Modelling the Influence of Weather Conditions on Traffic Accidents Using Machine Learning Algorithms

A. Mehairi[a] and L. Queiroz[a]

[a] Department of Chemical and Petroleum Engineering at the University of Calgary, Calgary, AB T2N 1N4 Canada

***Abstract*— Weather conditions, among several other factors, can affect the rate of road traffic collisions leading to significant financial and personal losses. Understanding the effect of various weather aspects on road safety enables effective management of such unfortunate events. This study utilizes unsupervised and supervised machine-learning algorithms to model the effect of weather variables on traffic accidents in Calgary, Canada. The results showed that classification algorithms perform poorly compared to regression models. Snowfall was found to have the most impact on accident rates followed by temperature and rainfall. The speed of maximum wind gust showed the least effect.**

I. Introduction

Road traffic accidents can have adverse effects not only on an individual level, but also at a societal level. Accident consequences range from increased travel times and emissions to severe injuries and casualties. Although the vast majority of accidents are attributed to human factors, mainly driver error, environmental factors could also have a major contribution [1]. Weather is one of the most obvious aspects of the environment that could affect road safety. According to Alberta Traffic Collision Statistics report, around 14% of fatal collisions and 16% of non-fatal injury collisions occur when slush, snow or ice was on the road [2]. Furthermore, weather conditions can also slow down emergency services and cost first responders valuable time, which can be critical when life–threatening injuries present. Thus, proactive planning for severe weather conditions is imperative to help reduce casualties and property damage resulting from traffic accidents.

Several recent studies have been conducted to investigate and model the influence of various weather conditions on road traffic accidents. In a review that covered studies between 1967 and 2005, Qiu and Nixon [3] concluded that snow can increase accident rate by 84% and injury rate by 75% while rain can increase accident rate by 71% and injury rate by 49%. They also indicated that wet, icy and slushy roads have an even more significant negative impact on road safety. Furthermore, Pennelly *et al.* [4] stated snow and rain, with snow showing the most impact, produced a higher risk of collision than strong winds and low visibility in the city of Edmonton in Canada. In the same study, it was also concluded that the combined effect of precipitation and another weather aspect (such as strong winds) resulted in higher accident risk. In terms of the effect of temperature, Saida *et al.* [5] reported that the rate of accident occurrence is greatest when road surface temperature is 0ºC. On the other hand, another study by Lobo *et al.* [6] suggested that temperature merely reflects other seasonal weather aspects and can hardly be the root cause of road crashes even though an increase of crash frequency at lower temperatures (less than 10ºC) and a decrease of crash frequency at higher temperature were observed in the same study. Contrary to the aforementioned studies, some studies concluded that wet or snowy weather may deter drivers from venturing onto the road and, thus, reducing the number of accidents due to the decreased traffic volume [7]. As a result, it appears that the influence of weather conditions on traffic accidents is highly dependent on driver behavior. The impact of weather on traffic safety could be minimum in the case where drivers either abstain from driving or drive cautiously in the presence of adverse weather conditions. Several modeling techniques were employed in these studies. For example, Brijs *et al.* [8] used Integer Autoregressive model (a discrete time series model) to model daily crash counts for two cities in the Netherlands using six weather conditions. Autocorrelation was noticed in the data and the use of appropriate statistical models is crucial to detect the effect of weather conditions. Poisson distribution along with regression were utilized to model hourly weather and accident data over the whole Netherlands and were able to conclude that the duration of the precipitation significantly impacts accident rates [9]. Keay and Simmonds [7] also used regression modelling but to determine traffic volume under various weather conditions in Melbourne, Australia and its metropolitan area. In addition, linear regression was employed by Lee *et al.* [10] to correlate traffic injuries to temperature, snow, fog, pressure, day of the week, holiday and date. The study concluded rain, snow and temperature were the most important factors.

