ALY6060: Decision Support & Business Intelligence

Assignment 5

Exploring Data Insights: K-Means Clustering and Neural Networks Analysis Using SPSS Statistics

Submitted to: Prof. Fatemeh Abkenari

Submitted by: Group 8

Abhilash Kumar Dikshit

Shamim Sherafati

Date: 12/10/2023

College of Professional Studies, Northeastern University Vancouver, Canada

I. Introduction:

The comprehensive statistical examination of property tax data using IBM SPSS Statistics sought to deeply investigate and comprehend diverse facets pertaining to property tax assessments in the Vancouver region until 2023. The primary focus was on extracting data insights through visualization, statistical analysis, and compelling storytelling. Leveraging a dataset encompassing 873,124 entries, the study probed into relationships between independent variables and dependent variables. Visualizations were employed to discern the influence of TAX_LEVY, REPORT_YEAR, and YEAR_BUILT on CURRENT_LAND_VALUE, CURRENT_IMPROVEMENT_VALUE, PREV_LAND_VALUE, and PREV_IMPROVEMENT_VALUE. The objective was to reveal latent patterns within property tax assessment variables, utilizing descriptive statistics, clustering, ANOVA, Multilayer Perceptron, Classification, and AUC techniques to predict land values of legal types based on property characteristics.

II. Analysis:

As depicted in the below bar chart, it illustrates the zoning classification in the Vancouver region along with past and present land values. The examination offers a thorough visualization of zoning attributes in conjunction with previous and current land values, offering insights into the urban landscape. This report delves into the observations derived from the dashboard components, portraying a clear picture of the correlation between zoning classifications and property values.

Descriptive Statistics

The FOLIO variable, representing unique identifiers, showcases a wide range, from 2.E+10 to 8E+11, with a mean of 4.99E+11 and a standard deviation of 2.496E+11. This indicates significant variability in the property tax data, with a concentration around the mean.

LAND_COORDINATE, denoting the geographic coordinates of properties, exhibits considerable diversity, ranging from 1963206 to 84531342. The mean of 49897988 with a standard deviation of 24958687.1 indicates a spread across different regions.

The variable FROM_CIVIC_NUMBER, depicting civic numbers of properties, has a mean of 866.30, suggesting a concentration around this value, with a standard deviation of 875.560, signifying variability in civic number assignments.

CURRENT_LAND_VALUE, representing the current assessed land values, shows a wide range, from 0 to 3568531000, with a mean of 1749660.86 and a substantial standard deviation of 100578195.3, indicating diverse property valuations.

Other variables like CURRENT_IMPROVEMENT_VALUE, TAX_ASSESSMENT_YEAR, PREVIOUS_LAND_VALUE, PREVIOUS_IMPROVEMENT_VALUE, YEAR_BUILT, BIG_IMPROVEMENT_YEAR, TAX_LEVY, NEIGHBOURHOOD_CODE, REPORT_YEAR.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
FOLIO	873484	2.E+10	8.E+11	4.99E+11	2.496E+11
LAND_COORDINATE	873484	1963206	84531342	49897988.80	24958687.1
FROM_CIVIC_NUMBER	427023	0	6705	866.30	875.560
TO_CIVIC_NUMBER	871240	1	31888	2389.22	1994.596
CURRENT_LAND_VALUE	860914	0	3568531000	1749660.86	10057195.3
CURRENT_IMPROVEMENT_VALUE	860914	0	876401000	451709.80	4766582.444
TAX_ASSESSMENT_YEAR	860914	2020	2023	2021.51	1.118
PREVIOUS_LAND_VALUE	850976	0	3488433000	1736914.94	10018672.3
PREVIOUS_IMPROVEMENT_VALUE	850976	0	652775000	424229.23	4279501.536
YEAR_BUILT	847604	1800	2022	1984.36	29.752
BIG_IMPROVEMENT_YEAR	847604	200	2022	1991.84	19.664
TAX_LEVY	861605	.00	9760300.00	8964.5822	64805.61572
NEIGHBOURHOOD_CODE	873484	1	30	16.55	8.943
REPORT_YEAR	873484	2020	2023	2021.51	1.118
Valid N (listwise)	408964				

Fig 1: Descriptive Statistics.

