Leadership in Sustainable Infrastructure

Leadership en Infrastructures Durables

Vancouver, Canada

May 31 – June 3, 2017/ *Mai 31 – Juin 3, 2017*

A DATA ANALYTICS SOLUTION FOR PREDICTING THE CONDITION OF ROADS USING THE MOST AFFORDABLE ATTRIBUTES

S. Madeh Piryonesi [1, 3], and Tamer E. El-Diraby [2]

[1] Ph.D. Student, Department of Civil Engineering, University of Toronto.

[2] Associate Professor, Department of Civil Engineering, University of Toronto.

[3] madeh.piryonesi@mail.utoronto.ca

ABSTRACT

The process of road condition assessment is costly and laborious. Many small municipalities have no sufficient financial resources to collect the distress data for their road network. On the other hand, larger municipalities cannot collect distress data for their entire road network due to the gigantic size of networks and limited budgets. Data analytics could be an efficient solution to this problem. Most municipalities already have collected some data as a part of their asset management program. This data could be utilized to learn models for predicting the condition of roads in the future. In this paper a decision tree is learned to predict the variations in the pavement condition index (PCI) of asphalt roads under no maintenance. The predictive attributes which were selected for model learning are as follows. The current condition of road, annual freeze index, total annual precipitation, annual average of minimum temperature, the functional class of road, pavement type, annual average of maximum temperature and the age of road. These attributes were chosen after interviewing asset management experts and also reviewing the asset management plans of ten small municipalities in Ontario and their data. The cost of data collection was one of the main considerations for selecting these predictive attributes. The machine learning model was learned based on the data retrieved from the Long-Term Pavement Performance (LTPP) database. The developed model successfully predicted the deterioration in the PCI of asphalt roads after three years.

Keywords: Road asset management, machine learning, data analytics, pavement condition index, data collection, LTPP.

1. INTRODUCTION

The deterioration of road network is not unique to Canada. It is a universal concern from Australia and United Kingdom to United States and Canada (Australian Government DIRD 2013); however, road asset management is still nascent in Canada. Mirza (2007) reported that the expanding infrastructure backlogs are going to become a crisis for most Canadian municipalities. According to the Canadian Infrastructure Report Card (2016) about one-third of Canadian infrastructure have a very poor, poor or fair condition. In the wake of such facts, road asset management gained a momentum in Canada over the last few years. Municipalities and other levels of government are looking for new tools and technologies to facilitate their asset management planning.

Due to a variety of reasons, understanding the condition of roads and their expected service life is crucial to municipalities and departments of transportation (DOTs). First, scheduling the remedial actions requires information about the condition of roads. Second, a road network with a poor condition results in a low customer satisfaction. Finally, understanding the condition of assets is necessary for budget planning, risk assessment and hence a successful asset management plan.

Different physical performance indicators (PIs) are used to assess the condition and remaining life of roads. Some of the most popular PIs used in estimating remaining life include pavement condition index (PCI), international roughness index (IRI) and present serviceability index (PSI). The main problem is that collecting data for the PCI, IRI or PSI needs both human and financial resources. Small municipalities usually do not have sufficient financial resources to conduct data collection frequently. Therefore, their condition data is usually outdated. In larger municipalities, the costs will be really high given the size of their networks. In such situations, a prediction model (developed through data analytics) can be highly useful in estimating performance indicators, hence, avoiding some of the laborious and expensive data collection work.

In this study, the pavement condition index (PCI) was chosen over IRI and PSI, because it is commonly used by municipalities and DOTs in North America. Specifically, Ontario municipalities mostly rely on the PCI for road condition assessment. The PCI is applicable to both cement and asphalt pavements. The value of the PCI varies between 0 and 100; a PCI of 100 represents the best possible condition, and a 0 is for the worst possible condition. The ASTM has devoted an entire guideline to the calculation of the PCI. ASTM D 6433 – 07 (2007), which is a standard practice for roads and parking lots pavement condition index survey, has explained the process of calculating the PCI in detail.

As explained by the ASTM D 6433 – 07 (2007) and the SP-024 (the MTO's guideline) calculating the PCI requires collecting distress data (Chong et al. 1982). As it was explained above, data collection is time consuming and costly. In this paper, we show how machine learning models could help municipalities in predicting the PCI of road sections using a number of attributes. These attributes have the lowest costs associated with their data collection. Therefore, the philosophy behind choosing them was not a simple mechanistic/engineering reasoning; the cost of data collection was one of the main considerations. The selected variables include the current value of the PCI, the freeze index, the total annual precipitation, the minimum annual temperature, the functional class of road sections, the pavement type, the maximum annual temperature and the age of road.