The aim of this study is to investigate the use of machine learning algorithms to model the influence of adverse weather conditions on road traffic accidents for the city of Calgary, Canada. Unsupervised and supervised machine learning algorithms are utilized to estimate the relationship between daily data of traffic accidents and five weather conditions, namely temperature, snowfall, rainfall, maximum wind gust and accumulated snow in the ground. Unsupervised learning using K-means, Mean Shift and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithms are used to pre-screen the data whereas supervised learning using nearest neighbors and decision tree analysis is used to classify the data. Linear regression is also employed to fit the data to linear models. Since this type of analysis has not been found in the literature for the city of Calgary, it is believed that such analysis is crucial in helping police and emergency services proactively plan for unit deployment to reduce emergency response times.

In the next section, we describe the data available for the study followed by a detailed description of the methodology. After that, we present the results and discuss them. Finally, we give our conclusions.

II. METHODOLOGY

A. Data description and treatment

This study is based on the data for the city of Calgary, Canada between December 6, 2016 to December 12, 2019 excluding the periods December 21, 2016 – February 7, 2017 and June 14, 2019 - September 15, 2019 during which the traffic accident data was not available. Calgary is located in the western province of Alberta about 80km east of the front ranges of the Canadian Rockies at 51°03'N and 114°04'W. It has a population of 1,285,711 as of September 2019. Calgary is well-known for frequent weather fluctuations as it has recoded snow every month of the year which stresses the importance of this study. The total number of samples is 948 with 5 attributes each.

Historical weather conditions data were obtained from the Government of Canada's website based on Calgary's International Airport weather station. The hourly data does not include precipitation data (rain or snow), so daily data that include both snow and rain fall were obtained instead. For each of the dates in the interval specified previously, mean temperature (°C), total snow (cm), total rain (mm), accumulated snow on the ground (cm) and speed of maximum gust wind (km/h). The choice of the attributes was based on the prior research conducted. Even though the literature shows that wind is not usually a significant factor, we examined the impact of wind in this study to determine if the combination of wind with other factors might have a significant effect. Table I shows the distribution of the weather data over the different seasons. It should be noted here that spring and summer is defined as the period between May 1st and August 31st while fall and winter is between September 1st and April 30th. Out of the 948 samples, only 172 samples contained snowfall and 168 samples contained rainfall. Thus, as evident from Table I, it is clear that the data for snow and rain are significantly skewed towards no snow and no rain fall which affected the modeling results as will be explained in the results and discussion section.

In addition, traffic incident data were obtained from the City of Calgary's website. It includes reported traffic incidents per the time reported. Although using hourly data might give accurate results when studying the effect of weather, the required weather data are only available in a daily basis. Thus, the traffic data were summed for each date in the interval under study. Also, this data contained non-crash data such as flashing traffic lights, road closures, train incidents and stalled vehicles. Thus, non-relevant incidents needed to be removed first. Furthermore, from incident description, we were able to extract useful data including number of single-vehicle, two-vehicle and multi-vehicle accidents as well as number of accidents occurred during morning rush hours (6:30 -9:30 AM) and evening rush hours (3:30-6:30 PM). Finally, number of accidents involving pedestrians and cyclists was also extracted for each day. Accordingly, the effect of weather conditions on each of these accident types can be modeled. Table II shows the distribution of the traffic accidents data. It is obvious that as seasons change from spring and summer to fall and winter, all types of accidents increase significantly which indicates some relation to weather conditions.

