UNIVERSITY OF NAIROBI
COLLEGE OF BIOLOGICAL AND PHYSICAL SCIENCES
SCHOOL OF COMPUTING AND INFORMATICS

COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS ON WEIGHBRIDGE DATA FOR OVERLOADED TRUCK PREDICTION (A CASE OF GILGIL WEIGHBRIDGE STATION)

BY:
BENEDICT S.O. BWIRE

A RESEARCH PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF MASTER OF SCIENCE IN COMPUTATIONAL INTELLIGENCE OF THE UNIVERSITY OF NAIROBI.

NOVEMBER, 2019

DECLARATION

Student

This project report is my original work and has not to the best of my knowledge been presented anywhere for the purposes of any academic award.

Signature: ________________________________ Date_______________________

BWIRE, Benedict Sibukho Ogombo

P52/6389/2017

Supervisor's approval

This project report has been submitted in partial fulfillment of the requirements for the Master of Science in Computational Intelligence of the University of Nairobi with my approval as the University Supervisor.

Signature: ______________________________ Date: _______________________

Dr. Elisha T. O. Opiyo

Senior Lecturer

School of Computing and Informatics

University of Nairobi

ACKNOWLEDGEMENT

I thank the Almighty God for the good health and resources He accorded me throughout my life and especially during the project period. I also remember my beloved family for their love and encouragement in the academic years that kept me going. To my supervisor Dr. Elisha Opiyo for whose advise without which I wouldn’t have completed this project. I want to thank the panelist prof Robert Obok whose guidance enabled me to enrich the project. And to all my classmates, with special respects to Charles Mutinda, Alexio Manyonde and Cedric Shikoli who shared many ideas and assistance during this project.

May our Great Lord bless you all abundantly.

ABSTRACT

Rapid increase of motorization, increased business and lack of alternative means for cargo transport leads to congestion at weighbridges resulting in delays in clearance at the axle-load control facility. Various weighbridge layouts have been recommended depending on the volume of trucks to be checked but this also have had challenges such as weighbridge breakdowns and unprecedented increase in traffic volumes. Machine learning algorithms that predict overloaded trucks with use of previous weighbridge data can contribute to the weighbridge traffic efficiency by identifying overloaded trucks that are then subjected to mandatory checks, clearing those predicted as within permissible limits. Prediction is a technique that generalizes the trends in the data that can then be applied to new instances. Data continually collected at the weighbridge facility was analyzed and wheel configuration, unit axle load and gross vehicle weight considered for prediction of overloaded trucks. The data was preprocessed and split in 2/3 and 1/3 partitions, and further in the ratio 10:1 of training and testing datasets respectively and loaded onto the Waikato Environment for Knowledge Analysis (WEKA). In this research performances of Random Forest, J 48 algorithm, Naïve Bayes, Multilayer Perceptron and PART algorithms were analyzed and a model of PART algorithm that had an accuracy of 73.1% deployed. 50 new unknown data instances were used to evaluate the model and a prediction accuracy of 88% was recorded.

Table of Contents

DECLARATION i

ACKNOWLEDGEMENT ii

ABSTRACT iii

LIST OF FIGURES vii

DEFINITION OF TERMS ix

CHAPTER 1: INTRODUCTION 1

Background 1

1.2 Problem Statement 5

1.3 Objectives 6

Specific objectives 6

1.4 Scope and limitation of the Study 6

1.5 Justification of the Study 7

1.6 Stakeholders 7

1.7 Organization of the Thesis 8

CHAPTER 2: LITERATURE REVIEW 9

2.1 Introduction 9

2.2 Permissible gross vehicle weights and unit axle loads 9

2.3 Weighbridge Selection, Layout, Installation and Operation 10

2.3.1Types of weighbridges 11

2.3.2Fixed Weighbridges 11

2.3.3 Single axle weighbridge 11

2.3.4 Axle unit weighbridges 12

2.3.5 Multi-deck weighbridges 12
2.3.6 Portable (Mobile) Weighbridges........ 13
2.4 Method of using the weighbridges........ 13
2.4.1 Static method 13
2.4.2 Dynamic method or weigh-in-motion (WIM)........ 13
2.5 Weighbridge Accuracy for Prosecution Purposes........ 14
2.6 Weighbridge Layout and Installation........ 15
2.6.1 Full Traffic Control Centre (FTCC)........ 16
2.6.2 Type 1 Traffic Control Center (TCC 1) 16
2.6.3 Type 5 Traffic Control Center (TCC 5) 17
2.6.4 Lay-by Control Center (LCC)........ 17
2.7 Type of weighbridge in relation to truck volumes and class of road 18
2.8 Weighbridge cost estimates 18
2.9 Weighbridge Data Collection........ 19
2.10 Machine learning 20
2.10.1 Supervised Learning 21
2.11 Classification Algorithms 22
2.11.1 Naïve Bayes Classifier........ 23
2.11.2 J 48 Algorithm 24
2.11.3 Multilayer perceptron........ 24
2.11.4 Random Forest Algorithm 25
2.11.5 PART Algorithm........ 27
2.12 WEKA 3.8 Toolkit........ 28
2.13 Confusion Matrix........ 30
2.14 Related work........ 31
2.14.1 Evaluation of weigh-in-motion axle load management system 31
2.14.2 Prediction of Axle Loads 32
2.14.3 Vehicle Overloading Management System........ 32
2.14.4 Algorithm application on Hypothyroid database for analysis........ 32
2.14.5 Algorithms for Intrusion Detection System 33
2.14.6 Student Retention System 33
2.15 The Gap........ 34
2.16 Conceptual Model........ 35

2.17: Overload Prediction Prototype 35
2.18 Summary of literature review 37
CHAPTER 3: METHODOLOGY 38
3.1 Introduction 38
3.2 Research Design 38
3.3 Business Understanding 38
3.4 Understanding Data 39
3.5 Data Preparation 39
3.6 Modelling 40
3.7 Evaluation 40
3.8 Deployment 41
CHAPTER 4: ANALYSIS, DESIGN AND IMPLEMENTATION 42
4.1 Introduction 42
4.2 Analysis of the Raw Database 42
4.3 Selection of Attributes 43
4.4 Analysis of the training Dataset 45
4.5 Design of the Model 46
4.6 Implementation 48
4.7 Evaluation of the prototype 50
CHAPTER 5: RESULTS AND DISCUSSION 51
5.1 The Performance of Machine Learning Algorithms 51
5.2 Comparison of classifiers 52
5.3 Discussion 56
CHAPTER 6: CONCLUSION AND RECOMMENDATIONS 59
5.1 Conclusion 59
5.2 Contributions 60
5.3 Future Work 60
REFERENCES 62
APPENDICES 65
Appendix 1: Evaluation of the Overload Prediction System on New Data Instances 65

LIST OF FIGURES

Figure 1: extract from Traffic Act on permissible axle loads for various wheel configurations 2
Figure 2: long traffic queue approaching Weighbridge 4
Figure 3: An overloaded truck 4
Figure 4: extracts from the Traffic Act, cap403 (revised 2018) 10
Figure 5: A single axle-weigher 12
Figure 6: Typical full traffic control center (FTCC) layout 16
Figure 7: Typical TCC1 facility 17
Figure 8: Typical layout of a TCC5 17
Figure 9: Typical layout of lay-by with HSWIM screening device 18
Figure 10: A Simple regression mapping function 21
Figure 11: Binary and Multi-class Classifications 22
Figure 12: two layer feed-forward neural network (MLP) 25
Figure 13: How Random Forest Algorithm works 26
Figure 14: how the PART algorithm builds a partial tree 28
Figure 15: WEKA user interface 29
Figure 16: fields of the confusion matrix 31
Figure 17: Conceptual Model 35
Figure 18: The overload prediction prototype 36
Figure 19: Process Chart for the overload prediction system 36
Figure 20: Proposed integrated layout for the prediction system the FTCC 37
Figure 21: Raw data captured continuously at Gilgil Weighbridge station 43
Figure 22: Attributes impact on prediction 44
Figure 23: arff file for loading on the WEKA 45
Figure 24: WEKA preprocessor visualization window for all attributes 46
Figure 25: WEKA output window for classification 47
Figure 26: System administrator interface for registering users 48
Figure 27: Data scientist's interface for training and prediction 49
Figure 28: System interface for predicting new instances of data 50
Figure 29: Classifier performance of accurately predicted instances 53
Figure 30: False positive rate 53
Figure 31: Precision 54
Figure 32: Recall 55
Figure 33: F-measure 55

LIST OF TABLES

Table 1: weights and Measures weighbridge accuracy 14
Table 2: Summary of Weighbridge types and methods of use 15
Table 3: type of weighbridge in relation to truck volumes and class of road 18
Table 4: typical weighbridge cost estimates 19
Table 5: Summary of the Related Work 33
Table 7: Variables in the database 43
Table 8: Performance of algorithms using 3355 and 1617 training and testing datasets respectively 51
Table 9: Algorithm performance at 4594 and 401 training and testing datasets respectively 52

GLOSSARY OF ACRONYMS AND ABBREVIATIONS

ALC-MIS	axle load control- management information system
ANN	Artificial Neural Network
COMESA	Common Market for Eastern and Southern Africa
CRISP-DM	Cross industry standard process for data mining
HSWIM	high speed weigh-in-motion
KeNHA	Kenya National Highway Authority
KeRRA	Kenya Rural Roads Authority
KURA	Kenya Urban Roads Authority
LSWIM	low speed weigh-in-motion
ML	Machine Learning
MLP	Multilayer Perceptron
NB	Naïve Bayes Algorithm
PART	Algorithm based on partial decision trees
RFA	Random Forest Algorithm
SSATP	Sub-Saharan Africa Transport Policy Program
WEKA	Waikato Environment for Knowledge Analysis
WIN	weigh-in-motion

DEFINITION OF TERMS

Axle weigher is a lightweight scale designed with two weigh pads to weigh the wheels of a single vehicle axle.

Check-weigh is the process of subjecting a truck to undergo weighing on a suitable weighing instrument for purposes of ascertaining conformity with the axle and gross overload control regulations.

High speed weigh-in-motion are WIM scales that can will weigh vehicles moving at speeds greater than 15km/h

Low speed weigh-in-motion are WIM scales that weigh vehicles at speeds less than 15km/h

Weigh-in-motion are scales designed to capture, display and record axle weights and gross weight and a vehicle passes over the scale platform without stopping

Weighbridge as per rule 105(b) of the weights and measures Act Cap 513 laws of Kenya means a weighing instrument for weighing loads carried on a vehicle where the vehicle is supported on rails or a platform either of which is linked to a system of levers or load-cells and whose capacity is 1000 kg or more.

CHAPTER 1: INTRODUCTION

This chapter deals with background of the study, the problem statement, objectives, scope and limitations of the study, the stakeholders and expected contribution of the study to society.

Background

Long queues, higher risk of passenger safety, increased fuel consumption and pollution are characteristic at all weighbridge stations in Kenya. The escalating number of vehicles that need to undergo the mandatory check-weighing has been occasioned by breakdown of the rail transport system over the years. Weighbridges are truck scales used to check loading of trucks so as to ascertain that they remain within limits permitted by vehicle manufacturer, to enable for billing the cargo and to keep within legal limits of weight permitted to be transmitted to the roads so as to protect transport infrastructure from deterioration and damage. Public weighbridges and axle weighers are for purposes of maintenance of infrastructure through enforcement of load limits for cargo trucks. Overloading is responsible for many road accidents and weakening and collapse of pavement. The indicative cost of overloading in COMESA Region is estimated at more than US$4 billion per year (Pinard M, 2010).

The Northern Transport Corridor, an international highway connecting Mombasa, Nairobi, Eldoret and Malaba to the neighboring Uganda, Rwanda and South Sudan, has multi-deck weighbridge stations at Mariakani, Mlolongo, Gilgil and Webuye to monitor and control overloading of trucks carrying transit goods (KeNHA, 2016). The highway is used by buses, cars, heavy commercial vehicles, trailers and semi-trailer trucks all of which are affected in one way or another by the weighbridges installed along the highway forcing all these vehicles to form part of the queue to the weighbridge even when most of them are not expected to be check-weighed. Side lanes have been constructed along the main highway to be used by the trucks as they approach the weighbridge platform but are also not adequate to eliminate the long queues by other vehicles that are not subject of the mandatory checks such as light commercial vehicles, cars and buses.

Congestion at the weighbridge installations continue to be witnessed as a result of increased volumes of trucks that must undergo mandatory check-weighing in bid to control axle-load for purposes of preserving road safety and infrastructure. Load limits are set by legislation in the Traffic Act so that trucks do not carry excess loads that are known to impact negatively on the life span of the road pavements and proper control of the vehicles.

Trucks exceeding 3.5tonnes are required to be check-weighed at weighbridge installations to ensure that they are of a weight not more than that permitted on the roads (Traffic Act Cap 403, 2015). Because of the large traffic involved, long queues are experienced causing big loss to the economy through waste of time resulting in increased cost of doing business. Measures by Government have not helped much to alleviate the traffic congestion at the weighbridge installations.

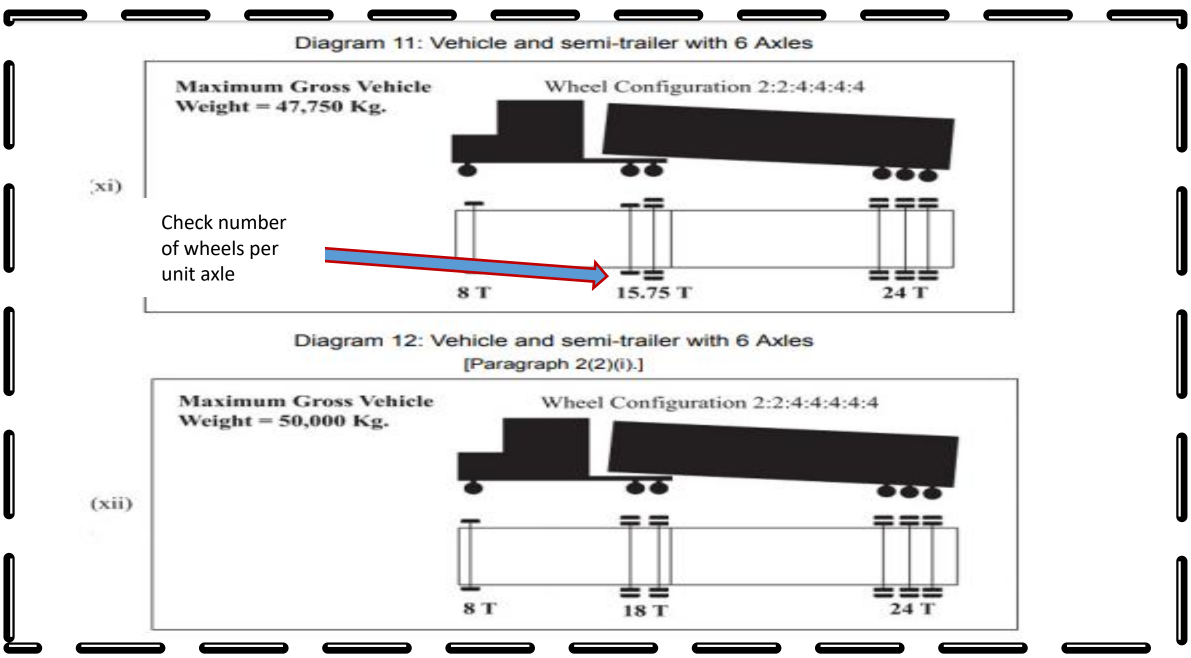

Figure 1: extract from Traffic Act on permissible axle loads for various wheel configurations

Road network infrastructure are expensive to construct and maintain and therefore efforts by authorities to preserve them need to be complemented. The weighbridges serve this purpose of indicating the gross weight of the truck and at times indicate the axle weight. Weighbridges are installed along major highways and their use involves a truck moving onto and stopping on the weighbridge platform, its weight indicated on the weighbridge indicator, recorded into auxiliary memory then the truck can leave the platform as directed by the weighbridge operator. This process in not instantaneous, it takes time leading to traffic snarl-up as vehicles have to wait for those ahead to them to be weighed before them.

The purpose of check-weighing is to determine vehicle loads and any vehicle with loads beyond the prescribed limits of gross weight or axle limit be prohibited from using the roads and is charged for the overload. Heavy trucks that are loaded within legal limits cause relatively less damage to road pavement structure compared to 60% damage occasioned by the overloaded trucks (Reddy R,). When roads are properly constructed and maintained well, transport costs will reduce; agriculture, tourism and industries will grow and increase employment opportunities and better lives for the citizens.

Overloading is addressed at the port of Mombasa by use of smart cranes that indicate the weight of cargo containers as the container is being hauled onto the truck. Such measure informs the transporter the weight of cargo being transported.

Cargo originating from rural agricultural set ups where weighbridges or other suitable weighing equipment are not available will not be of known weight and are likely to be loaded beyond capacity if profits are to be maximized.

Certain types of cargo for their nature tend to be unconsciously loaded beyond limit either by poor distribution so that the axle unit load is exceeded or by their compact nature they leave too much space in the truck luring loaders to add more cargo.

According to SSTATP guidelines, challenges have included lack of harmonization of axle load and gross weight limits among countries that are along a transport corridor resulting in many truck being detained at the holding yard to redistribute or reduce the cargo carried. Lack of trust in systems applicable in different countries have resulted in vehicles being weighed a few miles apart on either sides of their boundary.

Figure 2: long traffic queue approaching Weighbridge

Figure 3: An overloaded truck

The layout design of the weighbridge installation is dependent on the purpose, traffic capacity and installation capital. The weighbridge layout at Gilgil weighbridge station comprises of a 160tonne capacity multi-deck weighbridge on the Nakuru bound side and two low speed weigh-in-motion (LSWIM), one on either side of the highway for pre-screening overloaded trucks. This layout results from more transit cargo flowing from imports to the neighboring countries from international community an indication that they together with Kenya have lower exports.

