UNIVERSITY OF GHANA

PREDICTIVE MODELS FOR IDENTIFYING CRITICAL UNITS FOR INSPECTION IN A REGULATORY BODY

BY

FELIX DELA DJOKOTO

10598534

THIS THESIS IS SUBMITTED TO THE UNIVERSITY OF GHANA, LEGON IN PARTIAL FUFILLMENT OF THE REQUIREMENT FOR THE AWARD OF MPHIL STATISTICS DEGREE.

November 14, 2018

DECLARATION

I hereby declare that this submission is my own work towards the award of the MPhil. degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree of the university, except where due acknowledgment had been made in the text.

FELIX DELA DJOKOTO ....

Student Signature Date

(10598534)

Certified by:

DR. RICHARD MINKAH ....

Principal Supervisor Signature Date

Certified by:

DR. LOUIS ASIEDU ....

Co-Supervisor Signature Date

DEDICATION

I dedicate this work to my loving father, Mr Christopher Yao Djokoto.

ACKNOWLEDGEMENT

This work has been by the grace of God, and for that I will always be grateful. He has kept me growing stronger in every facet of life.

My sincerest gratitude goes to Dr Richard Minkah and Dr Louis Asiedu, my supervisors, for continually guiding and helping me throughout this research work. I greatly appreciate their advice and support.

Also, I would like to thank every member of my family for their support throughout the duration of my study, especially my parents Mr and Mrs Djokoto who have been a pillar of strength and support throughout my education. I surely cannot forget my aunt Mrs Faustina Lawson and family for their sacrifices. It is deeply appreciated.

I also thank the Chicago Department of Public Health for making the data available online for public use. My gratitude also goes to the developers of the statistical programming packages SPSS, R and MATLAB.

Finally, I say thank you to all lecturers at the faculty and colleagues at the department of Statistics, University of Ghana.

ABSTRACT

Routine inspections are conducted at various food establishments that yield large data sets, which capture attributes useful for data mining algorithms to predict critical violations. Critical violations related to food establishments cause serious public health problems, which may happen as result of unhygienic environment, leading to food contamination. This study presents predictive models to detect critical violations in food establishments by employing Logistic Regression (LR), Support Vector Machine (SVM) and K-Nearest Neighbour (KNN). A database from the City of Chicago data portal that contained food inspections from 2011 to 2014 was used. In the preliminary analysis, Principal Component Analysis was utilised and ten (10) relatively relevant variables, that are independent of each other, were selected from twenty-eight (28) to be used as inputs in the models. In the family of the SVM, several kernels were used and the optimal model selected was based on the performance measures Receiver Operating Characteristic (ROC), sensitivity and specificity. The optimal model of the KNN was also selected based on the same performance measures. The out of sample classification accuracies for the LR, SVM and KNN classifiers were 92.7872%, 92.7873% and 92.6650% respectively. The performances of the models showed no large marginal differences in classification accuracies; however, the SVM model appears to provide a better discrimination ability as compared to the LR and KNN.

CONTENTS

LIST OF ABBREVIATION

AMA ..... Accra Metropolitan Assembly

ANOVA ..... Analysis of Variance

AUC ..... Area Under Curve

ANN ..... Artificial Neural Networks

CFIA ..... Canadian Food Inspection Agency

CDC ..... Center for Disease Control

CDPH ..... Chicago Department of Public Health

CHAID ..... Chi-squared Automatic Interaction Detection

CCNND ..... Class Conditional Nearest Neighbor Distribution

CART ..... Classification & Regression Trees

ECG ..... Electricity Company of Ghana

DFPA ..... Discriminative Function Pruning Analysis

FDA ..... Food and Drugs Authority

fmGA ..... fast-messy Genetic Algorithm

GLM ..... Generalised Linear Models

GASVM ..... Genetic Algorithm Support Vector Machine

GSA ..... Ghana Standard Authority

HACCP ..... Hazard Analysis Critical Control Point

IG ..... Information Gain

KNN ..... K-Nearest Neighbour

KMA ..... Kumasi Metropolitan Assembly

LSSVM ..... Learning Vector Quantisation

LR ..... Logistic regression

MR ..... Magnetic Resonance

MCC ..... Matthews Correlation Coefficient

MLE ..... Maximum Likelihood Estimator

NIST ..... National Institute Of Standards And Technology

NN ..... Neural Network

NRA ..... National Restaurant Association

NPA ..... National Petroleum Authority

OR ..... Odds Ratio

OVA ..... one-vs-all

OP-KNN ..... Optimally Pruned K-Nearest Neighbors

OC ..... Output Coding

PIM ..... Partition Index Maximization

PCA ..... Principal Component Analysis

PNDCL ..... Provisional National Defense Council Law

QUEST ..... Quick, Unbiased, Efficient, Statistical, Tree

RBF ..... Radial Basis Function

ROC ..... Receiver Operating Characteristic

SOFM ..... Self-Organising Feature Map

SPSS ..... Statistical Package for the Social Sciences

SVM ..... Support Vector Machine

SVM RFE ..... Support Vector Machines Recursive Feature Elimination

USA ..... United States of America

WDNN ..... Weighted Distance Nearest Neighbor

WHO ..... World Health Organisation

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Inspections into many organisations/establishments are routinely done to ascertain that due diligence to rules and regulations of a particular domain are followed. Here, critical units are critical violations found in food establishments which can lead to food related threats to the health of patrons, like food contamination (Murphy, DiPietro, Kock, & Lee, 2011). Therefore, a predictive model in this setting helps to predict the probability of detecting food establishments with critical violations.

For the safety of patrons of food establishment, some recognised bodies like the Canadian Food Inspection Agency (CFIA), Chicago Department of Public Health (CDPH) and Ghana Food and Drugs Authority (FDA), are legally empowered to enforce or implement food-related rules and regulations. In light of this, several food establishments are inspected so that the duration of exposure of unsafe food establishments to patrons is reduced. For example, the Chicago Department of Public Health has a database that captures the inspections done. In this study, the algorithms Logistic Regression (LR), Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) are used to develop models that can be used to prioritise inspections by detecting the riskiest food establishments, based on food inspection data sets from the City of Chicago open data portal.

This chapter is composed of Section 1.1, discussing the background of the study; Section 1.2, exploring the problem statement; Section 1.3, highlighting objectives; Section 1.4, showing the significance of the study; Section 1.5, presenting scope of the study; Section 1.6, outlining some limitations of the study and Section 1.7, covering organization of the study.

1.1 Background of the study

In a world of an ever-increasing accumulation of data, the field of statistics and computer science offer several algorithms to analyse these large data sets. An algorithm here means sets of computations and problem-solving approaches that learn from a data to produce a model. Algorithms such as Neural Network (NN), K-means, Support Vector Machine, K-Nearest Neighbour, Classification & Decision Trees (CART) and Naive Bayes are employed on large data sets in clustering, classification and regression problems. These algorithms, known as data mining algorithms, are used to sift through large data sets to look for patterns and relationships to get understanding that can be translated to make informed decisions.

With respect to large data sets and food safety, routine inspections are conducted at various food establishments that yield large data sets that capture attributes useful for data mining algorithms. For example, in 2014, Chicago Department of Public Health inspected more than 15,000 restaurants with less than three dozen inspectors, which meant every inspector was responsible for approximately 417 food establishments. Looking at the huge number of inspections required of the inspectors to complete, an automatic inference can be drawn to mean increase in exposure time of patrons to critical violation-ridden food establishments. The cost and irredeemable exposure time involved make prioritising food establishments, based on a validated model, a necessity. Schenk et al. (2014) used logistic regression in this regard to prioritise inspections based on food inspection data from the City of Chicago open data portal. However Logistic Regression (LR) has some deficiencies as compared to Support Vector Machine and K-Nearest Neighbour. Conventionally, Logistic Regression (LR) seeks to fit a model as best as it can, even on a training data set with outliers which may lead to misclassification (Pochet & Suykens, 2006). Again, LR cannot detect potential non-linear structures in a set of observations. That is to say, a non-linear

relationship would require a non-linear discriminant/decision boundary for a better performance. Therefore, this study seeks to use SVM and KNN to assess the performance of these algorithms, relative to the LR, in predicting critical violations in food establishments so as to help prioritise inspections.

SVM is an algorithm that generates a mathematical function able to classify linearly or non-linearly separable data into two distinct categories (Vapnik, 1998). The main thrust of SVM in cases of classification is to present a model capable of precisely predicting the label of classes of a test data (new points), having already learned a training data. SVM has many desirable traits like accuracy, robustness and effectiveness (Golub et al., 1999; Wang & Huan, 2011) even in non-linear separable problems. It again exhibits greater generalization ability (Thome, 2012) i.e the ability of SVM to perform similarly on training data and any new future data set.

K-Nearest Neighbour is a non-parametric algorithm (Duda, Hart & Stork, 1973) conceptualised by Fix and Hodges (1951), which has its base on the natural intuition to classify new cases or objects by finding its nearest training examples. It is an instance-based learning algorithm, that can be seen as classified by majority vote based on a distance metric, to determine the closeness of a new object or a feature vector in the fold of training examples. KNN is robust to noisy training data (i.e meaningless data that may be as a result of data corruption or inaccurate recording) and very effective with large training data sets (Kalakuntla, 2017). It is an effective method that is relatively easy to execute (Bhatia, 2010).

Data mining algorithms have performed extremely well in areas of bankruptcy prediction and fraud detection (Kumar, Krovi & Rajagopalan, 1997; Nagi et al., 2008), facial recognition (Swiniarski, 2000) and database marketing (Brachman & Anand, 1996). Many of such techniques/algorithms have been used in various fields; however, there are lingering questions on their relative performance which make some algorithms suitable to certain problem terrains than others (Becker,

2001). In this study, LR, SVM and KNN are applied on the food inspection data set from the City of Chicago open data portal, to compare their performance in correctly identifying any critical violations in a food establishment.

In this study, the three algorithms suit the classification of food establishments into those with critical violations or those without. The primary process through which the CDPH detects a critical violation in a food establishment is through inspections which may arise through new license inspections, voluntary calls from concerned citizens and daily routine inspections. This would require a large workforce of inspectors, which is logistically costly. The application of SVM and KNN to the field of food inspections will be greatly valued, as it provides an alternative approach to identify an unsafe food establishment by analyzing data sets on the inspections.

This study seeks to present predictive models for detecting critical violations that can help make informed decisions, based on a historical data on food inspections.

1.2 Problem Statement

Food safety is an established requirement backed by laws in many countries for every food establishment to follow, and it is enforced by inspections into all food establishments. According to statistics from the World Health Organisation (WHO), about 600 million people fall ill and 420,000 die every year as a result of consuming unsafe food ("Food safety", 2017). In Ghana, the habit of patronising food establishments is increasing particularly due to changing lifestyles and modernization (Monney, Agyei & Owusu, 2013). Wilson et al. (1997) revealed caterers cause about 70% of every bacterium related food poisoning. Therefore, there is the need for routine inspections of food establishments to ascertain safe and good environments for operation.

In Ghana, the Food and Drugs Authority (FDA), backed by Food and Drugs

Act PNDCL 305 of 1992, is required to enforce food policies and ensure safety and wholesomeness of food across the country (Ababio & Lovatt, 2015). With the growing number of food establishments, prioritising inspections based on the riskiest restaurant would traditionally depend on the wit and experience of the inspector. The traditional routine inspection is not only time-consuming and revenue draining, but also stalls the detection of these unsafe food establishments. This could ultimately prove to be dangerous to the health of patrons. Schenk et al. (2014) performed a study in Chicago using Logistic Regression to help prioritise inspections, in order to reduce the exposure time of patrons to unsafe food establishments. The model realised identification of critical violations about 7.44 days earlier over a 60-day evaluation period relative to the normal daily inspections.

Arguably, limited research has been done in the area of comparing and evaluating the performance of SVM and KNN using different parameter settings on finding critical violations at food establishments. In light of this, supervised learning algorithms, logistic regression, SVM and KNN will be compared to bring forth the better one which is able to identify these unsafe food establishments early, with respect to the previous study.

Prioritising inspections based on effective model is of utmost importance. Therefore, this study focuses on developing and validating mathematical models that can be used to detect critical violations which will help prioritise inspections.

1.3 Objectives

The main objective of this study is to develop a validated set of models to detect food establishments with critical violations. Other specific objectives to be considered are:

- To identify and select suitable predictor variables that can be used as inputs in the predictive models.

- To examine the effects of parameter settings on the K-Nearest Neighbour (KNN) and Support Vector Machines (SVM) algorithms.
- To develop models that are able to detect critical violations in food establishments.
- To compare Logistic regression (LR), K-Nearest Neighbour (KNN) and Support Vector Machines (SVM) to determine the highest detection rate.

1.4 Significance of the study

This study is of utmost importance to many food safety organisations around the world (like FDA, Ghana) as it will help sensitise them to use data mining algorithms on their large accumulated inspections data to prioritise inspections. This will help reduce the cost and time involved in their traditional routine inspections. Literature on such study is arguably non-existent in Ghana; therefore, it will benefit the academia and the research community. Stakeholders in the food establishment sector will be better informed to invest in using data mining algorithms to draw statistically backed inferences that will make food establishments safer, which would inure to benefit the Ghanaian populace.

1.5 Scope

This study used data collected from the City of Chicago data portal, in the periods September 2011 to March 2014, and September 2014 to October 2014. The secondary data was on food inspections conducted at various food establishments in the City of Chicago and it contained thirty-seven variables (features or attributes). The algorithms Logistic Regression, Support Vector Machine and K-Nearest Neighbour were employed in finding the model with a relatively high dectection rate. A defined grid parameter setting was used in selecting the best model of the SVM and KNN. All the analysis were made possible by MATLAB

(R2017a), R and SPSS.

1.6 Limitations

Below are some of the limitations encountered in the course of the study.

- A defined grid search was used due to the processing speed of the computer used for parameter tuning
- Several efforts were made to acquire data from regulatory bodies like the Food and Drugs Authority, National Petroleum Authority (NPA) and Electricity Company of Ghana (ECG), but none yielded a real dividend in the kind of data appropriate for this study. The issue of privacy of the customer or entity was a challenge, particularly for the ECG to provide the full data needed for this study. Hence the non usage of local data.

1.7 Organisation of the study

The following is how the thesis is organised: Chapter one introduces the study whiles also giving the background, the problem statement, objectives and significance of the study. Chapter two explores relevant literature on food safety and related works on the use of Support Vector Machine and K-Nearest Neighbour. Chapter three provides a detailed explanation of the methods of analysis. Chapter four provides the analysis of data, model building and the discussion of results. Chapter five covers the conclusions and recommendations from the study.

CHAPTER 2

LITERATURE REVIEW

The early identification of critical violations in food establishments is a priority for every food safety organisation since it poses serious health implications for patrons of these unsafe food establishment. Making sound arguments statistically to detect critical violations early in a food establishment is a growing area in research.

This chapter presents the relevant literatures related to the study that brings to fore useful facts and findings realized by previous researchers. It begins with food safety in food establishments followed by various approaches used in reducing the risks. Some probable variables that predict an unsafe food establishment are explored while also discussing SVM and KNN, and comparisons with the logistic regression. Some relevant application of predictive models to inspections are also outlined. Finally, the measure of performance of the algorithms are explained.

2.1 Food safety in the food establishment industry

Food safety can be said to be the tradition of handling, making and storing food where the best possible means are used to reduce danger of people contracting food-borne illness. Foodborne illness occurs when food consumed is infested with pathogens like viruses, bacteria, parasites or by toxic chemicals (Torgerson et al., 2014). Foods can be rendered unsafe at any time from production to the point of consumption but law suits are more likely to be directed at food establishments in the event of an outbreak than any other player in the chain (Buzby, Jenkins & Fernandez, 2001).

Notwithstanding the rise in global consciousness of food-borne diseases, food

safety remains seemingly disregarded (World Health Organisation [WHO], 2015). One contributing factor is the dearth of precise data on the full scope and cost of food related illness that would propel the allocation of funds by policy makers (WHO, 2015). Conducting inspections in food establishments is an effective way of gathering data on food safety. Food establishments provides solution to individuals who are not able to regularly make their own foods at home. Food prepared at such places are done on large scale which automatically mean the involvement of many hands. In such situations unintentional contamination of food may occur and the spread would have serious health implications for patrons and the country as a whole when outbreaks occur (Omaye, 2004). Factors such as improper handling, preparation and storing of food, and unrestrained environments where pathogens like bacteria are easy to spread (Fielding, Aguirre & Palaiologos, 2001) account for critical violations in food service establishments. Food safety organisations like Canadian Food Inspection Agency (CFIA), Chicago Department of Public Health (CDPH) and Ghana Food and Drugs Authority (FDA) are particularly interested in finding critical violations in the food service establishment as early as possible since any delays would have health and financial implications for patrons and industry players. (Knight, Worosz & Todd, 2007). A chronicle of events of food poisoning from pathogens like *E.coli* and *Campylobacter* have made food safety a top priority for governments across Europe. Outbreaks from food establishments occurs locally, regionally, nationally and internationally. Notable of them is the *E.coli* outbreak in the USA at "Jack in the Box" restaurant where 700 people suffered foodborne illness and 4 children dying as a result of taking contaminated meat bought at the chain of "Jack in the Box" restaurant (Golan et al., 2004). This resulted in an estimated cost of $160 million to Jack in the Box restaurant as a result of reduction in sales and law suits by their patrons (Fielding et al.,2001). In 2014, there was cholera outbreak in Ghana with Greater Accra region as the epicenter. The reported number of cases and deaths were 28,975 and 243 respectively (WHO, 2015). Statistics from

the FDA Ghana in 2013 showed that about 77% of the total food borne illness can be traced to food establishments (Ghana Standards Authority [GSA], 2013). Reflecting on such outbreaks calls for stringent measures to prevent re-occurrence. In light of this, food safety organisations through local health trained professionals perform routine inspection as a control mechanism to decrease or eradicate entirely the risk factors related to foodborne diseases (Reske et al., 2007). The various ways identified to reduce risk of contracting foodborne disease at food establishments are outlined in the next section.

2.2 Approach to reduce food safety risks

The implementation of food safety in food establishments is still a challenge because of the large number of people involved. For example, employees of 13 million working at 980,000 restaurants oversee 190 million foods served daily in the USA (National Restaurant Association [NRA], 2012). This presents various opportunities for food prepared to be contaminated. Therefore, the implementation of the Hazard Analysis Critical Control Point (HACCP) system is highly recommended. HACCP, is a process control system that detects where hazards might happen in the food production process and sets strict measures to avoid the hazards from happening (WHO, 2015). It has placed more responsibility on the industry players to ensure protection of the consumer from foodborne illness. With a set goal to decrease or eradicate food safety risks, one shared approach is conducting routine inspections by trained professionals at every food establishment to look for any violations of the food safety regulation like the Food and Drugs Act PNDCL 305 of 1992. This activity however is focused on preventing imminent foodborne disease outbreak. The inspection of food establishments is not just to reduce risks of an outbreak but to have a database which can help in further studies to mitigate future violations. Routine inspections alone will not suffice as the Center for Disease Control (CDC) in the USA, estimates almost half of the 9 million people who suffer

foodborne illnesses every year, contract them from restaurants (Center for Disease Control [CDC], 2013). Therefore, there is the need for food establishments to take responsibility to help reduce foodborne illnesses. Following the HACCP principles helps establish a well monitored environment at the food establishment site. The principles of the HACCP are; conducting a hazard analysis, fixing critical control point, setting critical limits, setting monitoring point, acting on corrective measures, setting verification procedures and creating databases of all documentation (WHO, 2015). Several studies in Ghana revealed locally owned businesses have limited regulatory systems of food safety and low level of education on food safety among food handlers in Accra and Kumasi. (Ababio & Adi, 2012; Ababio et al., 2012; Feglo & Sakyi, 2012 ; Tomlins et al., 2002). Another study revealed, both food handlers and patrons interest in nearness, appearance of the environment or vendor and price clout their core responsibility of good hygiene practices (Rheinlander et al., 2008). Agyei-Baffour et al. (2013) has established existing food safety rules are integrated in some HACCP principles but sentisation among food operators is low in Ghana . In Ghana, FDA through Ghana Tourist Board, local government authorities like KMA and AMA sensitise food operators on the food safety guidelines and the HACCP principles through training (Agyei-Baffour, Sekyere & Addy, 2013). For example, according to the 2016 FDA report, 5430 street food vendors, travelers and market women were trained on food safety and 7983 pupils were also educated on food safety and hygiene (Food & Drugs Authority [FDA], 2016).