TABLE I. WEATHER DATA CHARACHTERISTICS FOR THIS STUDY

Weather Attributes	Spring and Summer			Fall and Winter		
	Min	***Mean***	***Max***	***Min***	***Mean***	***Max***
Temp., °C	-1.4	15.3	26.2	-28.2	-1.4	20.3
Snow, cm	0	0.03	5.7	0	0.7	32.8
Rain, mm	0	1.4	41.7	0	0.2	12.1
Accum. Snow, cm	0	0.02	5	0	3.8	31
Wind, km/h	10	45	96	0	39	91

TABLE II. TRAFFIC DATA CHARACHTERISTICS FOR THIS STUDY

Accident Type	Spring and Summer				Fall and Winter			
	Total	***Min***	***Mean***	***Max***	***Total***	***Min***	***Mean***	***Max***
Single-Vehicle	585	0	2	7	2152	0	3	29
Two-Vehicle	2639	2	9	2	6759	0	10	35
Multi-Vehicle	577	0	2	8	1754	0	3	18
Pedestrian/ Cyclist	149	0	1	3	441	0	1	6
Morning Rush Hour	296	0	1	7	1521	0	2	37
Afternoon Rush Hour	738	0	3	15	2403	0	4	27
All	3921	3	14	27	11060	1	17	57

B. Unsupervised learning

Unsupervised learning algorithms are used to identify patterns in unlabeled data, *i.e.* the output classes or labels are not known [11]. The lack of preset labels for the output in unsupervised learning can be advantageous as the algorithm might identify patterns that have not been previously considered [12]. This feature prompts us to use unsupervised learning first to eliminate any weather features that resulted in poor clustering. Three different algorithms were utilized to conduct this analysis; K-means, mean shift and DBSCAN. K-means is a clustering algorithm that uses the sum-of-squares criterion to separate the samples in k clusters with similar properties [13]. The algorithm chooses centroids that minimize the inertia, *i.e.* the some of the squared difference between any sample and the mean of the corresponding cluster. Even though, K-means converges fast, it requires the predetermination of the number of clusters and yields non-repeatable results. On the other hand, mean shift algorithm is another clustering algorithm that seeks high-density data points' areas by shifting each data point to the average of the data points in its neighborhood [14]. Although mean shift does not require the number of clusters and guarantees convergence, it is not a good fit for large datasets. DBSCAN is an improvement on the mean shift algorithm as it is highly scalable. DBSCAN has the ability to identify and ignore outliers[15].

K-means, mean shift and DBSCAN were used to perform clustering on the total number of accidents using temperature, snowfall, rainfall and speed of maximum wind gusts. In order to test the performance of these algorithms, the total number of accidents where classified into 3 classes; low (<=15), medium (>15 and <20) and high (>20). Accordingly, performance parameters including homogeneity, completeness and Fowlkes-Mallows score (FMI) were evaluated for each algorithm and combination of attributes. Cross validation technique was also employed to avoid over-fitting using 80% of the data for training and 20% for testing.

C. *Supervised learning*

Unlike unsupervised learning, supervised learning aims to train the algorithm with data samples with the corresponding classification already assigned [11]. Supervised learning algorithms can be classification or regression algorithms. In this study, K-Nearest neighbors (KN neighbors) and decision tree algorithms were used as classification algorithms whereas linear regression was used as a regression algorithm. KN neighbors predicts and assigns a classification to an unclassified sample point based on a predefined number of training samples closest in distance to the unclassified point [16]. In terms of the decision tree classification, each node represents an attribute, each link represents a decision and each leaf represents an outcome. Both KN neighbors and decision tree can also be used to regression, but they were only used for classification in this study. Several models for linear regression were evaluated using one or a combination of features.

KN neighbors and decision tree analysis was used to perform classification on total number of accidents using temperature, rainfall and snowfall. The accuracy of each model was computed to evaluate performance. Similar classes to the unsupervised learning were used in this analysis as well. Finally, linear regression was performed fitting 22 models on the total number of accidents using various combinations of attributes (temperature, snowfall, rain and wind). The least squared error was used as a measure to evaluate performance. In both classification and regression, cross validation was also utilized using 50% of the data for training and 50% for testing.

III. RESULTS AND DISCUSSION

A. *Unsupervised learning results*

As stated previously, the dataset was tested using unsupervised learning algorithms (K-means, mean shift and DBSCAN) to evaluate the influence of each attribute in terms of data segregation to identify patterns. Four different combinations of features were attempted and the results are presented in Table III. It can be seen that unsupervised learning has failed to identify any trends in the data as all the performance indicators are closer to zero than 1. The reason is associated with the lack of sufficient rain and snowfall data within the dataset as demonstrated using the histogram distribution in Fig.1.

