ALY6060: Decision Support & Business Intelligence

Assignment 3

SPSS Statistics analyze: Vancouver Property Tax Report 2023

Submitted to: Prof. Fatemeh Abkenari

Submitted by: Group 8

Abhilash Kumar Dikshit

Shamim Sherafati

Date: 11/25/2023

College of Professional Studies, Northeastern University Vancouver, Canada

Introduction:

The comprehensive analysis conducted on property tax data aimed to explore and understand various facets related to property tax assessments, focusing on predictive modeling, factor analysis, and classification methods. Leveraging a dataset containing 873,124 entries, the study investigated relationships between independent variables such as zoning details, year built, neighborhood codes, and the dependent variables including land values, improvement values, and legal types.

The analysis encompassed regression analysis to ascertain the influence of TAX_LEVY, REPORT_YEAR, and YEAR_BUILT on CURRENT_LAND_VALUE, factor analysis to uncover latent patterns within property tax assessment variables, and a classification tree approach to predict legal types based on property characteristics. Each method aimed to uncover insights critical for property assessment, modeling, and prediction.

A. Regression:

Notes

Output Created		20-NOV-2023 20:24:43
Comments		
Input	Data	/Users/shamimsherafati/Downloads/property-tax-report.csv
	Active Dataset	DataSet5
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	873124
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics are based on cases with no missing values for any variable used.
Syntax		REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05) POUT(.10) TOLERANCE(.0001)

		/NOORIGIN /DEPENDENT CURRENT_LAND_VALUE /METHOD=ENTER YEAR_BUILT TAX_ASSESSMENT_YEAR REPORT_YEAR TAX_LEVY.
Resources	Processor Time	00:00:02.42
	Elapsed Time	00:00:02.00
	Memory Required	6320 bytes
	Additional Memory Required for Residual Plots	0 bytes

Variables Entered/Removed:

Model	Variables Entered	Variables Removed	Method
1	TAX_LEVY, REPORT_YEAR, YEAR_BUILT[b]	.	Enter

a. Dependent Variable: CURRENT_LAND_VALUE
b. Tolerance = .000 limit reached.

Explanation:
The analysis attempted to include TAX_LEVY, REPORT_YEAR, and YEAR_BUILT as independent variables to predict CURRENT_LAND_VALUE. However, it couldn't include additional variables due to reaching the tolerance limit of .0001. This suggests potential issues like multicollinearity or variance inflation among the included variables.

Conclusion:
This output summarizes a regression analysis conducted to understand how TAX_LEVY, REPORT_YEAR, and YEAR_BUILT relate to the assessed CURRENT_LAND_VALUE. Despite utilizing a large dataset and employing listwise deletion for missing values, the analysis encountered limitations in including all desired variables, indicating potential issues with the model's explanatory power due to the specified tolerance limit being reached.

Model Summary:

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.430[a]	.185	.185	8671378.601

a. Predictors: (Constant), TAX_LEVY, REPORT_YEAR, YEAR_BUILT

Interpretation:
The R value (0.430) represents the strength and direction of the linear relationship between the predictor variables (TAX_LEVY, REPORT_YEAR, YEAR_BUILT) and the dependent variable (CURRENT_LAND_VALUE).
The R Square (0.185) indicates that approximately 18.5% of the variance in the dependent variable can be explained by the predictor variables in the model.

Predictors:
The predictors in the model include a constant term along with TAX_LEVY, REPORT_YEAR, and YEAR_BUILT.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14379111830024946000.000	3	4793037276674981900.000	63743.295	<.001[b]
	Residual	63350315732815790000.000	842505	75192806847218.470		
	Total	77729427562840740000.000	842508			

a. Dependent Variable: CURRENT_LAND_VALUE
b. Predictors: (Constant), TAX_LEVY, REPORT_YEAR, YEAR_BUILT