Installations that have the weigh-in-motion (WIM) have a great extend improved traffic flow but also has challenges during high peaks where queues build up and trucks have to wait for long time before they are screened.

The initiative for overload control is to facilitate trade and obviate undue advantage by the unscrupulous transporters from gaining excess profits at the expense of fast deteriorating roads. To do better trade, cargo transporters must take minimal time at the weighbridges and continue with their journey to deliver cargo to destination.

In this paper, we propose an optimal traffic management control involving machine learning solutions where historical data, a combination of real-time information where new-age algorithms are manipulated to predict suspected overloaded trucks, directing them to screening lanes and allowing others to proceed and so improve performance for seamless traffic flow and proficiently manage weighbridge assets. Computers today can perform tasks without explicitly programming them to in a process of machine learning. The learning by a computer begins by analyzing data and looking for patterns in the data that help in future prediction without help of human intervention. It is a key component in intelligent information systems for generalization as opined by Holmes G. et al (1994).

Ondiek M. (2017) opines that use of big data in the telecommunication enhanced quality of service, minimized congestion and was used to manage fraud and avoid losses.

The growing focus on the development of intelligent network systems and use of machine learning algorithms will assist the traffic management and result in reduced congestion and traffic snarl up. These classification algorithms will result in helping service providers at the various weigh bridge stations to provide efficiency of the current transportation system, estimate the transport models, and predict future network scenarios. Data analytics tools such as WEKA is a tool that can be used for data mining and visualization. This tool uses machine learning algorithms to manipulate datasets that are loaded onto the toolkit to make classifications or predictions on new data after training on previous or historical data.

1.2 Problem Statement

Rapid increase of motorization, increased business and lack of alternative means for cargo transport leads to congestion at weighbridges. It is mandatory for all trucks exceeding 3.5 tonnes to be checked at axle load control facilities for compliance with load limits prescribed in law. Various weighbridge layouts have been recommended depending on the volume of trucks to be check-weighed but this also has had challenges such as breakdown and unprecedented increase in traffic volumes. This leads to waste of time at weighbridge facilities resulting in increase in the cost of doing cargo transportation. There is need to screen the trucks and one way is to come up with predictive measures. In this case we look for a way of predicting whether a truck may or may not overloaded. Machine learning algorithms that predict overloaded trucks with use of previous data can contribute to the weighbridge traffic efficiency. Such tools when used will identify trucks that can then be signaled to enter a side lane leading to a static weighbridge without interfering with traffic flow on the main highway. Lack of machine learning tools for prediction of overloaded trucks subject to mandatory check-weighing by axle-load control institution has compounded efficiency at the weighbridge installation.

1.3 Objectives

The main objective of this research is to develop and evaluate a prototype system that predicts overloaded trucks based on the best performing machine learning algorithm.

Specific objectives

1. To determine the current axle-load control operations at weighbridge facilities.
2. To investigate the performances of the machine learning algorithms such as Random Forest Algorithm, Multilayer Perceptron, Naïve Bayes, J48 algorithm and the PART algorithms on weighbridge data.
3. Design the model for prediction of overloaded trucks based on the promising machine learning algorithm.
4. Implement a prototype for the overload prediction system.
5. Evaluate the prototype system using independent data

1.4 Scope and limitation of the Study

This study was limited to static weighbridge operations at Gilgil station located along the Northern Transport Corridor in Nakuru County. The data used was input manually with bias to overloaded trucks that were checked on the static weighbridge. Trucks that could be redistributed to within legal limits were treated as OK and those that could not were treated as EXCESS. The study was also limited to compare only five machine learning algorithms on the WEKA toolkit. In this study data concerning special release of overloaded trucks was not shared to the researcher.

1.5 Justification of the Study

The Kenya Government through KeNHA is responsible for the management, development, rehabilitation, and maintenance of Class A, B and C roads. For other classes of roads, various authorities have been established by legislation to manage, develop, rehabilitate and maintain them. KURA does the national urban trunk roads. KeRRA is mandated to develop, construct and maintain rural road network (Kenya Roads Act, 2007). All these authorities have the general mandate that include overload control and roads maintenance. A hooping ksh 176.75 billion was allocated for Road Transport Program in the Kenya national budget estimates in the Financial year 2016/2017 (Kinuthia J. et al) a big raise from ksh. 40.54 billion allocated the previous year. The ultimate goal of spending so much money on maintenance of roads and even having weighbridges installed along major roads is to facilitate trade and investments. Weighbridges are designed to allow for maximum flow of traffic (Odula V, 2016), yet this has not been achievable at Gilgil weighbridge facility as long queues continue to be witnessed causing increase in the cost of road cargo transportation. The increase of cargo haulage using road transportation results from the breakdown of railway transport and increased business between Kenya, its neighbors and the international community. The weigh-in-motion scales have done little to alleviate this problem. Breakdowns are also common at this facility and a maintenance company technical staff are always on standby to attend to the weighbridge when faults occur.

A lot of these expenses will be reduced by use technology in ICT. Efficient traffic management involving classification algorithms in machine learning where historical data, a combination of real-time information with new-age algorithms can be manipulated to predict overloaded trucks to improve performance for seamless traffic flow and proficiently manage weighbridge assets.

1.6 Stakeholders

The stakeholders that will be affected by the system are as listed below

1. The ministry of transport and infrastructure officials responsible for overload control. They are responsible for budget of maintenance of the road infrastructure. Low budget will be realized with the use of the system to replace physical weighbridges. They interact with the system by providing weighbridge data that will be used in the prediction of new instances.

2. Traffic police seconded to the weighbridge installation. More efficient service will be provided with the system automation. This category of users will query the system for trucks that are suspected to be overloaded or where the category is not known.
3. Transporters and Traders. When they are able to interact with the system, they will be able to predict which trucks are likely to fail compliance requirements and so take corrective measures that will save them of hefty court fines they are subjected to when found guilty of non-compliance with axle load control regulations.

1.7 Organization of the Thesis

In this chapter we have covered the background of the study, problem statement, objectives, scope, justification and contribution of the research to the stakeholders. In chapter 2, analysis of the literature review is done and highlights of the research methodology is covered in the chapter 3. Analysis, Design and implementation are detailed in chapter 4. Results and discussion of the project are highlighted in chapter 5, then conclusion and recommendations in chapter 6.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter will discuss permissible axle loads, weighbridge selection, layout, installation and operation; machine learning classification algorithms, WEKA 3.8 tool, classification algorithms namely Naïve Bayes, J48, Multilayer Perceptron, PART and Random Forest in their support for prediction of overloaded truck and related work.

2.2 Permissible gross vehicle weights and unit axle loads

The Traffic Act of Kenya cap 403 (2018) prescribes the maximum permissible Gross Vehicle Weight for all vehicles allowed on the Kenyan roads depending on their wheel configuration. The wheel configuration is classified according to the type of steering, the number of wheels per axle and the number of axles a vehicle has. The classification is as follows; 2*,2A,3*,3A,4*,4A,5*,5D,6*,6A,6C,6I,6G,7*; where the digit represents the number of axles; the alphabet represents the arrangement of wheels on the axles and the asterisk is used to denotes vehicles with single wheels that are larger than usual (super-single) in the axles. Each configuration has a maximum permissible load for both the axles and the gross vehicle load. A vehicle with two axles has a maximum of 18,000kg and one of three axles, 26,000kg as prescribed in the Traffic Act. The highest is 56,000kg which is a vehicle and a draw-bar trailer with a total of 7 axles(Figure 4 below). (Kenya Traffic Act, 2018). Configuration 2* stands for a vehicle with two axles, both axles fixes to a single wheel each; 2A has a single wheel each for the front (steering) axle and double wheels for the rear axle.

2. (1) The maximum weight which may be transmitted to the road in the case of a vehicle fitted with pneumatic tyres, whether laden or unladen, shall not exceed—

(a)	by way of any single steering axle whether controlled by drawbar or driver operated steering mechanism	8,000 kg.
(b)	by way of any single axle having four or more wheels fitted with conventional tyres	10,500 kg.
(c)	by way of any single axle having two wheels fitted with super single tyres	8,500 kg.
(d)	by way of any driver operated double steering axles having two wheels fitted with conventional tyres	12,000 kg.
(e)	by way of any drawbar controlled double steering axles having four wheels fitted with conventional tyres	16,000 kg.
(f)	by way of any tandem axle group each having two wheels fitted with Super single tyres	16,000 kg.
(g)	by way of any tandem axle group each having four wheels fitted with conventional tyres	18,000 kg.
(h)	by way of any triple axle group each having two wheels fitted with Super single tyres	22,500 kg.
(i)	by way of a triple axle group each having four wheels fitted with Super single tyres	24,000 kg.

Provided that in the case of any axle or axle group (excluding a single steering axle) where one or more of the axles is having only two wheels fitted with conventional tyres, the appropriate maximum allowable load figure given in this subparagraph shall be reduced by twenty-five per cent.

(2) Notwithstanding subparagraph (1), the maximum total weight of any vehicle or combination of vehicles fitted with pneumatic tyres, whether laden or unladen, shall not exceed the following—

(a)	vehicle with two axles	18,000 kg.
(b)	vehicle with three axles	26,000 kg.
(c)	vehicle and semi-trailer with a total of three axles	28,000 kg.
(d)	vehicle with four axles	30,000 kg.
(e)	vehicle and semi-trailer with a total of four axles	36,000 kg.
(f)	vehicle and draw bar trailer with a total of 4 axles	36,000 kg.
(g)	vehicle and semi-trailer with a total of five axles	44,000 kg.
(h)	vehicle and draw bar trailer with a total of five axles	44,000 kg.
(i)	vehicle and semi-trailer with a total of six axles	50,000 kg.
(j)	vehicle and draw bar trailer with a total of six axles	52,000 kg.

(3) The maximum total weight of a vehicle fitted with solid tyres, lawfully on a road under these Rules, shall be seventy-five per cent of the maximum weight specified for a similar type of vehicle with pneumatic tyres under subparagraph (1) or (2).

(4) Not more than seventy-five per cent of the laden weight of any motor vehicle (other than a motorcycle) shall be transmitted to the road by any two wheels of the vehicle.

Figure 4: extracts from the Traffic Act, cap403 (revised 2018)

2.3 Weighbridge Selection, Layout, Installation and Operation

The choice of weighbridge will be determined by the purpose it will serve, the cost involved in the procurement and installation, the site it is to be located and the user institution (Katahira, et al, 2015).

The weighbridge used in the context of this thesis, for control of overloading tendency among transporters. It may be used to determine payload among transporters and traders. It may be used as a tool to collect cess in the county governments.

2.3.1Types of weighbridges

There are several types of weighbridges that can be used in the management of control of overweight as have been discussed in the Sub-Saharan Africa Transport Policy Program (SATPP), Working Paper No. 90 (2010) and in the Road Transport Safety and Axle Load Control Study in Nepal (2015). They include fixed and portable weighbridges. Strategy of their use depends on their construction, accuracies desired and ease of operation.

2.3.2Fixed Weighbridges

These are weighbridges that have their platform permanently fixed on the road surface or side of the road. They are expensive to install and have limited placement and coverage. They are easy to operate and allow for cargo off-loading and redistribution. These weighbridges are of the following specifications;

2.3.3 Single axle weighbridge

These are small platform scales that weigh an axle at time and the operator has to add up the individual axle readings to determine the vehicle gross weight (Figure 5 below). Their small size makes it possible to move them to different locations where they can be installed in pre-constructed recesses. They are slower to complete weighing of a truck.

Figure 5: A single axle-weigher

2.3.4 Axle unit weighbridges

These are single-deck weighbridges with platform typically 3.2 m x 3 m to 3.2 m x 4 m supported on a 4-load cell weighing mechanism capable of weighing any axle unit of a truck (single axle, tandem or tridem unit) at an instance. They are faster and more accurate than the single axle weighbridges.

2.3.5 Multi-deck weighbridges

The platform of the multi-deck comprises of a number of decks each supported by its own weighing mechanism (typically 4 load cells, one at each corner). This weighbridge is capable of weighing a multi-axle heavy vehicle in one operation. It is faster and more accurate than the single axle and the axle unit weighbridges

2.3.6 Portable (Mobile) Weighbridges

These are weighbridges other than those permanently fixed on the road surface of road side used for unscheduled check-weighing of vehicles on any section of the roads. These weighbridges comprise of wheel scale platforms which are places on the road surface, the wheel of axles weighed, the total axle load obtained by summing the wheel loads (Katahira et al, 2015).

The scales will not give accurate results unless the leveling mats or rams are used to align the levels of all axles. The scale can also be put in a specially constructed pit in a lay-by.

Mobile weighbridges can be transported and used in any location, so wide coverage area. They are easily damaged while in use or during transportation and need frequent calibration. Police cooperation is necessary to stop and direct vehicles to the weighbridge.

2.4 Method of using the weighbridges

There are two methods of vehicle weighing; static and dynamic. Both methods are applicable to fixed and mobile weighbridges.

2.4.1 Static method

When static, the truck must move onto the weighbridge platform and stop, then weighment done before it moves away. There is higher accuracy, the scale can reliably be used for prosecution purposes but the operation is slow and time consuming

2.4.2 Dynamic method or weigh-in-motion (WIM)

For the dynamic type, the truck moves at a uniform specified speed as it passes over the weighbridge platform without stopping. The weigh-in-motion allows trucks to be weighed in the traffic flow without disruption (Reddy R, 2015). A WIM system measures the dynamic axle mass of a moving truck to estimate the corresponding static axle mass. It's effective for rapid monitoring of vehicles but has low accuracy and cannot be used for prosecution.

The WIM systems are divided into two, the High Speed WIM (HSWIM) for vehicle speeds greater than 15km/h, and Low Speed WIM (LSWIM) for vehicle speeds less or equal to 15km/h. WIMS have traditionally been used for screening of overloaded vehicles that need closer attention rather than enforcement purposes at or near static weighbridges.

2.5 Weighbridge Accuracy for Prosecution Purposes.

A weighbridge to be used for prosecution must be approved for such use by Director of Weights and Measures, and be verified and stamped by an inspector of weights and measures and be issued with a certificate of verification as proof of its accuracy. An extract of maximum permissible errors for various weighbridge capacities is shown in table 1 below.

Table 1: weights and Measures weighbridge accuracy

[Rev. 2012] **CAP. 513**

Weights and Measures

[Subsidiary]

Capacity of Instrument	Sensitivity when Fully Loaded		Maximum Permissible Error in Excess or Deficiency when Fully Loaded			
	On Verification	*On Inspection or Re-verification*	*On Verification*		*On Inspection or Re-verification*	
			Non-Self Indicating Type	Self or Semi-self Indicating Type	Non-Self Indicating Type	Self or Semi-Self Indicating Type
1	2	3	4	5	6	7
80 tonne	6 kg.	12 kg.	10 kg.		20 kg.	
100 tonne	6.5 kg.	13 kg.	11.5 kg.		23 kg.	
150 tonne	8 kg.	16 kg.	15 kg.		30 kg.	
200 tonne	9 kg.	18 kg.	19 kg.		38 kg.	
250 tonne	12 kg.	24 kg.	25 kg.		50 kg.	
300 tonne	15 kg.	30 kg.	30 kg.		60 kg.	
400 tonne	20 kg.	40 kg.	40 kg.		80 kg.	

Source weights and measures Act cap 513 laws of Kenya (2012)

Table 2: Summary of Weighbridge types and methods of use

Type of weighbridges / Method of weighing		Fixed weighbridges	Mobile weighbridges
		• Easy to operate • Minimum personnel • Cargo off-loading • High installation costs • Limited placement	• Wide coverage • Difficult site selection • High operating costs • Equipment easily damaged • Police cooperation • Traffic disruption
Static	• More precision • Accepted for legal enforcement • Slower	• Easiest to operate • Highest level of precision • Can weigh and register axle groups	• Lowest investment • Optimal for enforcement
Dynamic	• Rapid monitoring • Lower precision • Not acceptable for enforcement	• Rapid monitoring • Requires large installation • Requires careful direction of vehicles • Generally not accepted for enforcement	• Minimum disruption of commercial traffic • Lowest accuracy • Excellent for statistical moni-toring

Source: Sub-Saharan Africa Transport Policy Program (SATPP), Working Paper No. 90 (2010)

2.6 Weighbridge Layout and Installation

The weighbridge layout is subject to the following factors; Purpose of the facility, the volume of traffic expected at the facility, whether there will be screening of heavy vehicles, and prosecution of overloaded vehicles.

The following layouts are applicable depending on the needs for the weighbridge; Full Traffic Control Center (FTCC), Type 1 Traffic Control Center (TCC 1), Type 5 Traffic Control Center (TCC 5) and Lay-by Control Center (LCC).

