



**UNIVERSITY OF NAIROBI**  
**SCHOOL OF COMPUTING AND INFORMATICS**

**PREDICTIVE ANALYTICS AND BUSINESS INTELLIGENCE**  
**ADOPTION IN GENERAL INSURANCE (FOR CLAIMS**  
**MANAGEMENT)**

**BY**

**ROSEMARY A. ONYANGO**

**P52/65744/2013**

**SUPERVISOR**

**DR. ELISHA ABADE**

**OCTOBER, 2014**

**A research project submitted in partial fulfillment of the requirements of the Degree of Master of Science in  
Computational Intelligence at the University of Nairobi.**

## DECLARATION

This project, as presented in this report, is my original work and has not been presented for any other award in any other University.

**Name:** Rosemary Atieno Onyango

**Reg. No:** P52/65744/2013

**Signature:** \_\_\_\_\_

**Date:** \_\_\_\_\_

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

**Name:** Dr. Elisha Abade

**Sign:** \_\_\_\_\_

**Date:** \_\_\_\_\_

## **ACKNOWLEDGEMENT**

I thank God for walking me through this research project.

I also express sincere gratitude to my supervisor Dr. Elisha Abade for his input, honest guidance and assistance throughout the course of this research project.

Finally, I extend many thanks to my family and friends for their constant support that saw this project to its completion. God bless them all.

## **ABSTRACT**

In conducting their day-to-day businesses, insurance companies are faced with both critical and non-critical decisions mainly on fraud management, claims management, actuarial management, and customer relationship management. These decisions are best supported by analytics and business intelligence and are mainly driven by proper claims management. In order to remain competitive in the insurance industry, companies are being driven to attain new capabilities in this area (analytics) especially in Kenya. Analytics is becoming a required competency in the industry and promises to provide a competitive advantage to companies that invest in it.

Many claim executives want greater transparency into what drives operating costs, and more quantitative data about what factors determine claim outcomes (Keith, 2012). This research reviewed the adoption of analytics and business intelligence in the Kenyan insurance sector for claims management and proposed a cost-effective analytics and BI solution that can be used to manage the ever-changing business processes without having to procure commercial off-the-shelf analytics systems.

## TABLE OF CONTENTS

DECLARATION .....	i
ACKNOWLEDGEMENT .....	ii
ABSTRACT.....	iii
TABLE OF CONTENTS.....	iv
LIST OF FIGURES .....	vi
LIST OF TABLES.....	vi
LIST OF ABBREVIATIONS.....	vii
LIST OF TERMS.....	vii
1. INTRODUCTION .....	1
1.1. Background.....	1
1.2. Problem Statement .....	2
1.3. Research Objectives.....	4
1.4. Research Questions.....	4
2. LITERATURE REVIEW .....	6
2.1. Previous Work .....	6
2.2. Use of Analytics in Insurance Today.....	6
2.3. How This Study Was Different From What Has Been Done Before.....	8
2.4. Existing Solutions for Claims Analytics.....	9
2.5. Barriers to Adoption of Analytics and BI.....	9
2.6. Models Categories .....	10
2.7. Predictive Analytics .....	10
2.8. Predictive Modeling Process.....	11
2.9. Predictive Modeling Algorithms.....	11
2.10. Features of Predictive Modeling.....	12
2.11. Significance of the Study.....	13
2.12. Beneficiaries of the Research.....	13
3. METHODOLOGY .....	14
3.1. Application of Action Research.....	14
3.2. Why Action Research .....	15
3.3. Testing Strategy .....	15

3.4.	Anticipated Outcomes.....	15
3.5.	Expected Contribution .....	15
3.6.	Project Schedule.....	17
3.7.	Design and Implementation .....	17
3.7.1.	Survey Setup .....	18
3.7.1.1.	Proposed Participants.....	18
3.7.2.	Data Sources .....	18
3.7.3.	As-Is Analysis.....	18
3.7.4.	To-Be Analysis .....	19
3.7.5.	Data Collection and Preparation .....	21
3.7.6.	Data Analysis and Review .....	21
3.7.7.	Design and Implementation .....	21
3.7.7.1.	Model Development Framework .....	21
3.7.7.2.	Algorithm – Decision Trees .....	22
▪	J48.....	22
▪	Naïve Bayes .....	22
3.7.7.3.	Software Development Methodology .....	22
3.7.7.4.	Dataset Used .....	23
3.7.7.5.	Implementation Iterations .....	24
3.7.7.6.	Architectural Solution Design.....	31
3.7.8.	Testing.....	31
4.	RESULTS AND ANALYSIS.....	32
4.1.	Survey Results .....	32
4.2.	Prototype Results .....	40
4.3.	Research Evaluation.....	44
4.4.	Challenges and Limitations.....	45
4.5.	Recommendations and future work .....	45
4.6.	Conclusion .....	46
	REFERENCES .....	47
	APPENDIX.....	51