Initial Cluster Centers

The clustering process begins with the calculation of Z scores for Cluster 1 and 2. In the first iteration, we observe Z scores of 7.162 for Cluster 1 and 4.153 for Cluster 2. Subsequently, in the second iteration, convergence is swiftly achieved, with Z scores stabilizing at 0.000 for both clusters, signifying minimal changes in cluster centers.

Initial Cluster Centers

	Cluster	
	1	2
Zscore(FOLIO)	.24895	-1.49919
Zscore (LAND_COORDINATE)	.24895	-1.49919
Zscore (FROM_CIVIC_NUMBER)	-.35554	-.63764
Zscore (TO_CIVIC_NUMBER)	-.69700	2.42594
Zscore (CURRENT_LAND_VALUE)	22.80938	-.17397
Zscore (CURRENT_IMPROVEMENT_VALUE)	14.79284	-.09477
Zscore (TAX_ASSESSMENT_YEAR)	-.45759	1.33119
Zscore (PREVIOUS_LAND_VALUE)	24.36322	-.17337
Zscore (PREVIOUS_IMPROVEMENT_VALUE)	29.59779	-.09913
Zscore(YEAR_BUILT)	.05505	1.16423
Zscore (BIG_IMPROVEMENT_YEAR)	-.29680	1.38141
Zscore(TAX_LEVY)	47.34306	-.13833
Zscore (NEIGHBOURHOOD_CODE)	1.28085	-.73193
Zscore(REPORT_YEAR)	-.45895	1.32959

Iteration History[a]

Iteration	Change in Cluster Centers 1	Change in Cluster Centers 2
1	7.162	4.153
2	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 67.222.

Fig 2: Initial Cluster Centers and Iteration History.

Final Cluster Centers

Final Cluster Centers

	Cluster 1	Cluster 2
Zscore(FOLIO)	.24895	-.17821
Zscore (LAND_COORDINATE)	.24895	-.17821
Zscore (FROM_CIVIC_NUMBER)	-.35554	-.00931
Zscore (TO_CIVIC_NUMBER)	-.69700	-.30432
Zscore (CURRENT_LAND_VALUE)	23.53900	-.11036
Zscore (CURRENT_IMPROVEMENT_VALUE)	20.68018	-.04665
Zscore (TAX_ASSESSMENT_YEAR)	-.90479	.01790
Zscore (PREVIOUS_LAND_VALUE)	24.18016	-.11019
Zscore (PREVIOUS_IMPROVEMENT_VALUE)	28.63494	-.04563
Zscore(YEAR_BUILT)	.05505	.48452
Zscore (BIG_IMPROVEMENT_YEAR)	-.29680	.39461
Zscore(TAX_LEVY)	43.50381	-.09691
Zscore (NEIGHBOURHOOD_CODE)	1.28085	.27231
Zscore(REPORT_YEAR)	-.90608	.01648