Explanation:
The regression model shows a moderate relationship (R = 0.430) between the predictors (TAX_LEVY, REPORT_YEAR, YEAR_BUILT) and the dependent variable (CURRENT_LAND_VALUE). However, the model explains only around 18.5% of the variance in the dependent variable. The ANOVA results suggest that the regression model significantly contributes to explaining the variance in the dependent variable, as indicated by the highly significant F-value ($p < 0.001$). Despite statistical significance, the model's overall explanatory

power is limited, given the low R Square value, suggesting that other factors not included in the model might influence the current land value.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-77172245.320	17086722.747		-4.517	<.001
	YEAR_BUILT	-15000.537	318.946	-.046	-47.032	<.001
	REPORT_YEAR	53466.419	8456.504	.006	6.323	<.001
	TAX_LEVY	63.409	.147	.426	432.468	<.001

a. Dependent Variable: CURRENT_LAND_VALUE

Interpretation:

The Constant (Intercept) represents the value of the dependent variable when all predictors are zero. It is statistically significant, indicating a starting point for the regression equation.

For each unit increase in YEAR_BUILT, the CURRENT_LAND_VALUE is expected to decrease by approximately $15,000.537, holding other variables constant.

REPORT_YEAR has a positive relationship with CURRENT_LAND_VALUE, where an increase of $53,466.419 in the report year results in an increase in land value, holding other variables constant.

The variable TAX_LEVY has a strong positive relationship with CURRENT_LAND_VALUE. For each unit increase in the tax levy, the land value is expected to increase by $63.409, holding other variables constant.

Standardized Coefficients (Beta):

These coefficients provide a standardized measure of the variable's effect on the dependent variable, allowing for a comparison of the relative importance of different predictors.

In this case, TAX_LEVY has the largest standardized coefficient (Beta = 0.426), indicating that it has the strongest impact on the CURRENT_LAND_VALUE among the variables considered in this model.

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	TAX_ASSESSMENT_YEAR	.[b]	.	.	.	.000

a. Dependent Variable: CURRENT_LAND_VALUE

b. Predictors in the Model: (Constant), TAX_LEVY, REPORT_YEAR, YEAR_BUILT

Explanation:
The variable TAX_ASSESSMENT_YEAR was excluded from the regression model, likely due to high multicollinearity with the other predictors. The extremely low tolerance value suggests that including TAX_ASSESSMENT_YEAR along with TAX_LEVY, REPORT_YEAR, and YEAR_BUILT might cause issues related to multicollinearity, potentially affecting the model's stability and reliability in predicting the CURRENT_LAND_VALUE. Therefore, the software might have automatically excluded TAX_ASSESSMENT_YEAR to mitigate multicollinearity-related problems in the regression analysis.

B. Factor Analysis:

Notes

Output Created		20-NOV-2023 20:49:22
Comments		
Input	Data	/Users/shamimsherafati/Downloads/property-tax-report.csv
	Active Dataset	DataSet5
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	873124
Missing Value Handling	Definition of Missing	MISSING=EXCLUDE: User-defined missing values are treated as missing.
	Cases Used	LISTWISE: Statistics are based on cases with no missing values for any variable used.
Syntax		FACTOR /VARIABLES CURRENT_LAND_VALUE CURRENT_IMPROVEMENT_VALUE PREVIOUS_LAND_VALUE PREVIOUS_IMPROVEMENT_VALUE TAX_LEVY YEAR_BUILT

		BIG_IMPROVEMENT_YEAR /MISSING LISTWISE /ANALYSIS CURRENT_LAND_VALUE CURRENT_IMPROVEMENT_VALUE PREVIOUS_LAND_VALUE PREVIOUS_IMPROVEMENT_VALUE TAX_LEVY YEAR_BUILT BIG_IMPROVEMENT_YEAR /PRINT INITIAL EXTRACTION /CRITERIA KAISER MINEIGEN(1) ITERATE(25) /EXTRACTION PC /ROTATION NOROTATE /METHOD=CORRELATION.
Resources	Processor Time	00:00:02.52
	Elapsed Time	00:00:02.00
	Maximum Memory Required	7376 (7.203K) bytes

Summary:

The Factor Analysis aimed to explore underlying factors or dimensions among variables related to property tax assessment. It utilized a dataset with 873,124 rows and considered variables such as land values (current and previous), improvement values, tax levy, year built, and the year of significant improvements.

