UiT The Arctic University of Norway

Faculty of Science and Technology

Accident Prediction Model Using Machine Learning

ACCURACY OF PREDICTED MODEL

Sohel Mohammed Abdul

Master’s thesis in Technology and Safety in the High North   TEK-3901-1   June 2021

Abstract

Logistic regression is a predictive model machine learning algorithm that displays the results in a binary form, mostly used in prediction multivariable and as an advanced version of linear regression, where we used to predict the accuracy of our model. The current accident prediction model in Norway Tusi is a risk-based model used in predicting tunnel accidents. However, it has some problems and loose ends that could make predictions go wrong with minor deviations like human errors, availability of limited data or no data from similar tunnels, and more dependent on previous statics. Past research and this paper have shown the need to apply ML in accident prediction in road tunnels, and the study aims towards how the machine learning models can help overcome the problems and replace the traditional method.

For implementing the machine learning algorithm, the daily traffic data and accident data is gathered from the Norwegian public road administration for one of the existing tunnels of West Norway known as Eiksund tunnel for examining the correlations in the available variables. Analysis of testing results demonstrated why the machine learning model needs more study and expertise in computer language. Also, the calculation results shown that the narrow gap of study could have been considered.

It is recommended to understand the algorithm better since logistic regression is a powerful tool for predictive analysis, and there is also a need for better data gathering before the model is trained to strengthen the outcomes.

Keywords: logistic regression, tunnel accidents, risk analysis, machine learning

Table of content

Accident rate**:** The accident rate of other tunnels of equal cross section but not in the tunnel of same area. 6

One or two-way traffic**: The** allowable traffic movement inside the tunnel whether the moment of vehicles is allowed in both the direction or only in one direction. This helps in dividing the tunnel into small zones for easy calculations. The horizontal/vertical zones can be divided up to 11 as discussed in **(Norwegian Public Roads and Administration 2004)** and presented in the figure() below. 6

Average Annual Daily Traffic (AADT)**: AADT** is the total traffic volume of any given tunnel and is given by the total annual traffic divided by 365, that gives the total traffic in both directions. AADT helps in determining the category of any tunnel. But it requires 20 years of traffic volume after the tunnel is being opened for the public use this is because of the uncertainties of daily traffic according to the N500 *Norwegian Public Roads Administration Handbooks* **(Norwegian Public Roads and Administration 2004)** and AADT helps in choosing the class of the tunnel, from A to F as shown in Figure 3. 6

Length of the tunnel**: Tunnel** length helps in dividing the zones and also helps in defining the class of a tunnel.7Geometry**: Geometry** of the tunnel here defines the cross section measurements of the tunnel and both horizontal, vertical measurements of a tunnel are important as shown in the Figure 5. 8

Limitations of the Tusi model**: The** main drawback in using the Tusi model is that the accuracy of statistical data developed in the end depends directly to the data that we used from the tunnels similar on to the tunnel under investigation for better results and by any reasons if the consequence data is not available from other tunnels or past data is missing then other methods are to be adapted besides the Tusi model. 9

Machine learning**:** During the past decade, machine learning has been implemented in major working areas for examining the risk analysis example starting from the Industrial field to the medical fields. Different machine learning algorithms have been used in different purposes such as predicting the landslides, structural health monitoring of civil structures, and classifying a diseases, vehicle crashes and deep neural network in the project managements. 9

Tunnel specification and Case study**:** This paper tries to perform the ML to draw the accuracy of accident prediction in tunnels instead of the traditional method (Tusi model). For the study purpose we are using the data from the Eiksund tunnel as the case. The tunnel is considered one of the longest tunnels with a collective length of 7850m and first opened for public transportation use in 2008 and the main purpose of tunnel is to connect Ørsta, Unsteinvik and Volda counties for easy movement of traffic instead of ferry boats. The route is as shown in the Figure 9 below, the tunnel has two openings one has T8.5 profile towards the Steinnestranda to the bottom, and the other has a T11.5 profile towards the Eika side to the bottom of the tunnel as shown in the profile below in Figure 10. 12

Data collection (for accident prediction accuracy using machine learning algorithm): The data used is quantative, It has been gathered from the secondary source which is from the Norwegian road administration Statens vegvesen (Anon n.d.) and the tunnel Eiksund is maintained by More og Romsdal fylkeskommune which is in the north most part of west Norway. 16

Traffic data collection and saving process**:** This section is centered towards the explanation of how the traffic data is collected by Norwegian public road administration and

List of figures

1. Introduction: In the past decade use of vehicles have been increased drastically which led to the number of road incidents, one of the research paper focused on the fatal incidents of people driving towards work and it says at least 1 person out 41% of fatal road incidents involves (Nævestad, Phillips, and Elvebakk 2015). As the traffic increases the need for the safety inspection and accident prediction is needed on open roads, also in the tunnels and bridges.

Considering the methods of evaluation of accident prediction and risk evaluation there should be an advancement in the prediction system also because as the traffic increases the number of inspection also increases to check the minor deviation and improving them time to time.

But as the data increases and the analysis become complicated there are the high chances of not remembering all the variables it could be human error or any other reasons such as an audit conducted on the Norwegian tunnels shows that there are very low number of safety engineers for all 302 the country tunnels, according to the audit conducted on 11/06.2020 (Norwegin public road administration n.d.), Since Norway holding one of the largest number of tunnels in the world also it is a fact on the other hand country trying to achieve the Zero vision (Knapstad, Somme, and Njå 2019) which means to achieve the target of reaching zero major incidents on the roads.

Seeing the need for the risk analysis and accident prediction with low human involvement one of the machine learning algorithm holds the capability of predicting the accidents in less time known as Logistic Regression.

1.1. Background: Tunnels and bridges have increased the scope for using underlying areas by constructing the sub-sea tunnels and long suspension bridges. They made transportation easy on hilly regions and remote islands by tunneling. The significant uses of road tunnels are they reduce the travelling time and distance to reach the destination and small islands are connected to mainland through sub-sea tunnels.

In past few decades the tunnels have also replaced the long ferry travels by reducing the several hours of journey to minutes. All these factors brought the attention of the government and other businessmen in developing the infrastructure.

Considering environmental perspective, the tunnels and bridges also contributing by lowering the air pollution by cutting the travel distance and also saving the fuel. 'MSc. Anne Katrine Kalager' at The British Tunneling Society mentioned *"If we have surface challenges try underground solutions"* while specifying the technical transportation challenges in future tunnels of Norway. Tunnels in Norway are both deep and long.

According to (Amundsen and Ranes 2000) with over 900 tunnels the country holds a large number of tunnels inside the country compared to other places. The report from (Phillips and Meyer n.d.) titled Fatal road accidents involving people at work of 2005 to 2010 "*shows that in Norway at 41% of fatal road accidents and 36% of fatal accidents caused to employees driving work"*.

Road tunnels are often over-presented for the accidents. Comparison also shows that open roads the accident number are more but the intensity of destruction caused by the accidents inside the tunnels is the greatest risk, one can witness it after studying the famous incident of Mont-blanc in 1999 discussed by (Vuilleumier, Weatherill, and Crausaz 2002) the level of damage was huge, both personal and property and also other tunnel accidents.

On a keen observation it shows in general the extent of damage is proportional to the traffic trapped inside the tunnel meaning with the increasing traffic density the range of damage increases. The reason for more traffic in tunnels may varies for example in the busy time of day or the tunnels that constructed in the urban region have heavy inflow, can also be the tunnels connecting the industrial area increases the transportation of goods in heavy trucks one study says that there is a 20 to 40% of accidents in countries that has highly industrialized (Fort et al. 2010), the reasons can be many.

In such cases the accident prediction and traffic monitoring can help in avoiding incidents, the most traditional method of risk evaluation in Norwegian tunnels is the Tusi model which gives statistical data of risk prediction, the model has its own limitations which were discussed in section 2.1.

This paper tries to apply a Machine learning algorithm in predicting accidents in tunnels, the reason for focusing on the new method besides the existing one is due to the rapid improvement in technology because in the past decade machine learning had made its place in the safety field by the impressive outcomes.

Few of the applications of the machine learning are, it is used in predicting the landslides where the areal pictures were used and it is discussed by (Pham, Prakash, and Tien

Bui 2018), another application area is in medical diagnosis as shown in (Kononenko 2001), and speaking of the construction field machine learning algorithm helped in monitoring the structural health (Figueiredo et al. 2011), and one of the deep concepts but more useful deep neural network have been used to solve many problems such as safety on the construction site, modelling of the building occupancy and prediction of energy demand, others were discussed in (Doroshenko 2020).By seeing all the above applications one can easily make a pictorial representation to show the importance and spread of the machine learning applications as shown in Figure 1 below.

