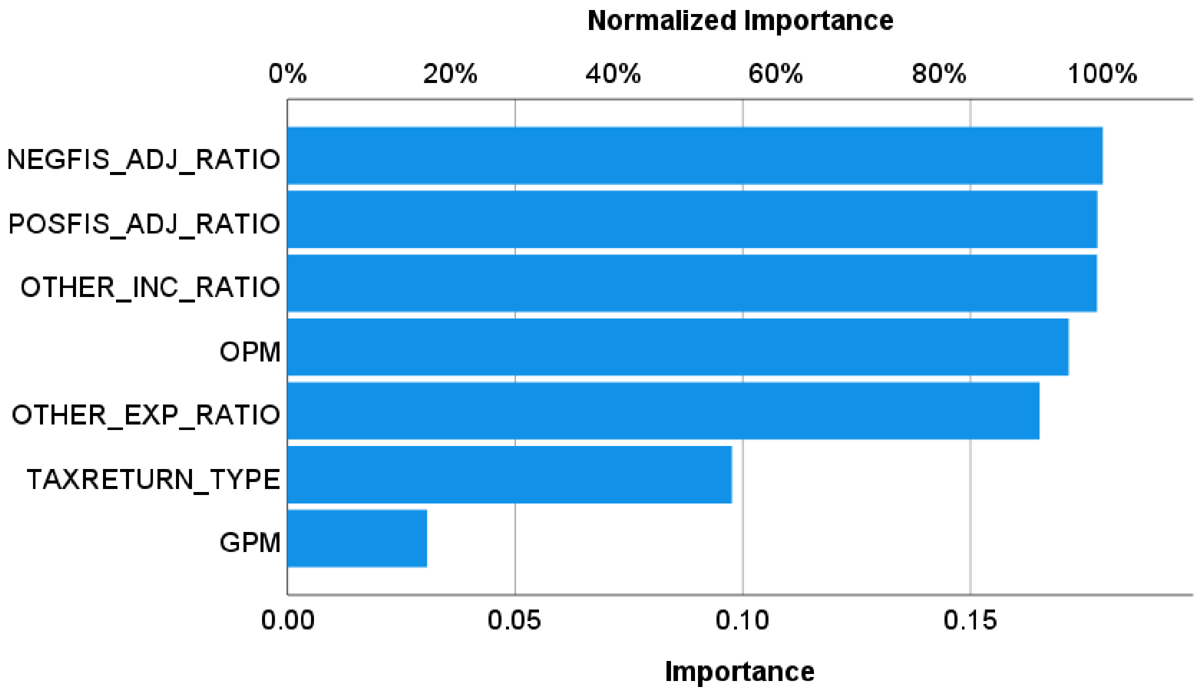
	Medium-large firms (IDR 50 billion to 100 billion)		Large firms (More than IDR 100 billion)	
	Importance	Normalized Importance	Importance	Normalized Importance
Annual tax return type	0.029	13.2%	0.036	17.4%
Gross profit margin (%)	0.045	20.1%	0.027	13.1%
Operating profit margin (%)	0.201	90.5%	0.195	93.9%
Other income ratio (%)	0.176	79.3%	0.178	85.6%
Other expense ratio (%)	0.222	100.0%	0.208	100.0%
Positive fiscal adj. ratio (%)	0.120	53.8%	0.166	79.9%
Negative fiscal adj. ratio (%)	0.207	93.0%	0.191	91.9%

Independent Variable Importance Graph

Small firms (n = 224,181)

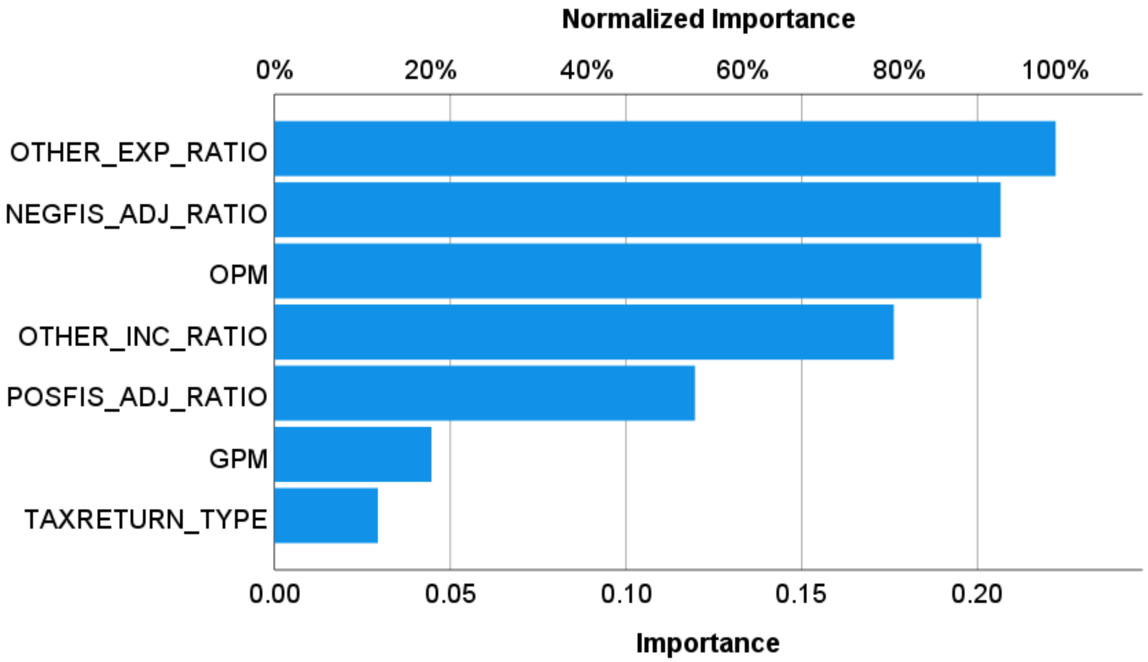


Medium firms (n = 166,171)