Apart from following the regulations spelt out for food safety at food establishments, some guidelines like owners or managers taking active role in the process of acquiring the foodstuffs, making stricter rules at premises and being responsive to the complains of customers can also help reduce food related illness.

2.3 Factors that affect food safety in food establishments

The main health problem in relation to food in the world is the issue of foodborne disease. As the population of the world increase, the responsibility of ensuring safety in our foods also become burdensome. In many developing countries, the challenge in food safety is more evident in the poor handling, preparing and storing of food, weak monitoring systems and lack of training for food handlers (Tessema, Gelaye & Chercos, 2014).

The issue of food hygiene is the most basic necessity for all food service establishments to follow. They involve the handling, preparing and storing of food. WHO, has outlined five practices essential in avoiding foodborne illness and these are to; keep food clean, isolate raw from cooked foods, cook food carefully, keep food at safe temperatures, and use safe water and raw materials (WHO, 2006). Flouting these essential practices affect food safety negatively. Almost all the five practices can be controlled in the food establishment (Arendt, Strohbehn & Jun, 2015). Practising these however have some challenges such as time constraints, inconvenience and inadequacy of resources as expressed by employees of a restaurant in a focus group discussion (Howells et al., 2008). In a bid to find the factors that have influence on the use of safe food handling manners by employees of restaurants, several studies concentrated on food safety knowledge, training, attitudes and motivation (Allwood et al., 2004; Lynch et al., 2003). A study by Cushman et al. (2001) showed that the length of stay in a particular restaurant by employees also affect their practice of personal hygiene because they identified part-time student employees practiced personal hygiene properly than the main employees.

Some environmental factors such as weather affect food safety, since an increase in temperature, increases the likelihood of spoilage of foods like the dairy products,

meat and fish. Other factors like physical damage of foods and prolong storage of food also affect the safety of foods. Food safety culture is the shared conduct or behavior of employer and employees on the handling of food in the environments of food establishments. Findings from a research showed that the creation of food safety culture is vital as this will let employees know how non-expendable food safety is to the establishment (Yiannas, 2008). Some food safety cultures in food establishments include support from administration, communication, and employees' attitudes and manners (Abidin, Arendt, & Strohbehn, 2014). Following a food safety culture creates atmosphere for food safety (under-girded by adherence to the food safety laws).

The challenge of facilitating food safety put huge obligation on food producers and handlers. No matter how small an outbreak may be, it can quickly grow into international emergencies because of how fast products move across. Therefore, deliberate effort is required from everyone (governments, private sector, industry players, customers etc.).

2.4 Reviewed literature on support vector machine

Support Vector Machine (SVM), a bunch of inspected learning methods introduced by Vladimir Vapnik, has shown to work effectively in regression and classification problems, and detection of outliers (Ekici, 2012). SVM is a useful machine learning method which attains top predictive accuracy by learning from a training dataset to develop an optimal hyperplane to classify data. It has gained popularity in pattern recognition and computer vision communities because of their good generalisation capabilities and high accuracy (Du et al. 2017). It is widely applied to many fields of study in the biological and other sciences like computational neuroscience, pharmaceutical data analysis and drug design and fraud detection (Arvey et al., 2012).

In separating two groups of observations by a hyperplane with high separating margin, two problems emerge, thus how well the separating hyperplane generalize and the computational challenge. In 1965, (Vapnik, 1982) gave a solution to the first part of the problem by presenting the best hyperplane for divisible classes. The optimal hyperplane used in this instance represents a linear decision function able to divide the points(vectors) into two classes by leaving a wide margin. Therefore, the support vectors are defined by the margin of widest separation between the two classes. The study posits that with a minimal quantity of support vectors in comparison to the size of the training set, there will be high generalization. The second problem still remained until 1992 where (Boser, Guyon & Vapnik, 1992) changed the process of operation. Also, a comparison was made between two vectors before transforming them non-linearly as this helped to make better decision surfaces. Hence the name support vector network, but this concept was extended to a much complex sphere of dealing with non separable data by using different kernels functions for better precision. The new learning machine became as powerful and widespread as the neural networks was birthed. Comprehensive information regarding the SVM classifier is available in (Cortes & Vapnik, 1995; Tsai et al., 2009).

The performance of an SVM invariably depends on the method to select the optimal count of features and how to fix the best kernel parameters (Frohlich & Chapelle, 2003).

2.4.1 Feature Selection

The process of feature selection helps to detect the certified predictive subgroup of fields in the available database so that the relevant (reduced) number is offered to the algorithm for furthur processing (Huang & Wang, 2006). This process helps to draw forth relevant information from the available data set which in turn reduces computational time (Huang & Wang, 2006). The choice of features affect various parts of the classification, like the algorithm's accuracy, computational time, the

training examples required and the related cost of the features.

Selecting a subset of features from a database is significant in SVM as it extracts the important information from the available data set to reduce the computation time. As stated by Yang and Honavar (1998), choosing a feature affects not only classification algorithm's accuracy, but also the time required for learning a classification function, costs related to the features, and the number of examples required for learning. Not many algorithms have been proposed for SVM feature selection in the literatures (Bradley, Mangasarian & Street, 1998; Bradley & Mangasarian, 1998; Weston et al., 2001; Guyon, Weston, Barnhill & Vapnik, 2002; Mao, 2004).

Mao, (2004) proposed a feature selection technique known as Discriminative Function Pruning Analysis (DFPA). The intuition behind DFPA technique is to learn SVM's classification function after the training data by first, making use of every input variable, followed by selecting the number of features through pruning analysis. The DFPA technique employed both wrapper and filter methods *i.e* it first uses the filter method to avoid training a huge number of SVM classifiers and later uses the wrapper method to assess the number of features selected based on the classifier's performance.

Bradley et al. (1998) proposed a mathematical programming technique that minimizes a concave function on a polyhedral set. Again, in a different study Bradley et al. (1998) uses another term to penalise the size of subgroup features. Weston et al. (2001) also presented a vector of two, that denotes the presence or non-existence of a feature and their best measures with a drive of finding the two vectors. Also, a real valued vector can be in such a way that the gradient descent approach will determine the best value of the two vectors and its matching subgroup feature. The approaches used in these studies, all assess features one by one. But Guyon et al., 2002 proposed another method for feature selection known as SVM Recursive Feature Elimination (SVM RFE) which assess

features collectively. The method removes irrelevant or redundant features and, also it entails few computations compared to the wrapper method. Therefore, a discriminative measure for selection of method for feature selection based on it's appropriateness should always be considered.

However, the Principal Component Analysis (PCA), is popular approach used in several studies (Guyon & Elisseeff, 2003; Song, Guo & Mei, 2010; Uğuz, 2011) to select/extract relatively important features that are independent of each other. In this study the PCA was employed in reducing the dimensionality of data before feeding it to the algorithms.

2.4.2 Kernel function selection

A Kernel is simply a function for measuring the similarity between two observations(Sahami & Heilman, 2006). The kernel function helps to map the data in input space into a high-dimensional feature space to fit an optimal hyperplane to separate the data into their respective class. The commonly used kernels for SVM classification are Polynomial kernel function, Radial basis function and Linear kernel function. Selecting a Support Vector Machine kernel can be tricky, since it typically depends on the distribution of the input values (z) of the training data set.

To solve support vector classifier problem, the inner products of observations is used instead of the observations themselves. Suppose the inner products of two observation are

$$\langle z_i, z_{i'} \rangle = \sum_{j=1}^{p} z_i, z_{i'j} \tag{2.1}$$

Linear support vector classifier can therefore be expressed as

$$f(z) = \beta_0 + \sum_{i=1}^{n} \alpha_i(z, z_i), \tag{2.2}$$

where α_i, $i = 1, 2...n$ are the number of parameters per training observation.

The number of inner products between all pairs of training observations thus $\binom{n}{k}$

are used to estimate the parameters $\alpha_1, ..., \alpha_n$ and β_0

Some kernels are explained below (James, Witten, Hastie, & Tibshirani, 2013).

- Linear kernel

  The linear kernel basically computes the likeness of a pair of observations using Pearson (standard) correlation. Using a linear kernel to train an SVM is generally much faster than with another kernel.

  Now suppose the inner product in (2.1) is generalised in the form

$$K(z_i, z_{i'}), \tag{2.3}$$

  where K is referred to as a kernel.

  The equation below is known as a linear kernel

$$K(z_i, z_{i'}) = \sum_{j=1}^{p} z_i, z_{i'j} \tag{2.4}$$

  Less parameters is therefore needed to optimise when training SVM using linear kernel.

- Polynomial kernel can be described as function showing vectors (training samples) of similar types in a feature space over polynomials of the actual variables, allowing learning of non-linear models. Replacing each instance of $\sum_{j=1}^{p} z_i, z_{i'j}$ with the following quantity yields

$$K(z_i, z_{i'}) = (1 + \sum_{j=1}^{p} z_i, z_{i'j})^d \tag{2.5}$$

  Equation (2.5) is known as the polynomial kernel of degree d, where d is a positive integer. Therefore when d $>$ 1, suggests moving from the linear realm into a higher-dimensional space. Combining the equation (2.5) with the support vector classifier yields a SVM with a polynomial kernel of the

form

$$f(z) = \beta_0 + \Big(\sum_{i \in S} K(z_i, z_{i'} \Big) \tag{2.6}$$

- Radial Basis Function(RBF) kernel RBF kernel also called the Gaussian kernel, is a well known kernel function used in SVM classification to draw completely non-linear hyperplanes. It takes the form

$$K(z_i, z_{i'}) = exp\Big(\gamma \sum_{j=1}^{p} (z_{ij} - z_{i'j})^2 \Big) \tag{2.7}$$

where γ is a positive constant that sets the "spread" of the kernel.

2.4.3 Theoritical review of SVM

Some significant contributions made to SVM and its accomplishments in several data mining works of diversified fields are surveyed below.
In biomedical sciences, Valentini (2002) used gene expression data and proposed approaches built on SVM of non-linear nature using Gaussian and polynomial kernels, and learning machines of Output Coding (OC) ensembles to separate a typical tissue from malicious tissue, categorize lymphoma of different types and also to examine the roles of collections of coordinately stated genes in processes of cancer-causing lymphoid tissues. The study showed that SVM has the ability to appropriately divide normal tissues from tumour ridden ones, and different types of lymphoma can be classified by using OC ensembles.

Yang et al. (2005) used three classifiers, Learning Vector Quantisation (LVQ), Self-Organising Feature Map (SOFM) and Support Vector Machines, to study the making of an innovative signal classifier for small interchanging refrigerator compressors by means of vibration and noise signals. A novel approach was proposed to identify goods at semi-finish stage in a spontaneous bulk produce of interchanging compressors for refrigerators utilised in homes. SOFM with LVM was found to exibit high accurateness and shows to be the finest technique

for categorising healthy and malfunctioning condition of small interchanging compressors.

Polat and G*ü*neş (2007) used Least Square Support Vector Machine (LSSVM) to develop a medical decision where LSSVM was used in detecting breast cancer. Evaluation was done to check the robustness of LSSVM using specificity and sensitivity analysis, classification accuracy, confusion matrix and k-fold cross validation method. Wisconsin Breast Cancer Diagnosis dataset was used in the study and classification accuracy of 98.53% was realised which suggest LSSVM can help in diagnosing breast cancer. The research proposed furthur exploration on large data set would yield increase in the accuracy level.

Chaplot et al. (2006) suggested a different method for classifying MR images by making use of wavelets as input to support vector machines and neural network self-organising maps. A dataset of 52 MR brain images was used whereby the data was separated into two groups as either normal or abnormal. The neural network self-organising maps and support vector machines achieved a good classification percentage of 94% and 98% respectively. In comparison, the SVM classifier showed a high classification rate than the self-organising map. The method was applied to only T2 weighted images at a specific depth inside the brain. The research proposed, exploring it on T1-weighted, proton density and different kinds of MR images where a software for a diagnostic system can be developed for identifying brain disorders like Alzheimer's, Parkinson's, Huntington's diseases etc.

Hong et al. (2008) in their work proposed a novel method to use SVMs integrated with one-vs-all (OVA) scheme and naive Bayes classifies in multi-class fingerprint classification systems. To train the OVA SVMs and naive Bayes classifiers, some indicative fingerprint features such as the FingerCode, singularities and pseudo ridges were used. The NIST-4 database was used to validate the method proposed and a classification accuracy of 90.8% was realised for the five-class classification

problem and 94.9%, for the four-class classification problem.

Zhang et al. (2008) presented a study that describes the effect of employing multi-words for text representation on potentials of text classification. To use the multi-words for text representation, two strategies based on the several semantic levels of the multi-words were developed. Strategy one involves a proposal to perform the multi-word extraction from documents with respect to its syntactical structure. Strategy two involves a combination strategy with respect to the subtopics of the general proposal for representation. The robustness of the classification performance was realised by using Information Gain (IM) method to rid the multi-word from the feature set. Finally, SVMs in linear and non-linear kernels were used respectively on a series of tasks with text classification. It was realized that the effect of utilising distinct representation strategies outweighs the effect of utilizing several kernels on classification performance. It illustrated furthermore the usage of individual words representation outclass any usage of multi-words representation. It confirmed SVMs power in text classification.

To support the claim of SVMs accuracy in classification problems several studies like Parikh et al. (2010) showed a classification accuracy of a new SVM proposed to be around 98%. The input features proposed by the new SVM fault-based classification algorithm to identify the fault phases were three samples of phase currents together with the zero-sequence current. Tests were carried out on a data set of 25,200 test cases to check the feasibility of the technique developed which indicated that the new technique proposed, is accurate and robust to any fault condition and variation in system. Este et al. (2009) used a new classification technique based on SVM, described an algorithm that permits the classifier to correctly perform with a few hundred training samples. The proposed classifier was tested on three sets of traffic traces and classification accuracy went above 90%. With reduced size of training data sets the study confirmed SVM classifier as very effective.

Support Vector Machine is well known to be effective in dealing with outliers. Qu and Zuo (2010) in their study proposed an algorithm for effective data cleaning, data processing and feature selection. The algorithm was based on SVM and random sub-sampling violation. Outliers and irrelevant features were determined based on measuring the misclassification rate whiles adopting the backward selection method of feature selection. Three data sets were used to test the performance of the data cleaning algorithm, which exhibited a good capability of detecting outliers in all the data sets. Again, Wu et al. (2014) in order to provide a solution to the problem of SVM being sensitive to noises or outliers for the training data set, used fuzzy methods on SVM. The proposed Partition Index Maximization (PIM) clustering based Fuzzy SVM (FSVM) algorithm showed more reasonable membership when used on five benchmark data sets. It also showed that the PIM-FSVM algorithm is more robust to noises which indicated that the algorithm is effective. Lo and Wang, (2012) used classifiers based on SVM to classify MR images by conducting experiment of twofold; one made up of phantom images, generated by a computer and the other made up of MR images. The SVM showed a better classification accuracy than the C-Means (CM) when the efficiency and feasibility of the methods were evaluated. The SVM again showed its mettle in being robust against noise.

Chou et al. (2014) in their study proposed a high performing hybrid artificial intelligence model to merge a fast-messy Genetic Algorithm (fmGA) with SVM aimed at early prediction of dispute propensity at the initial phase of public-private partnership projects. The fmGA optimises the parameter of SVM and the SVM helps to provide learning and curve fitting. Several classifiers were proposed when applying them on CART, QUEST, C5.0, CHAID and GASVM,the hybrid approach. Considering precision, accuracy, AUC and sensitivity of all the models, the GASVM produced the topmost overall performance measurement score of 0.871.

Rai and Yadav. (2014) introduced a new and efficient way for recognising

and extracting features from the iris by utilising both SVM and Hamming distance. The approach used distinct feature extraction techniques for Hamming distance and SVM based classifier which they supposedly claim will increase efficiency. The proposed method's accuracy proved to be excellent and successful computationally as the recognition rate on image databases CASIA and Chek were 99.91% and 99.98% respectively.

In summary, the results of reviewed literature suggest Support Vector Machine is an effective algorithm that has high predictive power particularly for regression and classification problems.

2.5 Reviewed literature on k nearest neighbour

The K nearest neighbor is a simple algorithm conceptualised by Fix and Hodges, (1951) which is based on the natural intuition to classify a new case or point by finding its nearest training examples. It is particularly used in regression and classification problems and considered as one of the earliest, accurate and simplest algorithm (Hamamoto, Uchimura & Tomita, 1997; Alpaydin, 1997). It can also be deduced from the KNN that, only identical instances have identical class labels (in classification) or similar target values (regression). A machine learning algorithm learns automatically from data and improve from experience, of which KNN is referred to as a lazy type, in the world of machine learning users. Algorithm like K nearest neighbor was made known to deal with pattern classification (Yang, 1999) likewise Support Vector Machine (SVM) (Japkowicz, 2000). Two choices that primarily affect the performance of the KNN algorithm is selection of K and the distance metric used (Latourrette, 2000).

2.5.1 Selection of K- value

A very important and sensitive parameter in the KNN algorithm is the choice of K since it has an effect on the performance of the classifier. Seldom do studies

explain the type of method employed to choose the KNN parameters. Sun and Huang (2010), in their quest to identify an optimal K brought forward an adaptive KNN algorithm to find a K that each training example can utilise to attain right class label. Having identified a limitation in the conventional KNN algorithm to find for each test example, similar number of nearest neighbours, the adaptive KNN algorithm was proposed and tested on numerous data sets. Empirical results from tests suggested the adaptive KNN algorithm is more effective than the conventional KNN algorithm.

Some published studies use varying choice of K, for instance Rosenfeld et al. (2008) made use of a KNN algorithm to predict the origin of cancer tissue using microRibonucleic acid profile. By considering limited varying values of K in the parameter space, $K = 3$ was the best K value. Similarly, Lu et al. (2005) considered this type of approach in selecting an optimal K parameter.

Guo et al. (2003) transformed the training set to a model that allot a group to a set of similar examples from available data set. The output model in this case is made up of the category of group and likeness of the far apart points in a particular group relative to the amount of points at centre of that group. In this work, the training data's size is reduced and the choice of K is determined automatically since the best value of K can be said to be the number of points in every group. The model was applied on six data sets to test its performance and it showed to be more accurate than the traditional KNN. Suguna and Thanushkodi (2010) in another study employed a Genetic Algorithm together with KNN in order to increase performance. The traditional KNN algorithm considers the entire training samples and selects k-neighbors, but this method employed GA to select k-neighbors right away and the distance is computed to appropriately classify the test samples. The combined algorithms was tested using five distinctive data sets and empirical results suggested an improvement in classification accuracy.