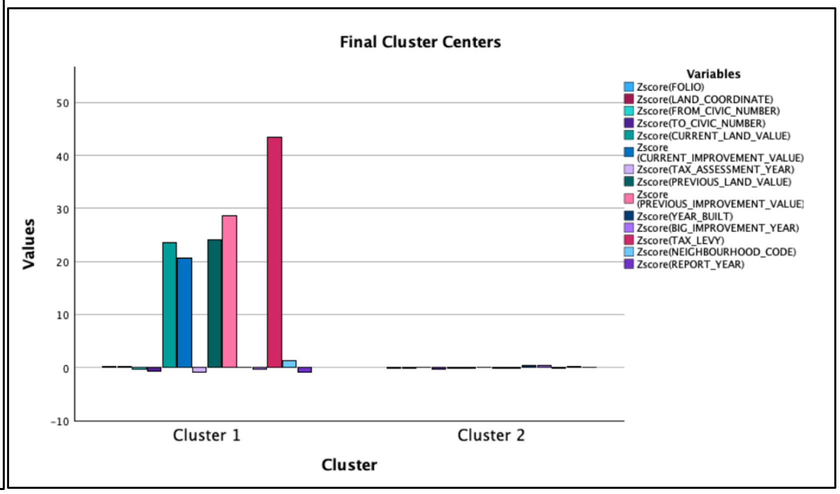


Fig 3: Final Cluster Centers

ANOVA

ANOVA results reveal that F tests should be utilized descriptively, given the intentional clustering to minimize differences among cases. With 408,962 cases in Cluster 2 and only 2 in Cluster 1, the observed significance levels lack correction for the minimized differences.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore(FOLIO)	.365	1	.983	408962	.371	.542
Zscore (LAND_COORDINATE)	.365	1	.983	408962	.371	.542
Zscore (FROM_CIVIC_NUMBER)	.240	1	.982	408962	.244	.621
Zscore (TO_CIVIC_NUMBER)	.308	1	.738	408962	.418	.518
Zscore (CURRENT_LAND_VALUE)	1118.579	1	.003	408962	334654.854	<.001
Zscore (CURRENT_IMPROVEMENT_VALUE)	859.199	1	.001	408962	614541.695	<.001
Zscore (TAX_ASSESSMENT_YEAR)	1.703	1	1.004	408962	1.697	.193
Zscore (PREVIOUS_LAND_VALUE)	1180.037	1	.004	408962	322722.246	<.001
Zscore (PREVIOUS_IMPROVEMENT_VALUE)	1645.143	1	.002	408962	1084474.670	<.001
Zscore(YEAR_BUILT)	.369	1	.251	408962	1.471	.225
Zscore (BIG_IMPROVEMENT_YEAR)	.956	1	.407	408962	2.349	.125
Zscore(TAX_LEVY)	3802.027	1	.003	408962	1327956.583	<.001
Zscore (NEIGHBOURHOOD_CODE)	2.034	1	1.118	408962	1.820	.177
Zscore(REPORT_YEAR)	1.702	1	1.003	408962	1.697	.193