2.6.1 Full Traffic Control Centre (FTCC)

This comprises of a full range of facilities to fully undertake overload control without disrupting the flow of high volumes traffic on either side of the road (Figure 6 below) and typically includes within its operations the following;

i. A high-speed weigh-in-motion (HSWIM) screening device in the main road.
ii. A low-speed weigh-in-motion (LSWIM) screening device to confirm vehicles suspected to be overloaded as indicated by the HSWIM in the screening lane; and
iii. A static multi-deck weighbridge for accurately weighing axle and axle unit loads and total vehicle or combination mass for prosecution purposes

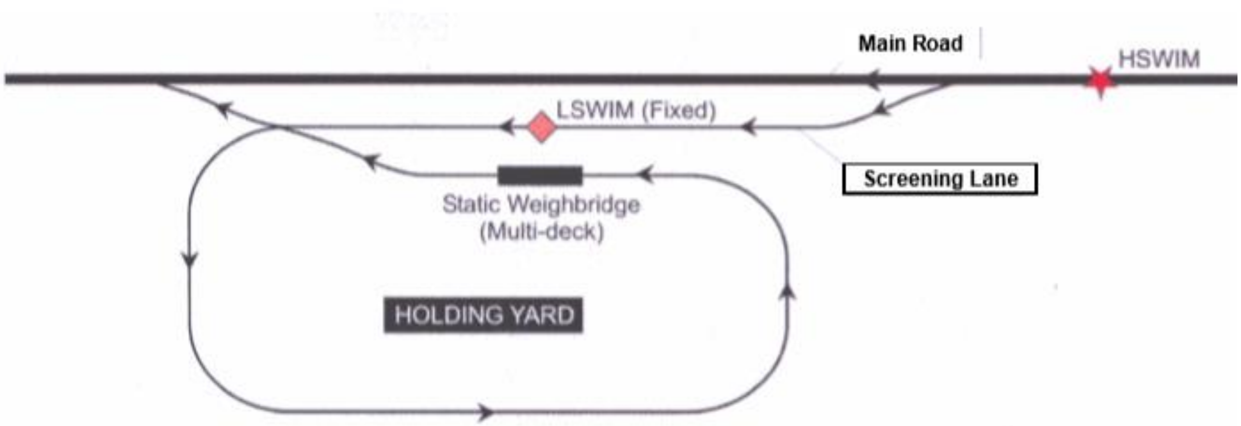

Figure 6: Typical full traffic control center (FTCC) layout

2.6.2 Type 1 Traffic Control Center (TCC 1)

This layout is similar to the FTCC only that it operates on one side of the road. The HSWIM is located in the screening lane. A major challenge is that vehicle travelling on the other side of the road that are identified by the HSWIM as overloaded must cross the opposing traffic to get to the multi-deck to be weighed (Figure 7 below).

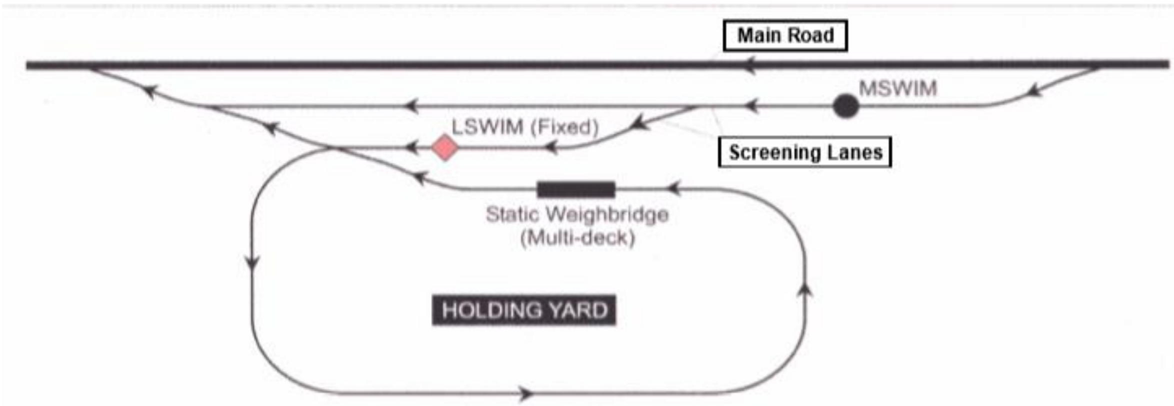

Figure 7: Typical TCC1 facility

2.6.3 Type 5 Traffic Control Center (TCC 5)

This system requires all heavy vehicles to leave the main road and cross over a LSWIM. Compliant vehicles will continue with their journey but suspected overloaded have to move to the static scale for weighing and prosecution (Figure 8).

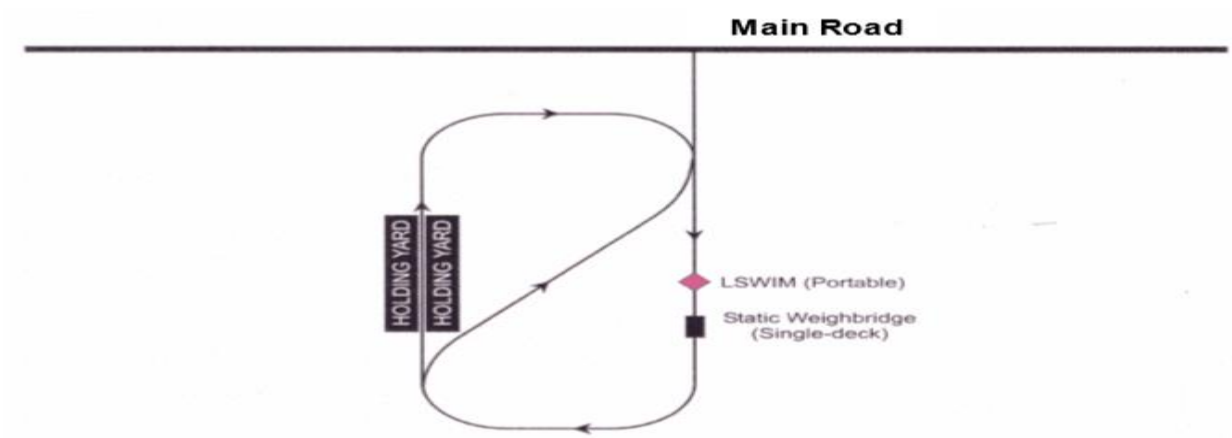

Figure 8: Typical layout of a TCC5

2.6.4 Lay-by Control Center (LCC)

The facility comprises a level concrete platform constructed alongside the road where the scale can be installed. It can be operated together with a HSWIM as a screening device (Figure 9).

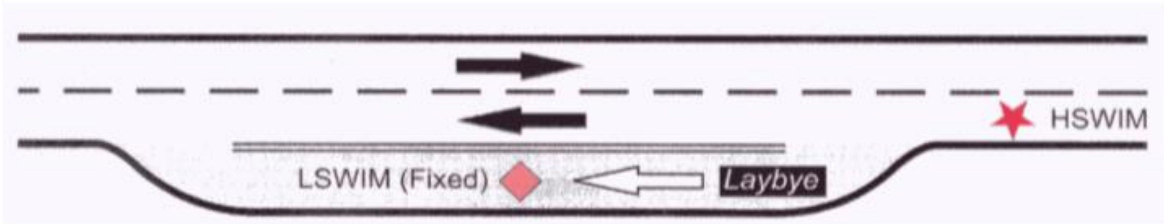

Figure 9: Typical layout of lay-by with HSWIM screening device

2.7 Type of weighbridge in relation to truck volumes and class of road

Gilgil weighbridge station is situated along a Class A road which requires multi-deck weighbridges on both sides of the road. Installation of weighbridges on Class A and B roads is relatively costly due to such requirements as type of scale, the use of pre-screening WIMs, large parking and stacking facilities and office accommodation to handle larger numbers of trucks and staff. The type of weighbridge and the volumes they are expected to handle are as provided in table 3 below.

Table 3: type of weighbridge in relation to truck volumes and class of road

Weighbridge type	*Traffic volumes (Heavy vehicles/day)*	*Road class*
Multi deck scale (both sides of freeway)	> 4 000	A
Multi deck scales (four decks)	1 000 – 4 000	B
Multi deck scale (four decks)	500 – 1 000	C
Axle unit scale	< 500	D
Single axle scale	< 500	D

2.8 Weighbridge cost estimates

Weighbridges are expensive equipment that require proper planning for their procurement and subsequent installation. Table 4 below show the estimate costs of various weighbridges

Table 4: typical weighbridge cost estimates

Weighbridge type	*Cost estimate (US$ - 2008)*
Single axle scale	0.4 to 1.0 million
Axle unit scale	0.4 to 1.0 million
Multi-deck scale (four decks)	1.0 to 2.0 million
Multi-deck scales (four decks)	2.0 to 4.0 million
Multi-deck scale (both sides of freeway)	6.0 to 8.0 million

2.9 Weighbridge Data Collection

Activities at the weighbridge involve weighing of the trucks and where a truck is found to be loaded beyond such limits as may be prescribed by legislation say the traffic Act, then punitive measures are taken against any such offender. Evidence would include information that would satisfactorily identify the offending truck, the goods carried, origin and destination of the goods; and the driver and his/her employer. The collection of this information will either be manually or electronically and is essential for effective monitoring of overload control operations.

Data will be entered into the computer by the weighbridge operator a process that is prone to typing errors. With advancement of technology incorporated in the weighbridge data collection has become more accurate and cost effective.

After collecting the data for period of time, analysis can be done to determine certain statistical overload controls and long term trends. Data analysis should create statistics and trends that will be of interest from various perspectives (strategic, management, administrative, financial, technical, etc.) of vehicle overloading control as discussed by Katahira et al (2015) in Road Transport Safety and Axle Load Control Study in Nepal.

As it is for many management systems data is collected for analysis so as to improve the operations at the weighbridge facility. Where the weighbridge records the various routes, an analysis of the degree of overload will inform routes prone to overloading. Weighing data with names of transporters that own a fleet of trucks will inform law enforcers of the most frequent offenders and award necessary ratings depending on the frequency of occurrences. Analysis of weigh data per commodity type will inform law enforcers to focus on such problem commodities and track their origin e.g. certain mines, quarries, etc (Pinard M, 2010).

2.10 Machine learning

It is the science of making computers perform tasks without explicitly programming them to. The computer learns through experience and improves without human assistance. The learning by a computer begins by observing data and looking for patterns in the data that help in future prediction without help of human intervention. It is a key component in intelligent information systems for generalization as opined by Holmes G. et al (1994). Data used by the computer for learning is the training data, and may have input and outputs. Using good quality data and algorithms, computers will built models. To test that learning has occurred testing data is used.

A computer is said to learn from experience if it performs a task in a better way after the experience. In traditional programming, input data and a program are run on a computer to give an output. In machine learning (ML), input and output data are fed onto the machine during training and the machine creates its own code that can be evaluated by testing using new instances of data. ML is used in many fields including health in the detection of cancer by observing cell images in slides. The ML model is trained on a large database of good quality image data of the cell slides, and can later be relied upon detect cancer cells when new instances are tested on it.

ML methods are either supervised, unsupervised or reinforced algorithms. When supervised, labelled examples with known inputs and target outputs are used to train the algorithm.

Unsupervised learning is used when the data is not labelled and the algorithm has to infer a pattern in the unlabeled data. The algorithm restructures the data into a pattern representing a class or a series with new features that may provide meaningful insights about the data.

In reinforced learning, the algorithm interacts with the environment producing actions and making necessary corrections as it learns. This case is similar to the unsupervised learning but the algorithm is allowed to have positive or negative feedback depending on the action taken. It is the learning by trial and error iterations with the environment.

2.10.1 Supervised Learning

The supervised machine learning model can then make predictions on new data inputs to provide the expected outputs. In supervised learning, we have input variable (x) and an output variable (y). The function y = f(x) that maps inputs to the outputs is the model that is built from the algorithm. This mapping function can then be used to for prediction of new instances of the input variable. The C4.5 program is an example of supervised learning algorithm that inputs a set of labelled data and generates a decision tree as output as insinuated by Vijayarani et al, 2011.

Supervised learning can further be split into two groups, regression and classification.

In regression, the attribute is a continuous value such as predicting the height of a person given the weight of the person. Linear regression is the simplest as it is a mapping function of a straight line.

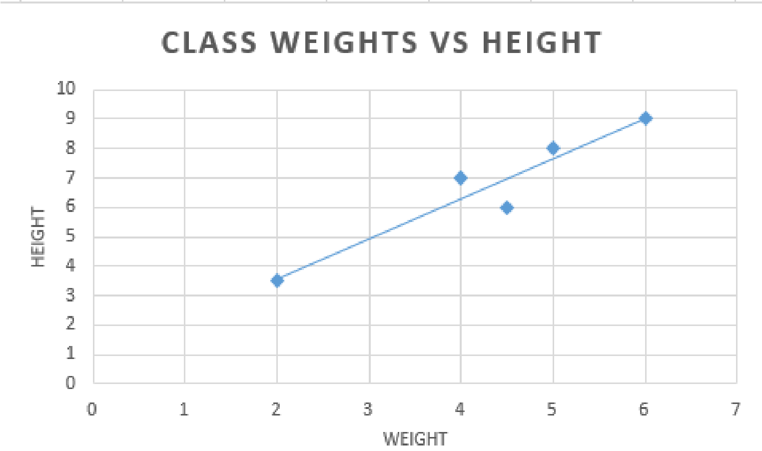

Figure 10: A Simple regression mapping function

Given a weight, the mapping function can be used to predict the height of a student in the class as illustrated in the Figure 10 above.

In machine learning, classification is a supervised learning approach where computer program learns from data and uses what it has learned to classify new instances.

2.11 Classification Algorithms

This is the process of arrangement of things into groups with observed similar parameters. In machine learning, it is the process of mapping the class to which a new set of categories belong on the that a training data set whose membership categories is known. Classification assigns items in a group to target class (Parsania et al, 2014). Classification is a technique used for prediction of members of a group with similar data instances (Vajayarani et al, 2011).

There are two types of classification, binary where the data is categorized into two distinct classes like say TRUE and FALSE or YES and NO. The second type of classification is the multi-class classification where the class categories are more than two. An illustration of the two classes is shown below in Figure 11.

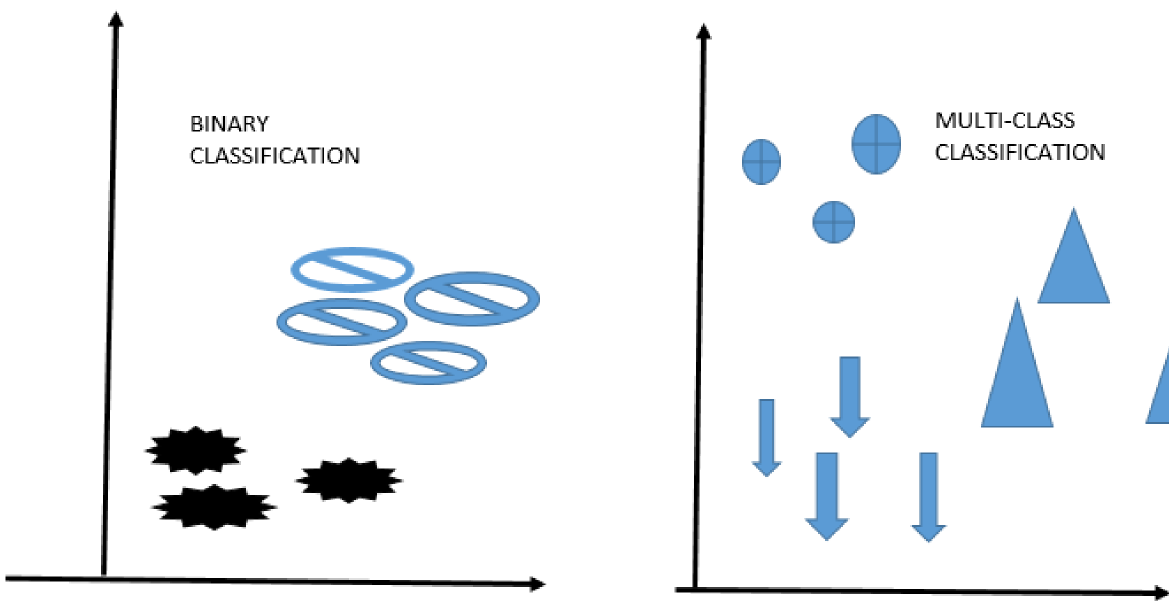

Figure 11: Binary and Multi-class Classifications

In machine learning, classification is a supervised learning approach where computer program learns from data and uses what it has learned to classify new instances. Machine learning algorithms learn a task like say classification using data that is representative of the population. Test data is presumably drawn from the same distribution with labels that are similar to the training data (Dredze et al., 2009).

The algorithm will classify whether a person is male or female, or whether the ball is red or black, whether a cell is cancerous or non-cancerous after analyzing attributes and their respective outcomes and using this learned knowledge to predict. The classification algorithm will learn from the categories into which the data has been labelled then it can categorize new instances with similarities into their related classes.

Classification algorithms in machine learning include Linear Classifiers: Logistic Regression, Naive Bayes Classifier, Support Vector Machines, Decision Trees, J48, PART, Random Forest, Neural Network and Nearest Neighbor.

2.11.1 Naïve Bayes Classifier

It is a machine learning algorithm based on the Bayes family and that uses probability of occurrence of an event to predict the next unknown instance given the prior probability of its likelihood to occur. Example in a box containing 5green and 2red balls, you are mostly likely to pick a green ball than a red one, because the probability of picking a green ball is 5/7 and that of picking a red ball is 2/7. It is reasonable to assume that a ball picked from the box is more likely to be a green ball because green balls have the higher chances of being picked.

The algorithm works with a prior probability and posterior probability. Take for example the data of whether to go out and play or not given weather conditions of outlook, temperature, humidity and wind condition is well computed using the naïve Bayes algorithm where the posterior probability is what we need to calculate given the prior probability that one can play 5times in 14days. Naïve Bayes is a probabilistic algorithm based on the Bayes Theorem that works on the principle of conditional probability of an event occurring given prior knowledge of conditions related to the event.

$P(A/B)= P(B/A)*P(A)/P(B)$

Where; $P(A/B)$ = conditional probability of A given B,

$P(B/A)$ = conditional probability of B given A,

$P(A)$ = probability of event A,

$P(B)$ = probability of event B.

The prediction is calculated from the theorem.

The Naïve Bayes algorithm has the advantage that it is a fast algorithm and can be used in real time prediction. It can work with increased predictor attributes and values. It can handle continuous and discrete data. It is simple on implementation.