Selecting a desirable K rests entirely on the data. A larger K greatly reduce

noise in classification but also fails to set clear boundaries among the classes and invariably a small K lead to a large variance in prediction. Therefore, K should be set at a point where it is large enough to reduce misclassification and small enough to let the K nearest cases to be close to the new case (Hassanat, Abbadi, Altarawneh & Alhasanat, 2014).

2.5.2 Distance Metric

In K-NN algorithm, the distance from a new point and nearest neighbours are very important in making predictions. Some popular choices of distance metrics are the Euclidean, Hamming, Manhattann and Minkowski distances. If c and d are vectors that has numeric attributes, thus $c = (c_1, c_2, ..., c_n)$ and $d = (d_1, d_2, ..., d_n)$,

- Minkowski distance

  These distance metrics have special relationship with Minkowski distance, namely the Manhattan distance, Chebyshev distance and Euclidean distance. Minkowski distance is given by:

$$D_{Mi}(c, d) = \sqrt[z]{\sum_{i=1}^{n} |c_i - d_i|^z}, \tag{2.8}$$

  where z is positive value. Once $z = 1$, then it turn into Manhattan distance. Again, once $z = 2$, then it turn into Euclidean distance. Chebyshev distance is a Minkowski distance modified where $z = \infty$. Also the i^{th} values of vector c and d are c_i and d_i respectively hold.

- Euclidean Distance is the root of the sum of squares of differences among the opposing quantities in vectors. Given by

$$D_E(c, d) = \sqrt{\sum_{i=1}^{n} |c_i - d_i|^2} \tag{2.9}$$

- Manhattan Distance is the sum of absolute differences among the opposing quantities in vectors. Given by

$$D_{Ma}(c,d) = \sum_{i=1}^{n} |c_i - d_i| \tag{2.10}$$

- Chebyshev distance is a measure of the distance among two vectors in a vector space where any difference between them is the largest on every coordinate dimension. Given by

$$D_C(c,d) = max_i|c_i - d_i| \tag{2.11}$$

- Hamming Distance is a metric that quantifies the number of bad fits among two vectors. It is usually a measure for nominal data and string analyses, and also suitable for numerical data. Given by

$$D_H(c,d) = \sum_{i=1}^{n} 1_{c_i \neq d_i} \tag{2.12}$$

2.5.3 Theoretical review of KNN

Some significant accomplishments in the application of K nearest neighbor in several data mining works in diversified fields are surveyed below.

In a bid to improve the performance of the nearest neighbor a method known as Weighted Distance Nearest Neighbor (WDNN)(Jahromi, Parvinnia & John, 2009) was proposed. It assigned non negative weights to each training instance to compensate for the nearest neighbour sensitivity to distance function whiles utilising all the training instance when generalising. The technique can be described as a great way of reducing instances in a training set.

In a related work, Liu and Chawla (2011) proposed a new KNN weighting strategy to tackle problems in traditional KNN that arise as a result of the presence of more class samples of one class than the other. The method uses class confidence

weights which involve utilising probability of attribute values with specified class labels to weight prototypes in KNN. The bias to the majority class is corrected and this translates onto improved performance.

Kriminger et al. (2012) also proposed using the geometric structure of data to lessen the influence of class imbalance on KNNs performance. The method is known as Class Conditional Nearest Neighbor Distribution (CCNND). Existing approaches make use of a type of sampling scheme or by applying error costs to check the imbalance in distribution of the classes. It was applied on data sets fetched from UCI Machine Learning Repository (imbalanced data sets) and real world oil pipeline data. The CCNND extremely performed better than traditional KNN.

Another work by Yu et al. (2010) proposed a method known as Optimally Pruned K-Nearest Neighbors (OP-KNNs) which proved to be competitive to advanced methods while reducing computational time. By making use of KNN as kernels to do regression, a one hidden-layer feedforward neural network is created. The method showed good performance while remaining relatively a simple model.

Some factors contribute to the performance of KNN, so Parry et al. (2010) considered varying factors like number of features, distance metric, number of neighbours, vote weighting and decision threshold which gave 463,320 KNN models. The models were validated using data on 478 neuroblastoma patients. With varying factors, the optimal KNN model was identified.

From the literature reviewed, other methods have being integrated into the traditional KNN in order to either reduce computational time or select the optimal K. However this study makes use of varied values of K to select the optimal model for the KNN algorithm.

2.6 Application of predictive models in inspections

A process of building, testing and evaluating models to predict the likelihood of an event to happen can be described as a predictive model. In most human institution, inspections are conducted to ascertain that rules and regulations are followed. Predictive models has being applied in various areas to help detect critical units. Some of these are surveyed below.

Schenk et al. (2014), used data on food inspections, 311 complains, crime data and others to build a predictive model able to detect critical violations in food establishments. The main purpose of the study was to prioritise inspections, so that the exposure time of patrons to unsafe food establishments is reduced. The logistic regression model was used and it able to identify approximately critical violations 7.44 days earlier over an evaluation period of two months.

In a similar study by Kassel (2017), Azavea (i.e. Geospatial technology company) partnered with government agencies and other private companies to build a model that can make inform decisions. Machine learning algorithms were used to build models that statistically foretells the probability that a building will fail an inspection based on historical data from the City of Philadelphia Department of Licenses and Inspections. The classification accuracy was the evaluation criteria for the model and it predicted 74.19% accuracy on the test set used. The study suggested more data-driven tools could be used to improve on the model developed.

The recent fire incident that happened at the Grenfell Towers in London, among others motivated the City of Pittsburgh's Bureau of Fire (PBF) to use its data on inspections, to develop predictive models to detect risky properties likely to experience fire. Historical data on fire incidents coupled with routine inspections on properties were used. Logistic regression, Ada Boost, Random Forest and Xg

Boost were the machine learning algorithms used to help prioritise inspections. The presence of an alarm or smoke detector was the predominant predictive feature among other variables in the model. The predictive model built, was able to statistically foretell the presence of a fire incident at a specific location. It performed better than the previous methods used to prioritise inspections (Smart Cities Initiative, 2018).

Also, Madaio et al. (2016) asked themselves the questions, "how do we help Atlanta Fire Rescure Department (AFRD) identify new properties that need inspections?" and "how do we help AFRD to prioritise their property inspections by fire risk?". In order to answer these questions, the algorithms, Random Forest, logistic regression and Gradient Boosting Tree SVM were used on historical inspections of properties with fire incidents. The Random Forest and SVM performed best. At the end 69 properties were flagged as high risk and 48 violations of the fire safety codes were found.

The Philadelphia Department of Licenses and Inspections (L&I) razed several buildings to the ground for concerns of possible collapse. Mosley and Steif (2018) utilised the data on previous demolitions to train model capable of predicting the likelihood a building collapse. The predictive algorithms Naive Bayes, logistic regression, Random Forest (RB) and Gradient Boosting Machine (GBM) were compared. The predicted properties flagged to be demolished by the algorithms were compared with actual properties razed down. The outcome saw a better performance from the GBM and RB. With sights on the predicted probabilities from these algorithms, the GBM found about $1,800$ parcels citywide, unsafe with probabilities greater than the 10% threshold set.

Though many data mining algorithms has been used in the literature reviewed, arguably little has been applied to detect critical violations in food establishments. Inasmuch as many studies (Palaniappan, Sundaraj & Sundaraj, 2014; Weinberger, Blitzer & Saul, 2006; Li, Zhang & Zhao, 2017) use fixed or single parameter setting

(for example $k = 1$ in KNN and $C = 1$ in SVM) for the SVM and KNN, this study made use of a grid with different values circled around the fixed parameters usually considered.

Predicting the likelihood of identifying critical units in various sectors is a difficult adventure to take therefore data scientists suggest, using data mining algorithms. Hence, the usage of SVM, LR and KNN in this study.

2.7 Comparing Logistic regression, K-Nearest Neighbour and Support Vector Machine

There are many supervised learning algorithms such as Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes K-Nearest Neighbour, random forest, decision tree, etc. Supervised learning algorithm focuses on building a model able to make predictions of the response values for a new dataset. The main challenge posed in supervised learning is the selection of the appropriate data mining algorithm for classification. One underlining criteria vital for the selection of an algorithm is based on characteristics of the training data set, such as the size, quality and nature. Duda (2001) established and the 'no free lunch' theorem (Wolpert, 1997) further reiterate that no particular classifier works optimally all the time but depends on the type of problem and the available data set. Depending on the data set, some questions such as the number of training examples, the features dimensionality and its independence are realised.

Parry et al. (2010) related KNN and logistic regression to defend the nonlinear classifiers used in its study of gene expression. KNN performed significantly better than logistic regression when their mean performance were compared using a Bonferroni adjusted significance level of 0.005. Again the Matthews Correlation Coefficient (MCC) showed that KNN performed significantly better than logistic regression based on the responses from some tumour cells in the data set. The logistic regression performed better than KNN in the comparison of gender.

Whereas linear classifiers, like logistic regression use a straight line to separate the feature space, nonlinear classifiers like KNN and Support Vector Machine have the chance to build further complex decision surfaces (Parry et al., 2010). LR, KNN and SVM well suit objects that can be classified into separate class labels. Kuramochi (2005) based on gene profiles saw that KNN performed better than SVM which supposedly has more complex structure. Logistic regression becomes disadvantageous to an algorithm that uses kernel functions to map input vectors into a high dimensional feature space like SVM for classification. One significant property of a hyperplane is that it is better fine-tune to enlist the details of data.

In a study by Joachims (1998), KNN, SVM and others algorithms were applied to Reuters data and KNN performed best among other methods which (Yang, 1997) confirmed the findings in a different study.

Rana et al. (2015) compared SVM, logistic regression, Naïve Bayes and KNN using an online data i.e UCI machine learning repository. As noted by the researchers, thus every algorithm performs totally different which is largely due to the parameter selection and available dataset. The SVM had a test accuracy of 93% and 68% on the diagnoses of breast cancer and the recurrence or non-recurrence of breast cancer data respectively. The regularised logistic regression had a test accuracy of 92.10% and 72% on the diagnoses of breast cancer and the recurrence or non-recurrence of breast cancer data respectively. Also, the KNN using Euclidean distance had test accuracies of 95.63% and 72% on the diagnoses of breast cancer and the recurrence or non-recurrence of breast cancer data respectively. But the training accuracy of both SVM and KNN (using Euclidean distance) were 100% for both sets of data and the regularised logistic regression was 93.54% and 80% for the data sets respectively. An SVM with the Radial Basis Function (RBF) gave the best outcome when the parameter values (γ is small and C large). The regularised logistic regression used performed better than the generalised logistic regression. The KNN showed to be the best for the

overall methodology.

The SVM is known for its accuracy (Meyer, Leisch & Hornik, 2003) and it is confirmed by the findings of Übeyli (2007), as SVM showed a classification accuracy of 99.5% compared with several types of Artificial Neural Networks (ANN).

All the three algorithms works well on large data sets but logistic regression and KNN are relatively simple to compute compared to SVM. SVM has many desirable traits like accuracy, robustness and effectiveness (Golub et al., 1999; Wang & Huan, 2011) even in non-linear separable problems. KNN is also robust to noisy training data (i.e meaningless data that may be as a result of data corruption or inaccurate recording) and very effective with large training data set (Lavanya & Divya, 2017). Considering the literature reviewed, SVM and KNN tend to be more desirable to contend with the performance of logistic regression.

2.8 Performance measures of the algorithms.

A model evaluation procedure is needed to help estimate how well a model generalise to a different future sample regardless of the choice of set of classifiers, the optimal tuning parameter or choice of different sets of features. However, evaluation metric is needed to pair with procedure so that model performance can be quantified. In classification, using the right performance metric to evaluate a learned classifier is fundamental to assess its quality. One way of evaluating a model's performance is basing it on statistical significance or confidence intervals. Another way is by using a metric for the model's evaluation (Ferri, Hernández-Orallo & Modroiu, 2009). Evaluating a classifier will depend on several factors like predictive accuracy, robustness, scalability, simplicity etc. Notwithstanding there are diverse measures to evaluate a classifiers performance(Mulak & Talhar, 2013) but ROC curve sensitivity, specificity and error rates were used in this study.

Sensitivity can be described as the percentage of the presence of a

condition/activity understudy which is rightly identified by the classifier. For example, the percentage of food establishments identified to have critical violation that truly identify as food establishments with critical violation. It is also called the true positive rate.

$$Sensitivity = \frac{number\ of\ true\ positives}{total\ number\ of\ positives} \tag{2.13}$$

Specificity can be described as the percentage of the absence of a condition/activity understudy which is rightly identified by the classifier. For example, the percentage of food establishments identified to be without critical violation that truly identify as food establishments without critical violation. It is also known as true negative rate.

$$specificity = \frac{number\ of\ true\ negatives}{total\ number\ of\ negatives} \tag{2.14}$$

$$error\ rate = \frac{number\ of\ true\ positives + number\ of\ true\ negatives}{total\ number\ of\ positives + total\ number\ of\ negatives} \tag{2.15}$$

where a true positive as defined by Mulak and Talhar (2013) is the positive tuples that are correctly classified as positive and true negatives are the negative tuples that are correctly classified as negative.

2.9 Summary

The focus of this chapter is on reviewed literature on food establishments, K-Nearest Neighbour and Support Vector Machine (SVM). It gives insight into food safety, its related risk factors and approaches to deal with it. Some works by researches on SVM and KNN are also reviewed as well as comparisons between LR, SVM and KNN.

CHAPTER 3

METHODOLOGY

This chapter describes the various methods that will be used in identifying critical violations in a food establishment. It outlines the feature selection technique used for selecting the attributes that influence the response variable (presence of a critical violation or not). It also explains the criteria for selection of the algorithms that will be employed thus, logistic regression, K-Nearest Neighbour and Support Vector Machine (SVM). A detailed description of the three classifiers and the performance evaluation of the models are also outlined in this chapter.

3.1 Preprocessing the data

In the beginning stages of analysing the data, certain processes has to be followed in order to remove relatively unimportant aspects of the data that might affect the accuracy of the algorithm. All the processes is collectively referred to as preprocessing the data.

3.1.1 Data formatting

Here, data formatting involves prepping the data before any further analysis. With the set objectives in mind, the data is subjected to data formatting which involves cleaning, by treating the missing data (i.e. removal or applying imputation algorithm) and normalising them (i.e rendering data to be in the same range). In this study, missing data was removed since the data was large enough to accommodate the removal of 0.006% of the whole dataset.

The data used contained many variables with different measurement scale so normalisation was done to render the various features to have the same scale.

3.1.2 Feature Extraction

Feature extraction is the process of capturing a subset of features (attribute) from the actual feature set while maintaining interpretation and focus on the goal of the analysis.

Feature extraction is a dimensionality reduction technique that is very vital in providing more information about the data such that only relevant features are included in the training process. There is the likelihood that every feature contain as much information but it is in order of importance or usefulness (Nilsson, Peña, Björkegren & Tegnér, 2007). The goal of feature selection is to remove irrelevant features i.e invariably identifying relevant features in order reduce the likelihood of over-fitting to noisy data, decrease computational time and space needed to run the algorithm. Discussions on the importance on selecting every relevant feature is expounded more by Guyon and Elisseeff (2006).

This study uses Principal Component Analysis (PCA) to reduce the number of features by inextricably extracting the relatively useful features.

Principal Component Analysis (PCA)

The method Principal Component Analysis (PCA) also known as "Hotelling transform" used in data processing transforms features into a set of features which are a linear combination of the original features and these new features are called principal components. PCA is an orthogonal linear transformation that transform data to a new coordinate system such that the greatest variance by some projection of the data lies on the first principal component. The first principal component explains the greatest variability found in the data and every component that follows from the second to the last explains the remaining variability. Only the first few components are considered and interpreted because of the amount of variance they explain. Conventionally, the defining character between components is that they are uncorrelated.

The following highlights the method involved in the PCA assuming the data is n-dimensional;

- Subtract the mean from each of the data dimension. This makes the mean of the data become zero.
- Calculate the covariance matrix. The dimension of the covariance matrix will be an $n \times n$
- Calculate the eigen vectors and the eigenvalues of the covariance matrix as they convey useful information concerning the data.
- Select components and create feature vector. This step welcomes in the dimensionality reduction.
- Form the new data set with features retained in the data.

The goal of PCA is to reduce the dimensionality of the huge data set and also detect new important underlying features in order to explain a certain phenomenon. The eigen analysis is a technique used in Principal Component Analysis. The sums of squares and cross products is used in determining the eigen vectors and eigenvalues of a square symmetric matrix. The relationship between the principal components, eigen vectors and eigen values is such that the eigen vector linked with the greatest eigenvalue has the same direction like the first principal component. It follows that the second eigen vector linked with the second greatest eigenvalue shows the second principal component's direction.

The Principal Component Analysis conveys information that concerns not only the varying patterns in features but also the relationship that exist between features (Qi & Luo, 2015). The final display of PCA analysis, presents components that have different degrees of correlation with the observed features. Jolliffe (2011) provides further explanation on the Principal Component Analysis.

3.2 Logistic regression

Logistic regression is a special type of regression used in predictive analysis where the probability of a dichotomous outcome is modeled based on one or more predictors (numerical or categorical) by means of a logistic function. The model seeks to estimate the probability that an event occurs for any randomly selected observation against the probability that the event does not occur, hence very useful in classification. It explains the relationship between a response/dependent variable and one or more explanatory/independent variables.

Depending on the response variable, logistic regression may be binomial or multinomial. With binomial logistic regression the observed outcome can take on only two categories, a typical example is a "yes" or "no". In multinomial logistic regression the observed outcome has more than two possible categories.

Logistic regression is an example of a Generalised Linear Models (GLM), i.e broad class of models which includes linear regression, ANOVA , Poisson regression, etc. The GLM also originates from a family of distribution known as the exponential family.

3.2.1 The Logistic Regression Model

Consider a data of n independent observations $y_1, y_2, y_3, ...y_n$ and treating the i^{th} observation as a realisation of a random variable Y_i. Assume that the Y_i has a Bernoulli distribution with parameter θ, where $\theta = P(x = 1)$.
A typical probability function (p.d.f) of the Bernoulli distribution is of the form

$$p(x|\theta) = \begin{cases} \theta^x(1-\theta)^{1-x} & \text{for } x = 0 or 1 \quad 0 < \theta < 1. \\ 0 & \text{otherwise} \end{cases}$$

In the exponential form we express the pdf as

$$\begin{aligned} p(x|\theta) &= exp\Big\{log\Big[\theta^x(1-\theta)^{1-x}\Big]\Big\} \\ &= exp\Big[xlog\theta + (1-x)log(1-\theta)\Big] \\ &= exp\Big[xlog\frac{\theta}{(1-\theta)} + (1-x)log(1-\theta)\Big] \end{aligned} \tag{3.1}$$

The general form of the single exponential family of distributions is stated as

$$g_X(x|\vartheta) = g(x)exp(\vartheta.Q(x) - B(\vartheta)) \tag{3.2}$$

Therefore comparing, equations 3.1 and 3.2

$\vartheta = log\theta/(1-\theta)$, $Q(x) = -log(1-\theta)$ and $g(x) = 1$ The natural parameter can be rearranged as,

$$\theta(x) = \frac{1}{1+e^{-(X_i\beta)}} \tag{3.3}$$

the logistic regression model is equation (3.6). In order to fit logistic regression certain assumptions have to be met. This is outlined in the next subsection.

3.2.2 Assumptions of Logistic Regression

Assumptions of the logistic regression model is as follows;

(i) The dependent variable, Y_i is binomially distributed. Therefore there is no need for the Y_i to be normally distributed, but assumes a particular distribution from the exponential family.

(ii) The cases or data, $Y_i, Y_2, ..., Y_n$ are independently distributed.

(iii) A linear relationship is assumed between the logit of dependent and independent variables; $logit(\theta) = \beta_0 + \beta_i X_i$.