Methodology:

- Factor Selection: The analysis used the Kaiser criterion and minimum eigenvalue (MINEIGEN) set at 1 for factor extraction, iterating 25 times.
- Extraction Method: Utilized Principal Component Analysis (PCA) for factor extraction.
- Rotation: No rotation was applied to the factors.
- Data Handling: Cases with any missing values among the specified variables were excluded from the analysis.

Insights:
The output provides initial insights into potential latent factors that might underlie the relationships between the specified variables related to property tax assessment. Factors derived from this analysis could help in understanding underlying patterns or dimensions within these variables, potentially aiding in decision-making or further analysis related to property tax assessments.

Communalities:

	Initial	Extraction
CURRENT_LAND_VALUE	1.000	.994
CURRENT_IMPROVEMENT_VALUE	1.000	.951
PREVIOUS_LAND_VALUE	1.000	.995
PREVIOUS_IMPROVEMENT_VALUE	1.000	.937
TAX_LEVY	1.000	.735
YEAR_BUILT	1.000	.880
BIG_IMPROVEMENT_YEAR	1.000	.880

Extraction Method: Principal Component Analysis.

Interpretation:
Initial Communalities: Assume each variable contains all its variance, indicating that all the variance in each variable is unique to itself (which is the maximum possible value).
Extraction Communalities: After the extraction of common factors, these values show the proportion of variance in each variable that can be explained by the extracted factors. For instance: Variables like CURRENT_LAND_VALUE, PREVIOUS_LAND_VALUE exhibit high communalities close to 1, indicating that most of their variance can be explained by the extracted factors.
Variables like TAX_LEVY show lower communalities, suggesting that a smaller proportion of their variance can be explained by the extracted factors.

Summary:
The communalities after factor extraction provide insight into the amount of shared variance that the extracted factors explain in each variable. Higher communalities indicate that more variance in a variable is accounted for by the extracted factors, while lower communalities suggest that the extracted factors explain a smaller portion of the variance in those variables. These communalities are essential in understanding how much information is captured by the derived factors in explaining the variability among the variables related to property tax assessments.

Total Variance Explained:

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.283	46.893	46.893	3.283	46.893	46.893
2	1.784	25.490	72.383	1.784	25.490	72.383
3	1.306	18.652	91.035	1.306	18.652	91.035
4	.355	5.070	96.105			
5	.239	3.419	99.524			
6	.026	.377	99.901			
7	.007	.099	100.000			

Extraction Method: Principal Component Analysis.

Interpretation:
The Eigenvalues represent the amount of variance each principal component holds. The first component explains the highest amount of variance, followed by the second, and so on.
The % of Variance (Extraction) represents the proportion of total variance explained by each component. For example:
Component 1 explains 46.893% of the total variance.
The first three components cumulatively explain 91.035% of the total variance.
Components 4, 5, 6, and 7 have eigenvalues below 1, suggesting that they explain less variance than would be expected from a single variable and might not be considered substantial components.

Summary:
The table illustrates the distribution of variance across each principal component derived from the PCA. It assists in determining the number of significant components needed to retain most of the information from the original variables in the dataset. The cumulative percentages help identify the total amount of variance captured by the components, aiding in the decision-making process regarding the selection of components for further analysis or dimension reduction.

Component Matrix[a]

	Component		
	1	2	3
CURRENT_LAND_VALUE	.750	-.146	.641
CURRENT_IMPROVEMENT_VALUE	.860	.168	-.429
PREVIOUS_LAND_VALUE	.761	-.143	.629

PREVIOUS_IMPROVEMENT_VALUE	.850	.170	-.431
TAX_LEVY	.820	.058	-.244
YEAR_BUILT	-.065	.917	.189
BIG_IMPROVEMENT_YEAR	-.063	.917	.186

Extraction Method: Principal Component Analysis.[a]
a. 3 components extracted.