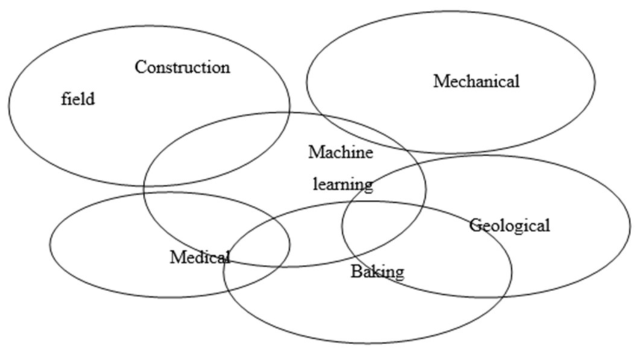


Figure 1 Machine learning in other fields

The basic reason for machine learning to become more diverse/flexible is 'the data availability' since after the introduction if cloud storage almost all the data is available but in the scattered form. The only work that has to be done or needed is to gather them and pre-process/tune them to train the data workable accordingly.

Machine learning has also proved its efficiency in performance and the flexibility of its use in almost every field and this is one of reasons for being more popular.

1.2. Goal/Hypothesis: The aim of this thesis is to test one of the Machine learning algorithms known as Logistic regression by predicting the past accidents that had already happened in the Eiksund tunnel and extracting the accuracy level of prediction in the end, also giving concrete improvements that open the scope of future study for machine learning in tunnels.

In result section 3, the prediction of accidents is carried out and the limitations are discussed in discussion section 4, areas of improvement are mentioned in the conclusion section 5.

1.3. Previous studies: In recent years the application of machine learning in the risk assessment area has been improved and several studies have made in an efficient ways to reduce the risk by predictive analysis. One of the reasons for using machine learning algorithm is because it reduces the human involvement meaning low human errors, due to its ability of learning from the past incidents/accidents (discussed in the section 2.2), like in the injury classification by (Marucci-Wellman, Corns, and Lehto 2017) shows the human-machine learning combo approaches in maximizing the accuracy rate for the machine designated codes providing necessary filtering for manual inspection.

Another study on the motor vehicle crash prediction using machine learning algorithm (Support vector model) shows better outputs than the traditional risk evaluation method and the study done by (Li et al. 2008)**.** Further a literature review on '*Application of machine learning on risk assessment in engineering*' provided the required information and motivation in writing the paper, which is included in the article published by Norwegian Institute NTNU and written by (Hegde and Rokseth 2020).

2. Methodology: The following section presents a structured discussion on one of the most practised methods in Norwegian tunnels for accident prediction known as Tusi model and it is being discussed in section 2.1, the machine learning prediction model known as logistic regression is being defined in section 2.2 followed by their limitations and scope, applications in different areas of engineering. Once we get finished with the basic definitions (of Tusi and ML algorithm) the tunnel specifications were explained in section 2.3, and the section 2.4 and section 2.5 are used to draw the details of how data in the tunnel is collected and explanation for selection of logistic regression model.

The paper follows the step wise procedures in processing the data to predict the accuracy of tunnel accidents as shown the figure below. Once the data is gathered the selectin of algorithm is a difficult task since machine learning has many different algorithms, in this thesis the algorithm selected was logistic regression. After the selection of appropriate algorithm pre-processing of data can be done where all the impurities and redundancies are removed to process the program with no difficulties.

Once the data is pre-processed the data splitting can be started by dividing the dependent and independent variables for final testing, training the data. After the data is trained and fitting it to predict the accuracy is done during this process if any error is noticed it can be tuned by pre-processing and modelled accordingly for better accuracy.

After getting the satisfactory results the codes can be used in the real world projects and save the time also by getting the efficient results. The step wise procedure is being discussed in the result sections with the Figure 2 and codes used.

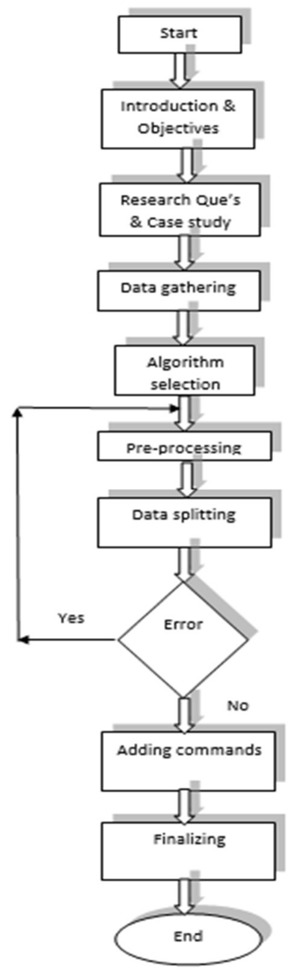


Figure 2 Flow chart of data processing

2.1. **Tusi model:** In Norwegian tunnels, the traditional method for accident prediction is the Tusi method as we can see in this article review of different methods of risk analysis

(Anon n.d.). The goal of the Tusi model is straightforward, basically it is to estimate the number of personal injuries and the frequency of the accidents, material damage in accidents, incidents where the vehicles get stopped in the tunnels (for an indefinite period, making the other vehicles difficulty to drive through) and fire break in vehicles (independent of reasons), also fire in the tunnel. The model produces the statistical data in the end as accident frequencies and their consequences in terms of the number of fatalities.

Data needed for the Tusi model in developing the statistical output of the prediction of accidents and their frequency of occurrence is as followed.

a) **Speed limit**: The maximum allowable speed inside the tunnel that a vehicle can drive.

b) **Accident rate:** The accident rate of other tunnels of equal cross section but not in the tunnel of same area.

c) **One or two-way traffic:** The allowable traffic movement inside the tunnel whether the moment of vehicles is allowed in both the direction or only in one direction. This helps in dividing the tunnel into small zones for easy calculations. The horizontal/vertical zones can be divided up to 11 as discussed in **(Norwegian Public Roads and Administration 2004)** and presented in the figure() below.

d) **Average Annual Daily Traffic (AADT):** AADT is the total traffic volume of any given tunnel and is given by the total annual traffic divided by 365, that gives the total traffic in both directions. AADT helps in determining the category of any tunnel. But it requires 20 years of traffic volume after the tunnel is being opened for the public use this is because of the uncertainties of daily traffic according to the *N500 Norwegian Public Roads Administration Handbooks* **(Norwegian Public Roads and Administration 2004)** and AADT helps in choosing the class of the tunnel, from A to F as shown in Figure 3.

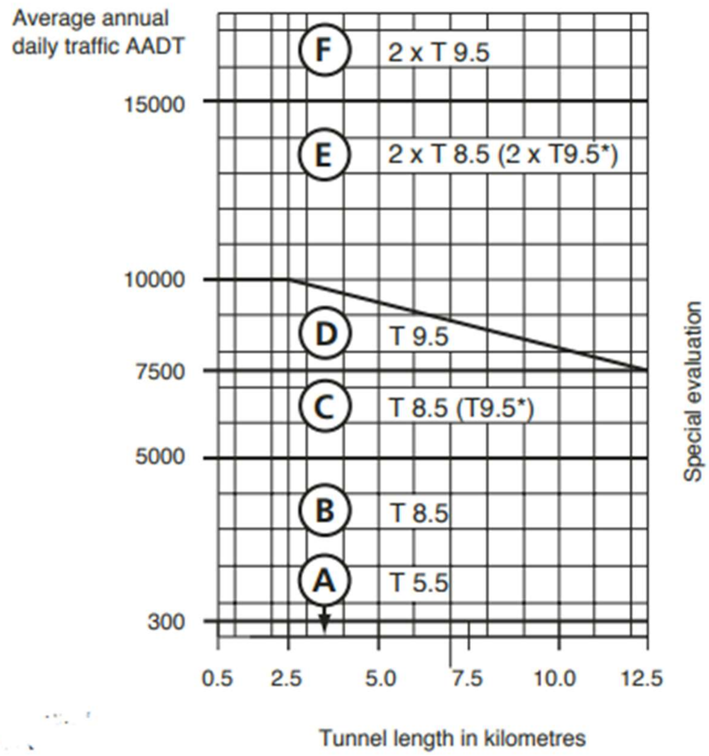


Figure 3 Tunnel categories depending on length and AADT

e) **Length of the tunnel:** Tunnel length helps in dividing the zones and also helps in defining the class of a tunnel.