2. THE STATUS QUO OF ASSET MANAGEMENT AND THE GAPS

Although a deterioration model is an integral part of asset management plans, a recent study reveals that most small municipalities in Ontario do not incorporate a deterioration model in their asset management analyses. Those who paid attention to deterioration models were the larger municipalities with more experience in asset management; but they mostly depend on deterministic deterioration curves to predict the condition of their assets (El-Diraby et al. 2017). These deterioration curves have a number of pitfalls. First, they are deterministic; the user cannot acquire any probabilities to incorporated them in risk analysis. Second, they only rely on the value of the PCI. It means that the curve predicts the PCI of the road in the future based on the PCI in the current year. In other words, these curves overlook other attributes of road such as age, pavement type, functional class, traffic and the climatic attributes.

More sophisticated probabilistic deterioration models are available in the literature. Markovian models are an example of these probabilistic models (Kleiner 2001). Despite the deterioration curves, Markovian models have an advantage of incorporating probabilities. Notwithstanding their capability, they have serious issues; they often disregard the history of deterioration and the previous maintenance actions (Neves and Frangopol, 2005; Piryonesi and Tavakolan 2017). Furthermore, they may require longitudinal data that is not easily to collect (Ens 2012). Considering these limitations, machine learning and artificial intelligence (AI) algorithms can be an alternative solution. Machine learning and AI tools have become quite popular in engineering (Provost and Fawcett 2013; Woldesenbet et al. 2015; Moghaddam et al. 2016; Lagzi et al. 2017). Specifically, in the domain of civil engineering, the implementation of data analytics algorithms to predict the condition of roads is not rare in the literature. For instance, Lou et al. (2001) used neural networks to predict the variations of crack index of asphalt roads in short term. The neural networks have a good learning capability; however, large amount of data is needed for their training and calibration (Ens 2012). Furthermore, they are hardly intelligible for some decision makers in the municipal sector.

Nowadays, various machine learning and data mining algorithms are available for no charge. They have a high learning capability. That is why they are being used in different disciplines of engineering and science. However, choosing the type of algorithm to solve a specific problem is not trivial. Therefore, careful consideration must be given to choosing the most relevant algorithms.

In this paper a decision tree is learned to predict the deterioration in the PCI of asphalt roads. There are a number of reasons behind choosing a decision tree. First, there is no pre-requisite or assumption for implementation of a decision tree. In other words, there is almost no limitation on the type of attributes that are used to train a decision tree, whereas this is not the case for other algorithms. For instance, the attributes to train a naive Bayes classifier must

be independent from each other; or using a linear regression model without a number of assumptions is invalid. Second, decision trees are very intelligible and easy to interpret. Third, they have a good learning capability, and more importantly are easy to implement and reuse for new data. Despite other classifying algorithms such as k-nearest neighbor (KNN) or Bayes classifier, decision trees result in an explicit model, which can easily classify new examples. For other algorithms, such as KNN and naive Bayes classifier or even clustering algorithms, every new data point must be assessed based on the existing data (Provost and Fawcett 2013; Najibi et al. 2017).

The objective of this paper is to learn an intelligible machine learning model that could be used by small municipalities with small financial resources. Hence, a decision tree is learned using the most affordable attributes that most municipalities have data about them. Furthermore, the developed model requires data for only one year. Therefore, municipalities with no longitudinal data can benefit from this models as well. Using such analyses, municipalities not only can predict the condition of their road network, but also can identify the informative data in decision making and use the results of the models to update their GIS maps, which are common these days (Woldesenbet et al. 2015; Moradi et al. 2015).

3. DATA RETRIEVAL

Every data analytics model needs a training set and a test set. For this paper, data was extracted from the long-term pavement performance (LTPP) database which is available online (LTPP InfoPave 2016). With several hundred tables, the LTPP database is the world's largest and most comprehensive pavement performance database; its data is not only comprehensive, but also has a high quality, and most importantly, is available for no charge. The LTPP data is regularly updated every six month. The number of pavement test sections that are monitored in the LTPP program is more than 2,500 (FHWA 2016). They include both asphalt and Portland cement concrete, and are located in the US and Canada. After understanding the structure of data, the meaning of each field and table and also the table relations, the data was retrieved using SQL-based queries. After downloading data, Microsoft SQL Server Management and Microsoft Excel were used for data preparation.