TABLE III. HOMOGENIETY, COMPLETENESS AND FMI SCORES FOR UNSUPERVISED LEARNIG MODELS

Indicators	K-Means	Mean Shift	DBSCAN
Mean Temperature vs Total Snow			
Homogeneity	0.0199	0.0673	0.0375
Completeness	0.0284	0.0372	0.1789
FMI	0.4797	0.3284	0.6429
Mean Temperature vs Speed of Gust			
Homogeneity	0.0155	0.1255	0.1388
Completeness	0.0232	0.0358	0.0399
FMI	0.4655	0.1385	0.1442
Mean Temperature vs Total Rain			
Homogeneity	0.0188	0.0482	0.005
Completeness	0.0265	0.025	0.0483
FMI	0.4834	0.2935	0.6433
Total Snow vs Speed of Gust			
Homogeneity	0.0099	0.0477	0.0041
Completeness	0.0152	0.0216	0.0163
FMI	0.476	0.2374	0.6213
Mean Temperature vs Total Snow vs Speed of Gust			
Homogeneity	0.0099	0.0477	0.0041
Completeness	0.0152	0.0216	0.0163
FMI	0.476	0.2374	0.6213

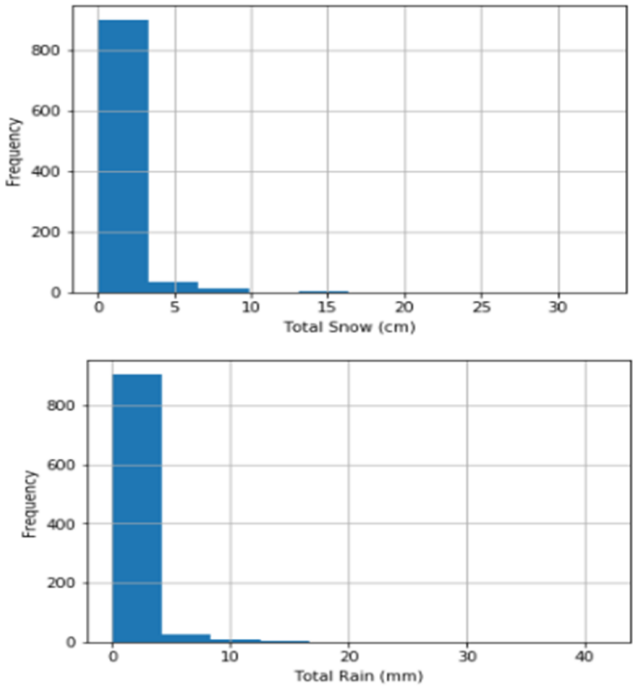

Figure 1. Total Snow (cm) and Total Rain (mm) histogram distribution

B. *Supervised learning results*

1) *KN nearest neighbors classification*

KN nearest neighbors classification was used to model the influence of ten different combinations of four weather attributes (temperature, snow, rain and speed of maximum wind gust) on the total number of accidents. The accuracy for each model was evaluated for 200 values of k (number of

neighbors) using both uniform and distance weights. As can be seen from Table IV, all ten clustering attempts resulted in low accuracies (50-60%) regardless of the number of neighbors chosen.