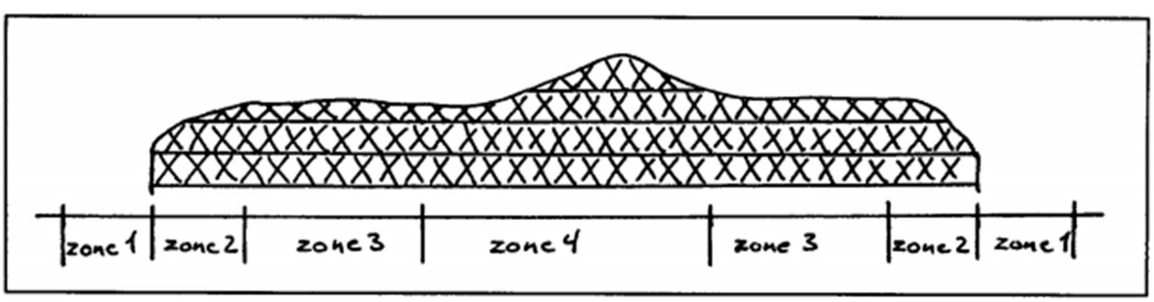


Figure 4 Zone division in the tunnels

f) **Geometry:** Geometry of the tunnel here defines the cross section measurements of the tunnel and both horizontal, vertical measurements of a tunnel are important as shown in the Figure 5.

The cross-section of Eiksund tunnel is T8.5 on one end and is best suited for tunnels with 2-way traffic, categories C and B, also for each tube in the tunnel category E. Whereas the tunnel cross-section T11.5 is on the other side of Eiksund and is best suited depending on the requirement for the 3-lanes or an extremity lay-by in tunnel categories B, C and E. The tunnel cross-sections will also give place for two lanes and a separate foot-traveler and cycle path separated with a concrete guardrail as discussed in the N500 hand book.

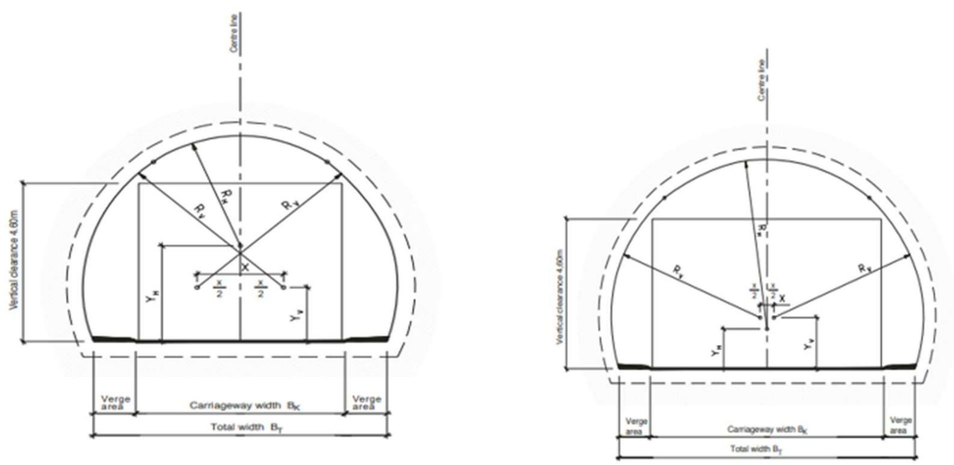


Figure 5 Geometric measurements of tunnels T 4m to T 8.5m and T 9.5m to T 12.5m

Cross-section	Total width B_T	Carriage-way width B_K	Centre point wall radius X	Centre height wall radius Y_V	Wall radius R_V	Centre height lining radius Y_H	Lining radius R_H
T4	4.0	3.0	-	-	-	1.33	2.40
T5.5	5.5	3.5	3.40	1.77	4.79	3.17	2.59
T7	7.0	5.0	2.06	1.57	4.79	2.78	3.20
T8.5	8.5	6.5	0.40	1.77	4.79	1.98	4.50
T9.5	9.5	7.0	0.44	1.57	4.79	1.22	5.20
T11.5	11.5	9.5	2.60	1.77	4.79	- 0.26	7.20
T12.5	12.5	10.0	3.44	1.57	4.79	- 0.46	7.45

All measurements in metres

Figure 6 Geometric specification of tunnels

g) **Limitations of the Tusi model:** The main drawback in using the Tusi model is that the accuracy of statistical data developed in the end depends directly to the data that we used from the tunnels similar on to the tunnel under investigation for better results and by any reasons if the consequence data is not available from other tunnels or past data is missing then other methods are to be adapted besides the Tusi model.

The time taken for the evaluation of final results is directly proportional to the data required to process meaning that as the traffic in the tunnel increases the data gets more and more and the analysis has to be done in short intervals, which may directs to the chances of human errors (avoiding minor deviations).

2.2. Machine learning: During the past decade, machine learning has been implemented in major working areas for examining the risk analysis example starting from the Industrial field to the medical fields. Different machine learning algorithms have been used in different purposes such as predicting the landslides, structural health monitoring of civil structures, and classifying a diseases, vehicle crashes and deep neural network in the project managements.

This paper deals with the application of machine learning (ML) in foretelling accidents in tunnels. According to (Samuel 1959) ML is *”Programming the computer to learn from its past experiences is referred to as Machine learning”,* meaning in this paper we teach the program from the pattern of accidents occurred in the past in the Eiksund tunnel and train the model to predict the occurrence or non-occurrence of accidents in future and extrace the accuracy of predictions.

ML is basically divided into two main streams supervised machine learning and unsupervised machine learning according to (Maglogiannis 2007) “*Every case in a dataset used by the ML algorithm is represented using some set of features, these features may be continuous, categorical or binary. If cases are given with known labels to the corresponding correct outputs then the learning is called Supervised Learning and Unsupervised machine learning if the cases are not labelled”.*

The accident prediction in this paper was done using the supervised machine learning algorithm known as Logistic Regression (LR). LR is a sub-branch of machine learning algorithm that is used for classification problems and the nature of LR is a predictive analysis that based on the concept of probability.

It has a complex cost function *"Cost function is used to minimize the error and provide accurate results"* (Pant 2019) that makes it different form of Linear regression and thus used in multivariable regression problems as shown in Figure 7which is known as Logistic function/sigmoid function discussed by (Pant 2019) *"To outline prognosticated values to the probabilities, we apply the Sigmoid function. It helps in mapping the real values into the value between 0 and 1. In machine learning, the sigmoid function is used for mapping prediction to probabilities."* the hypothesis keeps the cost function lie between 0 and 1 according to (Peng, Lee, and Ingersoll 2002) meaning the final prediction will be binary either 0 or 1.

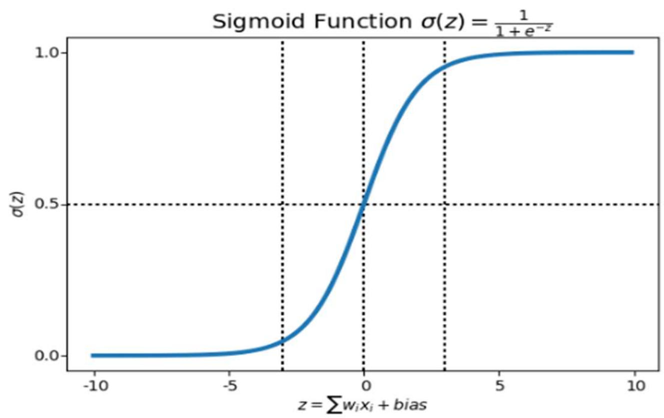
The geometric representation of sigmoid function has curves in the ends which makes the outcomes lies between 0 and 1, it is slightly different from the linear function which has a straight line. Meaning the ends are more curved which makes it look like 'S' as shown in the picture (sigmoid).

Hypothesis limits for Logistic regression $0 \leq h\theta(x) \leq 1$ and the mathematical expression is [$\sigma(Z) = \sigma(\beta_0 + \beta_1 X)$].

Cost function for Logistic regression $J(\theta)$:

$-\log(h\theta(x))$ if y=1 (accident) and

$-\log(1-h\theta(x))$ if y=1 (no accident)



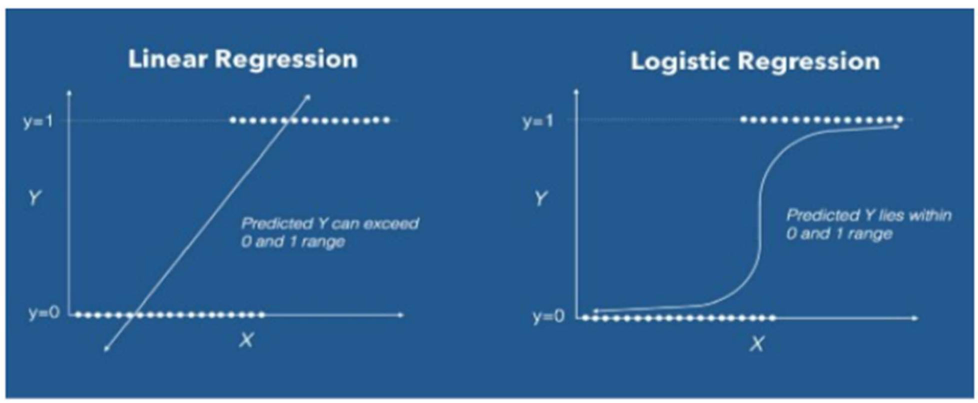


Figure 7 Linear Regression & Logistic Regression

LR final results are marked using decision boundary according to (Lee 2005) *"It is a threshold limit where the boundary of variable changes"* for example in our case occurrence of accident 1 and not occurring of an accident is 0 and the set limit threshold is 0.5 above which result shows the occurrence of an accident 1 and output lower than 0.5 shows there are low chances of accident 0 now this change in a decision is known as the decision boundary.