4. ATTRIBUTES AND DATA PREPARATION

4.1. Generating the PCI values to train the model

As it was mentioned above, this paper intends to predict the change in the PCI in the short term. Therefore, the training set must include the PCI values. Despite the variety of available data in the LTPP database, it does not include the PCI value for road sections. The lack of the PCI data in the LTPP database is a serious issue, but not a fatal one; fortunately, the PCI value is calculable from the distresses and the geometry of road sections. The process of calculating the PCI from road distresses is standardized and documented by the ASTM. The first step to generate the PCI is to extract the distress data. Therefore, the distress data was retrieved from the online platform of the LTPP (LTPP InfoPave 2016) using SQL-based queries. In the next step, a Python program was developed to generate the PCI from the distresses according to the ASTM methodology. For this purpose, all deduct value graphs and correction curves were digitized and mathematically represented. After finding the mathematical function of curves, they were implemented using a Python program. The required steps for digitization and formulae extracting will be explained for one of the graphs.

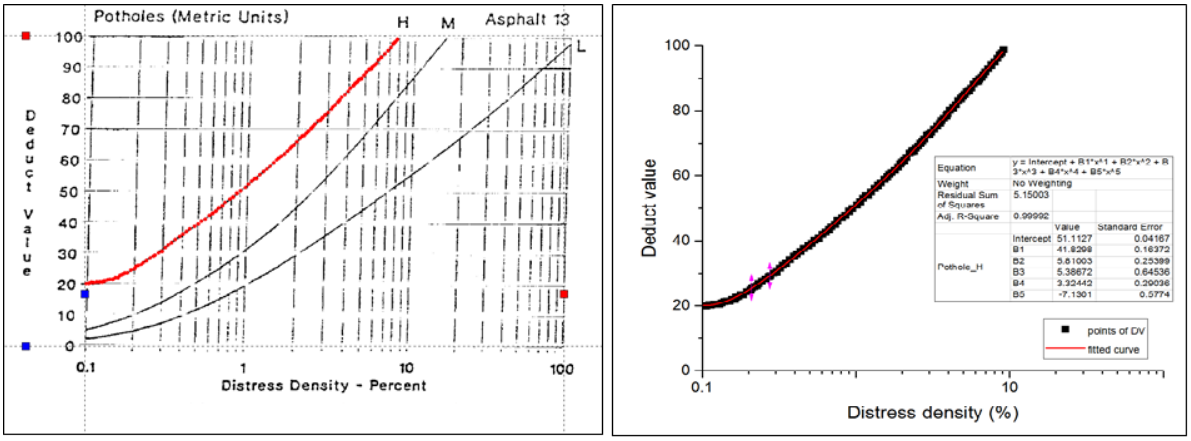

Figure 1. Digitizing the graphs for deduct value calculation (high severity potholes, metric units); graph is adapted from the ASTM D 6433 – 07 (2007)

Figure 1 (left) shows the curves proposed by ASTM D 6433 – 07 for calculating deduct values of potholes with different levels of severity. As shown on the left side of Figure 1, first, we picked 112 points on the curve, which are shown by red points. The points are then drawn on a scatter chart with a logarithmic scale on x axis. A polynomial curve was fitted to the points, and as it is seen on the right side of Figure 1, the coefficient of correlation was very close to 1. Similarly, mathematical functions were extracted for other severity levels and, also, for all other deduct value and correction curves. Altogether, thirty-one deduct value versus distress density curves and eight correction curves were digitized and imbedded in a spreadsheet and a Python script to calculate the PCI automatically. Considering the coefficients of Figure 1 (right), the formula for calculating the deduct value of potholes at a high severity will be as follows:

$$y = -7.130\text{x}^5 + 3.324\text{x}^4 + 5.387\,\text{x}^3 + 5.81x^2 + 41.830x + 51.113; \quad R^2 = 0.9999$$

where y is the deduct value, x is the logarithm of distress density and R^2 is the coefficient of correlation.

4.2. Attributes

Predictive attributes were chosen after a literature review, interviewing domain experts and checking the collected data by 10 small municipalities in Ontario. After delving into the literature and eliciting visions from the experts, a number of attributes were chosen as candidate attributes. In the next step, a final list of attributes was prepared by analyzing the candidate attributes. The analysis included visualization and finding the possible correlations using Microsoft Excel and Rapidminer. The latter is an open source tool for data analytics (Rapidminer 2016).