(iv) Errors does not need to be normally distributed but must be independent.

(v) The assumption of homoscedasticity is not a necessity.

(vi) Maximum likelihood estimation (MLE) is preferred to the ordinary least squares (OLS) for parameter estimation, which rests on large-sample approximations.

(vii) It usually requires large sample data.

3.2.3 Odds and Odds Ratio

The odds of a dependent variable can simply be described as the ratio of the probability of an event happening to the probability of the event not happening. Thus,

$$Odds = \frac{P(event\ \ happening)}{P(event\ \ not\ \ happening)} = \frac{\theta}{1-\theta}$$

where,

$P(Y = 1) = \theta$ is the probability of an event occurring

and

$P(Y = 0) = 1 - \theta$

The ratio of two odds is defined as Odds Ratio (OR). Also represented by,

$$OR = \frac{odds\ \ of\ \ event\ \ happening}{odds\ \ of\ \ event\ \ not\ \ happening}$$

Defining odds of one event as θ_0 and another as θ_1,

then

$$OR = \frac{\left(\frac{\theta_0}{1-\theta_0}\right)}{\left(\frac{\theta_1}{1-\theta_1}\right)}$$

For a simple logistic regression model with an independent variable, the odds ratio is explained mathematically as,

$$OR = \frac{odds(x+1)}{odds(x)} = \frac{\left[\frac{\theta(x+1)}{1-\theta(x+1)}\right]}{\left[\frac{\theta(x)}{1-\theta(x)}\right]} = \frac{exp[\beta_0 + \beta_1(x+1)]}{exp[\beta_0 + \beta_1 x]} = exp\beta_1 \qquad (3.4)$$

This exponential relationship with the odd ratio suggests that the odds that a particular characteristic is existent is multiplied by $exp(\beta_1)$, for every unit increase in X. Having identity as an odds ratio means the odds do not change with time. Thus,

- Odds increase if $\beta_j > 0$, then exp (β)>1
- Odds decrease if $\beta_j < 0$ then exp (β)<1

3.2.4 Parameter estimation of logistic regression coefficients

The estimation method that was used in this research is the maximum likelihood estimation. This method is a routine procedure for obtaining estimators for unknown parameters from a set of data. Firstly, a function known as likelihood function has to be established. This function can be described as the probability of observing the actual data conditioned on the values of the parameter. The maximum likelihood estimate is the value of the parameter that maximises the likelihood function over the entire ring-space of the parameter. Thus, it is the parameter value that is most likely in the light of what has been observed. Estimates from MLE is known to have some enviable properties such as efficiency, consistency, invariance and asymptotic normality. Maximum Likelihood Estimator is preferred because it utilises every information about the parameters found in the data and it's comparatively highly flexible. (Denuit et al., 2007).

Assume a probability distribution is defined by a parameter α. If the likelihood

function $L(\alpha)$ of observations (Z_j) could be created from the distribution having probability density $f(z)$, the likelihood can be defined as the product of the individual likelihoods over the observations. If the data is a vector $Z = (Z_1, Z_2, ..., Z_n)$ with parameter vector $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$ defined on a multi-dimensional parameter space from an unknown population with pdf $f(Z, \alpha_1, \alpha_2, ..., \alpha_n)$. The likelihood for each model given by

$$L(Z_1, Z_2, ..., Z_n|\alpha_1, \alpha_2, ..., \alpha_n) = L(Z|\alpha) = \prod_{i=1}^{n} f(Z, \alpha_1, \alpha_2, ..., \alpha_n) \tag{3.5}$$

For a binary logistic regression model, the likelihood function,$L(\alpha)$ can be expressed as,

$$L(\alpha) = \prod_{i=1}^{n} \alpha_i^{y_i} \Big[1 - \alpha_i\Big]^{1-y_i}$$

The maximum likelihood estimate is the value of α that maximises the likelihood function. However, it is more expedient to use log likelihood rather than the likelihood (Geyer, 2003). The log likelihood is expressed as

$$lnL(Z|\alpha) = ln\left[\prod_{i=1}^{n} f(Z_i, \alpha)\right] = \sum_{i=1}^{n} lnf(Z_i, \alpha) \tag{3.6}$$

Also,

$$L(\alpha) = lnL(Z|\alpha) = \sum_{i=1}^{n} \Big\{y_i ln\alpha_i + (1 - y_i)ln(1 - \alpha_i)\Big\} \tag{3.7}$$

The MLE is obtained by finding the derivatives of $InL(Z|\alpha)$ with respect to α and equating it to zero. Thus,

$$\frac{\partial ln(\alpha)}{\partial \alpha}\bigg|_{\alpha=\hat{\alpha}} = 0 \tag{3.8}$$

3.2.5 Testing the Goodness – of – Fit

The term goodness of fit is very useful in comparing the observed sample distribution with the expected probability distribution. It involves assessing a random sample from an unknown distribution to test the null hypothesis that

the unknown distribution function is actually from a known distribution.

The procedure for determining the Goodness–of–Fit is by stating a hypothesis, calculating the test statistic and then computing the probability of finding data which have a greater value of this test statistic than the observed value. If the hypothesis is true, the probability is known as the confidence level. In assessing the model fit, some of the techniques that will be employed are below.

Deviance and likelihood ratio tests

Several software algorithms employ the deviance rather than the log-likelihood function as the basis of convergence when using GLM to estimate logistic models (Lovric, 2011). Deviance measures the lack of fit to a data and it is computed by making a comparison between a given model and a saturated model.

Also expressed as,

$$D = -2ln\left(\frac{likelihood\ \ of\ \ fitted\ \ model}{likelihood\ \ of\ \ saturated\ \ model}\right) \tag{3.9}$$

Using equations (10) and (12), the deviance for a logistic regression model can be stated as

$$D = 2\sum_{i=1}^{n}\left\{y_i ln\left(\frac{y_i}{\alpha_i}\right) + (1-y_i)ln\frac{(1-y_i)}{(1-\alpha_i)}\right\} \tag{3.10}$$

Inside the brackets is the quantity known as the likelihood ratio. A test of such nature is termed the likelihood ratio test. Saturated model is a model having a theoretically perfect fit. Obtaining smaller values gives an indication of a better fit since deviation from the saturated model is small.

Assessing upon a chi-square distribution, will give an indication of a good model fit when non-significant chi-square values are obtained which suggests a lot of the variance is explained and, importantly, that little remains unexplained.

The value of D with and without the explanatory variables needs to be compared in order to assess the significance of the predictors in the equation. Therefore,

the model deviance is subtracted from the null deviance. Thus

$$\begin{aligned} D_{null} - D_{model} &= \Big[-2lnL(\alpha)_{null} - (-2)lnL(\alpha)_{saturated} \Big] - \\ &\Big[-2lnL(\alpha)_{fitted} - (-2)lnL(\alpha)_{saturated} \Big] \\ &= \Big[-2lnL(\alpha)_{null} \Big] - \Big[-2lnL(\alpha)_{fitted} \Big] \\ &= -2ln \left[\frac{Likelihood\ \ of\ \ the\ \ null\ \ model}{Likelihood\ \ of\ \ the\ \ fitted\ \ model} \right] \end{aligned}$$

To assess the significance of any individual predictor the likelihood ratio test and the Wald statistic is preferably used.

The deviance test is central to the likelihood test since it tests the significance of the difference between the likelihood ratio's of the fitted model and null or reduced model. Consider the null hypothesis;

$$H_0 : \beta_j = 0. \ \ i = 1, 2, ..., p$$

The statistic for the likelihood ratio test is expressed as

$$-2ln \left(\frac{L_{null}}{L_{fitted}} \right) = -2 \Big[In(L_{null}) - In(L_{fitted}) \Big]$$

A chi-square statistic is obtained with this log transformation of the likelihood functions.

Another statistic that can be used is the Wald statistic. The Wald test is achieved by comparing estimate of maximum likelihood of the slope parameter,$\hat{\beta}_j$ to an estimate of its standard error.Under the null hypothesis

$$H_0 : \beta_i = 0. \ \ j = 1, 2, ..., p$$

The subsequent ratio follows a standard normal distribution.The Wald statistic

(W_i) is expressed as

$$W_i = \frac{\beta_j}{SE(\beta_j)} \sim N(0,1) \tag{3.11}$$

The Wald statistic however has some limitations, thus for large coefficient the standard error becomes inflated thereby reducing the Wald statistic value (Manard, 1995) and it is also inclined towards biasedness with a sparse data. Generally the likelihood ratio test is preferred over the Wald test.

McFadden's pseudo-R squared

A different approach to assess the effectiveness of a regression model is by measuring the strength of the relationship among the independent variable(s) and the outcome. The McFadden's pseudo-R squared is one of many versions founded on the log-likelihoods for the null model and full estimated model. Other version that can be used are Hosmer & Lemeshow's R^2 and Nagelkerke's R^2, Cox and Snell R^2. McFadden's R squared measure is defined as

$$R^2_{McFadden} = 1 - \frac{In(L_{full})}{In(L_{null})} \tag{3.12}$$

where L_{full} denote the likelihood from the current fitted model and L_{null} denote the likelihood from the null model The value obtained ranges from 0 to 1. These type of statistics can be suggestive on their own, but very useful when comparing competing models for the same data.

3.2.6 Confidence Interval Estimation

The confidence interval simply shows how accurately with which a sample statistic estimates a population parameter, given random sample size,N and the significance level, α. Usually the confidence interval for this slope is built from the Wald statistic. A $(1-\alpha)\%$ two-sided confidence interval for β_1

$$\hat{\beta}_1 \pm Z_{1-\alpha/2} SE(\hat{\beta}_1) \tag{3.13}$$

where $SE(\hat{\beta}_1)$ represents the standard error of a model-based estimate of the respective estimator of the parameter and $Z_{1-\alpha/2}$ is the upper $100(1-\alpha/2)\%$ point from the standard normal distribution.

3.3 Support Vector Machine

Support Vector Machine (SVM)(Vapnik, 1998) is an effective method that has a solid theoretical foundation and also has the ability to learn automatically from data and improve from experience(also termed as machine learning). SVM identifies a maximum margin function that divides a large set of observations into two categories where every observation is a point in a multidimensional space of feature measurements. It is known for its high prediction accuracy which is as result of learning from the training set (Meyer, Leisch & Hornik, 2003) to produce an optimal hyperplane which significantly simplifies classification and regression problems. High robustness and generalisation ability of SVM with a small number of samples are some of its many admirable features. Thus, SVM considers minority support vectors together with the complexity and learning ability of the model to define the final optimal hyperplane.

There are two problems to deal with when using SVM are; by what method to select the optimal count of features and how to fix the best kernel parameters. Both problems influence each other (Frohlich and Chapelle, 2003), hence finding optimal number of features and kernel parameters should simultaneously occur.

Selecting a subset of features from a database is significant in SVM as it extracts the important information from the available data set to reduce the computation time. As stated by Yang and Honavar (1998), the number of features affect not only a learned classification algorithm's accuracy but the time needed for learning a classification function, the cost associated with the features and the number of examples needed for learning. A few algorithms have been proposed for SVM feature selection in the literatures (Bradley et al., 1998; Bradley & Mangasarian,

1998; Weston et al., 2001; Guyon et al., 2002; Mao, 2004).

Suppose a training data $\{x_i, y_i\}_i = n$ where every x_i represents a training element and $y_i \in \{+1, -1\}$ a matching class label. The goal of the SVM problem is to find a hyper-plane that divides the two categories or classes of points with the highest separation margin. The foundation of this technique is mapping the input vector unto a high dimensional feature space using non-linear transformation function. Since exact separation between the two categories is extremely difficult, some error allowance variables known as slack variables ξ_i are introduced in classifying the data that is difficult to separate linearly (Vapnik, 1995). The categorisation of the surface equation $\omega \cdot x_i + b = 0$ satisfies the equation below:

$$y_i[(\omega \cdot x_i + b)] \geq 1 - \xi_i, \quad i = 1, 2, ..., n \tag{3.14}$$

where

ω is a weight vector

b is the classification threshold.

$$\begin{cases} if \;\; 0 < \xi_i < 1 & x_i \text{ is accurately classified} \\ if \;\; \xi_i \geq 1 & x_i \text{ is wrongly classified} \end{cases}$$

Below is the objective function

$$\phi(\omega \cdot x) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n} \xi_i \tag{3.15}$$

where

$\frac{1}{2}\|\omega\|^2$ is minimize objective function

C is a regularization parameter

$C\sum_{i=1}^{n} \xi_i$ is a penalty function

Equation 3.15 can be fixed by the convex quadratic programming below.

$$max \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j (x_1 \cdot x_j), \;\; 0 \leqslant a_i \leqslant C, \;\; i = 1, 2, ..., n \qquad (3.16)$$

a_i is Lagrange multiplier. Assuming a^* is the best solution, then

$$\omega^* = \sum_{i=1}^{n} a_i^* y_i x_i \qquad (3.17)$$

A linear combination of support vector can be defined as the face of the generalized optimal classification. The optimal classification of the function is given as:

$$f(x) = sgn\Big[\sum_{i=1}^{n} a_i^* y_i (x_i \cdot x) + b^*\Big] \qquad (3.18)$$

where b^* is a threshold classification without omitting the constraint condition $a_i[y_i(\omega^* \cdot x_i + b) - 1] = 0$

In situations where there are no relations linearly between the outcome and the predictors, enlarging the feature space by making use of functions of the predictors like cubic, quadratic or even higher-order polynomial functions are considered to address the non linearity problem. This could lead up to an immense number of features which intends make computation not manageable. SVM allows the support vector classifier to enlarge feature space using kernels in a way that leads to efficient computations (James et al., 2013). This leads to the importance of the selection of a kernel function and feature selection procedure.

3.3.1 Parameter selection

The selection of appropriate parameters is important in improving the classification accuracy of the SVM. Selecting the parameters is a requirement before training the SVM model. The regularization parameter C and the parameters of the kernel function, γ in the Radial Basis Function (RBF) kernel are

some parameters that should be optimized. There exist so many approaches to the selection of parameters, they comprise cross validation method, particle swarm optimization, experience choice method, Bayesian method, gradient descent method, Genetic Algorithm (GA) based method etc.

The study made use of cross validation to find the best parameter. A parameter range should be identified before concluding on the best parameters, and if the range is small, then a huge deviation for optimal parameters will be produced.

3.3.2 Proposed Procedure

The following illustrates the proposed procedure for the SVM classification algorithm in this study :

- Make available the collected data and transform the data to be compatible with the SVM package.
- Perform basic scaling on available data
- Select which kernel function to use. The RBF, linear and polynomial kernels were considered.
- Use the cross validation and a defined grid search procedure to choose SVM parameters to use. The procedures was used to determine the finest parameter, thus C and γ, the parameter's of the RBF kernel and C with the linear kernel and also the C, d and γ
- Train the algorithm using the optimum parameters obtained Make use of the optimum parameter identified to develop training set.
- Evaluate on the test set.

3.3.3 Data Preprocessing

Categorical Feature

For the SVM package to be compatible with the data, transformations should be made to the data. First, each data case has to be denoted by a vector of real numbers. This means any categorical variable attribute has to be converted into a numeric data. Hsu et al,(2003) suggest making use of p numbers to denote an attribute of p-category. For instance, a category with three attributes like {amateur, medium, difficult} can be denoted by (1,0,0), (0,1,0), (0,0,1). In cases where the total values in a particular attribute is not so huge, this type of coding is suggested since it may bring more stability than considering a single number.

Scaling

Scaling is prerequisite in order to apply the SVM. Scaling allows all numeric ranges to be the same across to avoid dominance of one range over the other and it also make numerical calculations simple. This is important since the kernel values ideally hinge on the the inner products of feature vectors. Every attribute will be scaled linearly in the range [0,1] in both data (training and testing).

3.3.4 Model Selection

Conventionally, SVM uses a kernel that is a set of mathematical functions whose function is recognising data as input and transforming it into the required form. This study used the linear kernel and the RBF kernel.
Here C the regularisation parameter (cost parameter) measures the trade-off between maximising the width of the margin and minimisation of errors (Eitrich & Lang, 2006)

The RBF Kernel is a popular model selection method as it has the ability to accommodate cases where data sets are not linearly separable. It is able to map samples that are nonlinear into a higher dimensional space. The RBF has only two

parameters thus C (regularisation parameter) and ω (kernel parameter). These two parameters is vital in the SVMs performance since it can lead to under-fitting or over-fitting troubles when selected inappropriately. A suggested approach to deal with this is making use of grid-search and cross validation (Hsu, Chang & Lin, 2003). The clear objective is to identify the best choice of both C and ω so that the model can generalise to new data set. However, other kernel behave like the RBF Kernel, example is the sigmoid kernel in some particular parameters (Lin and Lin, 2003) and the linear kernel behave like the RBF Kernel (with C, γ) in special cases where linear kernel has a penalty parameter C (Keerthi and Lin, 2003). Apart its popularity, RBF Kernel has less the number of hyperplanes that influence the model selection's complexity than the polynomial kernel which relatively has higher and hence the fewer numeric difficulties in RBF Kernel informed the decision for it's selection. Thus, values of a polynomial kernel can approach infinity but the value of a RBF Kernel is between 0 an 1.

Owing to the fact that logistic regression was used by Schenk et al. (2014), a linear kernel was also used. The linear kernel do not have parameters to tune besides C which makes it relatively flexible. It performs best when the data is linearly separable and relatively takes less computational time to train the SVM as compared to the RBF kernel. It less likely to lead to overfitting as compared to the RBF kernel.

Notwithstanding all these facts, the polynomial kernel was included purposefully for comparison with the other kernels.

The proper kernel selection ultimately has influence on the accuracy of the SVM classifier (Asraf, Nooritawati & Rizam, 2012).

3.3.5 Cross-validation and Grid-search

The parameters of the kernels are already unknown and selecting the optimal parameters demands some type of model selection. In order to predict accurately

the testing data, the optimal C, γ and d has to be identified. Before a classifier can be termed as having a high or low prediction accuracy, the data is divided into training and test data where the trained classifier is used on the test data after training the classifier on the training data. The capability of any classifier is rated when it is able to predict accurately the test data, hence the use of cross validation. Cross validation can be described as the measure of how the classifier in a statistical research is able to generalize even to unknown data set by dividing the data into equal sub-data sets.

In a typical v-fold cross validation, an equal partition of v subsets is made out of the training set. When the classifier is trained on the $v-1$ subsets, the trained classifier is used to test the remainder of the v subsets. Hence, every case in the entire training set is predicted once, therefore cross-validation accuracy can be described as the proportion of the data set which are precisely classified. The cross-validation approach is used in order to avoid overly fitted data.
A 10-fold cross validation was used in this study. This method prevents the classifier from over-fitting In this setting the training data is randomly divided into subsets of 10 with equal size. Out of the ten(10) subsets, one (1) is held to test the model and the 9 remaining subsets are used to train the data. This process of holding out one subset and training the model on the remaining 9 is repeated such that each of the 10 subsets is used only once as the test or validation data.

A defined grid-search was applied in selecting the optimal C,d and γ using cross-validation. The grid-search tries all the sets of C,dγ values and anyone having the top cross-validation accurateness is selected. Hsu et al, (2003) suggest a practical way of using sequences of C and γ in an exponentially growth trajectory to determine appropriate parameters (for instance, $(C = 2^{-6}, 2^{-4}, 2^{-2}, ..., 2^{-12}, \gamma = 2^{-12}, 2^{-10}, 2^{-8}, ..., 2^{-4})$. However, this study used a defined grid. This was informed by the default tuning parameters in software package like MATLAB. The defined grid used had the values $0.25, 0.5, 0.75, 1, 1.25$ as the C in all the

kernels, $0.01, 0.015, 0.2$ as the γ for the RBF kernel. The γ and d in the polynomial kernel were define by the values $0.01, 0.015, 0.2$ and $1, 2, 3$ respectively.