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

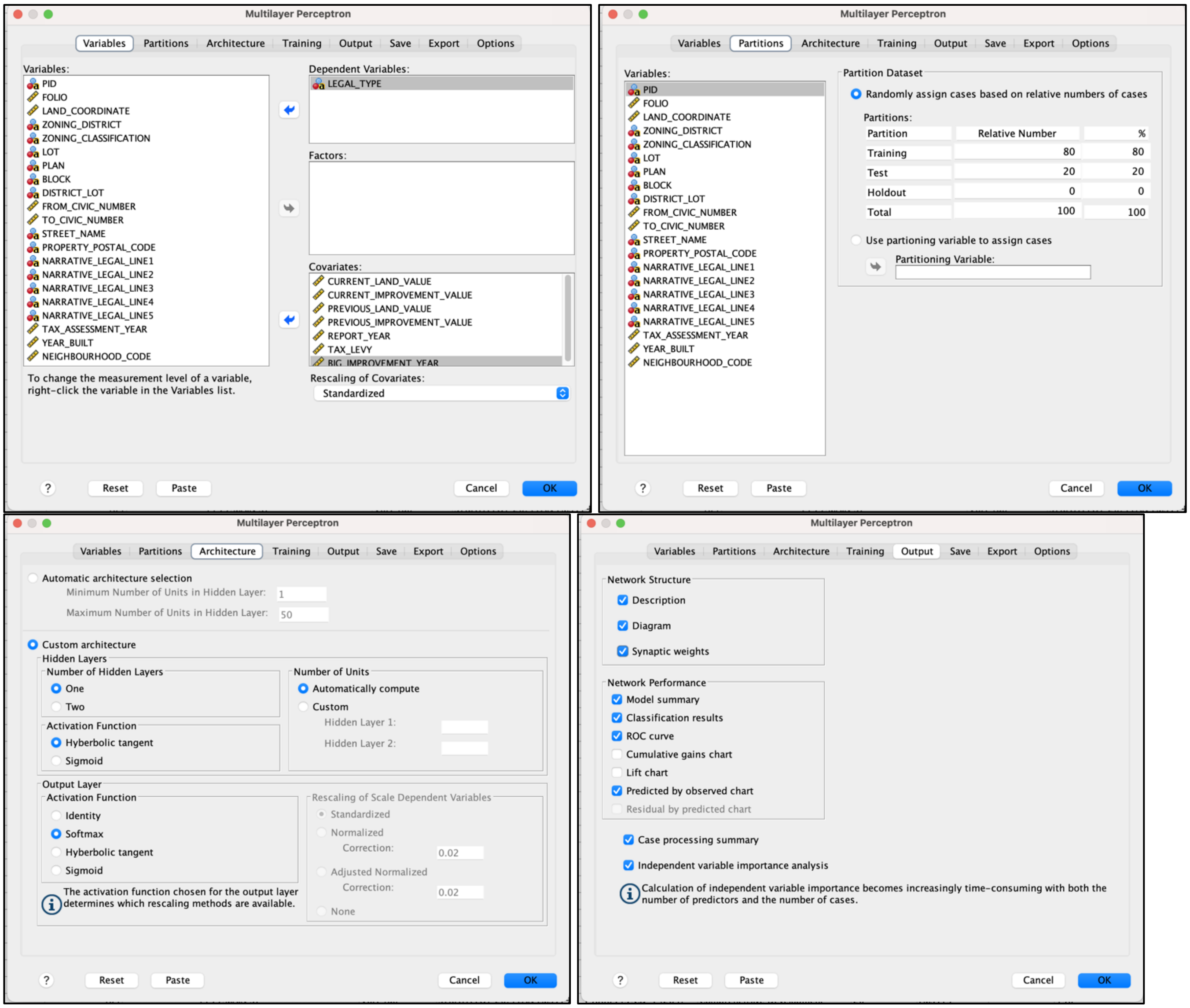
Number of Cases in each Cluster

Cluster	1	2.000
	2	408962.000
Valid		408964.000
Missing		464520.000

Fig 4: ANOVA

Multilayer Perceptron

The Multilayer Perceptron (MLP) model is structured with one hidden layer, employing hyperbolic tangent as the activation function. Training and testing involve an 80-20 split, with covariates including CURRENT_LAND_VALUE, CURRENT_IMPROVEMENT_VALUE, PREV_LAND_VALUE, PREV_IMPROVEMENT_VALUE, TAX_LEVY, BIG_IMPROVEMENT_YEAR. The model achieves convergence in two iterations.



Multilayer Perceptron

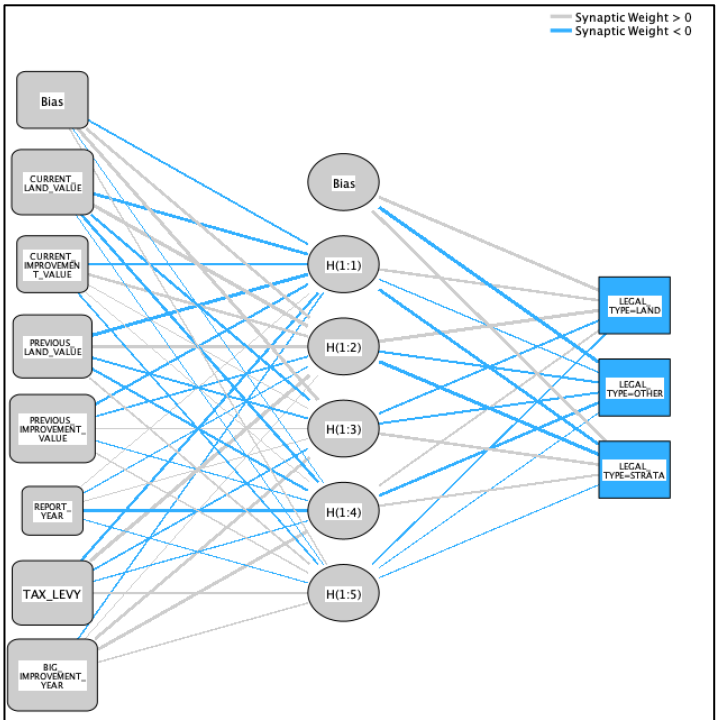
Case Processing Summary

		N	Percent
Sample	Training	672107	80.1%
	Testing	167335	19.9%
Valid		839442	100.0%
Excluded		34042	
Total		873484	