After checking the stored data in databases owned by 10 small Ontario municipalities and interviewing experts, we understood that the PCI has a key role in road condition assessment among Ontario municipalities. Hence, it should be bold in the model. Furthermore, we learned that most small Ontario municipalities do not have sufficient funding for data collection. Therefore, the cost of data collection for the selected attributes should be low enough to motivate them to embrace the model. Therefore, these attributes were not chosen based on mere engineering reasoning, and the cost of attributes was one of the main considerations. Table 1 shows the final attributes chosen to learn models.

Table 1. The attributes used to learn models and their description (the names are the same names used by the LTPP)

Field name	Description
PCI0	The initial value of the PCI or the value in the current year
AGE	Age of road
PAVEMENT_TYPE	Type of pavement
FREEZE_INDEX_YR	Calculated freeze index for year.
MAX_ANN_TEMP_AVG	Average of daily maximum air temperatures for year.
MIN_ANN_TEMP_AVG	Average of daily minimum air temperatures for year
TOTAL_ANN_PRECIP	Total precipitation for year

FUNC_CLASS	Functional class of road
PCI (target variable)	PCI represents the pavement condition index in after three years (as categorized by the ASTM)

Due to several reasons, a classification model is the most suitable type of model in this case. First, from a practical point of view, some municipalities present the results of their condition assessment as ordinal variables (i.e. good, fair, poor) rather than numeric variables. Also, the ASTM has a recommended discretization for the PCI numeric values. Second, some of the predictive variables such as pavement type or functional class are not continuous; hence their implementation in a decision tree is more convenient.

On the other hand, since classification models cannot deal with continuous values of the target variable, it must be discretized. As it was mentioned in Table 1, the future PCI, which in this paper is the PCI after three years, was chosen as the target variable. The PCI was discretized according to the ASTM PCI rating scale (Figure 2).

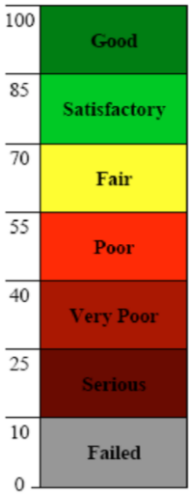

Figure 2. The classes of PCI according to ASTM D 6433 – 07 (2007); the same classes were used in this paper

5. LEARNING MODELS

A couple of models were learned using the 705 examples of road sections in the training set, and one of them was chosen. The objective of learning these models is to enable the municipalities to predict the PCI of their road sections after three years. The philosophy behind three years is that most municipalities in Ontario conduct a comprehensive survey on their road network every five years. Therefore, we wanted to make the missing interim information available to municipalities. A conceptual representation of the implemented model is shown in Figure 3. As shown in Figure 3, the user inputs the values of 8 attributes, which were discussed above, and gets the range of PCI after three years.