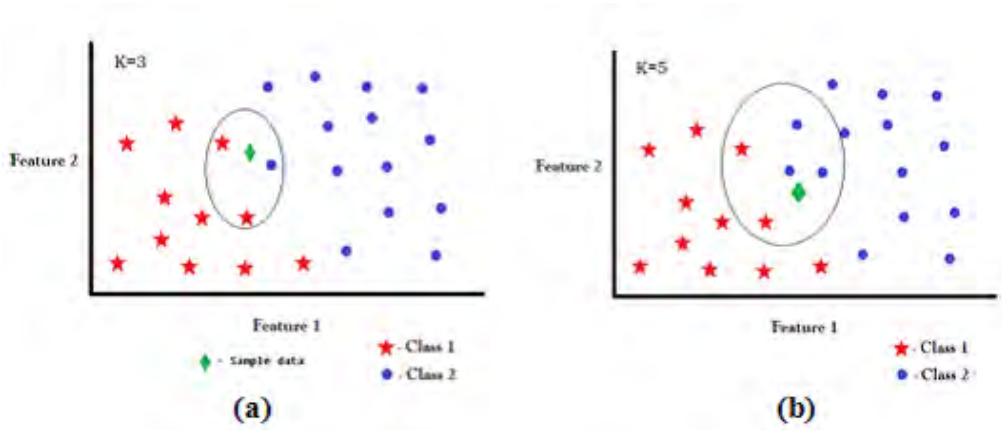
3.4 K-Nearest Neighbour

The nearest neighbor algorithm finds the class of undefined data point based on its nearest neighbor whose class is defined aforehand. It has been extensively used in recognition of patterns (Vaidehi, 2008; Xu, & Wu, 2008), ranking models (Geng, 2008), event recognition (Yang, 2000) applications and text or categorization (Elnahrawy, 2002). The algorithm is non-parametric in nature and its usually used for regression and classification. It is based on classifying cases based on their similarity. In the world of machine learning, this method was developed to identify patterns in a data without the requirement of exactly matching any stored case or pattern and it is considered by many as the simplest algorithm among all machine learning algorithms. It is an effective method that is relatively easy to execute (Bhatia, 2010). In simple terms, when similar or matching cases are nearer to each other and not similar or mismatching cases are far from each other, the similar cases are termed as "neighbours". The distance between two cases measure their dissimilarity. Assuming a new case is presented, the distance between existing cases and the new case is computed. Similar cases are classified and tallied and the new case is given a class based on the highest nearest neighbours the class has. The number of nearest neighbour, K is specified beforehand by making some considerations. The K, can be termed as a user-defined constant.

3.4.1 K - value selection

Selecting a desirable K rests entirely on the data. A larger K greatly reduce noise in classification but also fails to set clear boundaries among the classes and invariably a small K leads to a large variance in prediction. Therefore K should be set at a point where it is large enough to reduce misclassification and small

enough to let the K nearest cases to be close to the new case. Consider the following figure



The goal here is to classify or estimate the new case contingent on the number of nearest neighbour around it. If the green point is a new case to be classified, and considering both selection on the boundary in (a) and (b) with $K = 3$ $\&K = 5$ respectively. Again, considering the first case with k=3, the new case is said to belong to class 1 since the selected boundary has more class 1 nearest neighbours. Similarly in (b), the new case is said to belong to class 2 since the selected boundary has more class 2 nearest neighbours.

3.4.2 Training and testing of K-NN Classifier

In classification, the dependent variable is categorical therefore any point introduced is classified by majority voting. The choice of the K value greatly affect the quality of prediction. The K-Nearest Neighbors algorithm is mostly associated with regression and classification. In the feature spaces of both cases the input hold the closest training examples.

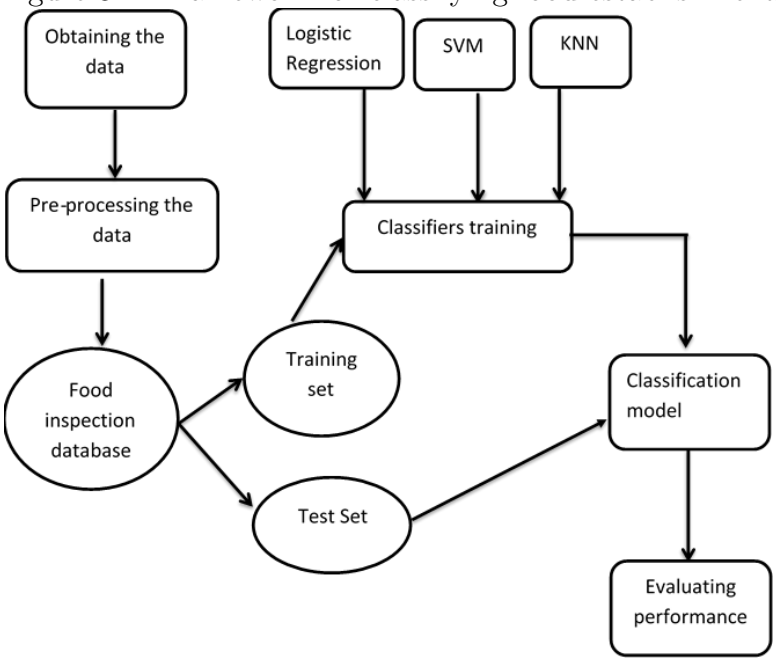
In this study, data on food inspections is partitioned into two thus training and test sets. The K-NN classifier is set to classify the data as either there is critical violation or not, by learning the already defined training data. The cross validation technique is then applied to check the accuracy of the K-NN classifier.

3.4.3 Steps

The following steps describes the use of the KNN classifier.

1. Preparing the data which involves:
    - Normalising the data, thus adjusting features such that inferences drawn are not distorted by variables with wide ranges.
    - Dividing the data into training and testing sets in order to assess the performance of the algorithm.
2. Storing the class labels and training samples by subjecting it to preprocessing
3. Classification is done by majority vote. Any new case/instance introduced follows these steps:
    - All distances between a test sample and a training sample is calculated using a distance metric (Euclidean was used in this study).
    - With a pre-defined value of k ranging from 1 to 50, the class of a new vector is determined.
    - A new case is assigned a class if among the K nearest nearest samples it is the most frequent class.
4. Assessing model performance.

Figure 3.1: Framework for classifying food establishments



The figure 3.1 describes the processes involved in classifying restaurants using the logistic regression, SVM and KNN algorithms. After obtaining the data, it is subjected to pre-processing, which involves formatting the data, cleaning, scaling and feature extraction. The data is then divided into train set and test set with 16697 and 1636 observations(restaurants) respectively. Each classifier learns from the training set and subsequently evaluated on the test set.

3.5 Criteria for selection of algorithms

There are many supervised learning algorithms such as Support Vector Machine (SVM), Logistic Regression (LR), *Naïve* Bayes K-Nearest Neighbour (KNN), random forest, decision tree, etc. Supervised learning algorithm focuses on building a model able to make predictions of the response values for a new dataset. The main challenge posed in supervised learning is the selection of the appropriate data mining algorithm for classification. When considering supervised learning algorithms, several factors like linearity, accuracy, training time, number of features, complexity and interpretability of the algorithm are the usual victims but a lot also lies in the data. One underlining criteria vital for

the selection of an algorithm is based on characteristics of the training data set such as the size, quality and nature.

The LR, SVM and KNN are considered in this study are all examples of data mining algorithm. The name data mining connotes sifting through large data sets to make meaningful inferences. Both predictive data mining algorithms have a bigger capacity to handle large and even noisy data as a study by Lavanya et al. (2017) showed their strength in big data analysis. The data used in this study is considerably large with 28 features and 18334 observations. It is obtained from the City of Chicago data repository which has a world class quality. Also the nature of the data makes it possible to use both algorithms since the data suits a classification overlaid problem. Also, with the study aiming to prioritise inspections based on a validated set of models, one vital characteristic that cannot be overlooked is the predictive power of the algorithm. Several studies show SVM and KNN have high predictive power (Meyer et al., 2003; Wang & Huan 2011; Alkhatib, Najadat, Hmeidi & Shatnawi, 2013) This informed the decision to use SVM and KNN.

3.6 Performance evaluation of the models

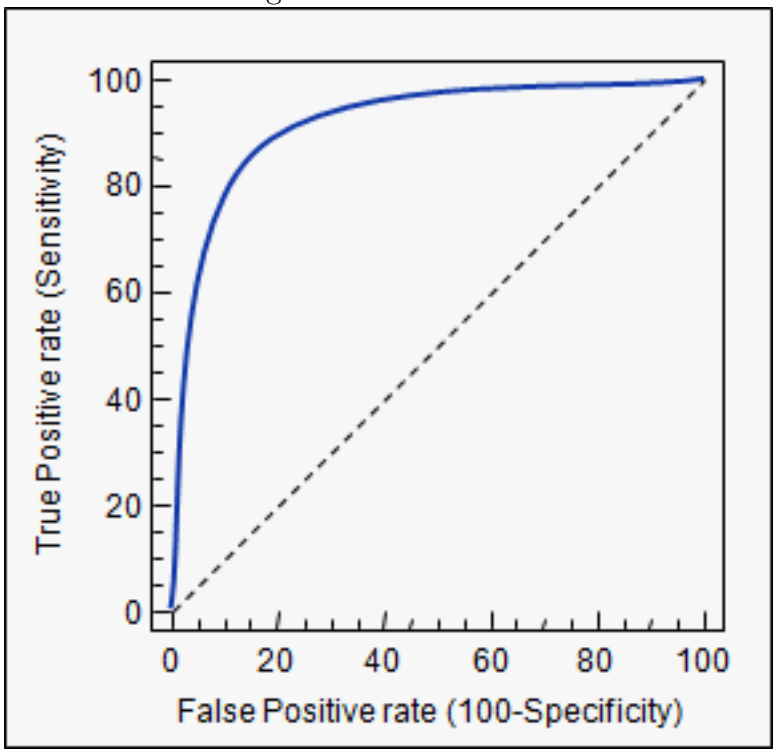
Evaluating a classifier will depend on several factors like predictive accuracy, robustness, scalability and simplicity, however there are diverse measures to evaluate a classifiers performance(Mulak & Talhar, 2013). This study will consider ROC curve specificity, sensitivity and error rates to assess the performance of the classifiers.

3.6.1 Receiver Operating Characteristic (ROC) curve

The ROC curve is a visual display of diagnostic test evaluation measure that plots the sensitivity (true positive rate) and specificity (false positive rate). Every point in the ROC curve denotes a sensitivity/specificity pair consistent with a

specific discriminative threshold. The area under curve is considered when one wants to determine how well a particular parameter can discriminate among two groups. The area under the ROC curve quantifies how accurate the model does in classifying members. We describe a test to have a perfect discriminative measure (100 % specificity and 100% sensitivity) if the ROC curve passes through the upper left corner of the curve below.

Figure 3.2: ROC curve



We describe a test to have a perfect discriminative measure (100 % specificity and 100% sensitivity) if the ROC curve passes through the upper left corner of 3.2

$$Sensitivity = \frac{number\ \ of\ \ true\ \ positives}{total\ \ number\ \ of\ \ positives} \tag{3.19}$$

$$specificity = \frac{number\ \ of\ \ true\ \ negatives}{total\ \ number\ \ of\ \ negatives} \tag{3.20}$$

$$error\ \ rate = \frac{number\ \ of\ \ true\ \ positives + number\ \ of\ \ true\ \ negatives}{total\ \ number\ \ of\ \ positives + total\ \ number\ \ of\ \ negatives} \tag{3.21}$$

where a true positive as defined by Mulak and Talhar (2013) is the positive tuples that are correctly classified as positives and true negatives are the negative tuples that are correctly classified as negatives.

Using a simple 2×2 table (Confusion matrix) to illustrate further,

Table 3.1: Confusion Matrix

	Event	No event
Event	Q	R
No event	S	T

$$Sensitivity = \frac{Q}{Q+S} \tag{3.22}$$

$$Specificity = \frac{T}{R+T} \tag{3.23}$$

The study also makes use of classification accuracy(Acc_{cl}) which is the portion of samples that are rightly classified (i.e addition of true positives and true negatives). A good model must be able to accurately classify every member of a dataset. It is evaluated by the formula:

$$Acc_{cl} = \frac{c}{m} \times 100 \tag{3.24}$$

where c is the portion of rightly classified samples and n, all the samples used.

3.7 Summary

The focus of this chapter is on the supervised learning algorithms (i.e logistic regression, K-Nearest Neighbour and Support Vector Machine (SVM))used. It

gives insight into how the various parameters are selected and estimated, and also sets out some definitions used. The criteria for selection of the supervised learning algorithms are also spelt out. The data considered satisfies all the classifiers which is very paramount in model building.

CHAPTER 4

DATA ANALYSIS AND DISCUSSIONS

This chapter presents the varied stages of analysing the data and explanation of findings in the study while supporting them with relevant literature. It focused on using the supervised learning algorithms, logistic regression, support vector machine and K Nearest Neighbour under varying parameter settings to detect critical violations at food establishments.

4.1 Data collection and description

This study used secondary data from the City of Chicago data portal. Some desirable features considered in selecting the data has a lot more to with the availability, quality and accessibility coupled with the dearth in literature on application of SVM and KNN on such data. So many sources of data and attributes were gathered and used in the development of the models. The data sets, contained information on food inspections, business licenses, detailed crime data, 311 complaints from patrons of food establishment such as sanitation and garbage complains, and finally weather data in Chicago from 2011 to 2014.
Similar data was used in the Schenk et al. (2014) which entailed the information on extracted variables. The binary response variable is whether or not there is the presence of critical violation. The independent variables are as follows;

Table 4.1: Description of variables

No.	feature	Description of feature
1	timeSinceLast	time since last inspection
2	pastCritical	history of previous risk level "critical" associated with each food establishment
3	patSerious	history of previous risk level "serious" associated with each food establishment
4	criticalCount	number of risk level type "critical" at food establishment on inspection
5	seriousCount	number of risk level type "serious" at food establishment on inspection
6	ageAtInspection	business license age at the time of inspection
7	BlueInsp	A feature identifier for first sanitarian cluster (Blue)
8	BrownInsp	A feature identifier for second sanitarian cluster (Brown)
9	GreenInsp	A feature identifier for third sanitarian cluster (Green)
10	OrangeInsp	A feature identifier for fourth sanitarian cluster (Orange)
11	PurpleInsp	A feature identifier for fifth sanitarian cluster (Purple)
12	YellowInsp	A feature identifier for sixth sanitarian cluster (Yellow)
13	license_insp	presence of license for consumption
14	package_goods	sale of packaged goods
15	pastFail	previous failure at inspection
16	tobacco_sale	licensed for sale of tobacco over the counter
17	PubAmusement	presence of public space of amusement
18	burglary	recent burglaries
19	sanitation	recent sanitation complaints
20	garbage	recent garbage cart requests
21	precipIntensity	average expected intensity of precipitation
22	regBisLic	regulated business license
23	filling_station	nearness to filling station
24	catLiqLic	caterers liquor license
25	mobFoodLic	mobile food license
26	temperatureMax	daily high temperature on the day of inspection
27	windSpeed	wind speed on the day of inspection
28	humidity	humidity on the day of inspection

The data considered only regular canvass inspections and inspections as a result of a complain. An observation in the model represents every regular inspection. A food establishment based on the expectation of their food handling practices are classified as risk one, two or three. A food establishment is assigned a "risk one (High)" category if it handles the preparation of food and ingredients directly or heating and cooling of food. A food establishment is assigned a risk three (Low) category if it deals in already packaged and non perishable foods. A risk two (Medium) suggests the food establishment engages in both i.e already packaged goods and direct handling of food and ingredients. In a year, the number of times inspections are directed at a particular food establishment is determined

by the food handling practices of the establishment. Risk one establishments are frequented twice a year by food inspectors, risk two facilities are checked at least once a year and risk three facilities, once every other year.

The data also contained individual sanitarian inspectors who were grouped into clusters such that each cluster were assigned a colour coded name in order to hide their identities.

Schenk et al. (2014) used data from January 2011 to January 2014 as the training data set and evaluated the classifier on data from September 2014 to October 2014. Sufficient time space was allowed between the training data set and the test data set (evaluation set) in order to decrease likely correlation between both periods. In order to make a case for the comparison of the three classifiers relative to the work by Schenk et al. (2014), similar data had to be used. Available data only made it possible to use data from September 2011 to March 2014 as training data set and September 2014 to October 2014 as the test set.

4.2 Research Design

The study followed these steps as the plan in designing this research:

1. The raw data set was preprocessed which involves transforming it to the desirable format by filtering, scaling and removing missing data. It was then divided into training and test data sets.

2. The Principal Component Analysis (PCA) was primarily used to extract the relatively relevant features.

3. The SVM kernel functions namely linear, Radial Basis Function (RBF) and polynomial were employed using a defined range of tuning parameters.

4. Different K values were used in the K-Nearest Neighbour algorithm to determine the optimal model.

5. A 10-fold cross validation was used to prove the reliability of the outcome of the models, where repetitively nine sets are used for training and the remaining used for testing such that each run uses different sets.

6. All the models were evaluated on the test data set and compared based on their classification accuracy.

4.3 Preliminary analysis

This section focuses on the preparation of the data for model building. In order to feed the correct data to the algorithms (LR, SVM and KNN), the data has to be prepared adequately for correct analysis. The data first has to be in the right format, so that only relevant attributes are included and this is achieved by subjecting it to preprocessing. In prepocessing the data, it was formatted to be in the right shape and data cleansing was done by removing missing data (i.e instances where the data are incomplete). The final data used in the model comprised of 16697 inspections from September 2, 2011 to March 31, 2014 (two and half years) as training set and 1636 inspections from September 2, 2014 to October 31, 2014 as test set. The division of the data in this format is informed by Schenk et al. (2014). The data was normalised, so that the range of the exploratory variables will have the same scale (Quackenbush, 2002). The data was analysed to select the best features relevant to the study and this is explained in the next subsection.

4.3.1 Extraction of features

In the preliminary analysis of the data, principal component analysis was used to select the best features relevant to the study. The results from the principal component analysis showed two selected principal components. The scree plot showed a significant reduction in eigen value and leveled-off on the third component and this is generally regarded as the criterion for identifying

the number of components to interpret (See appendix 5.1 for the scree plot). Also, only two components had eigen values greater than 1, hence only two components were retained. The table below shows the correlation between the principal components and the actual features.

Table 4.2: Rotated Component Matrix

	Components	
	1	2
pastSerious	0.841	
pastFail	0.813	
timeSinceLast	-0.608	
pastCritical	0.547	
ageAtInspection		
license_insp		
humidity		
windSpeed		
temperatureMax		
PubAmusement		
OrangeInsp		
BrownInsp		
mobFoodLic		
sanitation		0.638
burglary		0.632
garbage		0.623
tobacco_sale		0.452
BlueInsp		0.357
seriousCount		0.352
package_goods		
filling_station		
GreenInsp		
criticalCount		
YellowInsp		
catLiqLic		
regBisLic		
PurpleInsp		
precipIntensity		

Extraction Method: PCA

Table 4.2 shows only values greater than 0.3 (Samuels, P. 2016; Field, 2013). These values are the farthest from zero in either direction and it represent the features that are strongly correlated with the principal components. Four (4) of

the original features highly correlated with the first principal component whiles the second principal component correlated with six (6) of the original features. It showed 10 features as relatively important to be retained out of the 28 features.