Gradient decedent: In order to minimize the cost function we use gradient decedent function, here we need to be more careful about the step size that we are adapting because large values can lead to bounce back in another direction and small value takes more number of iterations. Shown in the figure (gradient decendent) (Peng et al. 2002)

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update all θ_j)

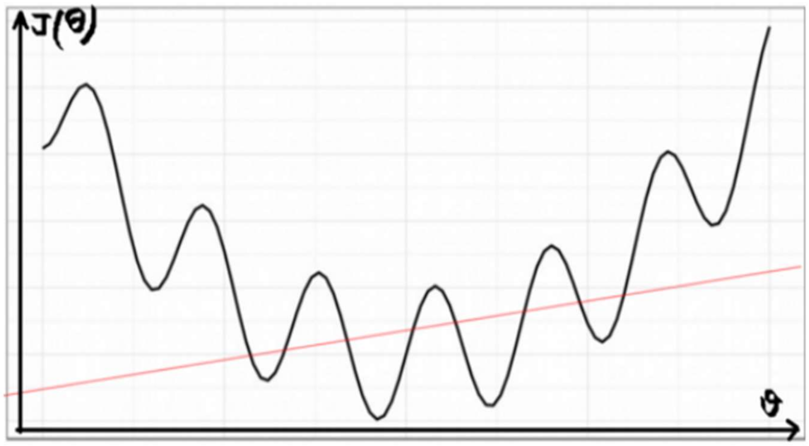


Figure 8 Gradient decedent graph

2.3. Tunnel specification and Case study: This paper tries to perform the ML to draw the accuracy of accident prediction in tunnels instead of the traditional method (Tusi model). For the study purpose we are using the data from the Eiksund tunnel as the case. The tunnel is considered one of the longest tunnels with a collective length of 7850m and first opened for public transportation use in 2008 and the main purpose of tunnel is to connect Ørsta, Unsteinvik and Volda counties for easy movement of traffic instead of ferry boats. The route is as shown in the Figure 9 below, the tunnel has two openings one has T8.5 profile towards the Steinnestranda to the bottom, and the other has a T11.5 profile towards the Eika side to the bottom of the tunnel as shown in the profile below in Figure 10.

Figure 9 Eiksund tunnel route.

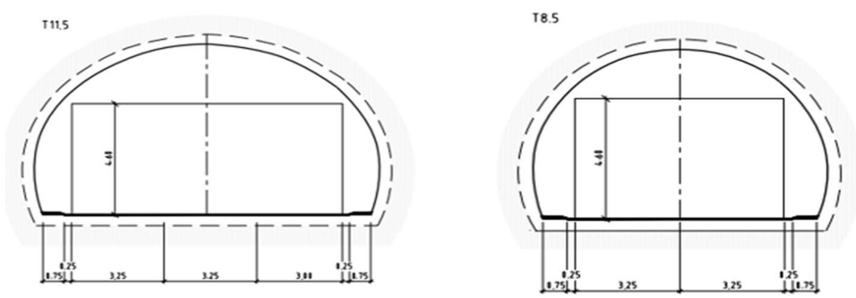


Figure 10 Figure(respective profiles of openings Eika T11.5 and Steinnestranda T8.5)

Since the tunnel is connecting the island to the mainland it passes under the sea, fjords and there is also a small bridge in connection. The tunnel is also marked as one of the steep road tunnels with a maximum slope gradient of 9.6% but the sigh board indicates 10% and also recorded one of the deepest sub-sea tunnels in the country with a maximum depth of 272m, presented in the Figure 11 below . Due to the presence of deep fjords as we get into the tunnel the road gets steeper and the maximum slope gradient reaches approximately 10% which is double compared to any other European road tunnels.

Speaking about the traffic volume in the tunnel compared to the Oslo tunnel the density of volume is low, this was the reason in not gathering the big data of traffic. Transportations trucks in the Eiksund tunnel passes usually and it has an annual daily traffic volume(ADT) of 3130 vehicles/day from the report taken in 2018 out of all the vehicle traffic volume there are 7% of truck movement in the tunnel.

The ADT is more than the expected traffic during the construction period, according to the Norwegian tunnels handbook N500 English edition April 2004 the Eiksund tunnel falls under B-class (can be seen in picture(ADT)) since the length is greater than 3000m and annual traffic data volume (ADT=3010) lies in between 300 and 5000.

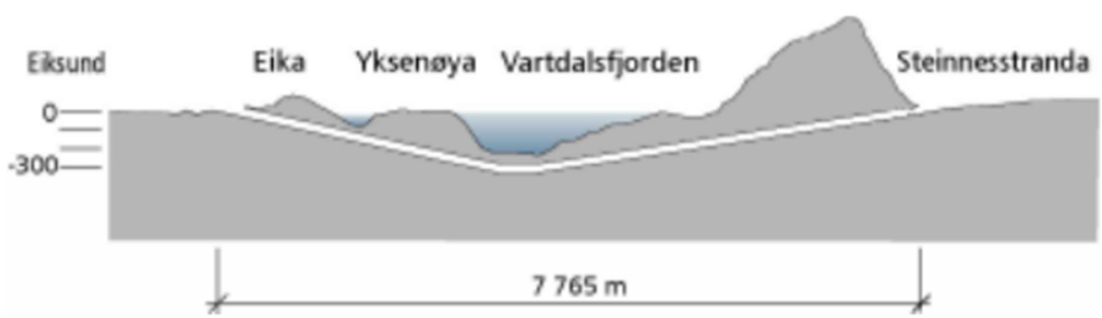


Figure 11 Longitudinal cross section of tunnel (showing length, openings and depth)

Figure below shows the traffic data from past 10 years meaning the traffic data (ADT) on the monthly basis since the opening year 2008 of the tunnel to year 2018.

År	Jan	Feb	Mars	April	Mai	Juni	Juli	Aug	Sept	Okt	Nov	Des	ÅDT	ÅDT tunge	% tunge
2008			1500	1623	1719	1758	1567	1679	1756	1723	1576	1469	1595	124	7,8 %
2009	1513	1639	1761	1736	1847	1993	1631	1760	1794	1766	1688	1524	1721	126	7,3 %
2010	1509	1669	1745	1783	1831	2038	1745	1920	1999	1853	1876	1567	1794	134	7,5 %
2011	1629	1822	1812	1844	2043	2108	1756	2034	2091	1939	1891	1777	1895	146	7,7 %
2012	1878	1931	1981	1895	2142	2233	1836	2137	2168	2249	2113	1805	2031	138	6,8 %
2013	1964	2155	2167	2224	2375	2453	2115	2338			2168	1942	2206	155	7,0 %
2014	2121	2270	2278	2387	2530	2886	2582	2811	2927	2820	2608	2350	2547	195	7,7 %
2015	2517	2664	2833	2872	2925	3135	2663	2996	3070	2917	2774	2524	2824	207	7,3 %
2016	2485	2784	2890	2998	3134	3275	2942	3068	3256	3019	2890	2597	2944	207	7,0 %
2017	2615	2843	2843	2860	3160	3368	2873	3196	3219	3080	2925	2560	2962	198	6,7 %
2018	2858	2997	3098	3294	3378	3431	3013	3330	3258	3163	3110	2629	3129	212	6,8 %

Figure 12 Monthly traffic volume.