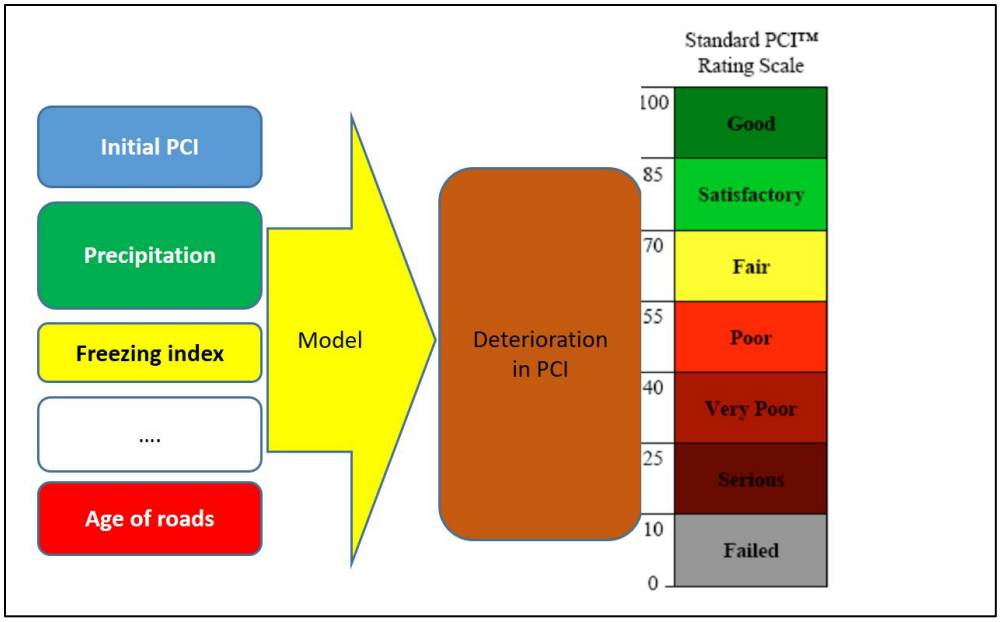

Figure 3. A conceptual representation of the implemented model

5.1. Learned decision trees

Most analyses were conducted using Microsoft Excel and Rapidminer. In addition to machine learning and data mining algorithms, Rapidminer has a good capability in data visualization and presenting the correlation. Visualization of data and investigating all correlations is highly useful in the initial stages of model learning. Before selecting the aforementioned attributes and developing the models, data was imported to Rapidminer to investigate the potential correlations among different fields.

A couple of decision trees were learned from the 705 examples of road sections: the default decision tree of Rapidminer and a C4.5 decision tree. Both models were tested multiple times with similar number of examples and similar parameters; the C4.5 decision tree showed a higher accuracy. For instance, Figure 4 compares the accuracy of two decision trees both learned from 400 examples, but with 6 different sets of parameters. It demonstrates that decision tree II (i.e. C4.5) is consistently outperforming its rival in all six cases. As a result of similar observations, the C4.5 was chosen over decision tree I.

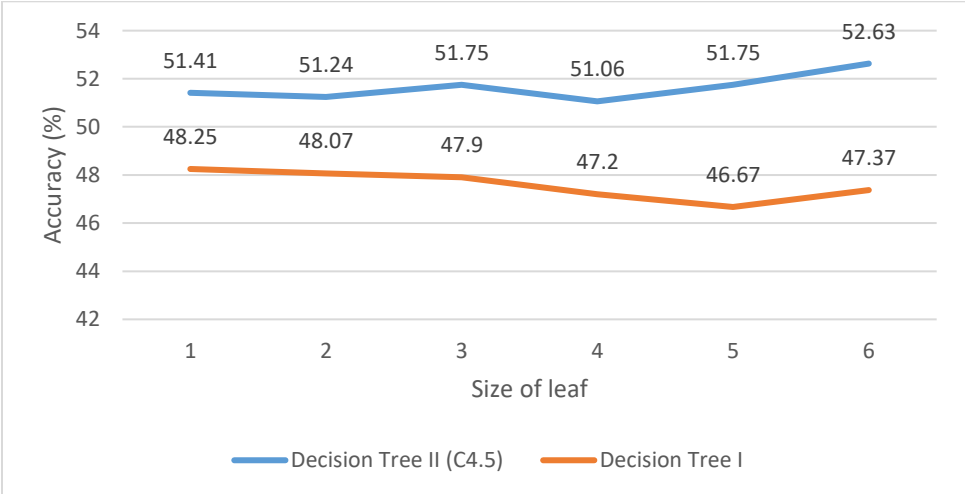

Figure 4. Accuracy of two different decision trees trained with 400 examples but with different leaf sizes

Figure 5 shows the C4.5 decision tree learned from the prepared examples. C4.5 is one of the most popular machine learning algorithms around the world (Wu et al. 2008). Rapidminer can easily handle this algorithm through an extension named Weka (W-J48). In addition to its intelligibility, C4.5 is a conservative algorithm, which enables it to simply avoid major pitfalls such as overfitting. Since the tree of Figure 5 is not easily readable given its large size, the right side of the tree is recreated in Figure 6 for more convenience.