2.11.2 J 48 Algorithm

This algorithm is renamed by WEKA project team after the C4.5 algorithm invented by Ross Quinlan. C4.5 was an improvement of the original ID3 algorithm that had challenges of dealing with missing data, normalized values in attributes and tendency to overfitting. It works on the concept of information gain and entropy. As a decision tree algorithm, the J48 builds decision trees from a set of training data. The resultant decision tree can then be used to predict unseen data to test the algorithms ability to generalize. The J48 has additional features such as decision trees pruning, derivation of rules and continuous attribute value ranges. In this algorithm, classification is done recursively until every leaf is pure.

2.11.3 Multilayer perceptron

This is a feed-forward artificial neural network algorithm that comprises of multiple layers; an input layer, a hidden layer and an output layer. Neural networks are an attempt to mimic the biological neural system of the brain (Patterson, 1996). The multilayer perceptron is a supervised learning algorithm that employs back-propagation for training.

By an activation function the weighted inputs are mapped onto the outputs. As illustrated in Figure 12 below, Inputs I_1 and I_2 are fed into the neural networks hidden layer nodes H_1 and H_2 to produce outputs O_1 and O_2. To train the algorithm, these outputs are compared with the target output and errors found are back-propagated towards the input layer thereby updating the weights W_{o12}, W_{o11}, W_{o21} and W_{o22} and thereafter the weights W_{h10}, W_{h20}, W_{h11}, W_{h21}, W_{h12} and W_{h22}. The process is repeated on and on with the purpose of reducing the output error from the target to a minimum. The number of correctly classified instances is the output of the model.

After training, new inputs can then be fed into the system and the predicted outputs will compare favorably with the model, the accuracy of performance similar with the result of the training. Neural networks are used when there is known information and there is need to infer some unknown information based on the known. For example, based on a loan applicant's income, previous history, etc. it's possible to predict whether it will be risky to give or deny the loan. The neural network will have been trained on previous data that contained customer information of those that repaid their loans and those that had issues paying the loan.

The neural network is trained using back propagation algorithm which is a supervised learning algorithm devised by Rumelhart et. al,(1986). The algorithm uses the labelled data i.e known inputs mapped to known outputs, to adjust the bias weights and thresholds in the network nodes in order to reduce errors in the prediction compared to the actual output. The model can then by deployed to make predictions on unknown function that relates to the input data to produce desired outputs, (Fausett, L. 1994).

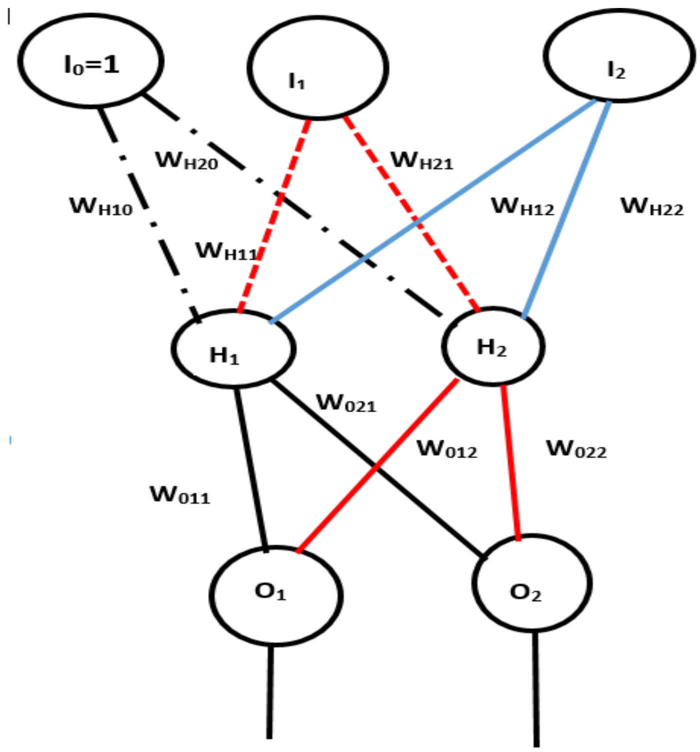

Figure 12: two layer feed-forward neural network (MLP)

2.11.4 Random Forest Algorithm

Random forest algorithm is a supervised learning algorithm used for classification and regression. It is a collection of decision trees with random subsets of features that then built smaller trees. The sub-trees are combined and a decision made based on averages of the results. Random Forest which is a learning algorithm that makes predictions by combining decisions from a sequence of decision trees. The best among the predictors in a node is chosen randomly using variance/mean square error and the out-put is determined by a simple majority vote (Khan M, 2018).

The Random Forest Algorithm works efficiently on large databases and has high accuracy. It maintains accuracy even when a large portion of the data has missing values by estimating the missing values. It does not suffer from overfitting since it creates random subsets of features then builds smaller trees of these subsets and combines the trees to get the simple majority. It has minimal training time. RFA is used in the banking sector to predict customers who have ability to pay to be given loans. It's used in medicine to get correct combination of components required in the dispensing of drugs. It's used in the diagnosis of diseases such as cancer.

In the example in the figure 13 below a dataset containing fruits was manipulated the Random Forest Algorithm using the three tree algorithms each outputting a fruit, the simple majority being the orange.

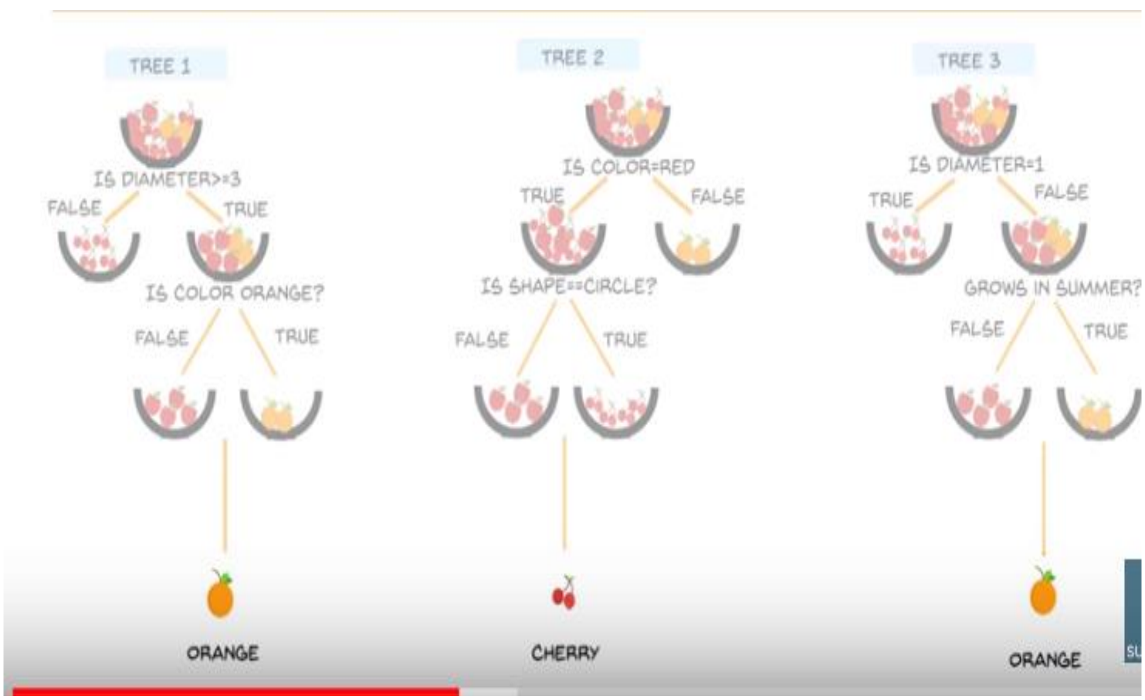

Figure 13: How Random Forest Algorithm works

2.11.5 PART Algorithm

PART is an algorithm based on partial decision trees. It is a rule induction algorithm that combines the C4.5 and the RIPPER algorithm and avoids their respective challenges. It is more accurate than the RIPPER, it is efficient and does not suffer from slow performance exhibited by the C4.5 algorithm on pathological data (Eibe F. et al,1998). The algorithm learns the rules by generating partial decision trees repeatedly using the separate and conquer strategy. A partial decision tree is a decision tree with branches leading to undefined subtrees. The algorithm learns by building a partial C4.5 decision tree and creates a rule from the best leaf in every iteration, producing sets of rules where new instances of data will be assigned to category of the rule in the first matching rule in the list (Figure 14 below). In absence of a successful match, a default is applied. Proposed by Eibe F. et al (1998), PART has two algorithms combined, the C4.5 and the RIPPER rule based machine learning algorithms (Mahajan et al, 2014).

C4.5 algorithm generates decision trees when labelled training data is input. It is similar to the ID3 algorithm (Vijayarani et al, 2011). At each node on the tree, the algorithm selects attributes with the highest information gain for the prediction.

Repeated Incremental Pruning to Produce Error Reduction (RIPPER), is an algorithm designed by Cohen. It has an outer and inner loop. The outer loop adds to the rule base one rule at a time while a condition is added to the current rule by the inner loop iteratively until all members of a class have been covered before proceeding to the next class repeating the process until all classes have been covered (Vaishali et al, 2014).

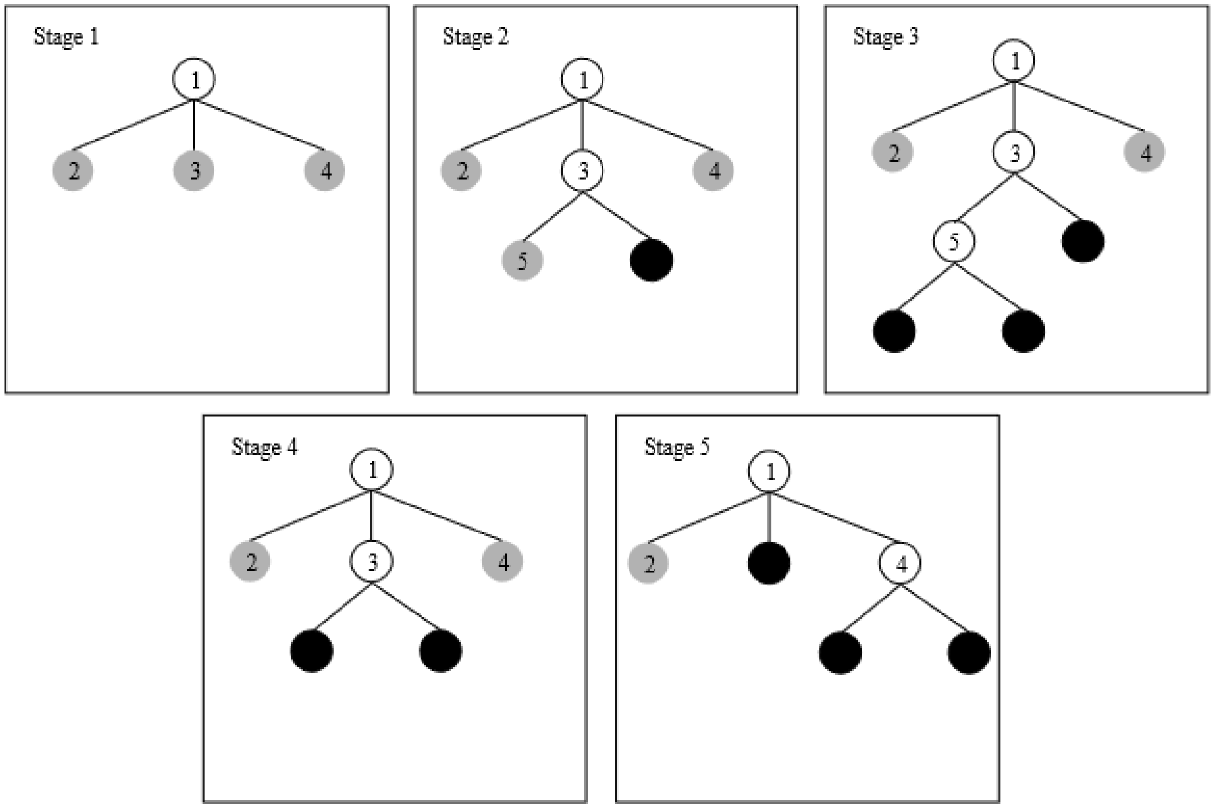

Figure 14: how the PART algorithm builds a partial tree

2.12 WEKA 3.8 Toolkit.

Waikato Environment for Knowledge Analysis (WEKA) is an open source machine learning software written in Java and can operate on any modern computing platform. It is an easy to learn free software with massive online courses to enable one to apply machine learning techniques for the many algorithms available on the WEKA tool kit.

The WEKA interface (figure 15 below) has several applications, the explorer, experimenter, knowledge flow work bench and simple CLI. In this project we intend to use the Explorer application.

The features available in the Explorer on the WEKA are comprehensive tools and include, Data Pre-processing, machine learning algorithms, methods of evaluation and environment to compare performance of the various algorithms. Clicking the data preprocessing button, various file formats such as ARFF and CSV can be imported. Data in URL and SQL datasets can also be read. Filters are the preprocessing tools of WEKA and are responsible for discretization, normalization and transformation of attributes. Other modules include Classification, Regression, Clustering, Association Rules, Attribute selection and Data Visualization.

The WEKA classifier has various families of classification algorithms that include; the Bayes, functions, lazy, trees and rules among others. Each family of algorithms has several algorithms e.g. in the family of functions are the multilayer perceptron, Logistics, SGD, SGD Text, and Simple Logistic. The Bayes family has Bayes Net, Naïve Bayes, and Naïve Bayes Multinomial Text. The classifier panel allows one to select a classifier of choice and execute it on the current data loaded on the WEKA. Test options available on the WEKA for this operation include, use of the training data, use of supplied test set, cross-validation, and percentage split. Performance of the classifier can be visualized in a classifier output window. Performance measures for every classification algorithm such as correctly classified instances, Recall, Precision and F-measures are displayed in the visualization window. This is together with the confusion matrix that summarizes the performance in a table.

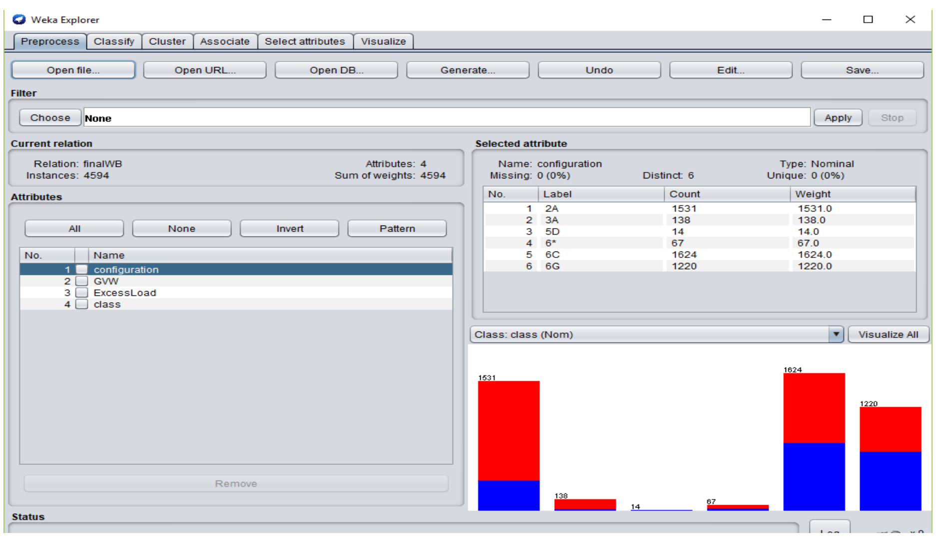

Figure 15: WEKA user interface

WEKA is a tool kit of choice as it provides as smooth transition to data science as it does not require the user to have deep knowledge of programming and coding. It can be used by people who are not specialists in data science. It is a free and available software that can operated on any platform. It provides many algorithms that are up-to-date for machine learning and data mining.

2.13 Confusion Matrix

Confusion matrix is a table that summarizes prediction performance of a machine learning classification algorithm. It allows visualization of how the algorithm performed. The rows in the table represent the instances of the actual class and the columns represent the prediction by the algorithm (Figure 15 below). The confusion is in the mislabeling one class as the other. The following performance measures can be done from the confusion matrix; Accuracy, Recall, Precision and F-measure.

(i) $$\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN}$$

How often is the overall prediction correct? A perfect classifier would have values in the True Negative and True Positive cell only. The other cells would be empty. It has an accuracy of 100%.

(ii) $$\text{Recall (TP rate)} = \frac{TP}{FN+TP}$$

How often does the classifier predict positive when it's actually true? This means that when it is actually true, how often does the system predict true?

Recall rate also called Sensitivity of the system. It is the number of positive predictions over the positive values in the data under test.

(iii) $$\text{Precision} = \frac{TP}{FP+TP}$$

When it predicts positive, how often is it actually true?

Precision is the number of true positives divided by number of true positives and false positives

(iv) $$\text{F-Measure} = \frac{2*(\text{Precision}*\text{Recall})}{\text{Precision}+\text{Recall}}$$

F-measure is the balance between Precision and Recall

(v). false positive rate is the proportion of cases incorrectly classified as positive. It is the possibility of raising a false alert.

FP rate= FP/FP+TN

	PREDICTED		
ACTUAL		NEGATIVE	POSITIVE
	NEGATIVE	TRUE NEGATIVE TN	FALSE POSITIVE FP
	POSITIVE	FALSE NEGATIVE FN	TRUE POSITIVE TP

Figure 16: fields of the confusion matrix

2.14 Related work

Various studies have been conducted in relation to prediction using data analytics and classification algorithms similar to what we intend to use for prediction of overloaded trucks. The aim is to optimize operations at weighbridge facilities.

2.14.1 Evaluation of weigh-in-motion axle load management system

Odongo G (2017) opines that delays and operational inefficiencies are frequent on the Kenyan weighbridges. The delays result from inspections for compliance such as axle load limits, valid load permits, driving licenses, driving under influence of alcohol and functional fire equipment. He conducted a study to investigate the effect of Origin and Destination on the likelihood of committing offences using the J 48 classification algorithm on the WEKA from data collected at

Webuye weighbridge station. The performance matrix of the model had F-measure at79.9%, True Positive Rate 82.7%, Recall 82.7% and Precision at 81.1%.