TABLE IV. KN NEAREST NEIGHBORS CLASSIFICATION RESULTS

Clustering with respect to	Max accuracy (%)		k for max accuracy	
	Uniform	*Distance*	*Uniform*	*Distance*
All attributes	58%	54%	20	50
Temperature	58%	50%	20	50
Rain	54%	53%	2	1
Snow	56%	56%	2	1
Wind	55%	49%	60	29
Temp. & Snow	58%	50%	13	75
Temp. & Rain	58%	51%	20	25
Temp. & Wind	58%	50%	25	25
Snow & Wind	58%	55%	20	20
Rain & Wind	55%	50%	50	3

Although rain and snow performed the best in terms of accuracy obtained using fewest number of neighbors, Fig.2 shows that no matter the number of neighbors the accuracy will stay the same for snow and rain only scenarios while it changes for other combinations. The reason for this is the small number of data points associated with rain and snow as explained in the unsupervised learning section. Since the nearest neighbors algorithm searches for the closest data points, the chance of the neighboring data point belongs to snow or rain is very small.

2) Classification tree results

Random forest was used to classify and predict the number of incidents based on the aforementioned weather parameters. The algorithm creates several decisions trees to fit a classification model. Based on the nature of the dataset for this study, random forest presented a unique opportunity to find a proper fit.

After modelling 1000 different decision trees, the accuracy of the algorithm was 43%, which means that the selected attributes do not have enough prediction strength. This emphasizes the importance of sufficient snow and rain data for classification algorithms to obtain meaningful results.

3) Linear regression

On top of the training and testing sets least squared errors, the global least squared errors (for the whole dataset) is also calculated for all 22 models and reported in Table V. The reason behind the constant term in all models is to capture the influence of non-weather factors on the total number of accidents.

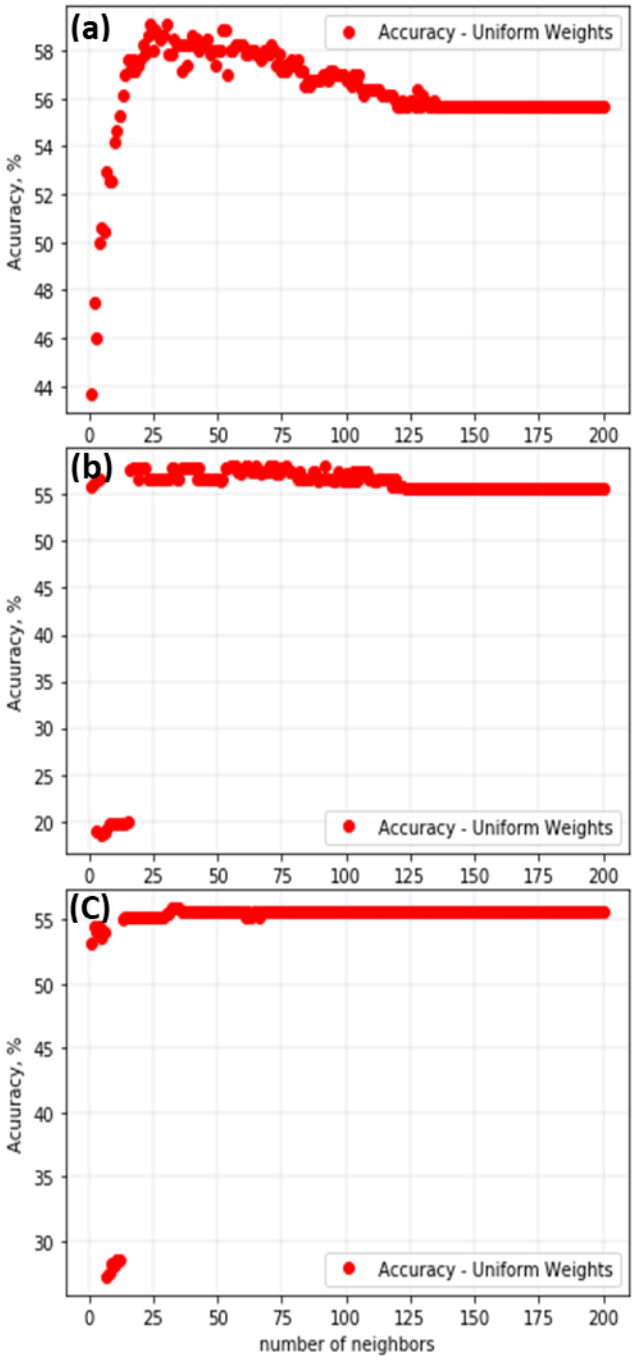

Figure 2. Accuracy vs number of neighbors for: a) All features, b) snow only and c) rain only