Network Information

Input Layer	Covariates	1	CURRENT_LAND_VALUE
		2	CURRENT_IMPROVEMENT_VALUE
		3	PREVIOUS_LAND_VALUE
		4	PREVIOUS_IMPROVEMENT_VALUE
		5	REPORT_YEAR
		6	TAX_LEVY
		7	BIG_IMPROVEMENT_YEAR
	Number of Units[a]		7
	Rescaling Method for Covariates		Standardized
Hidden Layer(s)	Number of Hidden Layers		1
	Number of Units in Hidden Layer 1[a]		5
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	LEGAL_TYPE
	Number of Units		3
	Activation Function		Softmax
	Error Function		Cross-entropy

a. Excluding the bias unit



Hidden layer activation function: Hyperbolic tangent

Output layer activation function: Softmax

Model Summary

Training	Cross Entropy Error	118003.715
	Percent Incorrect Predictions	6.5%
	Stopping Rule Used	Maximum number of epochs (100) exceeded
	Training Time	0:02:15.20
Testing	Cross Entropy Error	28974.427
	Percent Incorrect Predictions	6.3%

Dependent Variable: LEGAL_TYPE

Parameter Estimates

Predictor		Predicted: Hidden Layer 1 H(1:1)	H(1:2)	H(1:3)	H(1:4)	H(1:5)	Output Layer [LEGAL_TYPE =LAND]	[LEGAL_TYPE =OTHER]	[LEGAL_TYPE =STRATA]
Input Layer	(Bias)	-.698	1.613	2.643	-.074	.337			
	CURRENT_LAND_VALUE	-1.680	6.790	-1.097	-1.302	-.101			
	CURRENT_IMPROVEMENT_VALUE	-.927	1.351	.124	.110	-.606			
	PREVIOUS_LAND_VALUE	-2.819	4.881	-.813	-.941	.681			
	PREVIOUS_IMPROVEMENT_VALUE	-1.074	-.651	.040	-.196	.358			
	REPORT_YEAR	.085	-.196	.102	-1.122	-.181			
	TAX_LEVY	-1.096	8.440	-.619	-.235	.623			
	BIG_IMPROVEMENT_YEAR	-.402	.024	1.996	2.238	.412			
Hidden Layer 1	(Bias)						2.199	-4.047	2.090
	H(1:1)						1.254	-.275	-1.850
	H(1:2)						3.428	-.846	-3.478
	H(1:3)						-.843	-.844	1.571
	H(1:4)						.745	-1.560	.867
	H(1:5)						-.468	-.129	-.192

Fig 5: Multilayer Perceptron

Classification

Classification results for the Dependent Variable LEGAL_TYPE present observed and predicted values for LAND, STRATA, and OTHER categories.

Classification

Sample	Observed	Predicted LAND	Predicted OTHER	Predicted STRATA	Percent Correct
Training	LAND	268161	0	10257	96.3%
	OTHER	158	0	173	0.0%
	STRATA	32854	0	360504	91.6%
	Overall Percent	44.8%	0.0%	55.2%	93.5%
Testing	LAND	67029	0	2547	96.3%
	OTHER	20	0	39	0.0%
	STRATA	8010	0	89690	91.8%
	Overall Percent	44.9%	0.0%	55.1%	93.7%

Dependent Variable: LEGAL_TYPE

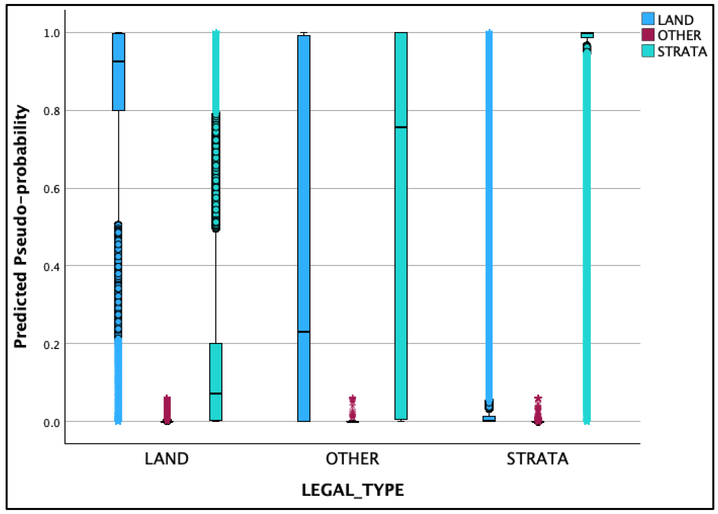
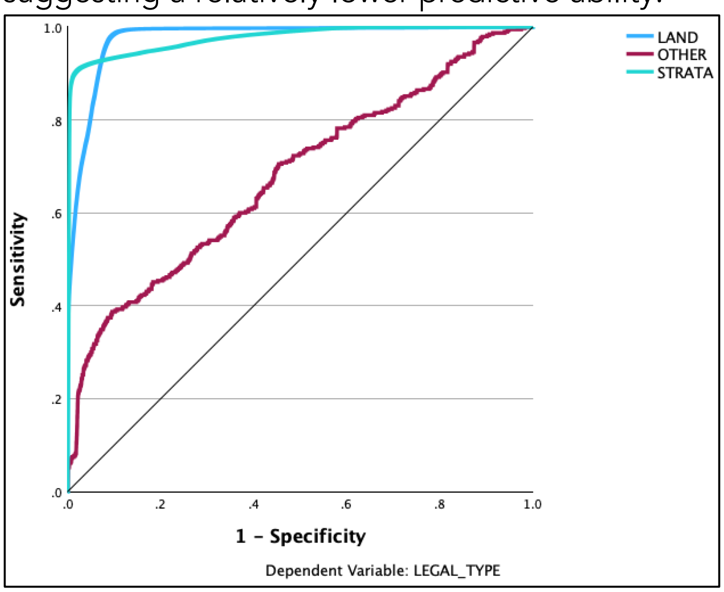