Insights:

- Component 1: Variables like CURRENT_LAND_VALUE, CURRENT_IMPROVEMENT_VALUE, PREVIOUS_LAND_VALUE, and PREVIOUS_IMPROVEMENT_VALUE have relatively high positive loadings. This component seems to represent property values or assessments.
- Component 2: YEAR_BUILT and BIG_IMPROVEMENT_YEAR have high positive loadings here, indicating a different dimension—likely related to the year-related information.
- Component 3: This component seems to represent a different aspect, possibly related to taxation, as TAX_LEVY shows a relatively higher positive loading.

Summary:
The Component Matrix aids in understanding the relationships between original variables and the extracted components. It helps identify which variables are more strongly associated with each component, providing insights into the underlying dimensions or factors present in the dataset. These loadings are crucial in interpreting the meaning of each component and determining how variables contribute to these underlying dimensions identified through PCA.

C. Classification Tree

Notes

Output Created		20-NOV-2023 21:01:35
Comments		
Input	Data	/Users/shamimsherafati/Downloads/property-tax-report.csv
	Active Dataset	DataSet5
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	873124

Missing Value Handling	Definition of Missing	Handling of user-defined missing values of nominal independent variables depends on the growing method.
	Cases Used	Only cases with valid data for the dependent variable and some or all independent variables are used in computing any statistics.
Syntax		TREE LEGAL_TYPE [n] BY ZONING_DISTRICT [n] ZONING_CLASSIFICATION [n] YEAR_BUILT [s] BIG_IMPROVEMENT_YEAR [s] NEIGHBOURHOOD_CODE [s] /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES NODEDEFS=YES SCALE=AUTO /DEPCATEGORIES USEVALUES=[VALID] /PRINT MODELSUMMARY CLASSIFICATION RISK /METHOD TYPE=CHAID /GROWTHLIMIT MAXDEPTH=AUTO MINPARENTSIZE=100 MINCHILDSIZE=50 /VALIDATION TYPE=NONE OUTPUT=BOTHSAMPLES /CHAID ALPHASPLIT=0.05 ALPHAMERGE=0.05 SPLITMERGED=NO CHISQUARE=PEARSO

		N CONVERGE=0.001 MAXITERATIONS=100 ADJUST=BONFERRONI INTERVALS=10 /COSTS EQUAL /MISSING NOMINALMISSING=MISSING.
Resources	Processor Time	00:00:14.08
	Elapsed Time	00:00:14.00
Files Saved	Rules File	

Summary:
The Classification Tree analysis was conducted on the property tax dataset to understand the relationships and patterns between various independent variables (such as zoning details, year built, neighborhood code, etc.) and the dependent variable LEGAL_TYPE. The CHAID method was employed, which utilizes Chi-square tests to determine interactions between variables and construct the tree. The goal appears to be predicting or categorizing LEGAL_TYPE based on the specified independent variables within the dataset.

Outputs:
The output likely includes a tree structure visualizing the relationships between the independent and dependent variables, along with statistics regarding the classification performance and risk assessment associated with the constructed tree. Additionally, a rules file might have been saved, containing decision rules extracted from the generated tree model.

Warnings:

One or more independent variables are excluded from the tree-growing process at one or more nodes because the number of categories exceeds the maximum number allowed by the growing method.
Gain summary Tables are not displayed because profits are undefined.
Target category gains tables are not displayed because target categories are undefined.