2.14.2 Prediction of Axle Loads

Data on trailers transporting single bundle and loose sugarcane and the respective payloads for each trailer was collected. Two algorithms were used in the prediction of axle loads induced by the sugarcane transport vehicles. A back-propagation neural network was trained with payloads and empty trailer axle loads as the inputs and the measured trailer and tractor rear-axle loads as the outputs. Linear regression was also employed on the data to relate the axle loads with the payloads of each trailer.

Kanali C. L. (1997) established that by ±5% residual error interval, both models attained over 85% prediction for trailer axle loads. Both algorithms also had 85% prediction for tractor rear-axle loads induced by loose sugarcane trailers. On the other hand the neural network achieved 70% compared with65% prediction by the statistical model for tractor rear-axle induced by single-bundle trailers.

2.14.3 Vehicle Overloading Management System

Weigh data from 11 permanent and 11 mini weighbridges was collected, initially manually and later automatically by computerized weighbridges, leaving out trucks that were within legal limits of axle load in South Africa over a period of time between 1988 and 1994. This data was analyzed at a central place by statistical models where transporters with higher tendency of overloading were identified and monitored (Nordengen et al,1995).

2.14.4 Algorithm application on Hypothyroid database for analysis

In this study Vaishali et al (2014) applied Naïve Bayes, BayesNet, PART, JRip and OneR algorithms on Hypothyroid database for analytical purposes to measure performance in accuracy, sensitivity, precision, false positive rate and f-measure. Considering majority of the parameters, PART was the best with correct predictions at 99.3, False Positive rate of 0.003 and precision of 0.945.

2.14.5 Algorithms for Intrusion Detection System

Two machine learning techniques were used in the security area to differentiate the normal traffic from intrusions as investigated in their study by Uziar B. et al (2017). Security continues to be compromised in the use of the internet and detection of intruders was explored. Performance of algorithms were compared and a detection rate tending towards 100% and 90% for the J48 and the Naïve Bayes respectively were obtained.

2.14.6 Student Retention System

Ngemu J, (2015) insinuates that students can drop out of College by factors that could otherwise be predicted. Various classification algorithms including Decision trees, Naïve Bayes, multilayer perceptron and Support Vector Machine were implemented on WEKA 3.8 toolkit on selected databases that contained variables that were suspected to influence student retention and predictions were achieved. The variables included course taken, KCSE grade, parents' education level and health among others.

J48 classifier was the best predictor with accuracy of 94.8% on 10-fold cross validation.

Table 5: Summary of the Related Work

DOCUMENT	**AREA**	**ALGORITHM**	**REMARKS**
Odongo G, 2017	Crime prediction using Origin and Destination weighbridge data	J48	TP rate of 82.7%, recall of 82.7%, precision of 81.1%
Kanali , 1996	Axle load prediction by sugarcane transport vehicles	1.Linear Regression 2. ANN	85% accuracy for both on axle load; 70% and 65% respectively for rear axle
Nordengen, 1995	Transporter tendency to overload	Statistical models	Identified and monitor targeted transporters

Vaishali, 2014	Algorithm analysis on hypothyroid data	PART	Correct prediction 99.3%, FP rate of 0.003, Precision of 0.945
Uziar B, 2017	Intrusion Detection System	1. J48 2. Naïve Bayes	100% accuracy 90% accuracy
Ngemu, 2015	Student retention	J48	94.8% accuracy on 10-fold cross validation

2.15 The Gap

Operations at the weighbridge have continued to be manual even during this era of computer technology with data mining and machine learning techniques for performance optimization in organizations. Many routine jobs are still being manned by human beings. Instead these would have been best handled by machines and skilled staff be left to handle more urgent matters for their trade. Use of machine learning techniques in the weighbridge facilities will eliminate the randomness exercised by weighbridge personnel when selecting which truck to go over the weighbridge during periods of congestion. During breakdown of the weighbridges, there is no alternative for the axle-load control and the overload prediction system will be the best available option for the control. It will further enhance efficiency of operations for all sections including management since visual tools will enable real-time decision making.

Machine learning algorithms have not been employed in axle load control to detect trucks that would otherwise be in contravention of the prescribed limits of weight permitted to be transmitted to the roads. From the related work, we chose to compare the Naïve Bayes, PART algorithm. J48, Multilayer Perceptron which is a neural network and introduced the Random Forest algorithm that shares a family of tree algorithm with J48 that had been used frequently in the related work.

2.16 Conceptual Model

The context diagram below serves as a high level design that is being proposed for prediction of overloaded trucks at weighbridge installations.

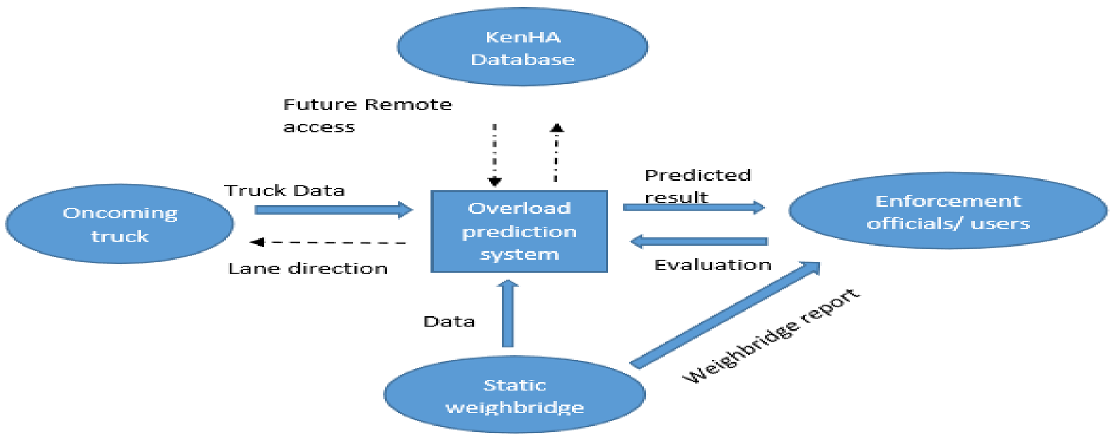

Figure 17: Conceptual Model

2.17: Overload Prediction Prototype

The prototype takes labelled input data that is data that has the target class labelled and trains using this data. New instances are introduces into the trained model and the output is the prediction of the class either excess or ok.

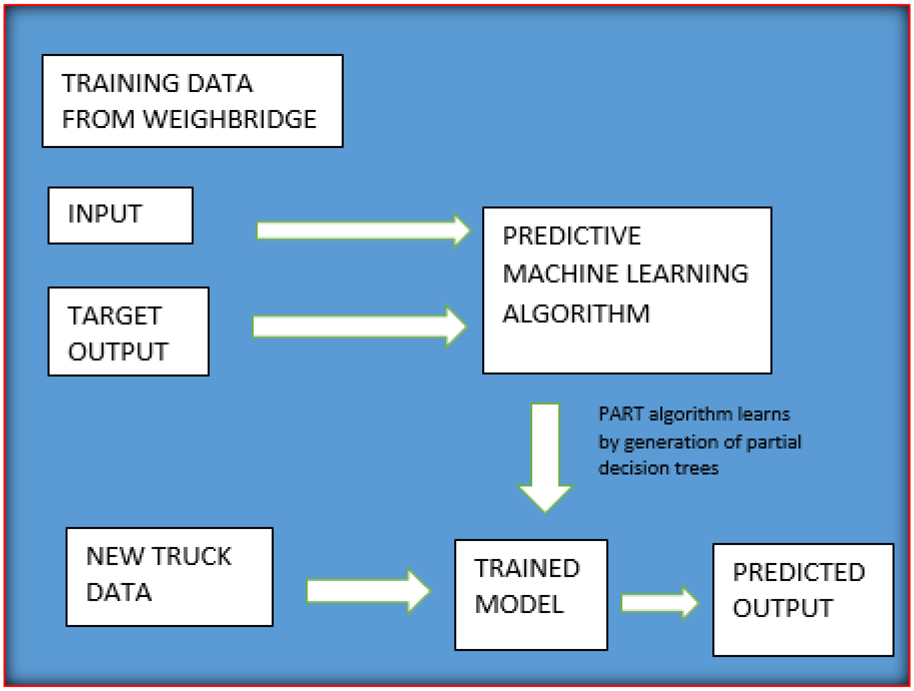

Figure 18: The overload prediction prototype

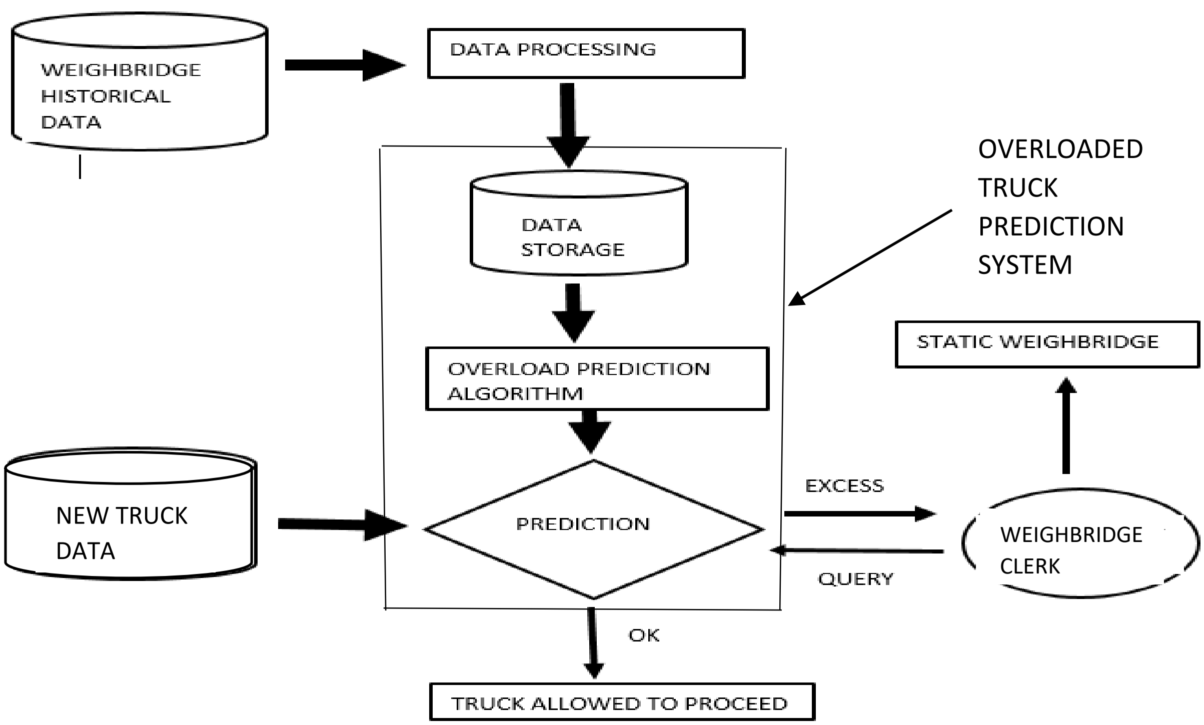

Figure 19: Process Chart for the overload prediction system

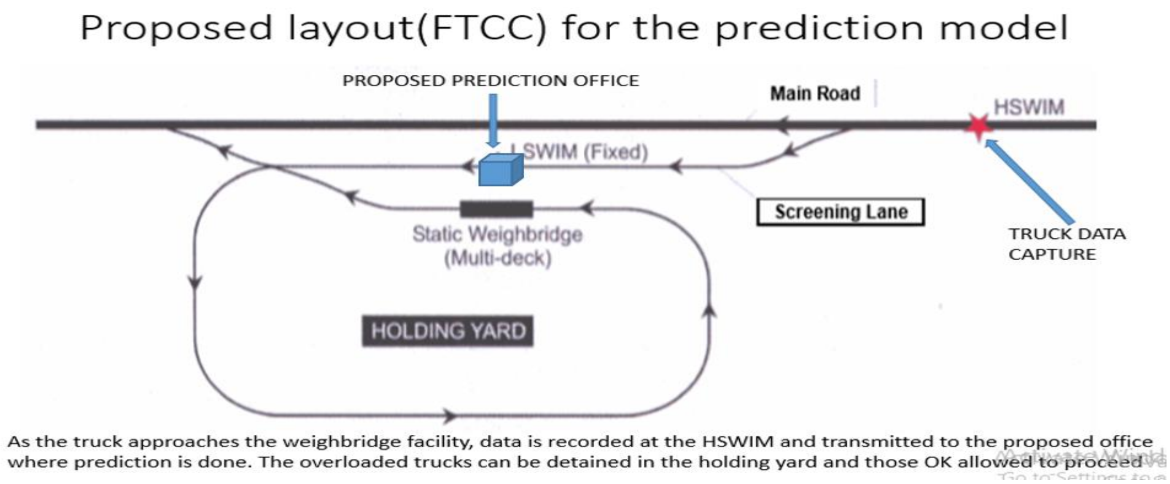

Figure 20: Proposed integrated layout for the prediction system the FTCC

2.18 Summary of literature review

We have covered aspects of weighbridge selection that is determined by the purpose for which the weighbridge is required, the layout that is governed by the class of road and number of vehicles that travel along the road. We mentioned the various combination of weighbridges that form various layouts including the lay-by layout that has a low speed weigh-in-motion unit axle scale.

Classification algorithms were emphasized as computer tools that would be employed in prediction of outcomes based on training used labelled data for training the algorithm then testing it for accuracy using unlabeled data.

Machine learning algorithm such as Multilayer Perceptron, J48, Random Forest, PART and Naïve Bayes were explored and explained.

The conceptual model explained how the overload prediction system will function.

CHAPTER 3: METHODOLOGY

3.1 Introduction

This research intended to develop and evaluate a prototype for prediction of overloaded trucks. We therefore collected historical weighbridge data from Gilgil weighbridge station that were record over a period of time. This data was preprocessed and stored in a database for analysis. The data was divided into training, testing and evaluation sets. Various classification algorithms were implemented on the WEKA toolkit to classify the data input into predicted outputs that could be compared with the actual outputs. The classifier that had the highest number of correct classification instances was built on Java and used as model for prediction of overloaded trucks system. Evaluation was conducted by introduction of different instances that also acted as new instances to the trained model. To achieve this, a suitable research design had to be adopted.

3.2 Research Design.

The research approach was based on a standard data mining methodology. The general research strategy was to use machine learning and data mining techniques to select weighbridge overload predictors out of some machine learning techniques such as PART algorithm.

We used CRISP-DM methodology, a data mining methodology that stands for cross industry standard process for data mining that describes the typical phases and the tasks involved and the relationship between the phases. Its choice is due to its acceptance as a standard model that can be applied to many fields. It provides the structured approach in planning a data mining project. In this model many tasks can be performed in different order and backtracking previous tasks is normal so as to repeat some tasks. The six phases involved in CRISP-DM are Business understanding, Data understanding, data preparation, Modelling, evaluation and Deployment. The way they were applied is described below.

3.3 Business Understanding

The weighbridge is a facility for controlling the axle-load of the vehicles travelling on the main highways. It ensures that overloading is eliminated or kept to a minimum. The heavy vehicles are therefore closely monitored by the facility to ensure compliance with a need for having acceptable loads on the roads. As a result, truck data is recorded every time a truck is weighed for compliance at the weighbridge. It's this data that we wish to analyze.

3.4 Understanding Data

The source data was collected and the important properties investigated. This was after correspondences with KeNHA headquarters Nairobi for permission to access collect and use weighbridge data from Gilgil Weighbridge station, a request which was granted. The researcher visited Gilgil and retrieved data from the weighbridge databases to a storage flash disc. This raw data was in the form of an excel sheet. A closer look at the raw data revealed the attributes and the instances available in the data. The attributes included weigh date and time, seq. No, Reg No, configuration, transporter, cargo, origin, destination, GVW excess, Axle/unit Excess, charged, last reweigh result and scale operator. The researcher had to interact with the ICT officer in charge and the weighbridge staff for clarification on what each attribute stands for and its relevance to the project. Verification of the quality of data was done to address issues of correctness of the data, completeness and whether there are missing values.

3.5 Data Preparation

Data preparation describes the activities performed on raw data in order to transform it into a form suitable in terms of format, quality and structure that will enable further processing, analysis and for prediction purposes. This phase starts by selection of the data by selecting the attributes and the records in the table. Data is cleaned so as to raise its quality to a level required by the WEKA toolkit. Attributes such as class with OK and EXCESS values were derived from last reweigh results column of the raw data that had values rectified and Excess.

These activities are time consuming and very tedious and are thought to consume 60%-80% of the time in a machine learning project. In this project, they included dealing with poorly recorded data, spellings, incorrect data, inconsistent, outliers and missing data.

Formatting of the data is important so that it best fits the machine learning algorithm. These algorithms are implemented on the WEKA so the data format has to suit standards compatible with the WEKA tool. The data collected was in excel files that had been recorded manually by weighbridge attendants who work on shifts all day long. Data was manually cleaned, spellings corrected and in the right case, instances with missing data were removed from the database. Outliers were also removed.

The data was split into training, testing and validation datasets using a randomization function to avoid chances of overlapping subsets of the training data in the evaluation data to assure proper testing takes place. The datasets were then saved as comma separated values file (csv), then opened in notepad. In notepad, more information was added in form of metadata with @ symbol preceding relation, attribute and data elements.