Fig 6: Classification

Area Under the Curve

AUC values indicate strong predictive performance, particularly for LAND and STRATA with scores of 0.976 each. The OTHER category exhibits a slightly lower AUC of 0.679, suggesting a relatively lower predictive ability.



Area Under the Curve

		Area
LEGAL_TYPE	LAND	.976
	OTHER	.679
	STRATA	.976

Fig 7: AUC

Normalized Importance provides insights into the contribution of variables to the model. Notably, BIG_IMPROVEMENT_YEAR holds the highest importance at 100%, followed by PREV_IMPROVEMENT_VALUE at 75.1%, and CURRENT_LAND_VALUE at 78.2%. These findings underscore the significance of these variables in predicting legal types based on property characteristics.

Independent Variable Importance

	Importance	Normalized Importance
CURRENT_LAND_VALUE	.161	78.2%
CURRENT_IMPROVEMENT_VALUE	.103	50.0%
PREVIOUS_LAND_VALUE	.146	70.8%
PREVIOUS_IMPROVEMENT_VALUE	.182	88.1%
REPORT_YEAR	.047	22.8%
TAX_LEVY	.155	75.1%
BIG_IMPROVEMENT_YEAR	.206	100.0%

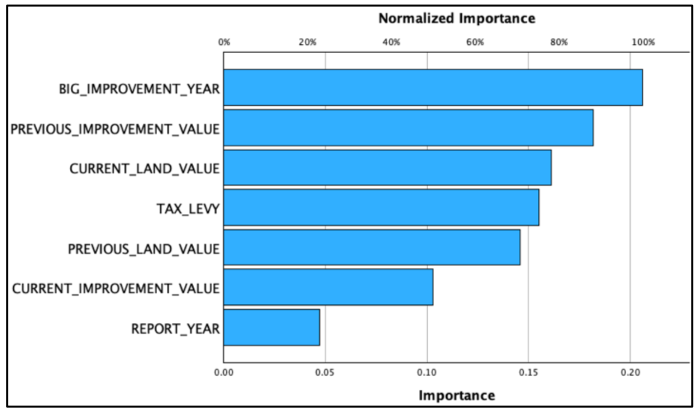


Fig 8: Independent Variable Importance

III. Conclusion:

This comprehensive analysis combines descriptive statistics, clustering, ANOVA, and advanced techniques like Multilayer Perceptron and Classification to derive meaningful insights from property tax data. The descriptive statistics offer a nuanced understanding of variable distributions, while clustering and ANOVA caution against hasty interpretations. The MLP model and subsequent classification demonstrate the robustness of the predictive framework, with AUC values providing a quantitative measure of performance. The identification of variable importance contributes to a refined understanding of the factors influencing legal types. This technical exploration lays a solid foundation for leveraging data-driven insights in business-oriented decision-making within the realm of property tax assessments.

IV. References:

Property tax report. (n.d.). https://opendata.vancouver.ca/explore/dataset/property- tax-report/table/?sort=-tax_assessment_year