Also the 10 variables retained had relatively high communalities. The communalities indicate the effect of each observed feature from all the factors related to it. In addition the variables that were eliminated also had the lowest communalities or the amount of variance explained as compared to the rest (see appendix, table 5.2). The table displays the relevant features retained from the bunch of available features:

Table 4.3: Features used in the model

Name of Feature	Description of feature
BlueInsp	A feature identifier for first sanitarian cluster
pastFail	Presence of previous record of failures
seriusCount	Number of serious violations
pastCritical	Presence of critical violation upon last visit
pastSerious	Presence of serious violation upon last visit
timeSinceLast	Time passed since last inspection
tobacco_sale	Licensed to sell tobacco
burglary	The intensity of burglaries (locally)
sanitation	Intensity of recent sanitation complains (locally)
garbage	Intensity of recent garbage cart requests (locally)

4.3.2 Descriptive statistics of the normalised data

Some descriptives of the features or attributes selected are illustrated in the tables below.

Table 4.4: Descriptive statistics of the normalised data

Variable	Mean	Standard deviation
past Critical	0.47340	0.12335
past Serious	0.47384	0.15287
serious counts	0.47938	0.18803
Time since last inspection	0.50813	0.22228
burglary	0.48679	0.19379
Garbage cart requests	0.48875	0.197096
Sanitation complains	0.48334	0.16088

From table 4.4, most of the features had relatively lower means with the feature "time since last inspection" having the highest among them suggesting most establishments are at the level of risk three where such establishments are inspected once every other year.

The data showed 1685 food establishments were licensed to sell tobacco whereas 16648 were not. The colour coded cluster of inspectors namely Blue, did 3323 inspections out of the total 18333. There were also 1609 record of previous failures out of the 18333 inspections. (See appendix table 5.1).

4.4 Logistic Regression (LR)

This section involves using the training set to fit the model and evaluating on the testing set.

4.4.1 Logistic regression model

Below is a summary of the model

Variables	Estimate (β)	Wald	Std. Error	z value	Pr(>|z|)	95% C.I Lower	for EXP(β) Upper
(Intercept)	10.801	463.929	0.330	32.717	$< 2e-16$ ***	-	-
pastFail	0.308	2.827	0.183	-1.681	0.093	0.950	1.948
pastCritical	-0.874	12.387	0.248	-3.519	$4.32e-4$ ***	0.256	0.679
pastSerious	-0.085	0.057	0.358	-0.238	0.812	0.455	0.1854
seriousCount	-20.666	1116.304	0.619	-33.412	$< 2e-16$ ***	0.000	0.000
timeSinceLast	-0.601	15.761	0.151	-3.970	$7.19e-05$ ***	0.407	0.738
BlueInsp	1.031	216.587	0.070	-14.717	$< 2e-16$ ***	2.443	3.215
tobacco_sale	-0.316	6.816	0.121	2.611	0.009 **	0.575	0.924
burglary	-0.568	10.045	0.179	-3.169	0.002 **	0.399	805
garbage	0.745	18.353	0.174	4.284	1.84e-05 ***	1.498	2.961
sanitation	-0.138	0.405	0.216	-0.636	0.525	0.570	1.331

Signif. codes: 0 '***', 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1

Table 4.5: Summary of the logistic regression model

From table 4.5 the variables, past failures, presence of serious violation on last visit and sanitations complains are not statistically significant at 5% level of significance. The variables, presence of critical violation on last visit, number of serious violations, time passed since last inspection, first sanitarian cluster, the intensity of local burglaries and intensity of recent garbage cart requests are statistically significant since their $p-$values is less than 0.05. The number of serious violations had the least $p-$value suggesting a very strong association between the number of serious violation and the probability of passing an inspection. Again, its coefficient having a negative sign suggests, with all variables

equal an increase in serious counts would lead to a likely failed outcome in inspection. Since the log odds is a function of the estimated coefficients, an increase in a variable like for example a licensed tobacco seller, increases the odds by $e^{0.316}$ (0.7290595). Considering the other independent variables in the model, the 95% Wald Confidence limit displays the confidence interval that suggests the true population odds is within the limits of the interval.

Also, an analysis of the deviance table (See Appendix, table 5.3) clearly shows that when each variable is added one by one, there is a drop in deviance. Moreover, the addition of *pastFail, pastCritical* and *pastSerious* causes a significant reduction in residual deviance. A big $p-$value shows that the null model accounts for more or less similar amount of variation. The results also indicates that all the variables significantly contribute to the model since the $p-$value is below 0.05.
The McFadden Pseudo R^2 was 0.6226629 which suggests the full model performed better than the null model.

The training time and processing speed of the logistic regression was 76.693 seconds and 13000 observations per second respectively (MATLAB R2017a) .

Evaluating the predictive ability of the model

In order to assess the predictive ability of the model, a new data set (test set) is used. Below is a confusion matrix to assess how the logistic model performed on the test set.

	Presence of Critical violation(Fail)	Absence of Critical violation(Pass)
Presence of Critical violation(Fail)	495	118
Absence of Critical violation(Pass)	61	1080

R output

From equations 3.22, 3.23 and 3.24,

$$Sensitivity = \frac{495}{495 + 61} = 0.8903$$

$$Specificity = \frac{1080}{118 + 1080} = 0.9015$$

$$Acc_{cl} = \frac{1518}{1636} \times 100 = 92.7873\%$$

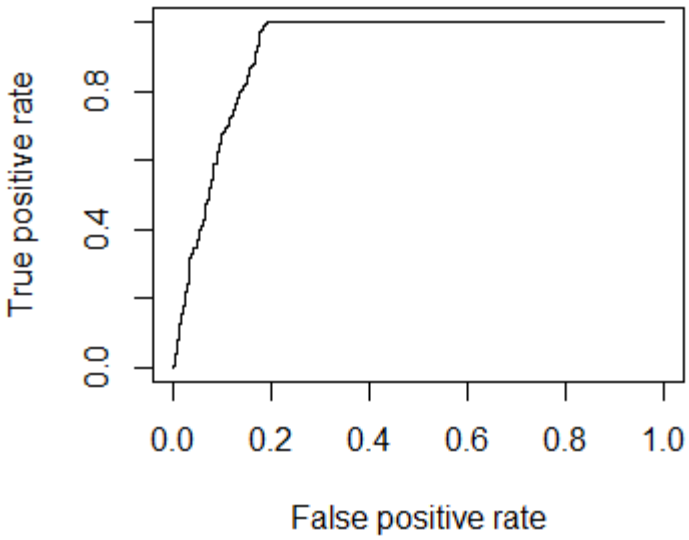


Figure 4.1: ROC of the logistic regression model

Figure 4.4.1 suggests the model has a high discrimination ability since the Receiver Operating Curve (ROC) curve goes very close to the top left corner. The value obtained for the area under the estimated curve (AUC) was 0.9199951. This shows that when any random pair of inspections are taken, one with critical violation and the other without critical violation, there is a 91.9995 % probability that the logistic regression model rightly ranks them as such.

The logistic regression model realised a prediction accuracy 92.7872%.

4.5 Support Vector Machine (SVM)

The SVM algorithm is a kernel based classifier specifically built for binary classification. The SVM uses kernel functions to map the training set to increase its semblace to a data set that can be linearly separated. It does so by increasing the data set's dimensionality. Three kernel functions were used, namely linear, Radial Basis Function (RBF) and polynomial.

4.5.1 Model selection for SVM

The kernels linear, Radial Basis Function (RBF) and polynomial was compared using their ROC, sensitivity and specificity measure. The study made use of 10 - fold cross validation to find the best parameters. The parameter range for the tuning parameters of each kernel was defined. The R package (caret), MATLAB (R2017a) was used in running these analysis. The computer used had the specifications, AMD E-300 APU with Radeon(tm) HD Graphics 1.30 GHz with installed memory of 4 GB.

Linear kernel

In the Linear kernel, the regularisation parameter also called cost function or the box constraint, C is the main tuning parameter. In defining the grid for the selection of the best C, as the tuning parameter for the kernel function, the values $0.25, 0.5, 0.75, 1, 1.25$ and 1.5 were considered in training the SVM. This stems from the usage of 1 by studies such as (Palaniappan et al, 2014) coupled with the fact that the default setting for software package like MATLAB for the choice of C is 1. Hence, the choice of this grid. This resulted in the creation of six models.

The following shows the results of the various tuning parameters considered.

Table 4.6: Resampling results cross the tuning parameters of linear kernel

C	ROC	Sens	Spec
0.25	0.8919498	0.7789694	0.9981073
0.50	0.8903240	0.7789694	0.9981073
0.75	0.8919143	0.7789694	0.9981073
1.00	0.8909931	0.7789694	0.9981073
1.25	0.8919220	0.7789694	0.9981073
1.50	0.8927530	0.7789694	0.9981073

Table 4.6 shows the various cost parameter considered and their sensitivity, specificity and ROC measures. All of the C values had the same sensitivity and specificity values which may be as a result of the closeness of each tuning parameter in defining the optimum cost parameter. It shows all the models have similar ability to correctly predict those with critical violations as those with critical violation and those without critical violations as those without critical violation. But their ROC measures, differentiate them since it compares the sensitivity and specificity across the range of the tuning parameters to be used in the model. ROC was used to select the optimal model using the largest value. The final value used for the model was C = 1.50.

The computational time for training and selection of the optimal C was 0.44891 hours (Source: R statistical package).

RBF kernel

With RBF kernel the default tuning parameters is the regularisation parameter, C and the kernel parameter, γ. In defining the grid for the selection of the best C and γ as the tuning parameters for the kernel function, the values $0.25, 0.5, 0.75, 1, 1.25, 1.5$ and $0.01, 0.015, 0.2$ were respectively considered in training the SVM. This creates 18 distinct models as each C value pairs all the γ values.

The following shows some of the resampling results cross the tuning parameters

(For the full table, see appendix table 5.4).

Table 4.7: Resampling results cross the tuning parameters of RBF kernel

γ	C	ROC	Sens	Spec
0.010	0.25	0.8956507	0.7789700	0.9981074
0.010	1.50	0.8935073	0.7789700	0.9981074
0.015	0.25	0.8962297	0.7789700	0.9981074
0.015	1.50	0.8978426	0.7790057	0.9980533
0.200	0.25	0.9028196	0.7897895	0.9763524
0.200	0.50	0.9029592	0.7877184	0.9822641

Table 4.7 shows the various regularisation and kernel parameters considered and their sensitivity, specificity and ROC measures. When $\gamma = 0.01$ for any value of C the sensitivity and specificity values remain the same but with different ROC. For the models with $\gamma = 0.015$ against $C = 0.25$ and 0.50 also showed similar sensitivity and specificity values. This may be as a result of the closeness of both parameters in defining the optimum C and γ. The sensitivity and specificity begin to differ when $\gamma = 0.015$ for $C = 0.75, 1.00$ & 1.50 and when $\gamma = 0.20$ for all values of C. The distinguishing measure was the ROC. ROC was used to select the optimal model using the largest value. The final values used for the model were $\gamma = 0.2$ and $C = 0.5$.

The computational time for training and selection of the optimal C was 18.048915 hours (Source: R statistical package).

Polynomial Kernel

With polynomial kernel the default tuning parameters are degree (d), cost (C) and scale. In defining the grid for the polynomial kernel function, the values $0.25, 0.5, 0.75, 1, 1.25, 1.5$ and $0.001, 0.01, 0.1$ were defined as the cost and scale parameters respectively whiles using degrees $1, 2, 3$ in training the SVM. All the parameters paired creates 27 distinct models.

The following shows some of the resampling results cross the tuning parameters (For the full table, see appendix table 5.5).

Table 4.8: Resampling results cross the tuning parameters of polynomial kernel

degree	scale	C	ROC	Sens	Spec
1	0.001	0.25	0.9124103	0.7789683	0.9981074
1	0.100	1.00	0.8893959	0.7789683	0.9981074
2	0.001	0.25	0.8837571	0.7789683	0.9981074
2	0.100	1.00	0.8843421	0.7792181	0.9974225
3	0.001	0.25	0.8803390	0.7789683	0.9981074
3	0.100	1.00	0.8890698	0.7839677	0.9903569

ROC was used to select the optimal model using the largest value. The final values used for the model were degree = 1, scale = 0.001 and C = 0.25.

The computational time for training and selection of the optimal tuning parameters was 15.678910 hours (Source: R statistical package).

4.5.2 Linear, RBF and Polynomial kernels

The tables below summarises the ROC, sensitivity and specificity of the three kernels based on the best tuning parameters:

Table 4.9: Receiver Operating Characteristic

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
linear	0.8688288	0.8818005	0.8910184	0.8927530	0.9023934	0.9281725
radial	0.8769418	0.8945710	0.9045632	0.9029592	0.9122074	0.9222675
Polynomial	0.8910859	0.9055932	0.9129130	0.9124103	0.9194117	0.9361986

Table 4.10: Sensitivity

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
linear	0.7468806	0.7665179	0.7821429	0.7789694	0.7910714	0.8196429
radial	0.7290553	0.7767857	0.7883929	0.7877184	0.8000000	0.8232143
Polynomial	0.7464286	0.7687500	0.7769847	0.7789683	0.7910714	0.8107143

Table 4.11: Specificity

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
linear	0.9954914	0.9972955	0.9981974	0.9981073	0.9990991	1.0000000
radial	0.9720721	0.9801802	0.9819820	0.9822641	0.9846812	0.9891794
Polynomial	0.9954955	0.9972973	0.9981974	0.9981074	0.9990991	1.0000000

From tables 4.9, 4.10 and 4.11, the polynomial kernel seems to have the advantage when it comes to the ROC and the specificity. Hence, comparisons were made by using the resampling approach for clearer display. Eugster et al. (2008) Hothorn et al. (2005) shows techniques for making decisions using resampling. Therefore, 50 resamplings were done and the figure below plots the kernels against their ROC.

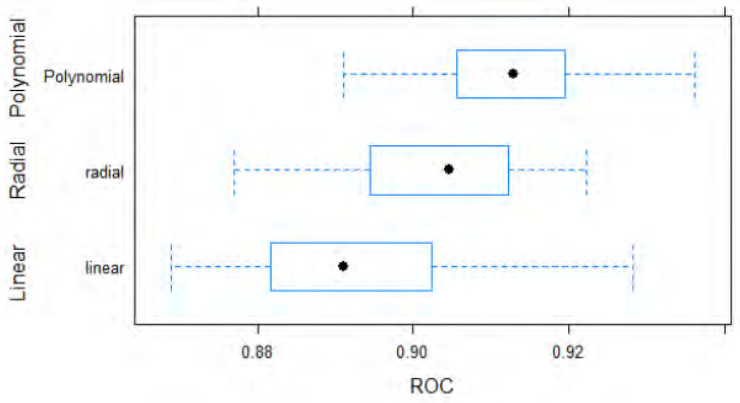


Figure 4.2: Plot of the three kernels Against ROC

Clearly, the polynomial kernel with the tuning parameters, degree = 1, scale = 0.001 and C = 0.25 is the optimal model.

4.6 K-Nearest Neighbour (KNN)

In the KNN algorithm, a new object is categorised by the greatest vote of its neighbours, that is the commonest(the number of neighbours) category amongst its nearest neighbours. The K is the number of nearest neighbours which is strictly a positive integer. In training the KNN classifier, the tune length was set to 20 where the best k was selected. Similarly, the studies (Lu et al. 2005; Zhu et

al. 2011) considered this type of approach in selecting an optimal k parameter. The Euclidean distance metric was used in this study.

Below are some resampling results across the k values. (See appendix table 5.6 for all the values)

Table 4.12: Resampling results cross the tuning parameter of KNN

k	ROC	Sens	Spec
5	0.9059139	0.8045366	0.9392038
11	0.9079729	0.7908606	0.9675018
19	0.9112177	0.7907181	0.9726389
25	0.9120057	0.7905747	0.9739367
31	0.9128466	0.7910390	0.9751260
37	0.9137709	0.7917529	0.9746935
41	0.9136882	0.7917172	0.9749281

Below is a plot of the accuracy of K against their ROC.

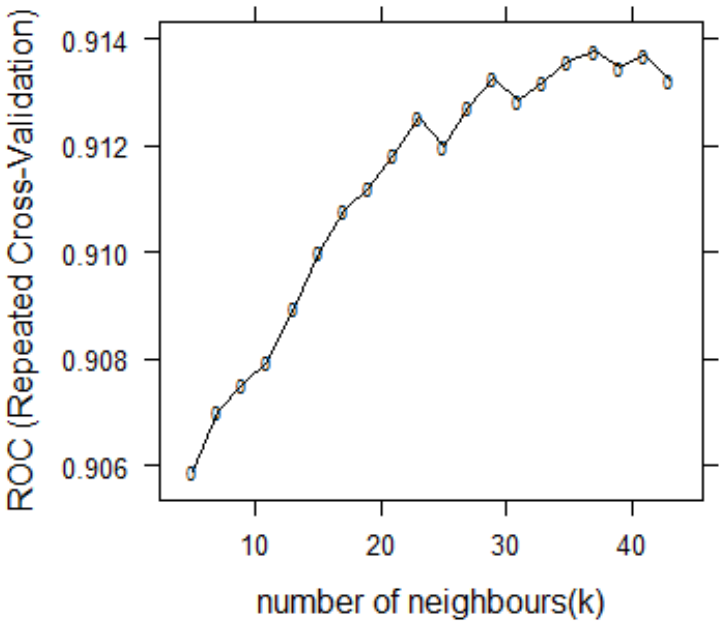


Figure 4.3: Comparison of Accuracy against K

Figure 4.3 displays the performance the number of nearest neighbour on the the training data. The ROC increases with an increase in the number of neighbours but it decreased when k was 31. It peaked again and the optimal k selected was 37.

The computational time for training and selection of the optimal tuning parameter K was 1.648078 hours (Source: R statistical package).

4.7 Comparing prediction performance of SVM models, Logistic regression (LR) and KNN

The optimal prediction performance of SVM models, Logistic regression and KNN were trained and tested on the same sets of data. A 10-fold cross validation method was used. The analysis was made possible by MATLAB R2017a software package.

Below are the results realised from both the SVM models.

Table 4.13: The performance of SVM models

Model (kernel)	C	γ	d	Prediction accuracy (%) Testing data	Training time	Predicted speed
Linear	1.50	N/A	N/A	92.7873	249.59 sec	4300 obs/sec
RBF	0.50	0.20	N/A	92.0538	1355.3 sec	1900 obs/sec
Polynomial	0.25	0.001	1	90.1589	1983.2 sec	250 obs/sec

As shown in table 4.13 the SVM classifier with the linear kernel obtained the maximum classication accuracy of 92.7873% amongst the RBF and the polynomial models. The linear kernel also had the fastest processing speed which translate into less computational time.

Table 4.14: The performance of KNN models

Number of nearest neighbours (k)	Prediction accuracy (%) Test accuracy
5	91.1369
11	92.2983
19	92.5428
25	92.5428
31	92.6650
37	92.6039
41	92.5428

Zhu et al. (2011) recommend using varied values of K to select the optimal K. Figure 4.14 shows all the models generalised well to the new test since they all had high test classification accuracy. The KNN classifier with $K = 31$ had a test classification accuracy of 92.6650% therefore having the highest generalisation capacity. Though, according to figure 4.3 and table 4.12, the optimal K measured by their ROC was 37 but after training and testing, $K = 31$ is the optimal model according to their classification accuracies. This means $K = 37$ did not generalise well to new data set.