3.6 Modelling

The datasets were loaded on the WEKA toolkit for analysis. To train the algorithms imbedded in the WEKA, the training dataset was loaded then the testing dataset selected and the classification algorithm selected from those available on the classifier tab. The classification algorithms involved were PART, Naïve Bayes, Random Forest, J48 and Multilayer Perceptron. The performance of the classifiers recorded. First the data that had been split in 2/3 and 1/3 for training and testing datasets respectively was evaluated from a population of 4972 instances. Further a set with enhanced number of training dataset of 4594 instances was trained with 401 as the testing data instances was also used . The best attributes for prediction were also selected that would enable real-time prediction as trucks approached the weighbridge station.

A model based on the best classifier algorithm was developed on the Java code. A test design was generated where data is split into training, testing and evaluation datasets. The model was built by running the modelling tool on the prepared datasets. The model was developed with a suitable interface for logging in and log out for the system administrator, a data scientist and a clerk each with level of permissions. The system administrator had the ability to register users, the data scientist with ability to train the model on whatever data that may be made available and also do predictions and the clerk with ability to use the system to only predict new instances of data. The system is designed as not be able to predict until training of the model has been successful and only registered users can access services.

3.7 Evaluation

The model was evaluated by subjecting it to new instances of data not known where the system generated predictions which were compared to the actual class labels of those instances. We assessed the degree to which the model meets our project objectives and determine if there are ways in which the model is not adequate. The model is tested on a real-life application introducing new unknown data instances to evaluate its ability to generalize.

3.8 Deployment

We propose that the model be deployed in future at the weighbridge facility by integration with the HSWIN, a high speed weigh in motion scale whose data will form the data input of an on-coming truck for the overloaded truck prediction system.

CHAPTER 4: ANALYSIS, DESIGN AND IMPLEMENTATION

4.1 Introduction

This chapter outlines the analysis of the raw data, selection of the attributes, design of the model, and implementation of the prototype.

4.2 Analysis of the Raw Database

The basic database in this study are the records of data extracted from the 180tonne weighbridge for a period of seven month from July 2018 to January 2019 at Gilgil Weighbridge station. This data was recorded by scale operators on duty operating in shifts 24hrs a day. It had the following columns or attributes; weigh date & time, Sequence No., Vehicle Registration No., Wheel Configuration, Transporter, Type of Cargo, Origin, Destination, GVW Excess, Axle/Unit Excess, Whether Truck Charged, last Reweigh Result and Scale Operator (figure 21 below).

The *weigh date and time* was a record of the instance when a truck having passed the weigh-in motion scale was directed to the static weighbridge for enforcement purposes. The *Sequence No.* is the order in which the trucks were weighed at the weighbridge and assigned a number in sequence for record purposes, the *Vehicle registration No*, is a number unique to every vehicle as issued by the Motor vehicle Registration Department. The *configuration* variable describes the number of axles on a vehicle truck as per the specifications by the ministry of transport. The numeral is the number of axles and the alphabet or asterisk being a special characteristic in tires either that the tire is bigger than normal or that a single tire was found in a specific axle.

Cargo was considered and some items identified for the analysis included clinker, cement and maize. Other attributes were *Origin* and *Destination* of the cargo being transported. These locations were between Mombasa and Congo.

The *gross vehicle weight* (GVW) is the total weight of the truck plus the goods exerted onto the road, and the *unit axle excess load* (axle excess) is the weight in excess of the limit permitted in the Traffic Act exerted on the road by a single or unit of axles for a vehicle of a particular class or configuration. The column for the *scale operator* is the name of the clerk attending to the weighbridge at a particular shift on a particular day.

The class variable is the *last reweigh result* that had variables rectified and excess. These variables in this study were converted to Ok and Excess respectively.

All the above parameters were analyzed and only some were considered for use in the prediction of the truck overloads.

	Weigh Date & Time	Seq. No.	Reg. No.	Configu-ration	Transporter	Cargo	Origin	Destination	GVW Excess	Axle/Unit Excess	Charged	Last Reweigh Result	Scale Operator
9	6/1/2018 0:07	KGN1801762F	KCF880V	2A	tri	assorted	Nairobi	Nakuru	0	1320	-	Rectified	Elizabeth_Kuria
10	6/1/2018 0:13	KGN18017631	KCA321C	2A	kenlink	Loose	Nairobi	Nakuru	0	900	-	Still in yard	Elizabeth_Kuria
11	6/1/2018 0:28	KGN18017639	KCB408K	6G	ggh	Cnt	Mombasa	Kpl	0	102	-	Rectified	Elizabeth_Kuria
12	6/1/2018 1:29	KGN1801764E	KCF497E	6C	hh	Loose	Mombasa	Kpl	0	100	-	Rectified	Elizabeth_Kuria
13	6/1/2018 1:40	KGN18017651	KCH098H	2A	delta	Loose	Nairobi	Eldoret	0	2080	-	AU Excess 1640kg (SR)	Elizabeth_Kuria
14	6/1/2018 2:09	KGN1801765F	KCM465Y	6*	ponty	Beer	Nairobi	Nakuru	0	595	-	Rectified	Elizabeth_Kuria
15	6/1/2018 2:14	KGN18017661	KCE390R	2A	fg	tiles	Nairobi	Kisumu	0	720	-	Rectified	Elizabeth_Kuria
16	6/1/2018 2:35	KGN18017666	KCD778S	6G	hj	Loose	Mombasa	Kpl	0	1062	-	Rectified	Elizabeth_Kuria
17	6/1/2018 2:39	KGN18017669	KCD972E	2A	logistified	Loose	Nairobi	Nakuru	0	640	-	Rectified	Elizabeth_Kuria
18	6/1/2018 3:01	KGN1801766B	KBV334K	2A	nakufleet	Loose	Nairobi	Nakuru	0	1440	-	AU Excess 2300kg (SR)	Elizabeth_Kuria
19	6/1/2018 3:29	KGN1801766E	KCJ272X	2A	eagle 097	Loose	Nairobi	Nakuru	0	140	-	AU Excess 140kg (No SR)	Elizabeth_Kuria
	6/1/2018 3:33	KGN1801	KBV870C	6C					0	120	-	Rectified	Elizabeth

Sheet1

Figure 21: Raw data captured continuously at Gilgil Weighbridge station

4.3 Selection of Attributes

Not all attributes in data were relevant for the prediction of overloaded trucks. The following attributes were considered to influence the prediction of overloaded trucks and included the truck wheel configuration, type of cargo, origin, destination, GVW excess and unit axle excess.

Every variable was analyzed and attributes with less than 20 instances ignored. The data had been fed manually and had no standard form, spellings and abbreviations made were sometimes not clear to the researcher and such were ignored. Below is the table 7 of the variables and their possible values that was considered for the training, testing and the validation datasets.

Table 6: Variables in the database

Variable	Possible values
Configuration	2A,3A,5D,6*,6C,6G

Cargo	AssortedGoods,Cement,Clinker,Container,hardware,Loose,Maize,Petrol
Origin	Kampala,Nairobi,Kitale,Mombasa,Nakuru
Destination	Kampala,Kisumu,Nairobi,Kitale,Mombasa,Nakuru,congo,Eldoret
Gross vehicle weight excess (GVW)	Numeric
Unit Axle Excess	Numeric
Class(outcome)	Ok,Excess

All the attributes were evaluated using Correlation Attribute Evaluator on WEKA to establish the best predictors and below are the attributes by ranking, Wheel Configuration taking the lead as the best predictor at 0.1806 followed by Gross Vehicle Weight (GVW) at 0.1381 and Unit axle excess load with 0.1232. Origin of the transit goods had the least impact on prediction as shown in the Figure 22 below.

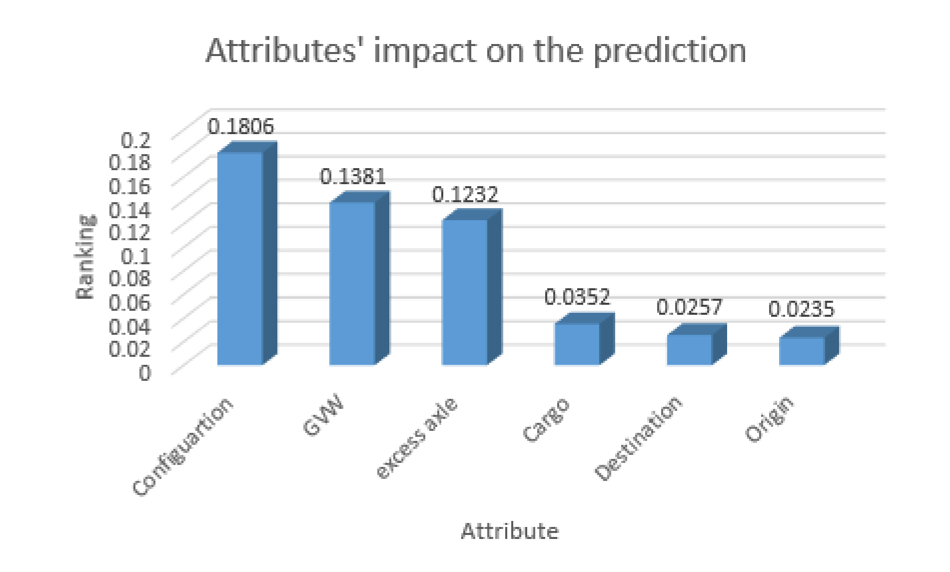

Figure 22: Attributes impact on prediction

Three attributes i.e. Wheel configuration, Gross Vehicle Weight and Unit Axle weight were preferred for the development of the prediction model as they had the highest impact on prediction.

The notepad (Figure 23 below) then saved as an attribute-relation file format (arff) for implementation on the WEKA a machine Learning Software. The ARFF file format is the format compatible with the WEKA tool.

```
new training with con - Notepad
File   Edit   Format   View   Help
@relation finalWB
@attribute configuration {2A,3A,5D,6*,6C,6G}
@attribute GVW numeric
@attribute ExcessLoad numeric
@attribute class {Ok,Excess}
@data
3A,640,880,Ok
6C,0,180,Ok
6G,0,1102,Ok
6C,0,640,Excess
6C,0,160,Excess
6C,0,40,Ok
6C,0,1020,Ok
6C,0,160,Excess
6G,110,600,Excess
6C,0,300,Ok
6G,0,2502,Ok
6G,0,460,Excess
6G,0,842,Ok
6C,0,500,Excess
```

Figure 23: arff file for loading on the WEKA

4.4 Analysis of the training Dataset

The training data was loaded on the WEKA toolkit and preprocessed by ensuring that it was in the format compatible with WEKA. Each attribute can be analyzed separately, for example in the figure 24 below, the red color represents the overloaded class and the blue trucks whose load was redistributed to within the permissible limits and allowed to proceed. In the configuration graph, vehicles with 2A value had a higher tendency of overloading. Vehicles of 6G configuration had a higher proportion of class OK. It can also be observed that vehicles of wheel configuration 2A, 6C and 6G were the most frequent trucks recorded at the weighbridge. Cement was the most important cargo that passed the weighbridge with a high of 2722 among all the cargo in the database. Nairobi was the favorite Origin of much of the cargo and Nakuru the Destination of choice. The Gross Vehicle Weight excess had a mean of 76.45kg and the Unit Axle Excess had a mean of 668.855kg. Of the data 1914 were labelled OK and 2680 labelled as EXCESS, and so the proportion of Excess being higher than that of the trucks labelled OK.

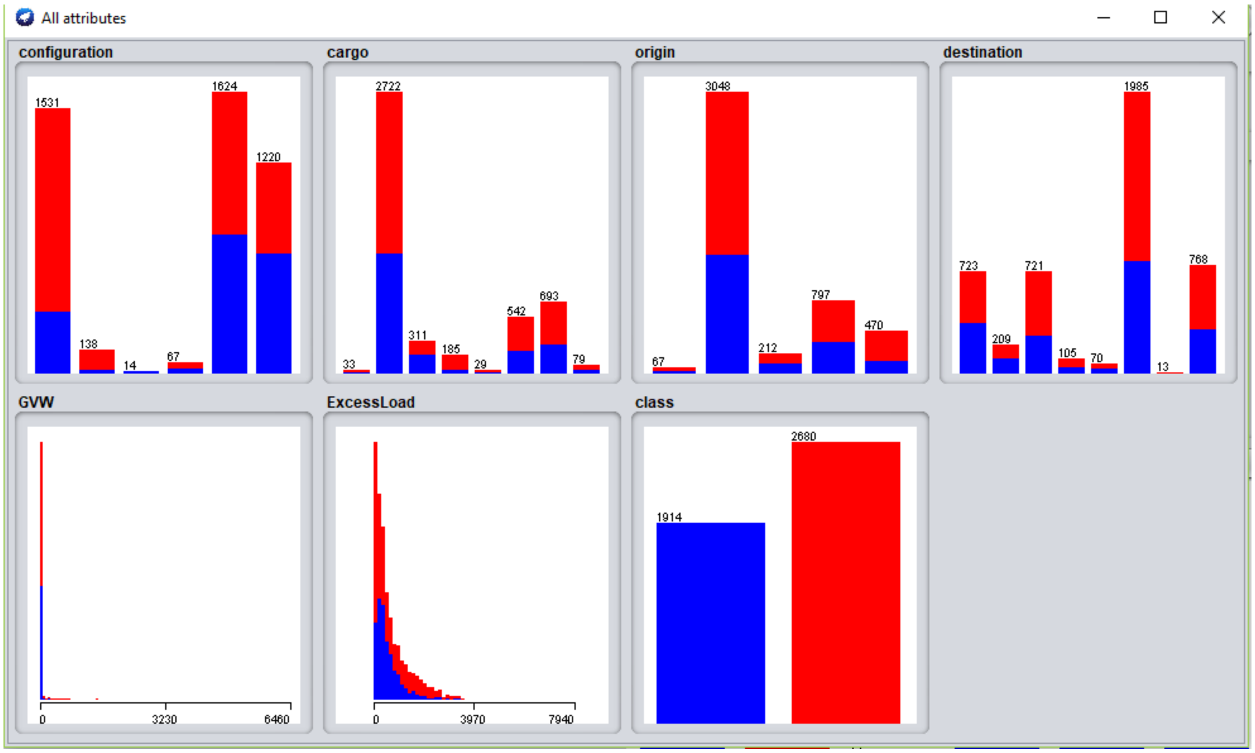

Figure 24: WEKA preprocessor visualization window for all attributes

4.5 Design of the Model

The algorithms discussed in the literature review were selected and implemented on the WEKA and each of them had their predictive performance evaluated. The following classification algorithms were deployed; the Random Forest Algorithm, the Naïve Bayes, the PART algorithm and the Multilayer Perceptron. For each algorithm, the training data was used to training and the testing dataset used to determine the accuracy in the generalization of the prediction model developed. The figure 25 below is the performance of the PART algorithm on the WEKA.

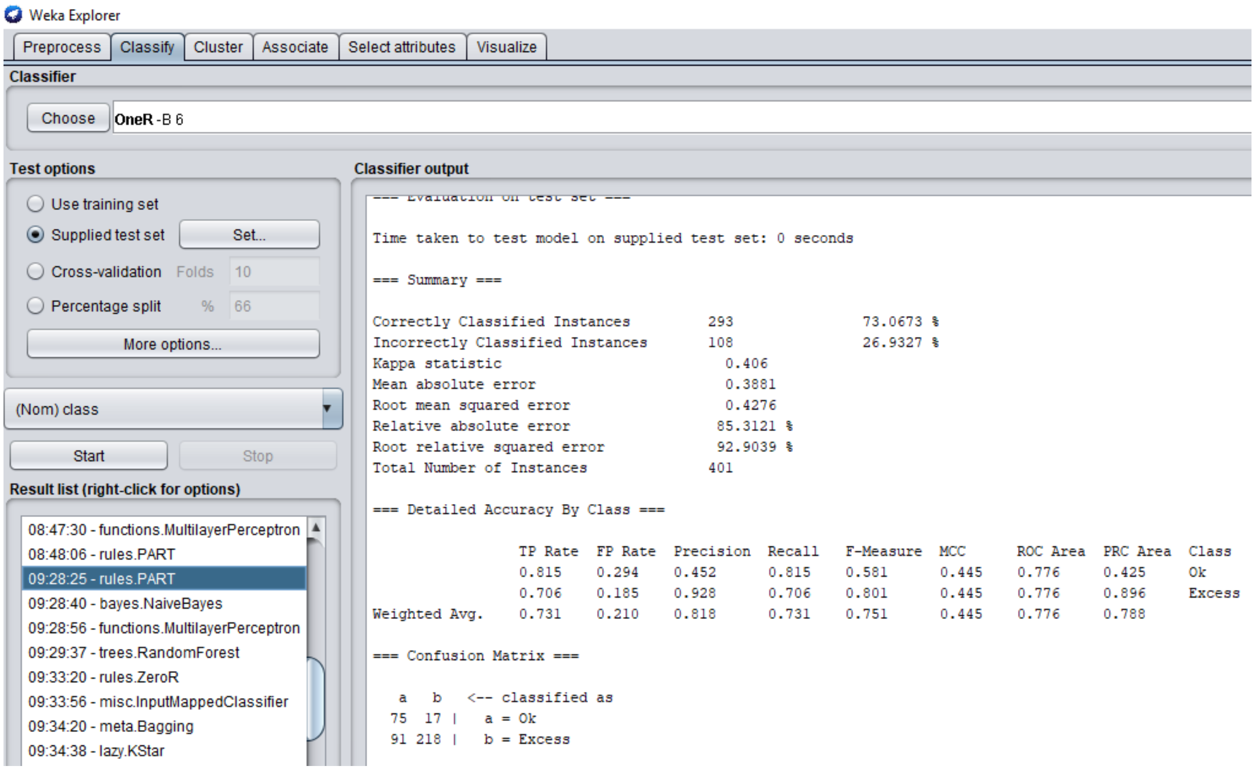

Figure 25: WEKA output window for classification.