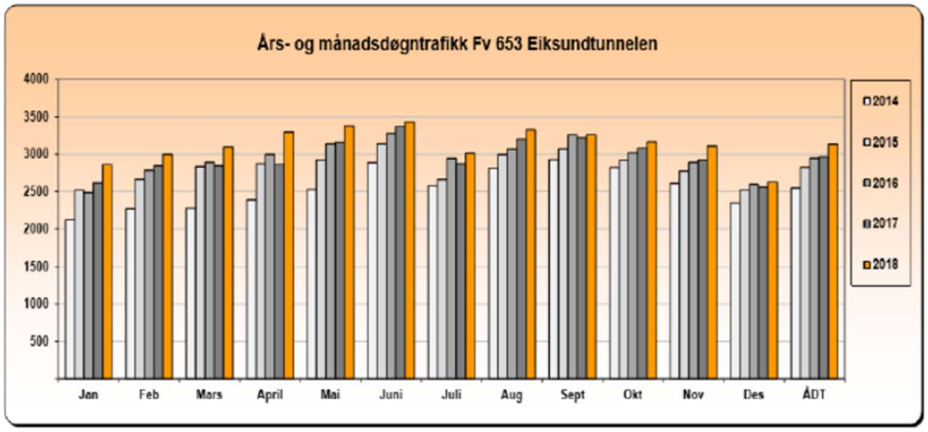


Figure 13 Bar graph of traffic data

Alternate route: The ferry service still exists and if in case th Eiksund tunnel is closed due to any reasons such as traffic block (vehicle stopping inside the tunnel), fire break in cars, fire in the tunnel safety system or closing of tunnel because of regular fire drills. There is an alternative route using the ferry switches E39 Festøy-Solavågen, take from the notes provided by the external supervisor (Olav).

According to the N500 Norwegian road administration hand book for tunnels the Eiksund tunnel lists in the class-B of tunnels on the basis of its daily traffic and length. Following figure gives the fire safety all other safety related systems according to the N500 handbook.

For the fire outbreak in the tunnel water supply is connected to the buildings towards the side of Eika, it can be used in emergency situations and in the emergency period/situations the time taken by the fire department of the tunnels to reach the spot of accident will be about 18 minutes to 19 minutes under ordinary conditions.

Volda fire department has been upgraded and has good experts to handle the task that involves safeguarding the fire preparedness in all tunnels associated. The data described above is taken from the risk analysis report of Eiksund tunnel which was shared by Olav Amund Myklebust co-supervisor and team leader electrical, tunnel and geosciences at Møre og Romsdal, which is shared in the appendix file.

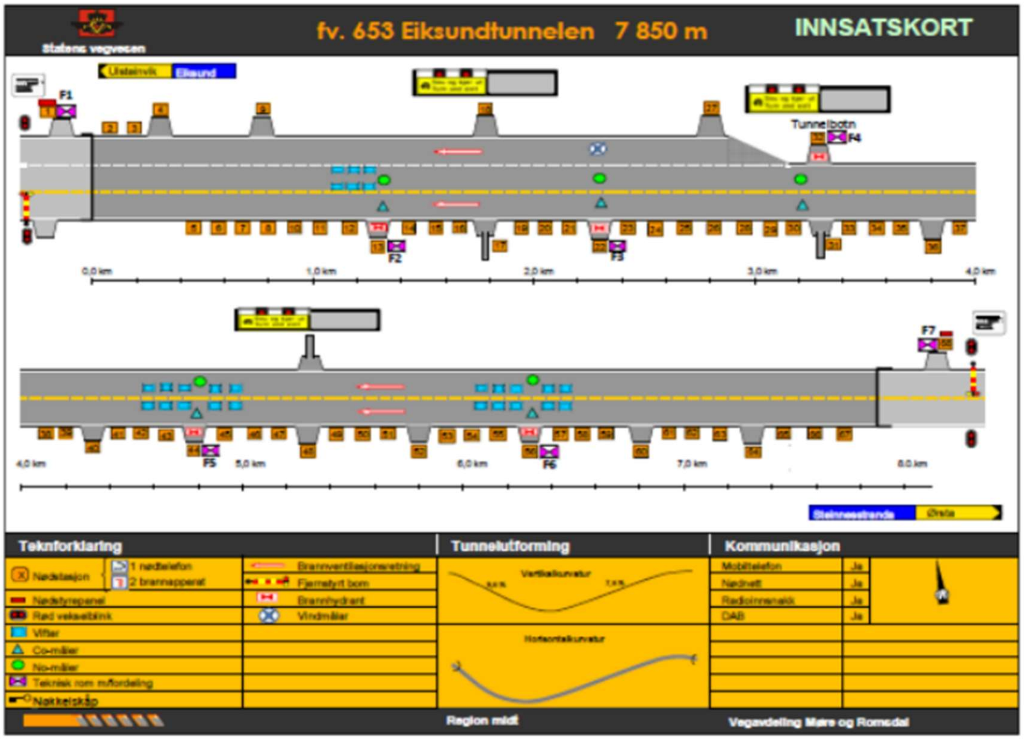


Figure 14 Tunnel safety system

2.4. Data collection (for accident prediction accuracy using machine learning algorithm): The data used is quantative, It has been gathered from the secondary source which is from the Norwegian road administration Statens vegvesen (Anon n.d.) and the tunnel Eiksund is maintained by More og Romsdal fylkeskommune which is in the north most part of west Norway.

The tunnel Eiksund is located in the Western part of Norway respective openings at Eika on one end and Steinnestranda on to the other end. The main reasons for using the secondary data in the paper is to see correlation in the patterns of accident occurred in the past, also the data was open to use for academic research purposes and the third reason is to reduce the time in gathering of data.

2.4.1. Traffic data collection and saving process: This section is centered towards the explanation of how the traffic data is collected by Norwegian public road administration and stored. According to the Norwegian Public Road Administration, the traffic data of the tunnels composes 'point measurements' and 'section measurement'. Meaning in point measurement we get the measurements from traffic registration stations on the state and county municipal road network as well as some municipal roads, and where as in the section measurements, we get the distance measurements of travel time between registration stations in and around the largest cities and on some main roads according to (Anon n.d.). Collectively both ‘Point measurement’ and ‘Section measurement’ helps in recording all the possible data of a vehicle.

The registration stations record both motor vehicles and also bicycles. The recorded details of a motor vehicle are the vehicle with which it enters the tunnel, length of the vehicle to differentiate that it was a car or a heavy vehicle, class of a vehicle and in each lane the distance between the vehicles.

The traffic data is filed using inductive loops, and their working is explained by(Asyrani et al. n.d.) in the wireless parking as shown in the Figure 15 below. When a vehicle rides above the loops, details are registered. The lane is constantly stating the tangible, physical location of the motor vehicle. The direction of the ride is said as the actual path of travel of the car or a truck. There can be instances of overtaking in the tunnel (because at some places int the tunnel there is also allowed for overtaking) and after crossing the mid-way of the road, the direction of travel shows opposite of the specified direction of the road, but nevertheless the correct, actual direction of travel.

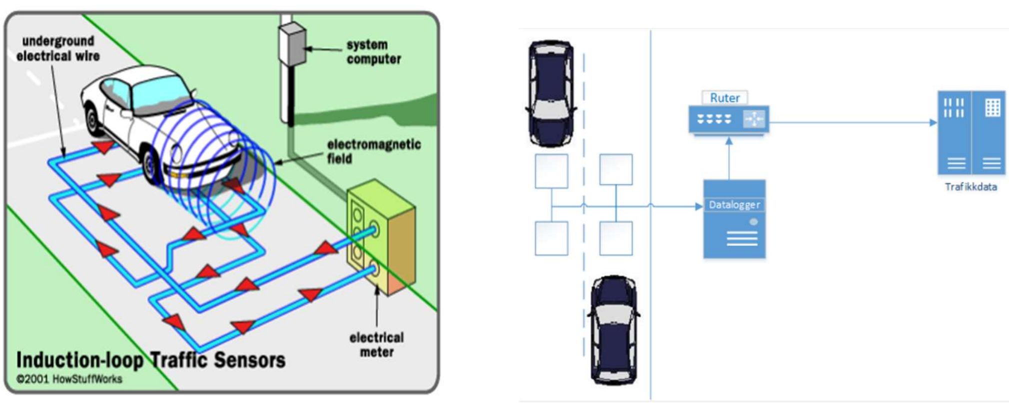


Figure 15 Inductive loop

The data gathered through inductive loops is transferred to the Norwegian Public Roads Administration and from there it takes about 2 to 3 hours for the data to get available on the traffic portal. The bicycle traffic can be registered by inductive loops and piezoelectric cables that are laid in the cycling paths as shown in theFigure 16. While a cyclist crosses the sensor, the speed is recorded, the registrations are conveyed in the same way as for motor vehicles.

Figure 16 Bicycle regestration point

Traffic registration stations hold one or more registration positions attached to them. Further, information on traffic registration points and associated traffic data can be extracted from the traffic data portal.

The regestration points for bicycles started from July 2015 and for motor vehicles the regestration started since November 2014 (Anon n.d.).

2.4.2. Data quality: Following section shows the quality of data that the Norwegian public road administration is collecting and also their conditions for registering a vehicle in the data.

2.4.2.1. Conditions for recording: All the vehicles entering the tunnel should follow certain limits (minimum and maximum) such as speed, length in order to get recorded by the system, which are discussed below.