Analyzing all 22 models enabled the determination of the most important weather factors influencing the total number of accidents in Calgary, Canada. It is clear that the two most important factors are temperature and snowfall. The speed of maximum wind gust has the least effect on the output similar to the reported conclusion by Pennelly *et al.*[4]. As can be seen from Table V, the best models that can describe the relation between weather attributes and the total number of accidents are 20 and 22. The difference between the two models is the inclusion of the rain factor. Although, there was no change in terms of the least squared error, which can be as a result of the scarce rain data within the dataset, the

inclusion of the rain factor might increase accuracy in periods of heavy rain as can be seen from Fig3. In Fig3., the model (22) is able to predict closely the number of accidents when rain fall is more than 10mm per day. Accordingly, model 22 was chosen as the best model for this study. A similar model was presented by Lee *et al.* [10] including other weather factors in addition to temperature, snowfall and rainfall. . The coefficients a, b, c and d for model 22 are -0.00988, 0.04885,0.00883 and 2.63959, respectively.

Fig.3 also shows that heavy rain, especially more than 20 mm, lead to an increase in the number of accidents. However, its effect is not as significant as temperature (Fig.4) and snowfall (Fig.5), which is contrary to the reported data by Qiu and Nixon [3]. The reason could be the nature of the data used in this study where only 18% of the data contains rainfall and only 1.6% contains rainfall more than 10mm.

TABLE V. LINEAR REGRESSION MODELS FITTED IN THIS STUDY

Model	Formula T=Temperature (ºC), S=Snow (cm), R=Rain (mm), W=wind (km/h), Y=total number of accidents	Least Squared Error
1	Y= aT+b	50,840
2	Y=aS+b	52,881
3	Y=aR+b	61,230
4	Y=aW+b	60,112
5	$Y=aT^2+b$	60,028
6	Y=aT+bS+c	49,418
7	Y=aT+bR+c	54,873
8	Y=aT+bW+c	54,913
9	Y=aS+bW+c	52,285
10	Y=aR+bW+c	60,126
11	Y=aT+bS+cW+d	49,488
12	Y=aT+bS+cR+d	49,453
13	Y=aT+bW+cR+d	54,968
14	Y=aT+bS+cR+dW+e	49,541
15	Y=aT*bS+c	52,344
16	Y=aT*bR+c	61,294
17	Ln(Y)=aT+bS+cR+dW+e	234
18	Ln(Y)=aT+b	244
19	Ln(Y)=aT+bS+cW+d	234
20	Ln(Y)=aT+bS+c	233
21	Ln(Y)=aT+bR+c	244
22	Ln(Y)=aT+bS+cR+d	233

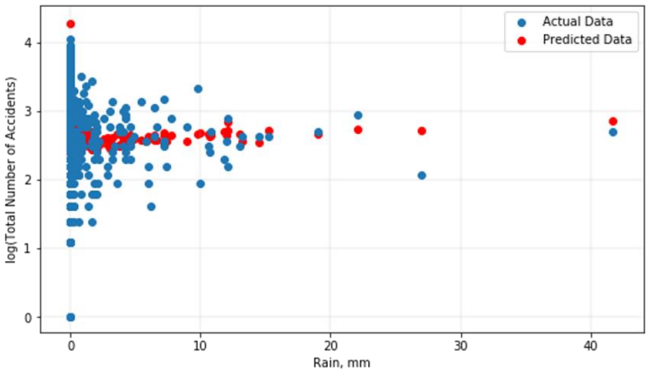

Figure 3. Rain vs ln(total accidents) using model 22

Since the temperature data was better distributed than snow or rain data, the model (22) was able to have good predictions. The model suggests that number of accidents increase as temperature decrease especially below 0ºC, which agrees with some literature reports [5]. However, logically speaking, the temperature physically cannot interfere with driving a vehicle. Thus, the temperature might be a mere reflection of other weather aspects [6], temporal effects and driver behavior.