The main focus is identifying the parameters that need to be considered when developing the system that will predict overloaded trucks as they approach the weighbridge installation. Using historical data, the system will in real-time analyze information fed by the operator into the computer or relayed by the HSWIM system located some distance as trucks approach the weighbridge and raise a flag for the prediction made. The researcher intends to identify users of this system and how they will use it after logging in. The data that should be input into the system and data to be output form the system. The requirements were then analyzed for validity and incorporation into the system to be developed.

The system and software design was prepared from the requirements. System Design helps in specifying hardware and system requirements and the overall system architecture. The researcher used WEKA in designing how system was trained using the data collected over a period of time at the weighbridge that included the truck configuration, type of cargo, origin, destination, GVW excess and unit axle excess to report of whether the truck was overloaded or not. The researcher then did a test strategy on what to test and how to test the system. Labelled data as either ok or excess was used to train and test the model. Only three of the attributes were selected for the model, i.e. wheel Configuration, Unit Axle excess load and Gross Vehicle Excess Weight. These attributes were automatically captured at the HSWIM system and gave real-time information for prediction. The model then used for prediction purposes.

4.6 Implementation

The truck overload prediction system was implemented in java with the front-end consisting of a user-friendly interface for the weighbridge clerk to log on and make entries for the overload prediction. The data scientist is responsible for the database and the training of the system. The system administrator responsible for managing users by registering the users who will log onto the system. The data scientist and the clerks are registered with their respective access codes together with the permissions available for them as illustrated in figure 26 below.

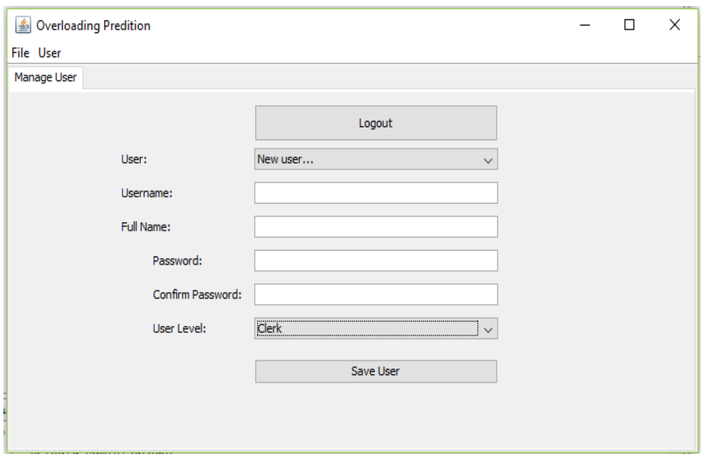

Figure 26: System administrator interface for registering users

The GUI has the option for prediction and data training where predictions can only be allowed when training of data has been successful. The system allows only the data scientist to train the model using the best predictor which was the PART algorithm as shown in the figure 27 below.

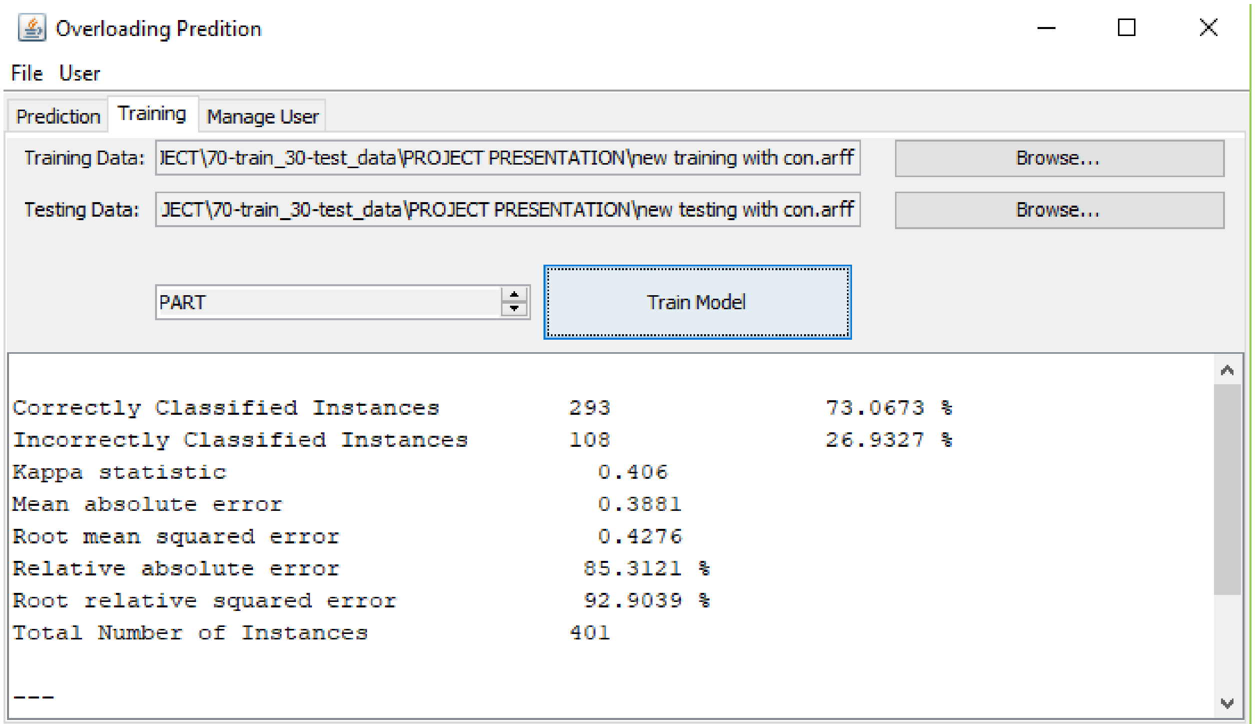

Figure 27: Data scientist's interface for training and prediction

Both the clerk and the data scientist can make predictions of new instance of data after training of the model and the figure 28 below is the GUI for prediction of the system.

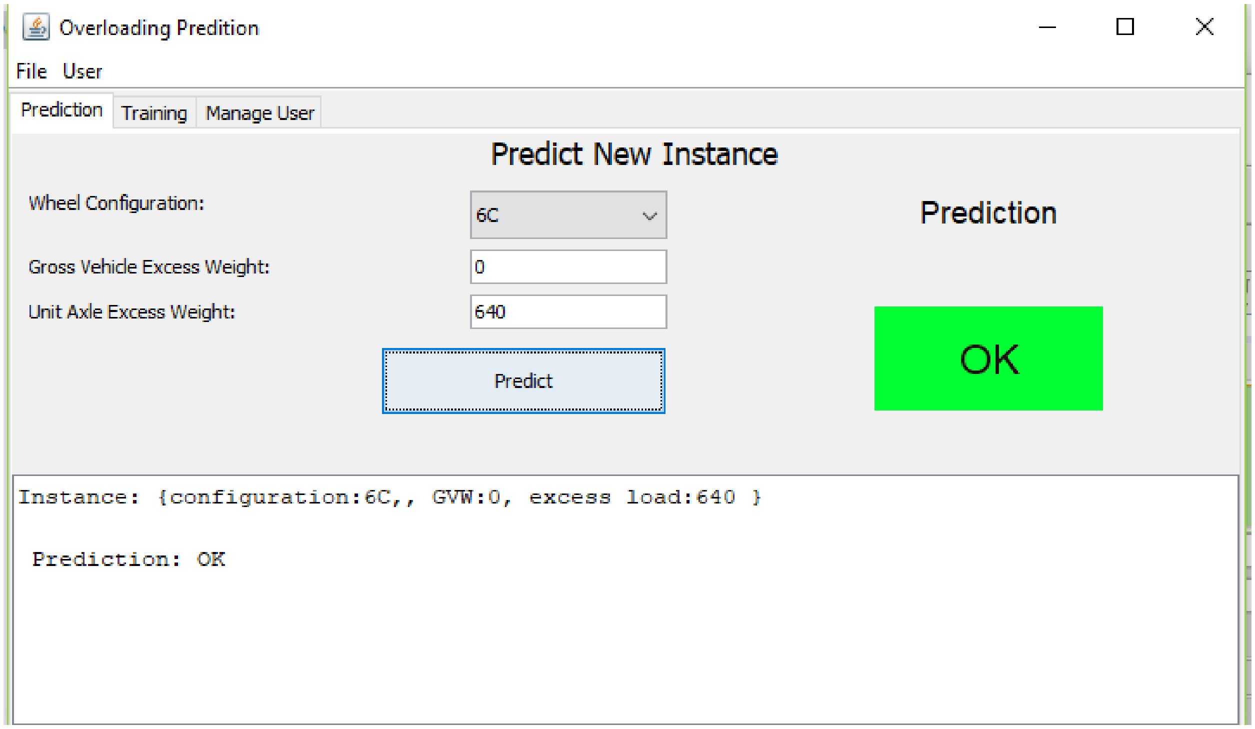

Figure 28: System interface for predicting new instances of data

4.7 Evaluation of the prototype

The code was developed using Java, the system then tested against the requirements to make sure that it solves the needs identified during requirements gathering. The evaluation criteria is to test the prototype as a black box where the researcher concentrated on whether particular inputs to the system will produce the expected outputs. All functional testing like introducing new data input to the system for prediction give outputs in a manner that reflected the accuracy, precision and sensitivity of the algorithm adapted for the prediction. Various attributes in the data were manipulated to evaluate the accuracy of the prototype to give accurate predictions using the new data sets.

The prototype was evaluated against the objective for which it was developed. It was evaluated on effectiveness for the system administrator to log on and register users. The System Administrator was able to register data scientists and clerks. The data scientist when logged in was capable of training the system using relevant data and could also input new instances of data for prediction. The Clerk when logged in had the rights to input new data instances for prediction only. The evaluation of the system was done by introduction of 50 new data instances as tabulated in appendix 1. Each instance of data was loaded onto the model and its prediction compared with the actual class label in which 44 instances were correctly classified giving an accuracy of 88%. This is an indication that the overload prediction system can be relied upon to deliver correct predictions to that accuracy.

CHAPTER 5: RESULTS AND DISCUSSION

This chapter compares the performance of the selected classification algorithms in this project and discusses the findings in relation to the project objectives.

5.1 The Performance of Machine Learning Algorithms

Classification of the preprocessed data was done using Naïve Bayes, Multilayer perceptron, PART and Random forest and implemented in WEKA 3.8. Each classifier was selected so as to compare their performance in terms of accuracy the prediction as they belong to various families of algorithms. Each classifier takes varying times to build a model that classifies inputs differently.

The data was split into 2/3 and 1/3 as training and testing datasets respectively. The training data had 3355 instances and the testing data 1617 instances. The performance of the various algorithms is as shown in the table 6 below. The Multilayer Perceptron was the best predictor with an accuracy of 71.6%.

Table 7: Performance of algorithms using 3355 and 1617 training and testing datasets respectively

ALGORITHM	% ACCURACY	FP RATE	PRECISION	RECALL	F-MEASURE
Naïve Bayes	62.0	0.33	0.67	0.62	0.62
Random Forest	67.4	0.379	0.668	0.674	0.668

PART	70.9	0.327	0.707	0.710	0.708
Multilayer Perceptron	71.6	0.317	0.713	0.716	0.714
J 48	69.8	0.32	0.70	0.698	0.699

The data was further split to increase the number of instances in the training in order that the algorithms learn better to enhance their performance levels. 4594 instances of the training data and a dataset of 401 instances were processed in the WEKA and the performance is as reported in the table 7 below where PART had an accuracy best of 73.1% as shown in the table 8 below.

Table 8: Algorithm performance at 4594 and 401 training and testing datasets respectively

ALGORITHM	% ACCURACY	FP RATE	PRECISION	RECALL	F-MEASURE
Naïve Bayes	49.3	0.212	0.789	0.494	0.512
Random Forest	69.6	0.274	0.786	0.696	0.719
PART	73.1	0.210	0.818	0.731	0.751
Multilayer Perceptron	60.3	0.202	0.802	0.603	0.632
J 48	70.3	0.18	0.826	0.703	0.727

49 new instances were used to validate the algorithms and PART gave 85.7%, Random Forest had 81.6%, J48 with 79.6, Naïve Bayes 51% and Multilayer Perceptron had 61.2% of correctly classified instances.

5.2 Comparison of classifiers

The four algorithms were implemented on WEKA Toolkit and below is their performance in details at 4594 instances of the training dataset and 401 in the testing dataset.

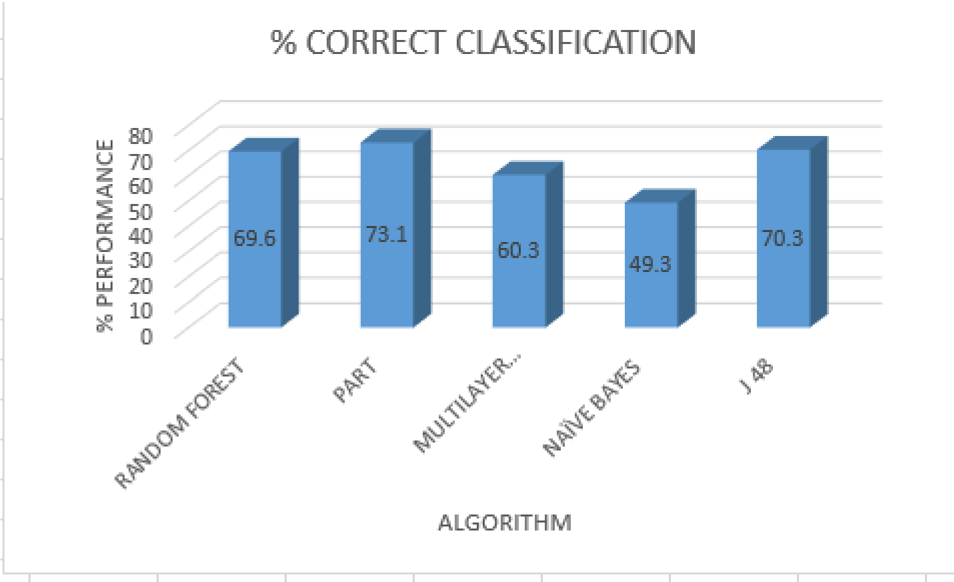

Figure 29: Classifier performance of accurately predicted instances

It is then evident from Figure 29 above that PART algorithm had the highest accuracy at 73.1% therefore being the best prediction algorithm for overloaded trucks prediction, followed closely by J 48 and Random Forest algorithm at 70.3 and 69.6% respectively. Naïve Bayes performed poorly at 49.3%.

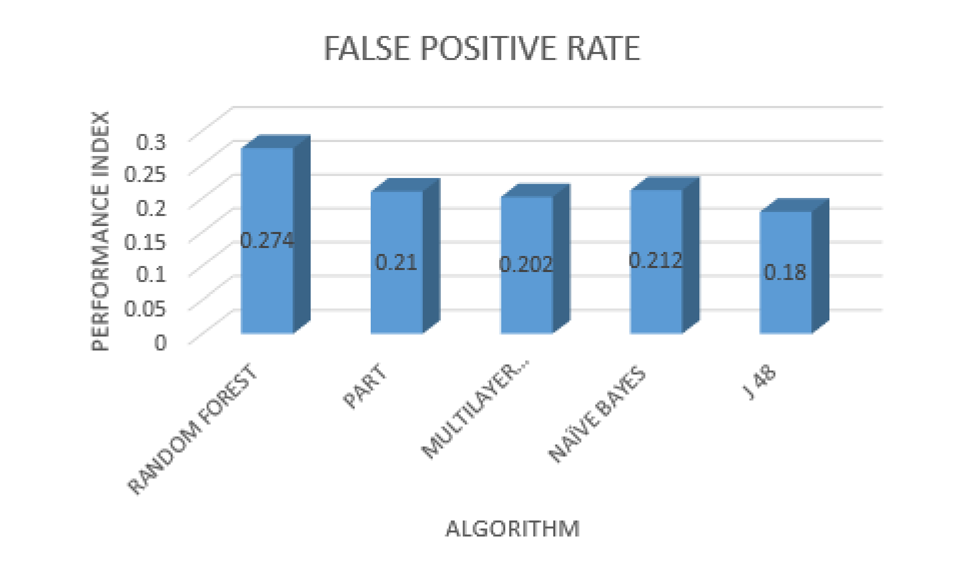

Figure 30: False positive rate

False positive rate is the proportion of negative cases incorrectly identified as positive. This is the probability of allowing an overloaded truck to proceed when it should have been detained for enforcement compliance reasons. J 48 algorithm had the best value of 0.18. Random Forest had the highest false positive prediction rate of 0.274 making it the worst algorithm (Figure 30 above).

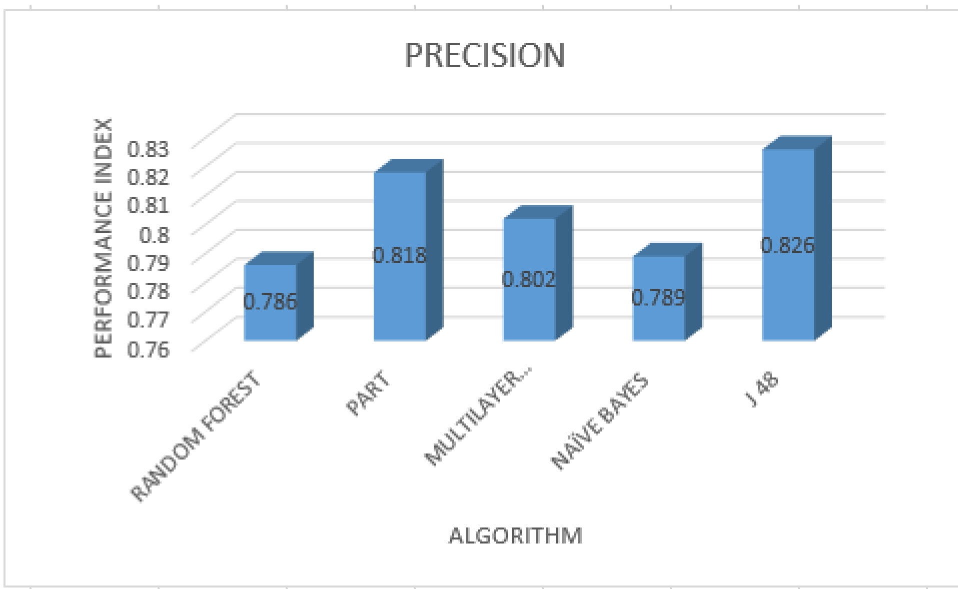

Figure 31: Precision

Precision is the ratio of the True positive to total predictions of positive (true positive plus false positive). J 48 had the highest precision value at 0.826 making it the preferred model among other algorithms, (Figure 31).