2.4.2.2. Length of vehicle: Vehicles that enters the tunnel in order to get registered in the data, the length of the motor vehicle should be measured in between 1meter and 27meters else entry is not recorded (meaning these are the upper and lower limits for the registering of a vehicle driving into the tunnel).

2.4.2.3. Speed of Vehicle: The vehicle speed is recorded by the measuring instrument, if the velocity of vehicle is between 7km/h and 300 km/h else are not registered.

2.4.2.4. Other reasons: In addition, other extra rules on registering are based on the suppliers quality parameters of the device or in other words the default rules on a machine which are designed by the company.

2.4.2.5. For bicycles: Regarding the data quality of the bicycle, registration depends on the devices if the device detects an object other than bicycle it marks as invalid, meaning the recorded picture must be recognized as the bicycle.

2.4.3. For checking the quality of data recorded: The mathematical equations used to evaluate the proportion of accepted registrations in any specific time interval with a range of quality data.

$$Speed\ quality\ rating = \frac{\text{number of registrations with valid speed measurement in the period}}{\text{total number of valid registrations in the period}} * 100$$

$$Longitude\ quality\ degree = \frac{\text{number of registrations with valid length measurement in the period}}{\text{total number of valid registrations in the period}} * 100$$

$$Classification\ quality\ grade = \frac{\text{number of registrations with valid classification in the period}}{\text{total number of valid registrations in the period}} * 100$$

Above equations gives the quality of data registered respectively for speed, length and classification between vehicles**.**

2.5. Algorithm selection:

Logistic regression: The calculations used and the notations in logistic regression used for the paper are discussed in this section, starting with the logistic regression graph it has the boundary line bent in the ends which makes it look like S (shown in figure()), this curve is known as the Sigmoidal curve (Kleinbaum and Klein 2010).

2.5.1. Why only Logistic regression & not the Linear regression or Tusi model:

First to start with is this thesis paper is to see the possibility and working of ML in predicting accidents, also ML had been improving rapidly in all the areas of engineering and medical due to its flexibility and efficiency.

However the other reasons for selecting the logistic regression over the linear regression and traditional risk analysis method (Tusi) are that the linear regression makes it difficult to judge fairly when the variables show an uneven pattern and the outcome final values can be displayed greater than 1 or less than 0 depending on how the data is distributed. In simple words it will be hard to predict the outcomes if the variables that do not follow a linear trend. Another reason for not adapting linear regression is that the errors has no normal distribution nor they are constant across the entire range of data. Logistic Regression does it by applying the Log transformation to the dependent variables (‘X’), as the LR model splits the data into two variables which are dependent variables and independent variables. 'X' variables are the which is being explained is called dependent or response, they are included in the data and the other variables used to explain or predict the outcomes are known as independent variables and mostly denoted by 'Y' according to (Hilbe 2009).

2.5.2. Data preparation: For all the projects related to programming or non-programming tasks there is a need for the collection of data and preparation of data. Data preparation is given the equal importance as the final results because if the data is pure and have no missing values the prediction gets fair. In last section we shortly discussed about the data collection and also the section focuses on the preparation of data.

Data preparations are given same weightage for accurate final results since the available data contains some impurities such as incomplete data in the big data gathering values often misses resulting in a weak prediction of problem encounters, noise or disturbances is when we have the outliers/errors in data and inconsistent data containing codes or names or both which makes challenging in training the data in future cases, it has been discussed by (Zhang, Zhang, and Yang 2003) in data mining.

To improve the final results, the raw data has to be preprocessed and analyzed to remove the unwanted errors by changing the symbols and codes, handling the missing data by adapting the values that fit best. Developing data includes the steps of data collection, data integration, data transformation, data cleaning, data reduction and data discretization, discussed by (Zhang et al. 2003).

The extracted data from Eiksundtunnel restrained some important measuring values and also embraced values of no importance other than this the file was in Norwegian language which made it difficult in translating to English and all the unnecessary values are dropped using the codes in python using (drop command) the values of Ventilation speed and accidents that happened in the tunnel were not included in the data so they have been added using the (fill-in command), which can be seen in the. CSV file attached.

The accident data is taken from the Norwegian road administration page for calculations (Anon n.d.) and ventilation speed is taken from the Eiksund file shared by Olav (co-supervisor).

2.5.3. Tunnel ventilation: The Eiksund tunnel has longitudinal mechanical ventilation with pairwise mounted ventilators in the tunnel roof. Each of the ventilators has an effect of 22 MW and can ventilate with the same effect in both directions.

There are a total of 28 ventilators in the tunnel which together are intended to generate and give an air velocity (fire ventilation) of about 2.5 m / s. According to NordFou's 2015 research project, it shows that tunnels of length greater than 3 km should have the ventilation of a 100MW fire that could provide a ventilation rate of approximately 5 m/s. The Eiksund tunnel will therefore be the basis for assessing whether the capacity of the ventilation is sufficient as it is dimensioned today but there has to be installing the strong ventilation system.

3. **Results:** The result section gives the brief concept of prediction, accuracy and step by step process of logistic regression. Here we have used the Jupiter notebook platform to run the machine learning algorithm Logistic regression to predict the traffic accidents in the tunnel. The steps followed in the process are:

I.**Data:** After selecting the case and finalizing the data to be used. Which, in this paper was traffic data of Eiksund tunnel (as discussed earlier about data gathering of the Eiksund tunnel in methodology section 2.4). The traffic data gathered from the 5 different accident scenarios, with time and date which were happened in the tunnel as shown in the figures() below and also the traffic data of no accident 01/03/2021 is taken for prediction. The data is taken from the Norwegian public road administration official site (Anon n.d.) where the traffic data is stored.