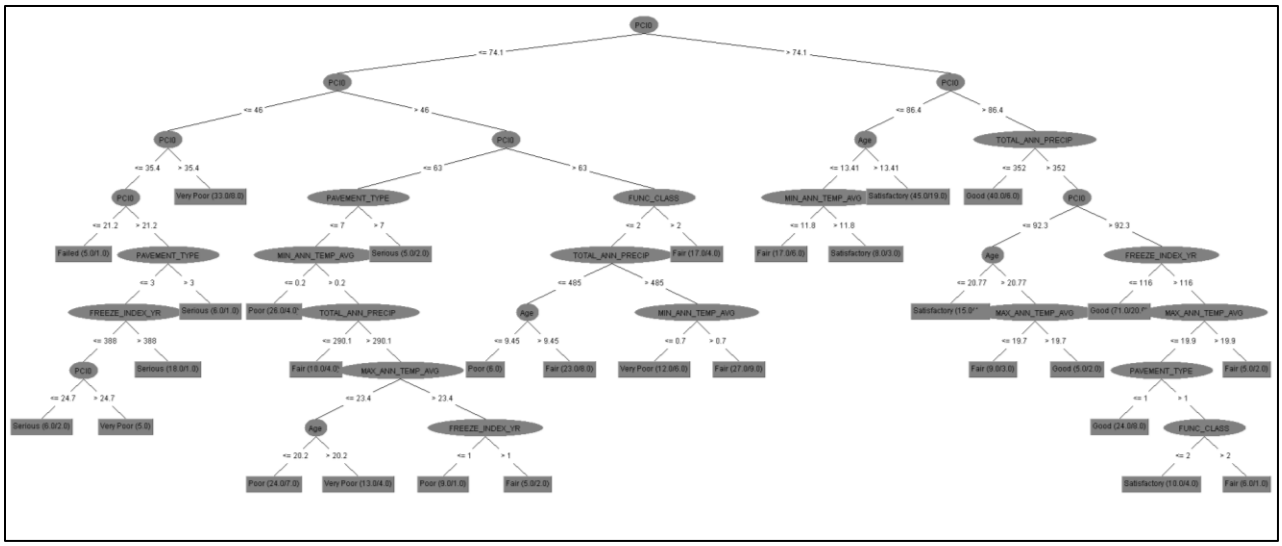

Figure 5. A C4.5 decision tree to predict the PCI after three years

The hierarchy of the attributes within a decision tree represents the informativeness of predictive variables. Therefore, according to Figures 5 and 6, the most informative attributes of the model are respectively as follows. The current value of the PCI (PCI0), the age of road, the total annual precipitation, the annual average of minimum temperature, the annual freeze index, the pavement type, the functional class of road and the annual average of maximum temperature. Reading decision trees manually is not always easy, especially when the depth of tree increases. For further convenience of the users, the decision tree of Figure 5 was implemented using MATLAB for practical uses. Therefore, users can input the characteristics of their roads and climatic data, and receive information about the deterioration of their roads after three years.