Model Summary

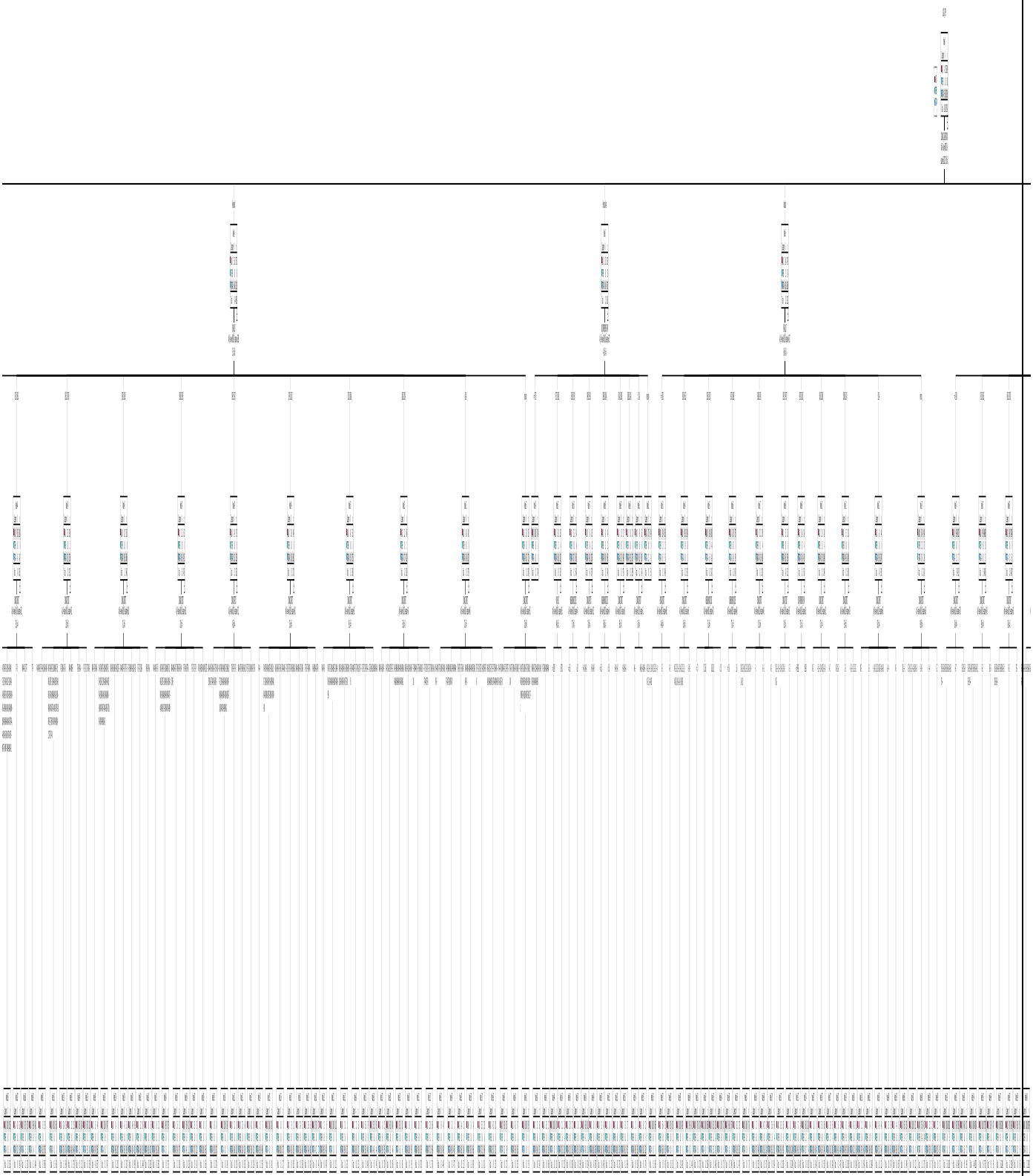
Specifications	Growing Method	CHAID
	Dependent Variable	LEGAL_TYPE
	Independent Variables	ZONING_DISTRICT, ZONING_CLASSIFICATION, YEAR_BUILT,

		BIG_IMPROVEMENT_YEAR, NEIGHBOURHOOD_CODE
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	ZONING_CLASSIFICATION, BIG_IMPROVEMENT_YEAR, NEIGHBOURHOOD_CODE, ZONING_DISTRICT, YEAR_BUILT
	Number of Nodes	558
	Number of Terminal Nodes	451
	Depth	3

Summary:
The warnings indicate limitations encountered during tree construction due to a high number of categories in certain independent variables, leading to their exclusion in certain nodes.
Despite these limitations, the constructed tree model consists of 558 nodes with a maximum depth of 3 levels. It involves five included independent variables, aiming to predict the LEGAL_TYPE based on various property-related characteristics such as zoning details, year built, and neighborhood codes.