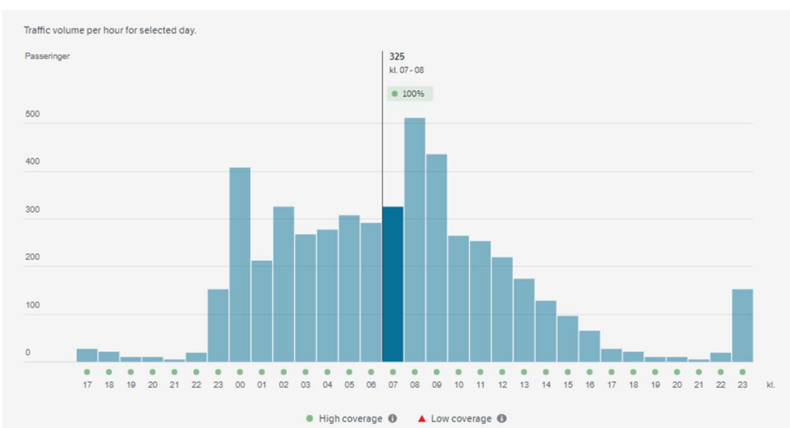


Figure 17 Data on 26/06/2018 at 07:00 & 8:00

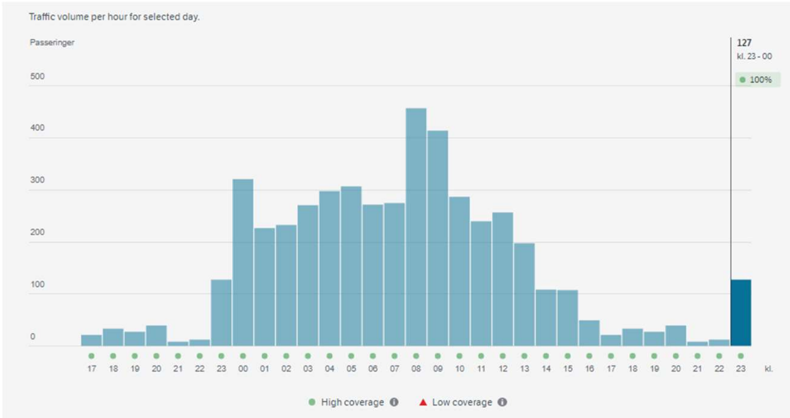


Figure 18 Data on 10/07/2018 at 23:00 & 00:00

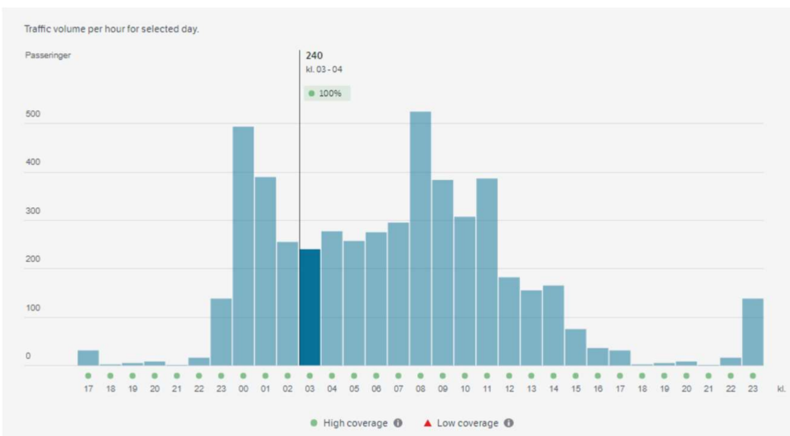


Figure 19 Data on 06/05/2019 at 03:00 & 04:00

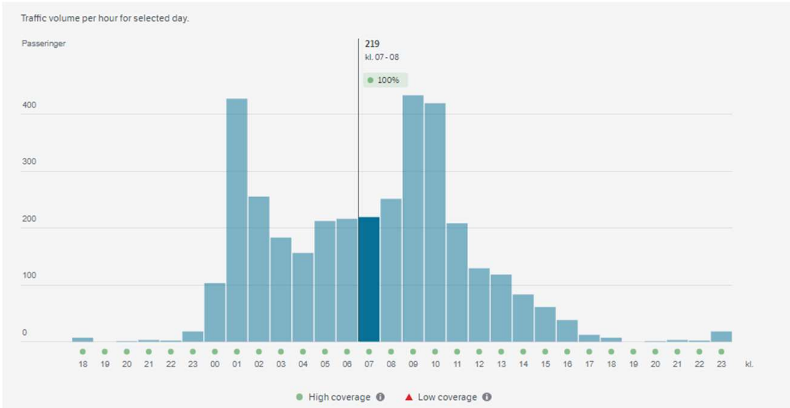


Figure 20 Data on 01/03/2021 at 07:00 & 8:00

To upload the data into the Jupiter notebook the pandas library was used in the command prompt as shown in the picture below, the 'pd.read_csv' command is used to load the data set. The other library used in further steps is 'NumPy' to sort the variables in a matrix format for easy dropping or adding of other variables. NumPy is also used for mathematical calculations these are the most commonly used libraries.

The traffic data is uploaded as data function (df) and after loading all the data to see/check the sample of data we have added a command prompt is used 'df.head' to see the first 5 rows and 'df.tail' to see the last 5 rows as listed below in the figure this was before pre-processing.

In [51]:
```
import pandas as pd
import numpy as np
```

In [4]:
```
df=pd.read_csv('Dailytraffic_Data.csv')
```

In [5]:
```
df.head()
```

Out[5]:

	Trafikkregistreringspunkt;Navn;Vegreferanse;Fra;Til;Dato;Felt;Volum;Dekningsgrad (%);Antall timer total;Antall timer inkludert;Antall timer ugyldig;Ikke gyldig lengde;Lengdekvalitetsgrad (%);< 5	6m;>= 5	6m;5	6m - 7
0	25819V2511481;Haddal;FV653 S1D1 m30;2019-01-01...	0;24;24;0;1;99	82;523;20;14;2;4;0;0;0	NaN
1	25819V2511481;Haddal;FV653 S1D1 m30;2019-01-01...	0;24;24;0;0;100	00;770;32;19;9;4;0;0;0	NaN
2	25819V2511481;Haddal;FV653 S1D1 m30;2019-01-01...	0;24;24;0;1;99	82;523;20;14;2;4;0;0;0	NaN
3	25819V2511481;Haddal;FV653 S1D1 m30;2019-01-01...	0;24;24;0;0;100	00;770;32;19;9;4;0;0;0	NaN
4	25819V2511481;Haddal;FV653 S1D1 m30;2019-01-01...	0;24;24;0;1;99	93;1293;52;33;11;8;0;0;0	NaN

```
In [6]: df.tail()
Out[6]:
```

	Trafikkregistreringspunkt;Navn;Vegreferanse;Fra;Til;Dato;Felt;Volum;Dekningsgrad (%);Antall timer total;Antall timer inkludert;Antall timer ugyldig;Ikke gyldig lengde;Lengdekvalitetsgrad (%);< 5	6m;>= 5	6m;5	6
3655	25819V2511481;Haddal;FV653 S1D1 m30;2021-01-01...	0;24;24;0;1;99	84;591;27;21;6;0;0;0;0	Na
3656	25819V2511481;Haddal;FV653 S1D1 m30;2021-01-01...	0;24;24;0;0;100	00;767;17;11;5;1;0;0;0	Na
3657	25819V2511481;Haddal;FV653 S1D1 m30;2021-01-01...	0;24;24;0;1;99	84;591;27;21;6;0;0;0;0	Na
3658	25819V2511481;Haddal;FV653 S1D1 m30;2021-01-01...	0;24;24;0;0;100	00;767;17;11;5;1;0;0;0	Na
3659	25819V2511481;Haddal;FV653 S1D1 m30;2021-01-01...	0;24;24;0;1;99	93;1358;44;32;11;1;0;0;0	Na

As the data was full of string variables and other unnecessary data that is of no use we pre-process it next steps.

II.**Applying the algorithm**: After uploading the data successfully we have applied the LR function to the program by using sikit learn library (sklearn), this library is used to load the LR and other big programs for predictive analysis.

The adding of function and performing it is shown in the figure below and the actual file is attached in the appendix folder.

```
In [266]: from sklearn.linear_model import LogisticRegression
          logmodel=LogisticRegression()
          logmodel.fit(X_train,Y_train)
Out[266]: LogisticRegression()
```

III.**Pre-processing the data**: The data gathered was in Norwegian language and there were many unwanted variables, so by using the dropping command we have remove them and there were also other steps which were not shown in the figures below but after changing the language, converting the strings to float and dropping unwanted columns the processing the data it looks like as shown below.

```
In [246]: df=df.drop(['LightingConditions'],axis=1)

In [247]: df=df.drop(['Conditions'],axis=1)

In [248]: df=df.drop(['LocationConditions','TireType'],axis=1)

In [249]: df=df.drop(['Date'],axis=1)

In [251]: df=df.drop(['MunicipalityName','MunicipalityName'], axis=1)

In [253]: df=df.drop(['CountyName'], axis=1)

In [254]: df
Out[254]:
```

	RoadWidthInMeters	SpeedLimitInKilometerPerHour	MunicipalityNumber	Year	ADT	CountyNumber	TemperatureInDegreeCel
0	10	80	1516	2009	1600	15	
1	7	80	1577	2012	2030	15	
2	7	80	1520	2018	3130	15	
3	7	80	1520	2018	3130	15	
4	7	80	1520	2019	3130	15	
5	10	80	1516	2021	3100	15	

```
In [255]: df.size
```

Once the unwanted data is removed now we fix the missing values (NAN) with the mean of columns, meaning the data is clean. Now the data is ready for splitting into testing and training which is discussed in the section below.

IV.**Testing and Training data:** After the data is being able to use to train, we separated the data into ‘X’ and ‘Y’ variables. Where ‘X’ are the known values, ‘Y’ is unknown variable and predicting ‘Y’ was our goal. Once the data is separated now split the data into training and testing sets. Here we have used the 80% for training the data and remaining 20% for testing the results.

```
In [264]: x = df.iloc[:, 0:11]
          y =df.iloc[:,-1]

In [265]: from sklearn.model_selection import train_test_split
          X_train,X_test,Y_train,Y_test=train_test_split(x,y,test_size=0.2,random_state=1)

In [266]: from sklearn.linear_model import LogisticRegression
          logmodel=LogisticRegression()
          logmodel.fit(X_train,Y_train)

Out[266]: LogisticRegression()
```

V.**Look for errors**: After the testing and training data had made and running the program will give the results, if errors are encountered then we go back to the pre-processing stage and try to improve the mistakes and follow the process again until we get the program run is successfully (as in our case) and if any errors are found then we need to pre-process our data.

VI.**Final result**: The predicted value shows there is no accident, meaning there are low chances of accident occurrence on that particular day and the accuracy of the developed model was 50%.

```
In [268]: Y_pred=logmodel.predict(X_test)
          print("Accuracy",(logmodel.score(X_test,Y_test)))
          from sklearn.metrics import confusion_matrix
          confusion_metrix=confusion_matrix(Y_test,Y_pred)
          print(confusion_matrix)

          Accuracy 0.5
          <function confusion_matrix at 0x0000023D47C5DC10>
```

The reason for the accuracy of model and the improvement suggestions are discussed in the discussion section.

4. Discussion: The paper concept was to try and implement one of the machine learning algorithms logistic regression to predict the accidents accuracy in the tunnel instead of the regular risk evaluation model Tusi to see and find better results with high accuracy.

Because as mentioned earlier in the introduction of the paper traditional model has some loose ends like time for the risk prediction is generally high and as the data increases the calculations gets complicated and other limitations discussed in the methodology section g).

The results indicates that the LR can be used as a prediction model for accidents in the tunnels even though the accuracy of prediction was low buy under a keen observations and by improving some of the areas where this research paper had failed to provide.