When analyzing the models presented in Table V, temperature showed the most significant effect on the least squares error followed by snowfall, *i.e.* removing temperature from the model leads to significant increase in the global least squared error. This result contradicts the results shown in Fig.4 and Fig.5 where snowfall has the biggest impact on the change of number of accidents similar to the reported results in literature [3], [4]. Similar to the rainfall data, snowfall data represents only 18% of the total data and only 6 (0.6%) samples contained snowfall more than 10 cm, which could explain why snowfall was not more significant than temperature in terms of the least squared analysis. Nevertheless, the model (22) closely predicted the number of accidents when snowfall is more than 10cm.

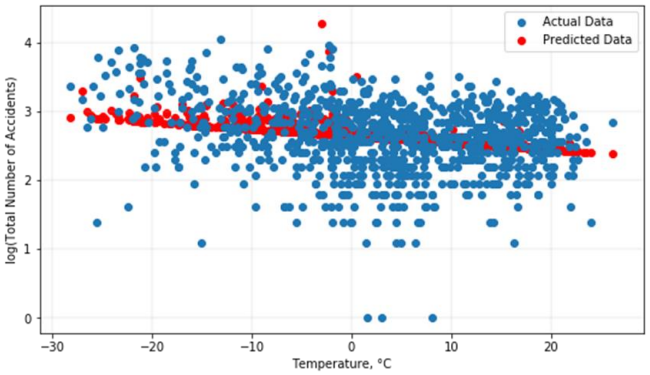

Figure 4. Temperature vs ln(total accidents) using model 22

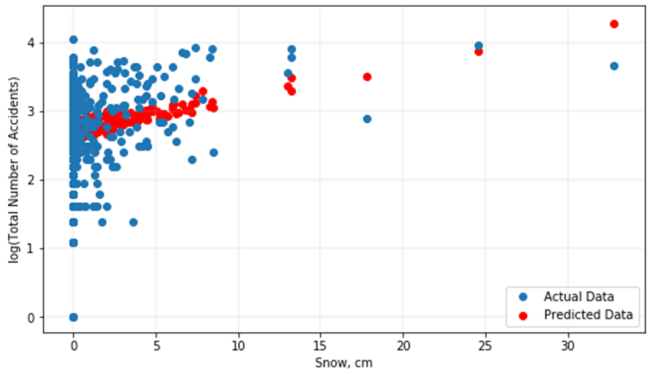

Figure 5. Snow vs ln(total accidents) using model 22

IV. Conclusion

This study presented the implementation of six machine-learning algorithms to model the influence of weather conditions on the number of traffic accidents in Calgary, Canada. It can be concluded that classification and clustering using unsupervised learning algorithms (*i.e.* K-means, mean shift and DBSCAN) as well as supervised learning classification algorithms (*i.e* K nearest neighbors and decision tree) are not suitable for studying daily weather and crash data. On the other hand, modelling using linear regression has shown good predictions of the actual data. The results show that snowfall is the most significant weather aspect followed by temperature. Rainfall showed less significance than snowfall and temperature. However, the speed of maximum gust had the least impact. Accordingly, a model was developed using temperature, snowfall and rainfall to predict the total number of accidents for the city of Calgary and can be used by emergency services to proactively plan resources allocations.

V. Future Directions

This study utilized daily weather data for weather aspects and traffic accidents. However, an hourly analysis of the effects could perhaps be more meaningful since snowfall and rainfall rarely continue non-stop for a full day. Thus, the most impact of weather on traffic accidents can be seen within the first three hours or so from any weather event [1]. After that, drivers behavior adjusts to the new conditions and caution is used as suggested by [7]. Daily analysis dilutes the impact of hourly weather events and thus the impact might not be captured. For example, if heavy snow falls in the morning, the number of accidents might be high during those hours, however, the total number of accidents for the whole day might not be higher than average due to adjustment of driver behavior.