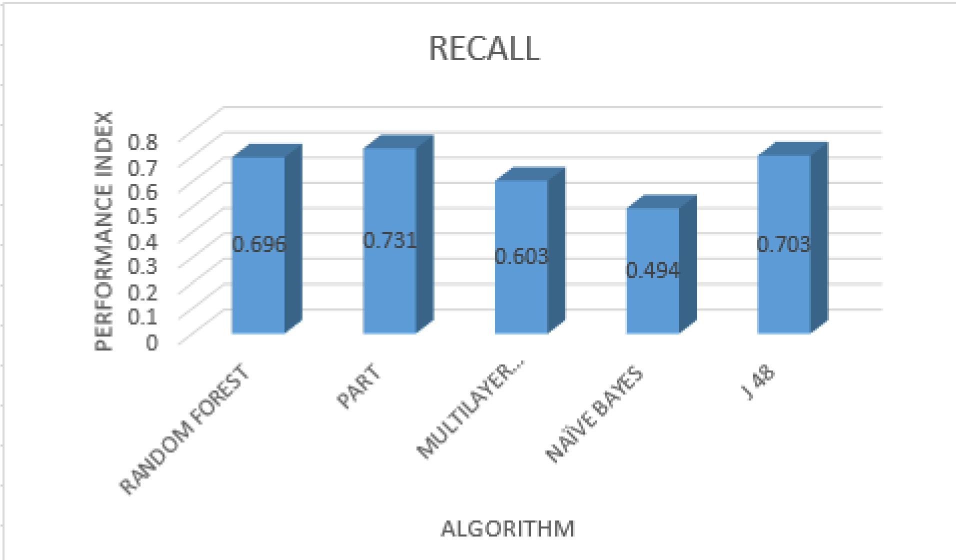

Figure 32: Recall

Recall as a metric is the number of positive predictions divided by the positive values available in the data. PART attained the best metric at 0.731 and so was the best performer for all the algorithms (Figure 32).

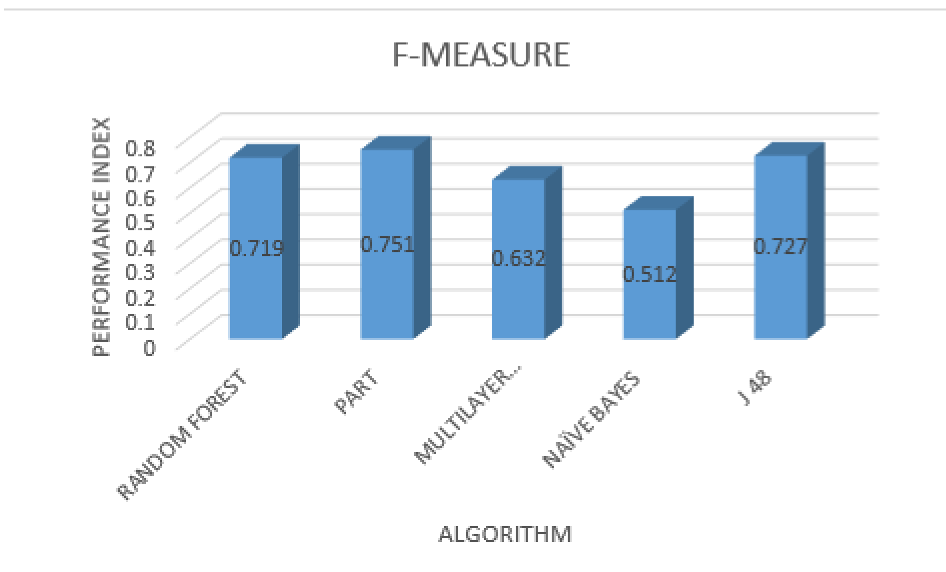

Figure 33: F-measure

F-measure is the balance between Recall and Precision and the higher the better the algorithm. PART at 0.751 was a better model than the other predictors. (Figure 33)

The PART algorithm had the best performance in correct classification of instances at 73.1%, recall with an index of 0.731 and F-measure of 0.751. The J 48 was algorithm of choice with performance matrix with False Positive rate of 0.18 and Precision of 0.826.

We opted to develop a model based on the PART algorithm.

The prototype was evaluated against the objective for which it was developed. It was evaluated on effectiveness for the system administrator to log on and register users entering the user name and password. The system administrator was able to register data scientists and clerks. The data scientist when logged in was capable of training the system using relevant data and could also input new instances of data for prediction. The clerks when logged in had the rights to input new data instances for prediction only. The evaluation of the system was done by introduction of 50 new data instances as tabulated in appendix 2. Each instance of data was loaded onto the model and its prediction compared with the actual class label in which 44 instances were correctly classified giving an accuracy of 88%.

5.3 Discussion

To determine the current axle-load control operations at weighbridge facilities, the operations at the weighbridge facility was discussed in the literature review where permissible loads for various wheel configurations were discussed. The data collected at the weighbridge was key in the project without which no project would have been realized. The type of weighbridges and their layout and installation also determined to guide on how best to integrate the system for overloaded trucks prediction.

We investigated the performances of the machine learning algorithms on weighbridge data by having the data preprocessed, split into training, testing and evaluation datasets and loaded onto the WEKA toolkit where classification algorithms; the Multilayer Perceptron, the Random Forest, J48, Naïve Bayes and the PART algorithms were implemented. Their performance measures were compared. The PART algorithm had the best performance in correct classification of instances at 73.1%, recall with an index of 0.731 and F-measure of 0.751. The J 48 was algorithm of choice with performance matrix with False Positive rate of 0.18 and Precision of 0.826, the details are on table 8.

We designed the model for prediction of overloaded trucks based on the promising machine learning algorithm in which the model was designed as to be able to train, be tested and make predictions on new instances of unknown data by generalization. The model had a user interface for a data scientist to train the model and also make predictions. A clerk can also login and can only input new instances of data for prediction as to whether the truck is OK or EXCESS. A system administrator can log in and register users.

We implemented a prototype for the overload prediction system by selecting the best predictor algorithm, the PART and deploying it in java code. The model was built by running the modelling tool on the selected training data. The model was designed as to enable training, testing and evaluation using the respective datasets and new unknown instances.

We evaluated the prototype system using independent data, when we used new instances of unknown truck data as input into the overloaded truck prediction system and an output of whether the truck is OK or EXCESS output. The evaluation of the system was done by introduction of 50 new data instances as tabulated in appendix 1. Each instance of data was loaded onto the model and its prediction compared with the actual class label in which 44 instances were correctly classified giving an accuracy of 88%.

CHAPTER 6: CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

In this study we built an overloaded truck prediction prototype with ability to predict overloaded trucks based on historical data captured at the weighbridge stations. A dataset of 3355 instances when used as training and a testing set of 1617 instances were deployed and of the classification algorithms used in this study, Multilayer Perceptron algorithm was found to classify instances more accurately at 71.6% than the J 48 at 69.8%, Naïve Bayes at 62.0%, PART at 70.9% and Random Forest classifiers with 67.4%. These algorithms were implemented on the WEKA platform, a data analytical tool that is open source. Another set of 4594 training and 401 testing data was used and the algorithms performed as follows; Multilayer Perceptron 60.3%, PART 73.1, Random Forest 69.6%, J 48 at 70.3% and Naïve Bayes 49.3%. This indicates that the PART algorithm continued to improve in correct classification as the data increased. Random Forest and J 48 algorithms also improved, but the reverse was true for the Naïve Bayes and the Multilayer Perceptron.

The PART algorithm had the best performance in correct classification of instances at 73.1%, recall with an index of 0.731 and F-measure of 0.751. The J 48 was algorithm of choice with performance matrix with False Positive rate of 0.18 and Precision of 0.826.

Use of weighbridges has been known for axle load control in order to save the road infrastructure from damage by the overloaded heavy commercial and transit goods trucks. The axle-load control also assists in safe driving. Overloaded trucks are dangerous to other road users as control of vehicle by drivers may be lost. Vehicles become unstable, breaks may fail and therefore accidents are likely to occur. The use of Overload Prediction System comes with automated prediction algorithms that will help avoid the unnecessary requirement of all vehicles weighing 3.5tonnes and above from congesting around axle load control weighbridges. Vehicles predicted as overloaded will be summoned to appear at the weighbridge. This therefore means that the time of travel along a route with an axle load control will tremendously be reduced resulting in reduced cost of cargo transportation.

It is evident that machine learning algorithms can be deployed in Overloaded Trucks Prediction System and be adopted by the National Highway Authority to improve on other functionalities provided at the station such as prediction of traffic flow during varying seasons for traffic management.

The System could also be used by transporter companies to subject their fleet to prior scrutiny that they may not be caught for non-compliance a situation that leads to hard earned revenue being lost in payment of heavy fines impost in courts of law.

Further research is proposed on the attributes so that better predictor attributes can be gathered in data collection that improve performance of the prediction algorithms to 95% level of confidence.

5.2 Contributions

This research contributes to the body of knowledge. This study has successfully applied machine learning techniques to the weighbridge data to support efficiency of operations at the facility. A system has been built for prediction of overloaded trucks in bid to optimize operations at weighbridge facilities. Use of the proposed overload prediction system enhanced effective axle load control. The prototype had a best correct classification at 73.1%. Use of ICT tools in the traffic management at weighbridge installations will ease congestion.

The use of this system will find domain in the axle-load control to ease congestion at weighbridge facilities and result in reduction of time wasters thereby reducing cargo transport costs. It will raise level axle load compliance as the notorious offenders will be isolated so being compelled to stop infringing the law. Consequently our roads will be safer, long lasting with little damage from overloaded trucks.

Human element that result in corruption by collusion between transporters and weighbridge attendants will be eliminated because all decisions are computer generated. The result will be lesser financing for roads maintenance and construction.

5.3 Future Work

The prototype had a best correct classification at 73.1% at the values of attributes selected. It had an accuracy of 88% on evaluation. It will be interesting to investigate the various ways of data capture that will allow for selection of weighbridge data attributes and algorithms that will enhance more accurate predictions of the overloaded trucks to a level above 95%. More accurate classifiers can also be adopted for the system development.

The system can be centrally managed and preserve the training part of the algorithm at a central server and have the user interface at the fleet management and the axle load control officials.

REFERENCES

1. Dredze Mark, Alex Kulesza, Koby Crammer (2009). Machine Learning: Multi-domain learning by confidence-weighted parameter combination.
2. Eibe Frank, Ian H. Witten (1998). Generating Accurate Rule Sets Without Global Optimization.
3. Fausett, L. (1994). Fundamentals of Neural Networks. New York: Prentice Hall.
4. Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy (1996). Advances in Knowledge Discovery and Data Mining.
5. Holmes Geoffrey, Andrew Donkin, Ian H Witten. (1994). WEKA: A Machine Learning Workbench
6. International Transport Forum's Corporate Partnership Board (2015). Big Data and Transport: Understanding and assessing options.
7. Japan International Cooperation Agency Padeco Co., Ltd. (2011). Study for the Harmonization of Vehicle Overload Control in the East African Community
8. John Kinuthia and Jason Lakin, (2016). Kenya: Analysis of the 2016/17 National Budget Estimates
9. Kanali C.L, (1997)Prediction of Axle Loads Induced by Sugarcane Transport Vehicles using Statistical and Neural-Network Models, College of Agriculture, Osaka Prefecture University, 1-1, Gakuen-cho, Sakai 593, Osaka, Japan.
10. Katahira & Engineers International (2015).Road Transport Safety and Axle Load Control Study in Nepal
11. Kattimani H D, Meghana N R, Nagashree B, Sahana Munegowda, Vijayalakshmi S. (2017). Vehicular Overload Detection and Protection
12. Khan Mohammad, Parwaiz S, Malik O.A, Pradhan D. (2018). Machine-Learning-Based Cyclic Voltammetry Behavior Model for Supercapacitance of Co-Doped Ceria/rGO Nanocomposite.
13. KMTI (2016). Republic of Kenya Ministry of Transport and infrastructure. Budget estimates.
14. Kenya Traffic Act cap 403(2018). Weights and Dimensions of Vehicles.

15. Kenya Weights and Measures Act cap 513 (2012). Rules relating to weighing and measuring equipment.
16. Mahajan Aditi, Ganpati Anita (2014). Performance Evaluation of Rule Based Classification Algorithms. Himachal Pradesh University Shimla, India
17. Ngemu Joseph Mutuku (2015). Business Intelligence Student Retention System for Higher Learning Institutions (A Case for Machakos University College).
18. Nordengen P. A. Hellens M. C. A (1995). System for Monitoring Overloaded Vehicles
19. University of Michigan Transportation Research Institute.
20. Odongo George O (2017). Using Information Gain to Evaluate Weigh-in-motion Axle Load Management Information System.
21. Odula Victor Odiwuor (2016). Assessment of Operations of Weighbridges in Kenya: A Case of Gilgil Weighbridge Station
22. Ondieki, Maureen Bosibori (2017). Effects of Big Data Analytics on Performance: A Case of Selected Telecommunication Firms in Kenya
23. Parsania V S, Jani N. N., Bhalodiya N H (2014). Applying Naïve bayes, BayesNet, PART, JRip and OneR Algorithms on Hypothyroid Database for Comparative Analysis, Gujarat-India.
24. Patterson, D. (1996). Artificial Neural Networks. Singapore: Prentice Hall
25. Pinard Michael Ian (2010). Guidelines on Vehicle Overload Control in Eastern and Southern Africa
26. Reddy Raj (2015). Analysis of overloading prevention system in trucks
27. Sub-Saharan Africa Transport Policy Program, Working Paper No. 90 (2010). Guidelines on Vehicle Overload Control in Eastern and Southern Africa. World Bank.
28. Uzair Bashir, Manzoor Chachoo (2017). Performance Evaluation of J48 and Bayes Algorithms for Intrusion Detection System.
29. Vijayarani S, Divya M (2011). An Efficient Algorithm for Generating Classification Rules, Bharathiar University, Coimbatore, Tamil Nadu, India

APPENDICES

Appendix 1: Evaluation of the Overload Prediction System on New Data Instances

EVALUATION OF THE OVERLOADED TRUCKS PREDICTION SYSTEM

s/no	Wheel configuration	GVW excess load	Unit axle Excess load	Actual Class label	Predicted Class	Correct Predictions
1	6C	0	120	Overloaded	Overloaded	✓
2	2A	0	1000	Overloaded	Overloaded	✓
3	3A	1460	260	Overloaded	Overloaded	✓
4	6C	0	120	Overloaded	Overloaded	✓
5	6G	0	220	Overloaded	Overloaded	✓
6	2A	0	1580	Overloaded	Overloaded	✓
7	6G	0	80	Overloaded	Overloaded	✓
8	6C	0	700	Ok	Ok	✓
9	6G	0	882	Ok	Ok	✓
10	6G	0	202	Overloaded	Overloaded	✓
11	2A	0	680	Overloaded	Overloaded	✓
12	6G	0	222	Overloaded	Ok	
13	2A	0	1620	Overloaded	Overloaded	✓
14	3A	1740	440	Overloaded	Overloaded	✓
15	6C	0	440	Overloaded	Ok	❖
16	2A	0	220	Overloaded	Overloaded	✓
17	6C	100	140	Overloaded	Overloaded	✓
18	3A	1980	680	Overloaded	Overloaded	✓
19	2A	0	1340	Overloaded	Overloaded	✓
20	6G	0	402	Ok	Ok	✓
21	2A	0	20	Overloaded	Overloaded	✓
22	2A	0	400	Overloaded	Overloaded	✓
23	2A	0	1020	Overloaded	Overloaded	✓
24	6C	0	60	Overloaded	Overloaded	✓
25	2A	0	1680	Overloaded	Overloaded	✓
26	6*	0	240	Ok	Ok	✓
27	6C	0	120	Overloaded	Overloaded	✓
28	6*	0	100	Overloaded	Overloaded	✓
29	6G	0	882	Ok	Ok	✓
30	2A	0	960	Overloaded	Overloaded	✓
31	6C	0	520	Overloaded	Ok	❖
32	2A	0	1440	Overloaded	Overloaded	✓
33	6C	0	2200	Ok	Ok	✓
34	2A	0	200	Overloaded	Overloaded	✓
35	6*	0	70	Overloaded	Overloaded	✓
36	6G	0	160	Overloaded	Overloaded	✓
37	6C	0	340	Overloaded	Ok	❖
38	6C	0	960	Ok	Ok	✓
39	2A	0	2500	Overloaded	Overloaded	✓
40	6G	0	82	Overloaded	Overloaded	✓
41	6G	0	1202	Overloaded	Ok	❖
42	6G	0	962	Ok	Ok	✓
43	6C	0	320	Ok	Ok	✓
44	2A	0	2620	Overloaded	Overloaded	✓
45	2A	0	400	Ok	Overloaded	❖
46	2A	0	700	Ok	Overloaded	❖
47	3A	560	600	Overloaded	Overloaded	✓

48	2A	560	2440	Overloaded	Overloaded	✓
49	2A	520	2140	Overloaded	Overloaded	✓
50	6C	0	1000	Ok	Ok	✓