One thing can be noticed by the result, there is a correlation of variable data taken from the tunnel accidents and the prediction, such as the condition vehicle condition, climate, vision as the values increasing and decreasing the result keep on changes.

In line with the hypothesis, even from the unsatisfied results we can still consider that the predictive model have a scope where it can be applied upon the collection of more valid data and by the change of methods of application in the machine learning algorithm.

In concern to the results comparing to the recent studies and the research the predictions were quite valid, but after viewing the final numbers there is clear gap which can be filled by studying and can be use the algorithm in tunnel safety systems for risk evaluations and by noting the accuracy model it partly supports the hypothesis of using the computerized solutions in accident prediction.

The study provides a new insight into the relationship between the manual evaluation of risk and the automated accident prediction on road tunnels.

Clearly there were few limitations in the experiment, it is beyond the scope of this study to a address the 100% accuracy of prediction with the available data and also the assumptions.

Further this research is required to establish whether to depend only on the machine learning algorithm or there should be involvement on human is needed.

The results could be improved in future studies by adding more historic data and the language was the barrier in reading the data. Since the basic idea behind using the machine learning model to read the past incidents and learn from them, the tunnel used as the case study has very low accident rate because of lower traffic flow. From the opening of tunnel 2008 to 2021 there were only 5 incidents recorded with one fatal accident, this factors made the prediction less accurate.

The improvements in this study can be made by better understanding of algorithm and adding more input data such as the tunnel maintaining authorities can add other factors like weather, precipitation, condition of the vehicle before accident occurred and the road steepness that directly impact the accident scenario.

5. Conclusion:

This research aimed in trying the implementation of computer based probabilistic model using machine learning algorithm for predicting the accidents in the tunnel as an alternative risk assessment tools, the aim was to fill the difficulties and drawbacks of classical method Tusi because the model had few loose ends that make calculations difficult.

It can be concluded that there is a narrow scope in application of logistic regression in tunnels for accident prediction, the research was expected that the predictions done by logistic regression would be more accurate than the Tusi method for contribution in achieving the Norway’s zero vision of accident.

The results indicates that the model can be reviewed again and there could be an improvement in the studying of algorithm, data gathering and training of the data. Besides the accident prediction, we can consider that the machine learning algorithm for other predictions also, since it has a broader applications. For example the quick clay liquification that had caused number of landslides and washing of surface fields have also been noted, the same type of clay had found under the tunnels that can be predicted using the machine learning algorithms.

Based on these conclusions the practitioner should consider the scope of machine learning in road tunnels, to better understand the implications of these results future studies could enrich the accuracy by gathering the real data.

References:

Amundsen, F. H., and G. Ranes. 2000. "Studies on Traffic Accidents in Norwegian Road Tunnels." *Tunnelling and Underground Space Technology* 15(1):3–11. doi: 10.1016/S0886-7798(00)00024-9.

Anon. n.d. "Harmonised Risk Assessment."

Anon. n.d. "Traffic Data." Retrieved May 6, 2021b (https://www.vegvesen.no/trafikkdata/start/om-trafikkdata).

Anon. n.d. "Trafikkdata." Retrieved May 1, 2021c (https://www.vegvesen.no/trafikkdata/start/eksport?datatype=HOUR&from=2021-05-01&lat=62.23536748290924&lon=5.914974877260503&registrationFrequencies=&to=2021-05-02&trafikanttype=vehicle&zoom=11).

Asyrani, Hamzah, Bin Sulaiman, Mohd Fareez, Bin Mohd Afif, Mohd Azlishah, Bin Othman, Mohamad Harris, Bin Misran, Maizatul Alice, and Binti Meor Said. n.d. *Hamzah Asyrani Bin Sulaiman et.al / International Journal of Engineering and Technology (IJET) Wireless Based Smart Parking System Using*.

Doroshenko, Anna. 2020. "Applying Artificial Neural Networks In Construction." *E3S Web of Conferences* 143:01029. doi: 10.1051/e3sconf/202014301029.

Figueiredo, Eloi, Gyuhae Park, Charles R. Farrar, Keith Worden, and Joaquim Figueiras. 2011. "Machine Learning Algorithms for Damage Detection under Operational and Environmental Variability." *Structural Health Monitoring* 10(6):559–72. doi: 10.1177/1475921710388971.

Fort, Emmanuel, Laure Pourcel, Philippe Davezies, Camille Renaux, Mireille Chiron, and Barbara Charbotel. 2010. "Road Accidents, an Occupational Risk." *Safety Science* 48(10):1412–20. doi: 10.1016/j.ssci.2010.06.001.

Hegde, Jeevith, and Børge Rokseth. 2020. "Applications of Machine Learning Methods for Engineering Risk Assessment – A Review." *Safety Science* 122:104492. doi: 10.1016/j.ssci.2019.09.015.

Hilbe, Joseph M. 2009. *Logistic Regression Models*. CRC Press.

Kleinbaum, David G., and Mitchel Klein. 2010. *Logistic Regression*. New York, NY: Springer New York.

Knapstad, Tonja, Morten Somme, and Ove Njå. 2019. "The Communication of Vision Zero Values in Tunnel Safety Management." Pp. 3009–16 in *Proceedings of* *the 29th European Safety and Reliability Conference (ESREL)*. Research Publishing Services.

Kononenko, Igor. 2001. "Machine Learning for Medical Diagnosis: History, State of the Art and Perspective." *Artificial Intelligence in Medicine* 23(1):89–109. doi: 10.1016/S0933-3657(01)00077-X.

Lee, S. 2005. "Application of Logistic Regression Model and Its Validation for Landslide Susceptibility Mapping Using GIS and Remote Sensing Data." *International Journal of Remote Sensing* 26(7):1477–91. doi: 10.1080/01431160412331331012.

Li, Xiugang, Dominique Lord, Yunlong Zhang, and Yuanchang Xie. 2008. "Predicting Motor Vehicle Crashes Using Support Vector Machine Models." *Accident Analysis & Prevention* 40(4):1611–18. doi: 10.1016/j.aap.2008.04.010.

Maglogiannis, Ilias G. 2007. *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*. IOS Press.

Marucci-Wellman, Helen R., Helen L. Corns, and Mark R. Lehto. 2017. "Classifying Injury Narratives of Large Administrative Databases for Surveillance—A Practical Approach Combining Machine Learning Ensembles and Human Review." *Accident Analysis & Prevention* 98:359–71. doi: 10.1016/j.aap.2016.10.014.

Nævestad, Tor-Olav, Ross Phillips, and Beate Elvebakk. 2015. "Traffic Accidents Triggered by Drivers at Work – A Survey and Analysis of Contributing Factors." *Transportation Research Part F: Traffic Psychology and Behaviour* 34:94–107. doi: 10.1016/j.trf.2015.07.024.

Norwegian Public Roads and Administration. 2004. *Road Tunnels: Standard.* S. l.: Norwegian Public Road Administration = Statens vegvesen.

Norwegin public road administration. n.d. "Sikkerhetskontrollørens Oppgaver - Tunnelsikkerhet - Vegtilsynet." Retrieved June 22, 2021 (https://vegtilsynet.com/tilsyn/tilsynsrapporter/sikkerhetskontrollorens-oppgaver-tunnelsikkerhet).

Pant, Ayush. 2019. "Introduction to Logistic Regression." *Medium*. Retrieved June 20, 2021 (https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148).

Peng, Chao-Ying Joanne, Kuk Lida Lee, and Gary M. Ingersoll. 2002. "An Introduction to Logistic Regression Analysis and Reporting." *The Journal of Educational Research* 96(1):3–14. doi: 10.1080/00220670209598786.

Pham, Binh Thai, Indra Prakash, and Dieu Tien Bui. 2018. "Spatial Prediction of Landslides Using a Hybrid Machine Learning Approach Based on Random Subspace and Classification and Regression Trees." *Geomorphology* 303:256–70. doi: 10.1016/j.geomorph.2017.12.008.

Phillips, Ross Owen, and Sunniva Frislid Meyer. n.d. "Kartlegging av arbeidsrelaterte trafikkulykker." 69.

Samuel, Arthur L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 3(3):210–29.

Vuilleumier, F., A. Weatherill, and B. Crausaz. 2002. "Safety Aspects of Railway and Road Tunnel: Example of the Lötschberg Railway Tunnel and Mont-Blanc Road Tunnel." *Tunnelling and Underground Space Technology* 17(2):153–58. doi: 10.1016/S0886-7798(02)00018-4.

Zhang, Shichao, Chengqi Zhang, and Qiang Yang. 2003. "Data Preparation for Data Mining." *Applied Artificial Intelligence* 17(5–6):375–81. doi: https://doi.org/10.1080/713827180.