Furthermore, when attempting to use machine learning algorithms, large number of data points is required to capture the impact of snow and rain. Using two or three year data might contain only few samples for rainfall and snowfall as seen in this study, thus, more data is required especially for classification algorithms. Finally, data normalization, which was not done for this study, might also improve the performance of the algorithms.

References

[1] J. Andrey and S. Yaga, “A TEMPORAL ANALYSIS OF RAIN-RELATED CRASH RISK,” *Accid. Anal. Prev.*, vol. 25, no. 4, pp. 465–472, 1993.

[2] O. of T. S. Alberta Transportation, “Alberta Traffic Collision Statistics 2016,” 2016.

[3] L. Qiu and W. A. Nixon, “Effects of Adverse Weather on Traffic Crashes Systematic Review and Meta-Analysis,” *Transp. Res. Rec. J. Transp. Res. Board*, pp. 139–146, 2008.

[4] C. Pennelly, G. W. Reuter, and S. Tjandra, “Effects of Weather on Traffic Collisions in Edmonton , Canada,” *Atmosphere-Ocean*, vol. 56, no. 5, pp. 362–371, 2018.

[5] A. Saida, M. Hirasawa, and N. Takahashi, “QUANTITATIVE EVALUATION OF THE RELATIONSHIP BETWEEN THE ROAD SURFACE CONDITIONS MEASURED BY CONTINUOUS TESTING VEHICLES AND THE RATE OF WINTER TRAFFIC ACCIDENTS,” in *18th International Conference Road Safety on Five Continents*, 2018, pp. 1–5.

[6] A. Lobo, S. Ferreira, I. Iglesias, and A. Couto, “Urban Road Crashes and Weather Conditions: Untangling the E ects,” *Sustainability*, pp. 1–13, 2019.

[7] K. Keay and I. Simmonds, “The association of rainfall and other weather variables with road traffic volume in Melbourne , Australia,” *Accid. Anal. Prev.*, vol. 37, no. 1, pp. 109–124, 2005.

[8] T. Brijs, D. Karlis, and G. Wets, “Studying the Effect of Weather Conditions on Daily Crash Counts Using a Discrete Time Series Model,” *Accid. Anal. Prev.*, vol. 40, no. 3, pp. 1–24, 2008.

[9] E. Hermans, T. Brijs, T. Stiers, and C. Offermans, “The Impact of Weather Conditions on Road Safety Investigated on an Hourly Basis,” *Transp. Res. Board*, no. May 2014, 2006.

[10] W. Lee *et al.*, “Does Temperature Modify the Effects of Rain and Snow Precipitation on Road Traf fi c Injuries ?,” *J. Epidemiol.*, vol. 25, no. 8, pp. 544–552, 2015.

[11] R. Sathya and A. Abraham, “Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification,” *Int. J. Adv. Res. Artif. Intell.*, vol. 2, no. 2, pp. 34–38, 2013.

[12] E. Oja, O. Simula, and J. Kangas, “Engineering applications of the self-organizing map,” in *IEEE*, 1996, no. 84, pp. 1358–1384.

[13] H. Bock, *Clustering Methods : A History of k -Means Algorithms*, no. 1. Springer Berlin Heidelberg, 2007.

[14] Y. Cheng, “Mean Shift , Mode Seeking , and Clustering,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, pp. 790–800, 1995.

[15] D. Birant and A. Kut, “ST-DBSCAN : An algorithm for clustering spatial – temporal data,” *Data Knowl. Eng.*, vol. 60, pp. 208–221, 2007.

[16] T. M. Cover and P. E. Hart, “Nearest neighbor pattern classification,” *IEEE Trans. Inf. theory*, vol. 13, no. 1, pp. 21–27, 1967.