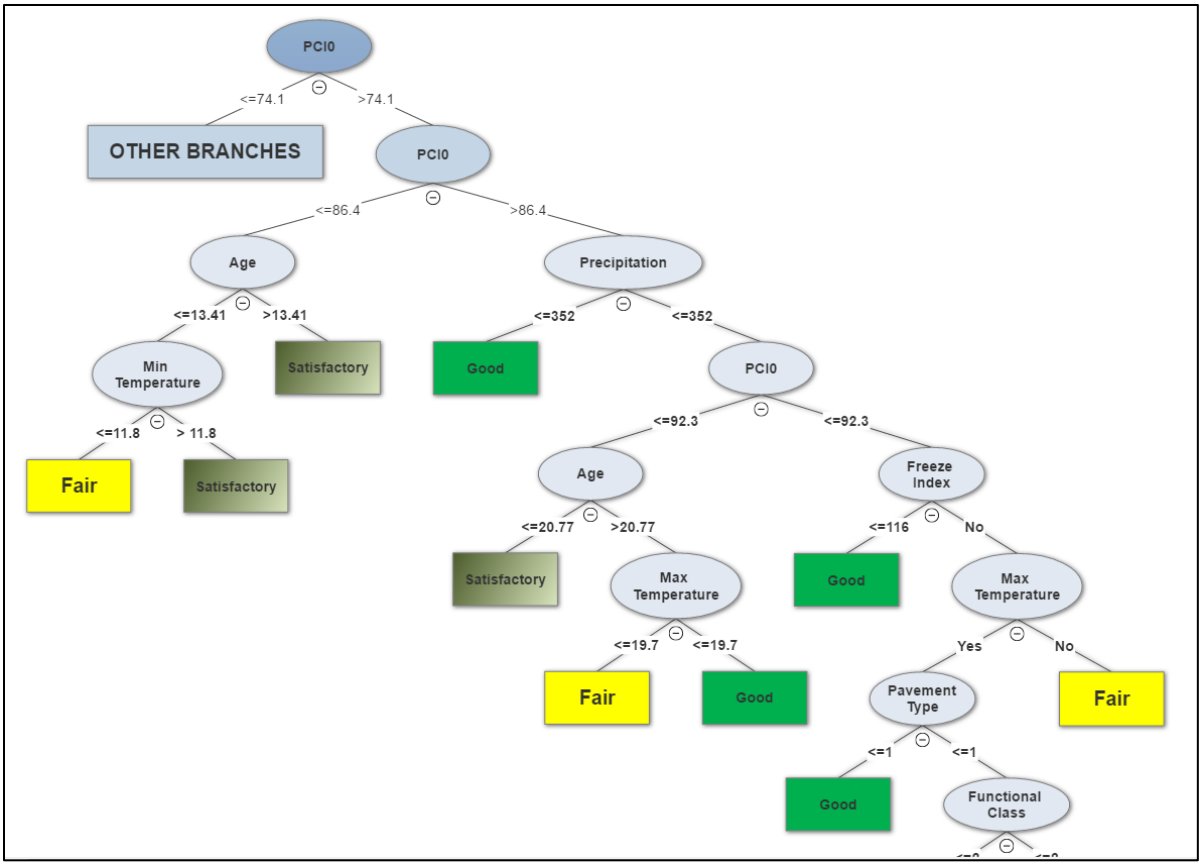

Figure 6. A snapshot of parts of decision tree of Figure 5; the tree can successfully differentiate different scenarios

5.2. The accuracy of the model

The accuracy of developed models was tested using a cross validation operator in Rapidminer. The steps for doing the cross validation was as follows. The training data was divided to 10 subsets. The model was trained based on 9 of them, and its performance was tested using the last one. This process was iterated 10 times, and every time the model was trained based on 9 subsets and tested on the remaining subset. Eventually, the accuracy of the model will be the average of these 10 iterations.

The overall accuracy of the learned decision tree trained by 705 examples was 58.2% ± 4%. It means that, on average, %58 of predictions done by this model are correct. This number is calculable by dividing the sum of diagonal elements of the confusion matrix to the number of all predictions (705). Note that in this case the odds of making a correct prediction by wild guessing is 1/7 = 14.3%, because the target variable has seven classes. The ±4% value represents one standard deviation of the accuracy.

5.3. The parameters of model

Changing the parameters of the model results in different sizes of decision tree and different accuracies. Generally, finding the parameters of a decision tree to maximize its accuracy is not trivial; some recommendations about the parameters of a decision tree are available in the literature (Lin and Chen 2012). A reasonable depth of tree that can differentiate between different scenarios is necessary; this depth should maximize the accuracy without overfitting. The question is which depth of the tree is the most suitable for this problem? This question was answered by trying different ranges of parameters. Therefore, different values of parameters were tested to find a tree with the highest accuracy that has a reasonable depth, which is endorsed by the domain experts. After testing the model with 33 pairs of parameters, the decision tree of Figure 5 with a minimal case of 0.25 and a pruning confidence of 5 was chosen. As it was mentioned above, it has an accuracy of 58.2% ± 4%.

6. CONCLUSION AND FUTURE RESEARCH

In this paper, a decision tree was learned based on eight attributes to predict the PCI of asphalt roads after three years. The attributes used to develop the decision tree were the most affordable and achievable attributes in Ontario. Small municipalities and DOTs in the early stages of asset management, with limited budgets and no historical data, can easily benefit from this model. The LTPP database was used for training the model, and Rapidminer was used for data analytics. A C4.5 was chosen to learn the tree; it was preferred over the default decision tree of Rapidminer due to a better performance. The decision tree could successfully predict the future PCI using several predictive attributes. We discovered that the order of informativeness of attributes is as follows. The current value of the PCI (PCI0), the age of road, the total annual precipitation, the annual average of minimum temperature, the annual freeze index, the pavement type, the functional class of road and the annual average of maximum temperature. The developed model was implemented using MATLAB, so that users can input data through an interactive interface and get the value of the PCI of their assets in three years.

The learned model was tested using a cross validation operator. The accuracy of the model was around 60%, which is satisfactory for these attributes. In the course of the future research, we will look into the possibility of adding more informative attributes to the model. These attributes include but are not limited to the traffic, the number of maintenances, the number of freeze/thaw cycles and pavement layers. Furthermore, incorporating the historical data in the attributes might increase the accuracy of the model as well. For instance, instead of inputting the traffic or precipitation of one year, the model could get these a measure of attributes over a specific number of years; however, the problem with such approaches is that they require a larger amount of data. Therefore, in the next phases of this research we will look into the cost of data collection for each of these fields, and do a cost-accuracy analysis for each model.