Table 4.15: The optimal prediction accuracy of LR, SVM and LNN

	LR	SVM	KNN
Accuracy	92.7872%	92.7873%	92.6650 %
Misclassification	7.2128%	7.2127%	7.3350%

Table 4.15 shows the prediction accuracies and misclassification rate of all the best models of the three classifiers. The SVM classifier with linear kernel, had the highest detection rate relative to the logistic regression and KNN models. In other words, the SVM classifier with linear kernel predicts more correctly which suggests that part of the variables have greater probabilities to predict critical violations in food establishments. The logistic regression followed closely with 92.7872% whereas the KNN model had 92.6650 %. However, there was only 0.0001 difference between the SVM classifier and logistic regression, which suggest how seemingly similar they are in classifying critical violations at food establishments. This also shows that the data is linearly separable which is furthur confirmed by the selected degree, $d = 1$ in the polynomial kernel.
The figure 4.4 clearly displays the classification accuracies and misclassifications rates of LR, SVM and KNN repectively.

Since KNN exercise a majority voting scheme, in the case where we have a data set that has p data-points, then always the p-nearest neighbour model will make use of every data point in the data set to categorise new instances/points. In the

H

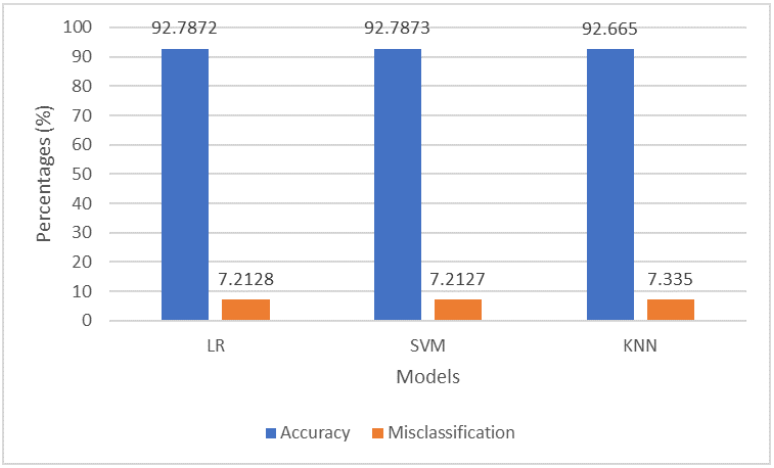


Figure 4.4: Accuracies and misclassification of LR, SVM and KNN

light of this when $k = 31$, only the nearest thirty-one data points are selected. An increase in k-value means an increase in the nearest neighbours which likely lead to a decrease in performance (Beyer, Goldstein, Ramakrishnan,& Shaft, 1999). This feature of KNN makes the case for KNN to contend especially with SVM.

Also comparing the computational time for training and selection of the optimal tuning parameters of SVM and KNN models, it suggests, the KNN is time consuming. Some studies (Bhaskar, Hoyle & Singh, 2006; Palaniappan, Sundaraj & Sundaraj, 2014) has reported on the computational complexity of the KNN algorithm, therefore leading to a high computational time. The logistic regression had the least computational time (76.693 seconds), therefore the logistic regression would be preferred when one is concerned about speed of the classifier. It took a little over thrice the amount of training time of the LR, for the SVM classfier to be trained (249.59 sec).

The logistic regression and the SVM without a kernel can be used interchangeably but when a kernel is used the SVM with kernels perform better (Kumar, 2018).This suggests the capability of an SVM, lies in the use of a kernel. The performance of the kernel also tells more of the type of data used in the study. The performance of the linear kernel again reiterate the fact that linear kernels performs better than the RBF kernel on a linearly separable data (Palaniappan

et al., 2014). Even under the constraints of the defined parameter settings of the SVM models, the SVM classifier with linear kernel had the highest detection rate.

4.8 Summary

This chapter began with a description of the data set and preliminary analysis to select the relatively relevant features by using the principal component analysis. The algorithms logistic regression, Support Vector Machine and K-Nearest Neighbour were utilised to classify food inspections data in order to predict whether an inspection would yield critical violation or not. The SVM classifier with linear kernel was the best as compared to LR and KNN models in identifying critical violations.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

This chapter presents a well trimmed version of the results of this study. It further makes conclusions that address the objectives of the study. Some suggestions and recommendations are also put forward.

5.1 Conclusion

The substantive issue of keeping our food establishments safe is a never-ending problem which needs every available resource to tackle. One effective way it can be checked is by conducting inspections into these facilities to make sure the stipulated rules and regulations are followed. Some recognised bodies like the Canadian Food Inspection Agency (CFIA), Chicago Department of Public Health (CDPH) and Ghana Food and Drugs Authority (FDA), are legally empowered to enforce or implement food-related rules and regulations. Food establishments fail inspections when critical violations are found. Critical violation are food related threats that affect the health of patrons like food contamination (Murphy et al. 2011).

Inspections of such nature yield huge size of data, and data mining algorithms like the Support Vector Machine (SVM), Logistic Regression (LR) and K-Nearest Neighbour (KNN) can help make predictions to make informed decisions. Previous studies used logistic regression in this regard to prioritise inspections based on food inspections data from the City of Chicago open data portal. However logistic regression has some deficiencies as compared to SVM and KNN. Therefore, this study used SVM and KNN to assess the performance of these algorithms relative to the logistic regression in predicting critical violations in

food establishments, so as to help prioritise inspections.

Using the available data set, the Principal Component Analysis (PCA) was used to extract 10 features that are independent of each other. These features were used as inputs in the predictive models. The linear, Radial Basis Function (RBF) and polynomial kernels of the SVM were trained on the training data using a 10-fold cross validation method to select the optimal models under varied parameter settings. Also, using varied K values, an optimal KNN model was selected when trained using the 10-fold cross validation method. The ROC, sensitivity and specificity were the performance measures used to select the optimal models. Similarly, the logistic regression model was trained using the 10-fold cross validation method. The logistic regression model realised the type of inspector, previous failures at inspections, the number of serious violations, licensed tobacco sellers, burglary, number of garbage cart requests, the intensity of local burglaries and the length of time since last inspection as the features that significantly influence the presence or absence of critical violations at food establishments. With a defined grid of the parameter setting, the SVM classifier with linear kernel and KNN model with $K = 31$ had higher accuracies relative to all the other models in the grid. Therefore, the logistic model, SVM model with linear kernel and KNN model with 31 nearest neighbours were able to detect critical violations in food establishments. The classification accuracies for the LR, SVM and KNN classifiers were 92.7872%, 92.7873% and 92.9025% respectively.

In conclusion, the findings showed that the SVM classifier with linear kernel marginally has a high generalisation capacity than LR and KNN models based on this data set. Therefore, prioritising inspections using the SVM model will marginally improve the previous work in Schenk et al. (2014) for early detection of food establishments with critical violations.

5.2 Recommendation

In relation to conclusions drawn from this study, these suggestions are made for future studies.

1. While this study used some defined parameter settings especially, for the SVM kernels, it remains therefore interesting for furthur studies to use some exhaustive parameter search approach.

2. Data on food inspections in Ghana should be properly collated and made available so that such studies can be readily conducted.

REFERENCES

Ababio, P. F., & Adi, D. D. (2012). Evaluating food hygiene awareness and practices of food handlers in the Kumasi metropolis. *Internet Journal of Food Safety, 14* (2), 35-43.

Ababio, P. F., Adi, D. D., & Commey, V. (2012). Food safety management systems, availability and maintenance among food industries in Ghana. *Food Science and Technology*

Ababio, P. F., & Lovatt, P. (2015). A review on food safety and food hygiene studies in Ghana.*Food Control, 47*, 92-97.

Abidin, U. F. U. Z., Arendt, S. W., & Strohbehn, C. H. (2014). Food safety culture in onsite foodservices: Development and validation of a measurement scale. *Journal of Foodservice Management and Education,* 8(1), 1.

Agyei-Baffour, P., Sekyere, K. B., & Addy, E. A. (2013). Policy on Hazard Analysis and Critical Control Point (HACCP) and adherence to food preparation guidelines: a cross sectional survey of stakeholders in food service in Kumasi, Ghana. *BMC research notes,* 6(1), 442. Addison-Wesley, Reading, MA, 1999.

Alkhatib, K., Najadat, H., Hmeidi, I., & Shatnawi, M. K. A. (2013). Stock price prediction using k-nearest neighbor (kNN) algorithm. *International Journal of Business, Humanities and Technology,* 3(3), 32-44.

Allwood, P. B., Jenkins, T., Paulus, C., Johnson, L., & Hedberg, C. W. (2004). Hand washing compliance among retail food establishment workers in Minnesota. *Journal of Food Protection, 67(12)*, 2825-2828.

Alpaydin, E. (1997). Voting over multiple condensed nearest neighbors. In *Lazy learning* (pp. 115-132). Springer, Dordrecht.

Andoh, A. H., Ackah, N. B., & Abbey, L. D. (2015). Let's adopt and implement the draft national food safety policy: feature article in The Ghanaian Times, Thursday, April 9, 2015.

Arendt, S., Strohbehn, C., & Jun, J. (2015). Motivators and barriers to safe food practices: observation and interview. *Food Protection Trends, 35*(5), 365-376.

Arvey, A., Agius, P., Noble, W. S., & Leslie, C. (2012). Sequence and chromatin determinants of cell-type–specific transcription factor binding. *Genome research, 22*(9), 1723-1734.

Asraf, H. M., Nooritawati, M. T., & Rizam, M. S. (2012). A comparative study in kernel-based support vector machine of oil palm leaves nutrient disease. *Procedia Engineering, 41,* 1353-1359.

Becker, S. (Ed.). (2001). *Data warehousing and web engineering.* (pp. 77-99) IGI Global.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999, January). When is "nearest neighbor" meaningful?. In *International conference on database theory* (pp. 217-235). Springer, Berlin, Heidelberg.

Bhatia, N. (2010). Survey of nearest neighbor techniques. *arXiv preprint arXiv:1007.0085.*

Bhaskar, H., Hoyle, D. C., & Singh, S. (2006). Machine learning in bioinformatics: A brief survey and recommendations for practitioners. *Computers in biology and medicine, 36*(10), 1104-1125.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152). ACM.

Brachman, R. J., & Anand, T. (1996). The process of knowledge discovery in databases. In Advances *in knowledge discovery and data mining* (pp. 37-57). American Association for Artificial Intelligence.

Bradley, P. S., & Mangasarian, O. L. (1998). Feature selection via concave minimization and support vector machines. In *ICML* (Vol. 98, pp. 82-90).

Bradley, P. S., Mangasarian, O. L., & Street, W. N. (1998). Feature selection via mathematical programming. *INFORMS Journal on Computing,* 10(2), 209-217.

Buzby, J. C., Frenzen, P. D., & Rasco, B. (2001). *Product liability and microbial foodborne illness.* US Department of Agriculture, Economic Research Service.

Centers for Disease Control and Prevention (CDC. (2013). Surveillance for foodborne disease outbreaks–United States, 2009-2010. MMWR. *Morbidity and mortality weekly report,* 62(3), 41

Chaplot, S., Patnaik, L. M., & Jagannathan, N. R. (2006). Classification of magnetic resonance brain images using wavelets as input to support vector machine and neural network. *Biomedical signal processing and control,* 1(1), 86-92.

Chou, J. S., Cheng, M. Y., Wu, Y. W., & Pham, A. D. (2014). Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification. *Expert Systems with Applications, 41*(8), 3955-3964.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning,* 20(3), 273-297

Cushman, J. W., Shanklin, C. W., & Niehoff, B. P. (2001). Hygiene practices of part-time student employees in a university foodservice operation. *The Journal of the National Association of College and University Food Services, 23*, 37-44.

Duda R. O., Hart P. E., and Stork D. G. (2001). Pattern Classification, *2nd ed. Wiley.*

Duda, R. O., Hart, P. E., & Stork, D. G. (1973). *Pattern classification* (Vol. 2). New York: Wiley

Eitrich, T., & Lang, B. (2006). Efficient optimization of support vector machine learning parameters for unbalanced datasets. *Journal of computational and applied mathematics, 196*(2), 425-436.

Ekici, S. (2012). Support Vector Machines for classification and locating faults on transmission lines. *Applied Soft Computing, 12(6)* , 1650-1658.

Este, A., Gringoli, F., & Salgarelli, L. (2009). Support vector machines for TCP traffic classification. *Computer Networks,* 53(14), 2476-2490.

Eugster M, Hothorn T, Leisch F (2008). "Exploratory and Inferential Analysis of Benchmark Experiments." *Ludwigs-Maximilians-Universitat Munchen, Department of Statistics,* Tech. Rep, 30.

Feglo, P., & Sakyi, K. (2012). Bacterial contamination of street vending food in Kumasi, Ghana. *Journal of Medical and Biomedical Sciences, 1*(1), 1-8.

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters, 30*(1), 27-38.

Field, A. (2013) Discovering Statistics using SPSS, 4^{th} edn. London: SAGE

Fielding, J. E., Aguirre, A., & Palaiologos, E. (2001). Effectiveness of altered incentives in a food safety inspection program. *Preventive Medicine*, 32(3), 239-244.

Fix, E., & Hodges Jr, J. L. (1951). *Discriminatory analysis-nonparametric discrimination: consistency properties.* California Univ Berkeley.

Food and Drugs Board. (1992). The food and drugs act. PNDCL 3058 1992. Retrieved from www.wipo.int/edocs/lexdocs/laws/en/gh/gh022en.pdf

Food and Drugs Authority. (2018) Retrieved from http://www.moh.gov.gh/foods-and-drug-authority/

Food safety. (2017). Retrieved from http://www.who.int/news-room/fact-sheets/detail/food-safety

Franco-Lopez, H., Ek, A. R., & Bauer, M. E. (2001). Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote sensing of environment, 77*(3), 251-274.

Ghana Health Service. (2012). In International Federation of Red Cross and Red Crescent Societies. Disaster Relief Emergency Fund (Dref) Final Report. Viewed 18/4/2013

Ghana Standard Authority. (2013). Personal communication with standard documentation department.

Golan, E. H., Roberts, T., Salay, E., Caswell, J. A., Ollinger, M., & Moore, D. L. (2004). *Food safety innovation in the United States: evidence from the meat industry* (No. 34083). United States Department of Agriculture, Economic Research Service.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science, 286*(5439), 531-537.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*(pp. 986-996). Springer, Berlin, Heidelberg.

Guyon, I., & Elisseeff, A. (2006). An introduction to feature extraction. In *Feature extraction* (pp. 1-25). Springer, Berlin, Heidelberg.

Guyon, I., & Elisseeff, A. (2003). *An introduction to variable and feature selection. Journal of machine learning research, 3*(Mar), 1157-1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3), 389-422.

Hamamoto, Y., Uchimura, S., & Tomita, S. (1997). A bootstrap technique for nearest neighbor classifier design. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19* (1), 73-79.

Hassanat, A. B., Abbadi, M. A., Altarawneh, G. A.,& Alhasanat, A. A. (2014). Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach. *arXiv preprint arXiv:1409.0919*

Hoak, J. (2010). The Effects of Outliers on Support Vector Machines.*Portland State University.*

Hong, J. H., Min, J. K., Cho, U. K., & Cho, S. B. (2008). Fingerprint classification using one-vs-all support vector machines dynamically ordered with naive Bayes classifiers. *Pattern Recognition,* 41(2), 662-671

Hornik, K., Meyer, D., & Karatzoglou, A. (2006). Support vector machines in R. *Journal of statistical software, 15*(9), 1-28.

Hothorn, T., Leisch, F., Zeileis, A., & Hornik, K. (2005). The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics, 14*(3), 675-699.

Howells, A. D., Roberts, K. R., Shanklin, C. W., Pilling, V. K., Brannon, L. A., & Barrett, B. B. (2008). Restaurant employees' perceptions of barriers to three food safety practices. *Journal of the Academy of Nutrition and Dietetics, 108*(8), 1345-1349.

Hsu, C.-W., Chang, C.-C., Lin, C.-J. (2004). A practical guide to support vector classification. Technical Report, Department of Computer Science and Information Engineering, National Taiwan University. Retrieved from http://www.csie.ntu.edu.tw/∼cjlin/papers/guide/guide.pdf.

Huang, C. L., & Wang, C. J. (2006). A GA-based feature selection and parameters optimization for support vector machines. Expert Systems with applications, 31(2), 231-240

Jahromi, M. Z., Parvinnia, E., & John, R. (2009). A method of learning weighted similarity function to improve the performance of nearest neighbor. *Information sciences, 179*(17), 2964-2973.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, pp.342-369). New York: springer.

Japkowicz, N. (2000). Learning from imbalanced data sets: a comparison of various strategies. In *AAAI workshop on learning from imbalanced data sets* (Vol. 68, pp. 10-15).

Jiang, P., Missoum, S., & Chen, Z. (2014). Optimal SVM parameter selection for non-separable and unbalanced datasets. *Structural and Multidisciplinary Optimization, 50*(4), 523-535.

Joachims, T. (1998, April). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning* (pp. 137-142). Springer, Berlin, Heidelberg.

Jolliffe, I. (2011). Principal component analysis. In *International encyclopedia of statistical science* (pp. 1094-1096). Springer, Berlin, Heidelberg.

Kalakuntla, P. (2017). Performance Analysis of kNN Query Processing on large datasets using CUDA & Pthreads: comparing between CPU & GPU.

Kassel, S., (2017). "Predicting Building Code Compliance with Machine Learning Models." Retrieved from https://www.azavea.com/blog/2017/09/21/building-inspection-prediction/.

Knight, A. J., Worosz, M. R., & Todd, E. C. D. (2007). Serving food safety: consumer perceptions of food safety at restaurants. *International Journal of Contemporary Hospitality Management,* 19(6), 476-484.

Kriminger, E., Príncipe, J. C., & Lakshminarayan, C. (2012). Nearest neighbor distributions for imbalanced classification. In *Neural Networks (IJCNN), The 2012 International Joint Conference on* (pp. 1-5). IEEE.

Kumar, A. (2018). *Machine Learning - When to Use Logistic Regression vs. SVM*

- *Reskilling IT*. Retrieved from https://vitalflux.com/machine-learning-use-logistic-regression-vs-svm/

Kumar, N., Krovi, R., & Rajagopalan, B. (1997). Financial decision support with hybrid genetic and neural based modeling tools. *European Journal of Operational Research, 103*(2), 339-349.

Kuramochi, M., & Karypis, G. (2005). Gene classification using expression profiles: A feasibility study. *International Journal on Artificial Intelligence Tools, 14*(04), 641-660.

Latourrette, M. (2000). Toward an explanatory similarity measure for nearest-neighbor classification. In *European Conference on Machine Learning* (pp. 238-245). Springer, Berlin, Heidelberg.

Lavanya, B., & Divya, B. (2017). Big data analysis using SVM and K-NN data mining techniques. *International Journal of Computer Science and Mobile Computing (IJCSMC), 6*(1), 84-91

Li, Z., Zhang, Q., & Zhao, X. (2017). Performance analysis of K-nearest neighbor, support vector machine, and artificial neural network classifiers for driver drowsiness detection with different road geometries. *International Journal of Distributed Sensor Networks, 13*(9), 1550147717733391.

Liu, C., Jiang, D., & Yang, W. (2014). Global geometric similarity scheme for feature selection in fault diagnosis. *Expert Systems with Applications, 41*(8), 3585-3595.

Liu, W., & Chawla, S. (2011). Class confidence weighted knn algorithms for imbalanced data sets. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 345-356). Springer, Berlin, Heidelberg.

Lo, C. S., & Wang, C. M. (2012). Support vector machine for breast MR image classification. *Computers & Mathematics with Applications, 64* (5), 1153-1162

Lovric, M. (2011). *International Encyclopedia of Statistical Science.* Springer pp. 591-756.

Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., ... & Downing, J. R. (2005). MicroRNA expression profiles classify human cancers. *nature, 435*(7043), 834

Lynch, R. A., Elledge, B. L., Griffith, C. C., & Boatright, D. T. (2003). A comparison of food safety knowledge among restaurant managers, by source of training and experience, in Oklahoma County, Oklahoma. *Journal of Environmental Health,* 66(2), 9.

Madaio, M., Chen, S. T., Haimson, O. L., Zhang, W., Cheng, X., Hinds-Aldrich, M., ... & Dilkina, B. (2016). Firebird: Predicting fire risk and prioritizing fire inspections in atlanta. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 185-194). ACM.

Monney, I., Agyei, D., & Owusu, W. (2013). Hygienic practices among food vendors in educational institutions in Ghana: the case of Konongo. *Foods, 2(3)*, 282-294.

Mosley, S., & Steif, K. (2018). *Urban Spatial.* Retrieved from http://urbanspatialanalysis.com/portfolio/proof-of-concept-using-predictive-modeling-to-prioritize-building-inspections/

Min, J. H., & Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert systems with applications, 28*(4), 603-614.

Murphy, K. S., DiPietro, R. B., Kock, G., & Lee, J. S. (2011). Does mandatory food safety training and certification for restaurant employees improve inspection outcomes?. *International Journal of Hospitality Management, 30*(1), 150-156.

Nagi, J., Yap, K. S., Tiong, S. K., Ahmed, S. K., & Mohammad, A. M. (2008). Detection of abnormalities and electricity theft using genetic support vector machines. In *TENCON 2008-2008 IEEE Region 10 Conference* (pp. 1-6). IEEE.

National Restaurant Association. (2012). *2013 Restaurant industry: Pocket factbook.* Retrieved from http://www.restaurant.org/pdfs/research/Factbook2013_LetterSize.pdf.

Nilsson, R., Peña, J. M., Björkegren, J., & Tegnér, J. (2007). Consistent feature selection for pattern recognition in polynomial time. *Journal of Machine Learning Research, 8*(Mar), 589-612

Omaye, S. T. (2004). *Food and nutritional toxicology.* CRC press. Boca Raton, pp. 163-173

Parikh, U. B., Das, B., & Maheshwari, R. (2010). Fault classification technique for series compensated transmission line using support vector machine. *International Journal of Electrical Power & Energy Systems, 32*(6), 629-636

Parry, R. M., Jones, W., Stokes, T. H., Phan, J. H., Moffitt, R. A., Fang, H., Shi L., Oberthuer., Fischer., Tong W., & Wang, M. D. (2010). k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *The pharmacogenomics journal, 10*(4), 292.

Palaniappan, R., Sundaraj, K., & Sundaraj, S. (2014). A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals. *BMC bioinformatics, 15*(1), 223.

Pochet, N. L. M. M., & Suykens, J. A. K. (2006). Support vector machines versus logistic regression: improving prospective performance in clinical decision-making. *Ultrasound in Obstetrics & Gynecology, 27*(6), 607-608.

Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing, 17*(4), 694-701.

Qi, X., & Luo, R. (2015). Sparse principal component analysis in Hilbert space. *Scandinavian Journal of Statistics, 42*(1), 270-289.

Qian, Y., Zhou, W., Yan, J., Li, W., & Han, L. (2014). Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery. *Remote Sensing, 7*(1), 153-168

Qu, J., & Zuo, M. J. (2010). Support vector machine based data processing algorithm for wear degree classification of slurry pump systems. *Measurement, 43*(6), 781-791

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature genetics, 32*, 496.

Rai, H., & Yadav, A. (2014). Iris recognition using combined support vector machine and Hamming distance approach. *Expert systems with applications, 41*(2), 588-593.

Rana, M., Chandorkar, P., Dsouza, A., & Kazi, N. (2015). Breast cancer diagnosis and recurrence prediction using machine learning techniques. *IJRET: International Journal of Research in Engineering and Technology eISSN,* 2319-1163.

Reske, K. A., Jenkins, T., Fernandez, C., VanAmber, D., & Hedberg, C. W. (2007). Beneficial effects of implementing an announced restaurant inspection program. *Journal of Environmental Health, 69*(9), 27-35.

Rheinländer, T., Olsen, M., Bakang, J. A., Takyi, H., Konradsen, F., & Samuelsen, H. (2008). Keeping up appearances: perceptions of street food safety in urban Kumasi, Ghana. *Journal of Urban Health, 85*(6), 952-964.

Rosenfeld, N., Aharonov, R., Meiri, E., Rosenwald, S., Spector, Y., Zepeniuk, M., ... & Lebanony, D. (2008). MicroRNAs accurately identify cancer tissue origin. *Nature biotechnology, 26*(4), 462.

Sahami, M., & Heilman, T. D. (2006). A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web* (pp. 377-386). AcM.

Salazar, D. A., Vélez, J. I., & Salazar, J. C. (2012). Comparison between SVM and logistic regression: Which one is better to discriminate?. *Revista Colombiana de Estadística, 35*(SPE2), 223-237.

Samuels, P. (2016). Advice on Exploratory Factor Analysis. 10.13140/RG.2.1.5013.9766

Schenk, T., Leynes, G., Solanki, A., Collins, S., Smart, G., Abright, B., Crippen, C., (2014). "Food Inspection Forecasting - City of Chicago." Retrieved from https://github.com/Chicago/food-inspectionsevaluation/blob/master/REPORTS/forecasting-restaurants-with-critical-violations-in-Chicago.Rmd .

Smart Cities Initiative (2018). Predictive Modeling of Building Fire Risk: Designing and evaluating predictive models of fire risk to prioritize property fire inspections. *Metro21 Research Publication.*

Song, Y., Huang, J., Zhou, D., Zha, H., & Giles, C. L. (2007). Iknn: Informative k-nearest neighbor pattern classification. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 248-264). Springer, Berlin, Heidelberg.

Song, F., Guo, Z., & Mei, D. (2010). Feature selection using principal component analysis. In *System science, engineering design and manufacturing informatization (ICSEM), 2010 international conference on* (Vol. 1, pp. 27-30). IEEE.

Suguna, N., & Thanushkodi, K. (2010). An improved k-nearest neighbor classification using genetic algorithm. *International Journal of Computer Science Issues, 7*(2), 18-21.

Sun, S., & Huang, R. (2010). An adaptive k-nearest neighbor algorithm. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on* (Vol. 1, pp. 91-94). IEEE.

Swiniarski, R. W. (2000). Data mining methods in face recognition. In *Applications of Artificial Neural Networks in Image Processing V* (Vol. 3962, pp. 52-60). International Society for Optics and Photonics.

Tarigan, A., Dewi Agushinta, R., Suhendra, A., & Budiman, F. (2017). Determination of SVM-RBF Kernel Space Parameter to Optimize Accuracy Value of Indonesian Batik Images Classification. *JCS, 13*(11), 590-599.

Tessema, A. G., Gelaye, K. A., & Chercos, D. H. (2014). Factors affecting food handling Practices among food handlers of Dangila town food and drink establishments, North West Ethiopia. *BMC public Health, 14*(1), 571

Thanh Noi, P., & Kappas, M. (2017). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors, 18*(1), 18.

Thome, A. C. G. (2012). SVM classifiers–concepts and applications to character recognition. In *Advances in Character Recognition*. InTech.

Torgerson, P. R., de Silva, N. R., Fèvre, E. M., Kasuga, F., Rokni, M. B., Zhou, X. N., ... & Stein, C. (2014). The global burden of foodborne parasitic diseases: an update. *Trends in Parasitology, 30*(1), 20-26.

Tsai, C. F., Hsu, Y. F., Lin, C. Y., & Lin, W. Y. (2009). Intrusion detection by machine learning: A review. *Expert Systems with Applications, 36*(10), 11994-12000.

Übeyli, E. D. (2007). Comparison of different classification algorithms in clinical decision-making. *Expert systems, 24*(1), 17-31.

Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems, 24*(7), 1024-1032.

Valentini, G. (2002). Gene expression data analysis of human lymphoma using

support vector machines and output coding ensembles. *Artificial Intelligence in Medicine, 26*(3), 281-304.

Vapnik, V. (1998). *Statistical learning theory. 1998.* Wiley, New York.

Vasickova, P., Dvorska, L., Lorencova, A., & Pavlik, I. (2005). Viruses as a cause of foodborne diseases: a review of the literature. *Veterinární medicína, 50*(3), 89-104. vector classification. Technical Report, Department of Computer

Wainer, J. (2016). Comparison of 14 different families of classification algorithms on 115 binary datasets. University of Campinas.

Wang, H., & Huang, G. (2011). Application of support vector machine in cancer diagnosis. *Medical oncology, 28*(1), 613-618.

Weinberger, K. Q., Blitzer, J., & Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems* (pp. 1473-1480).

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., & Vapnik, V. (2001). Feature selection for SVMs. In *Advances in neural information processing systems* (pp. 668-674).

Wilson, M., Murray, A. E., Black, M. A., & McDowell, D. A. (1997). The implementation of hazard analysis and critical control points in hospital catering. *Managing Service Quality: An International Journal, 7*(3), 150-156.

Witten, I. H., & Frank, E. (2001). Data Mining Carl Hanser. *München, Wien.*

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. IEEE *transactions on evolutionary computation, 1*(1), 67-82.

World Health Organization (2006). *Five keys to safer food manual.* Retrieved from http://apps.who.int/iris/bitstream/handle/10665/43546/9789241594639_eng.pdf;jsessionid=EF9630A5F47F6217D2A3363E91B597DB?sequence=1

World Health Organisation. (2015). "Situation Report on Cholera Outbreak in Ghana As of 12 April 2015 (Week 15)". Retrieved from https://reliefweb.int/report/ghana/situation-report-cholera-outbreak-ghana-12-april-2015-week-15

World Health Organization. (2014). WHO initiative to estimate the global burden of foodborne diseases: fourth formal meeting of the Foodborne Disease Burden Epidemiology Reference Group (FERG): Sharing New Results, Making Future Plans, and Preparing Ground for the Countries.

World Health Organization. (2015). *Food Safety: What you should know.* Retrieved from https://www.fda.gov/Food/GuidanceRegulation/HACCP/ucm2006801.html/

World Health Organization. (2015). WHO estimates of the global burden of foodborne diseases: foodborne disease burden epidemiology reference group 2007-2015.

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

Wu, Z., Zhang, H., & Liu, J. (2014). A fuzzy support vector machine algorithm for classification based on a novel PIM fuzzy clustering method. *Neurocomputing, 125,* 119-124.

Yang, B. S., Hwang, W. W., Kim, D. J., & Tan, A. C. (2005). Condition classification of small reciprocating compressor for refrigerators using artificial neural networks and support vector machines. *Mechanical Systems and Signal Processing, 19*(2), 371-390

Yang, J., & Honavar, V. (1998). Feature subset selection using a genetic algorithm. In *Feature extraction, construction and selection* (pp. 117-136). Springer, Boston, MA.

Yang, Y. (1997). An Evaluation of Statistical Approaches to Text Categorization School of Computer Science. *Carnegie Mellon University.*

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information retrieval,* 1(1-2), 69-90.

Yiannas, F. (2008). *Food safety culture: Creating a behavior-based food safety management system.* Springer Science & Business Media.

Yu, Q., Miche, Y., Sorjamaa, A., Guillen, A., Lendasse, A., & Séverin, E. (2010). OP-KNN: Method and applications. *Advances in Artificial Neural Systems, 2010,* 1.

Zhang, W., Yoshida, T., & Tang, X. (2008). Text classification based on multi-word with support vector machine. *Knowledge-Based Systems, 21*(8), 879-886.

Zhu, X., Zhang, S., Jin, Z., Zhang, Z., & Xu, Z. (2011). Missing value estimation for mixed-attribute data sets. *IEEE Transactions on Knowledge and Data Engineering, 23*(1), 110-121.

APPENDIX

Table 5.1: Descriptive of categorical features

Features		Frequency	Percentage
pastFail	Yes	16724	91.2
	No	1609	8.8
BlueInsp	Yes	15010	81.9
	No	3323	18.1
tobacco_sale	Yes	16648	90.8
	No	1685	9.2

SPSS output

Table 5.2: Communalities

	Initial	Extraction
pastSerious	1.000	0.721
pastFail	1.000	0.673
timeSinceLast	1.000	0.393
pastCritical	1.000	0.302
ageAtInspection	1.000	0.088
license_insp	1.000	0.037
humidity	1.000	0.021
windSpeed	1.000	0.018
temperatureMax	1.000	0.015
PubAmusement	1.000	0.014
OrangeInsp	1.000	0.005
BrownInsp	1.000	0.003
mobFoodLic	1.000	0.000
sanitation	1.000	0.408
burglary	1.000	0.400
garbage	1.000	0.390
tobacco_sale	1.000	0.236
BlueInsp	1.000	0.132
seriousCount	1.000	0.129
package_goods	1.000	0.085
filling_station	1.000	0.077
GreenInsp	1.000	0.050
criticalCount	1.000	0.055
YellowInsp	1.000	0.028
catLiqLic	1.000	0.010
regBisLic	1.000	0.009
PurpleInsp	1.000	0.006
precipIntensity	1.000	0.001

Extraction Method: PCA

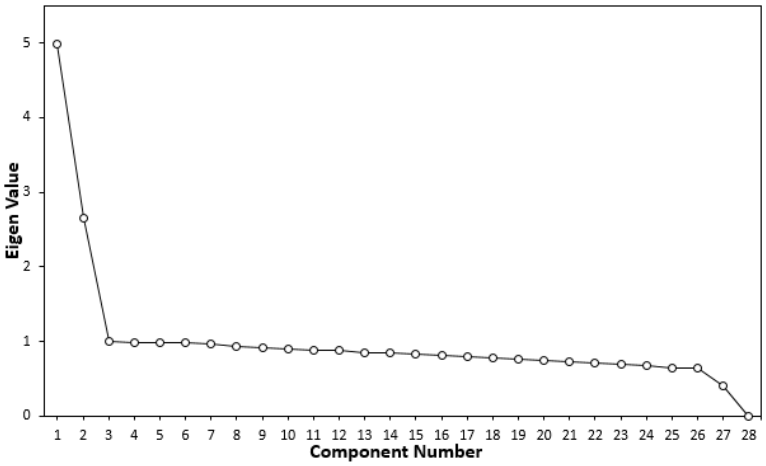


Figure 5.1: Scree plot

Table 5.3: Analysis of Deviance Table

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			16696	21304.4	
pastFail	1	44.8	16695	21259.6	2.139e-11 ***
pastCritical	1	31.1	16694	21228.5	2.475e-08 ***
pastSerious	1	13.0	16693	21215.5	0.0003094 ***
seriousCount	1	12934.0	16692	8281.5	< 2.2e-16 ***
timeSinceLast	1	16.9	16691	8264.5	3.861e-05 ***
BlueInsp	1	192.9	16690	8071.6	< 2.2e-16 ***
tobacco_sale	1	8.5	16689	8063.1	0.0035944 **
heat_burglary	1	5.4	16688	8057.8	0.0203608 *
heat_garbage	1	18.4	16687	8039.3	1.782e-05 ***
heat_sanitation	1	0.4	16686	8038.9	0.5255327

Signif. codes: 0 '***', 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1

Table 5.4: Resampling results cross the tuning parameters of the RBF kernel

sigma	C	ROC	Sens	Spec
0.010	0.25	0.8956507	0.7789700	0.9981074
0.010	0.50	0.8937628	0.7789700	0.9981074
0.010	0.75	0.8920192	0.7789700	0.9981074
0.010	1.00	0.8944685	0.7789700	0.9981074
0.010	1.25	0.8934944	0.7789700	0.9981074
0.010	1.50	0.8935073	0.7789700	0.9981074
0.015	0.25	0.8962297	0.7789700	0.9981074
0.015	0.50	0.8959974	0.7789700	0.9981074
0.015	0.75	0.8967519	0.7790057	0.9979812
0.015	1.00	0.8967626	0.7789700	0.9981074
0.015	1.25	0.8976418	0.7789700	0.9980173
0.015	1.50	0.8978426	0.7790057	0.9980533
0.200	0.25	0.9028196	0.7897895	0.9763524
0.200	0.50	0.9029592	0.7877184	0.9822641
0.200	0.75	0.9021967	0.7873969	0.9835799
0.200	1.00	0.9022667	0.7872896	0.9837600
0.200	1.25	0.9018410	0.7874679	0.9829307
0.200	1.50	0.9011346	0.7880042	0.9822819

R output

Table 5.5: Resampling results cross the tuning parameters of the polynomial kernel

degree	scale	C	ROC	Sens	Spec
1	0.001	0.25	0.9124103	0.7789683	0.9981074
1	0.001	0.50	0.8944628	0.7789683	0.9981074
1	0.001	1.00	0.8877146	0.7789683	0.9981074
1	0.010	0.25	0.8875028	0.7789683	0.9981074
1	0.010	0.50	0.8871005	0.7789683	0.9981074
1	0.010	1.00	0.8877925	0.7789683	0.9981074
1	0.100	0.25	0.8857886	0.7789683	0.9981074
1	0.100	0.50	0.8892116	0.7789683	0.9981074
1	0.100	1.00	0.8893959	0.7789683	0.9981074
2	0.001	0.25	0.8837571	0.7789683	0.9981074
2	0.001	0.50	0.8813082	0.7789683	0.9981074
2	0.001	1.00	0.8858293	0.7789683	0.9981074
2	0.010	0.25	0.8828022	0.7789683	0.9981074
2	0.010	0.50	0.8850110	0.7789683	0.9981074
2	0.010	1.00	0.8843879	0.7789683	0.9981074
2	0.100	0.25	0.8815593	0.7792540	0.9976748
2	0.100	0.50	0.8824951	0.7790397	0.9977651
2	0.100	1.00	0.8843421	0.7792181	0.9974225
3	0.001	0.25	0.8803390	0.7789683	0.9981074
3	0.001	0.50	0.8821149	0.7789683	0.9981074
3	0.001	1.00	0.8864385	0.7789683	0.9981074
3	0.010	0.25	0.8914778	0.7789683	0.9981074
3	0.010	0.50	0.8895648	0.7789683	0.9981074
3	0.010	1.00	0.8914105	0.7789683	0.9981074
3	0.100	0.25	0.8882463	0.7829675	0.9915826
3	0.100	0.50	0.8888486	0.7831101	0.9908797
3	0.100	1.00	0.8890698	0.7839677	0.9903569

R output

Table 5.6: Resampling results across different K values

k	ROC	Sensitivity	Specificity
5	0.9059139	0.8045366	0.9392038
7	0.9070431	0.7974663	0.9532986
9	0.9075180	0.7946809	0.9628878
11	0.9079729	0.7908606	0.9675018
13	0.9089746	0.7910396	0.9708910
15	0.9100148	0.7900036	0.9709632
17	0.9107945	0.7901466	0.9725854
19	0.9112177	0.7907181	0.9726389
21	0.9118572	0.7909674	0.9738287
23	0.9125478	0.7903965	0.9736844
25	0.9120057	0.7905747	0.9739367
27	0.9127539	0.7900749	0.9748920
29	0.9132674	0.7906103	0.9749457
31	0.9128466	0.7910390	0.9751260
33	0.9132206	0.7913244	0.9748561
35	0.9136069	0.7913961	0.9748199
37	0.9137709	0.7917529	0.9746935
39	0.9134619	0.7915386	0.9752886
41	0.9136882	0.7917172	0.9749281
43	0.9132412	0.7913602	0.9752164

R output