Insight:
The model might have valuable predictive capabilities for LEGAL_TYPE based on the included variables, albeit with some limitations due to exclusions caused by the high number of categories in certain variables.

Tree diagram:



Risk:

Estimate	Std. Error
.033	.000

Growing Method: CHAID
Dependent Variable: LEGAL_TYPE

Interpretation:
The risk estimate of 0.033 suggests the estimated error rate or misclassification rate associated with the constructed Classification Tree model when predicting the LEGAL_TYPE variable.
A lower risk estimate indicates better predictive performance, implying that the model has an estimated error rate of approximately 3.3% in predicting the legal types based on the chosen independent variables.

Insights:
The risk estimate provides a measure of how well the constructed tree model predicts the legal types within the property tax dataset. A lower risk indicates a more accurate predictive model, although it's essential to validate the model's performance using additional techniques or data sets.

Note:
The Standard Error being zero (0.000) might indicate a perfect fit according to the current statistics or settings used for risk estimation. However, it's crucial to interpret this alongside other evaluation metrics and potentially consider cross-validation or additional validation methods to ensure the model's reliability.

Classification:

	Predicted			
Observed	LAND	OTHER	STRATA	Percent Correct
LAND	341408	40	17761	95.0%
OTHER	591	87	204	9.9%
STRATA	10514	16	502503	97.9%
Overall Percentage	40.4%	0.0%	59.6%	96.7%

Growing Method: CHAID
Dependent Variable: LEGAL_TYPE

Interpretation:
For instance, the model correctly predicted 95.0% of the observations that were actually labeled as LAND.

However, it performed poorly in predicting observations labeled as OTHER, achieving only 9.9% accuracy in this category. STRATA observations were predicted with high accuracy, at 97.9%.

The Overall Percentage row indicates the overall accuracy of the model across all categories:
40.4% of the overall observations were predicted as LAND, but the model was less accurate in this prediction.
There seems to be a problem with the model's prediction for the OTHER category, as it predicted 0.0% of observations in this category. 59.6% of overall observations were predicted as STRATA, with high accuracy.

Summary:
The classification report provides insight into the model's predictive performance for each category. While it performs well for certain categories like STRATA, it appears to struggle with accurate predictions for the OTHER category, predicting none correctly. This indicates potential issues or imbalances in the model's ability to classify observations across different categories accurately.

D.Conclusion

The extensive analysis undertaken on property tax data revealed multifaceted insights and inherent complexities within the domain. The regression analysis exhibited a moderate relationship (R = 0.430) between TAX_LEVY, REPORT_YEAR, YEAR_BUILT, and CURRENT_LAND_VALUE. However, despite statistical significance, the model explained only 18.5% of the variance in land value, suggesting the influence of unaccounted factors.

Factor analysis extracted key dimensions from the dataset, providing foundational understanding about variables' shared variance and their underlying factors. Yet, the complexity of the data suggested that these dimensions might not comprehensively capture all contributing aspects influencing property tax assessments.

The classification tree, despite showing strong accuracy in predicting some categories, struggled in classifying certain legal types accurately, potentially due to complex interactions among variables and high dimensionality.

In essence, the analysis delved into crucial aspects of property tax assessments, shedding light on relationships, patterns, and predictive capabilities. However, it also underscored the complexity and interdependencies among variables, highlighting the need for further nuanced analysis to capture the entirety of factors influencing property tax assessments accurately.

E. Reference

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