7. REFERENCES

ASTM, D 6433–07. 2007. “Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys.” American Society for Testing and Materials-Edition.

Australian Government: The Department of Infrastructure and Regional Development 2013. “Local Government Infrastructure,” Available at <http://regional.gov.au/local/publications/reports/2003_2004/C4.aspx> (Feb. 06, 2017).

Canadian Infrastructure Report Card. Canadian Infrastructure Report Card: Informing the Future. 2016. Available at: <http://www.canadainfrastructure.ca/downloads/Canadian_Infrastructure_Report_2016.pdf>Accessed Jan. 16, 2016.

Chong, G. J., Phang, W. A., & Wrong, G. A. 1982. Manual for Condition Rating of Flexible Pavements-Distress Manifestations (No. Monograph).

El-Diraby, T. E., Kinawy, S., and Piryonesi, S. M. 2017. A Comprehensive Review of Approaches Used by Ontario Municipalities to Develop Road Asset Management Plans. *96th Annual Meeting of Transportation Research Board*, Transportation Research Board, Washington DC. No. 17-00281.

Ens, A. 2012. Development of a flexible framework for deterioration modelling in infrastructure asset management (MSc dissertation), University of Toronto.

Federal Highway Administration 2016. “Data Collection,” Available at: <http://www.fhwa.dot.gov/research/tfhrc/programs/infrastructure/pavements/ltpp/data.cfm> (Jul. 22, 2016).

Kleiner, Y. 2001. Scheduling inspection and renewal of large infrastructure assets. *Journal of infrastructure systems*, 7(4): 136-143.

Lagzi, S., Fukasawa, R., & Ricardez-Sandoval, L. 2017. A multitasking continuous time formulation for short-term scheduling of operations in multipurpose plants. *Computers & Chemical Engineering*, 97, 135-146.

Lin, S. W., and Chen, S. C. 2012. Parameter determination and feature selection for C4. 5 algorithm using scatter search approach. *Soft Computing*, 16(1): 63-75.

Lou, Z., Gunaratne, M., Lu, J.J. & Dietrich, B. 2001. Application of Neural Network Model to Forecast Short-Term Pavement Crack Condition: Florida Case Study. *Journal of Infrastructure Systems*, 7(4): 166.

LTPP InfoPave 2016. "LTPP InfoPave," Available at: <https://infopave.fhwa.dot.gov/> (Jan. 24, 2016).

Mirza, S. 2007. Danger Ahead: The Coming Collapse of Canada's Municipal Infrastructure. Federation of Canadian Municipalities, Ottawa.

Moradi, M., Delavar, M. R., & Moshiri, B. 2015. A GIS-based multi-criteria decision-making approach for seismic vulnerability assessment using quantifier-guided OWA operator: a case study of Tehran, Iran. Annals of GIS, 21(3), 209-222.

Moghaddam, M. H. Y., Moghaddassian, M., & Leon-Garcia, A. 2016. Autonomous Two-Tier Cloud Based Demand Side Management Approach with Microgrid. *IEEE Transactions on Industrial Informatics*.

Najibi, N., Devineni, N., & Lu, M. 2017. Hydroclimate drivers and atmospheric teleconnections of long duration floods: An application to large reservoirs in the Missouri River Basin. *Advances in Water Resources*, 100, 153-167.

Neves, L. C. and Frangopol, D. M. 2005. Condition, safety and cost profiles for deteriorating structures with emphasis on bridges. *Reliability Engineering and System Safety*, 89(2): 185-198.

Piryonesi, S. M., & Tavakolan, M. 2017. A mathematical programming model for solving cost-safety optimization (CSO) problems in the maintenance of structures. *KSCE Journal of Civil Engineering*, 1-10: DOI: 10.1007/s12205-017-0531-z.

Provost, F., & Fawcett, T. 2013. Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.".

Rapidminer 2016. <https://rapidminer.com/> (Dec. 5, 2016).

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

Woldesenbet, A., Jeong, H. D., & Park, H. 2015. Framework for integrating and assessing highway infrastructure data. *Journal of Management in Engineering*, 32(1), 04015028.