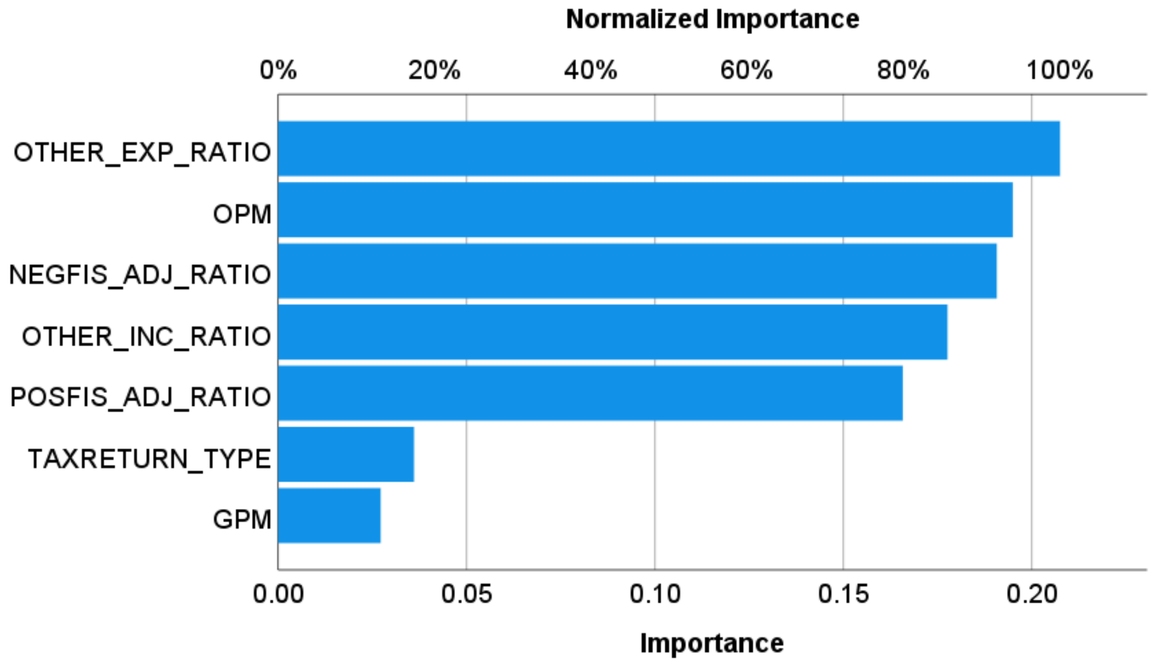


Independent Variable Importance Graph (2)

Medium-large firms (n = 61,572)



Large firms (n = 86,330)



Summary: Accuracy Rate, Predictors, and Areas of Concern

	Prominent predictors by firms' category			
	Small firms (IDR 5B to 15B)	Medium firms (IDR 15B to 50B)	Medium-large firms (IDR 50B to 100B)	Large firms (> IDR 100B)
Accuracy rate*	**89.1%**	**93.6%**	**93.1%**	**92.9%**
Annual return type	No	No	No	No
GPM	No	No	No	No
OPM	**Yes**	No	**Yes**	**Yes**
OIR	**Yes**	**Yes**	No	No
OER	No	No	**Yes**	**Yes**
PFAR	**Yes**	**Yes**	No	No
NFAR	No	**Yes**	**Yes**	**Yes**
Areas of concern within annual income tax return	Part 1c Form 1771-I, part 1e Form 1771-I, and part 5 Form 1771-I	Part 1e Form 1771-I, part 5m Form 1771-I, and 6e Form 1771-I	Part 1c Form 1771-I, part 1e Form 1771-I, and part 6e Form 1771-I	

Final thoughts: Unintended Consequence of ML

- GIGOLO → 'garbage in, garbage out, low outcome'
- Unintended or potentially harmful bias in the models
- User overreliance
- In search of fair and interpretable ML

Readings

Cook, T. R. (2020). Neural Networks. In P. Fuleky (Ed.), Macroeconomic Forecasting in the Era of Big Data: Theory and Practice

Haykin, S.S. (1999). Neural Networks: A Comprehensive Foundation, 2nd edition, Macmillan College Publishing, New York

International Business Machines (IBM). (2021). IBM SPSS Neural Networks 28, IBM Corporation

OECD (2016), Advanced Analytics for Better Tax Administration: Putting Data to Work, OECDPublishing, Paris

Ripley, B.D. (1996). Pattern Recognition And Neural Networks. Cambridge University Pres, Cambridge.

Wendler, T., & Gröttrup, S. (2021). Data Mining with SPSS Modeler: Theory, Exercises and Solutions. Springer Nature Switzerland

Thank you

Making an impact. Living a legacy.

arifin.rosid@pajak.go.id

arifin.rosid@ui.ac.id