## LIST OF FIGURES

Figure 1: A typical claims process and its decision points. ....	2
Figure 2: An illustration of the problem faced with little or no analytics in claims management .....	4
Figure 3: Analytics capability maturity against time in Kenya.....	7
Figure 4: Adoption and Use of analytics as at 2013 (SAP 2014) .....	8
Figure 5: Barriers to analytics and BI adoption. Source: MIT Sloan Management Review 2012 (SAP 2014) .....	9
Figure 6: Predictive Analytics .....	11
Figure 7: An Iterative approach to analytics and BI. ....	16
Figure 8: AS-IS Architectural Design.....	19
Figure 9: TO-BE High – Level Architectural Design.....	20
Figure 10: Extreme Programming Implementation Process .....	23
Figure 11: Process Flow and Implementation Iterations.....	30
Figure 12: High Level Architectural Design Details .....	31
Figure 13: Survey respondents’ organizations.....	32
Figure 14: Survey respondents' roles .....	33
Figure 15: Survey responses on analytics and BI target audience. ....	34
Figure 16: Survey responses on the frequency of need for analytics and BI reports.....	35
Figure 17: Survey responses on the level of urgency of analytics and BI reports. ....	35
Figure 18: Survey responses on the reporting history for analytics and BI reports. ....	36
Figure 19: Survey responses on how much is spent on analytics and BI investments. ....	38
Figure 20: Survey responses on the use of analytics and BI reports.....	38
Figure 21: Sample modeling results on attribute <i>IS_fraudulent_claim</i> using J48. ....	40
Figure 22: A model's tree visualization from the prototype.....	41
Figure 23: Sample results of a prediction run on attribute <i>IS_fraudulent_claim</i> .....	41
Figure 24: Sample results of a prediction run on attribute <i>IS_fraudulent_claim</i> in a chart. ....	42
Figure 25: Results of a prediction run on attribute <i>IS_PSV</i> in a chart. ....	42
Figure 26: Results of a prediction run on attribute <i>IS_PSV</i> . ....	43
Figure 27: Results of a prediction run on attribute <i>Cover_Maintained</i> . ....	44

## LIST OF TABLES

Table 1: Project Schedule .....	17
Table 2: Implementation Details.....	21
Table 3: Training Set Summary .....	23
Table 4: Modeling Attributes.....	25
Table 5: Confusion Matrix Structure .....	26
Table 6: Confusion matrix using J48 algorithm.....	26
Table 7: Confusion matrix using Naïve Bayes algorithm.....	26
Table 8: Model Results (based on models that resulted in the matrices on Tables 6 and 7).....	41

## LIST OF ABBREVIATIONS

1. **IT** – Information Technology.
2. **BI** – Business Intelligence. A set of theories, methodologies, architectures, and technologies that transform raw data into meaningful and useful information for business analysis purposes.
3. **WEKA** - Waikato Environment for Knowledge Analysis.
4. **ORM** – Object Relational Mapping
5. **RPC** – Remote Procedure Call
6. **PSV** – Public Service Vehicle
7. **AKI** – Association of Kenyan Insurers
8. **RBA** – Retirement Benefit Authority

## LIST OF TERMS

1. **Analytcs** - the discovery and communication of meaningful patterns in data.
2. **Insurance** - the equitable transfer of the risk of a loss, from one entity to another in exchange for payment.
3. **General insurance** - non-life insurance, including automobile and home owners policies, provide payments depending on the loss from a particular financial event.
4. **Predictive analytics** – a form of analytics that includes a variety of statistical techniques from modeling, machine learning, and data mining that analyze current and historical facts that are used to make predictions about future, or otherwise unknown, events.
5. **Analytical model** - A mathematical model into which data are loaded for analysis.
6. **Decision Tree**- In machine learning, a decision tree is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features (Breiman, 1984). It is used to decide what category an element, represented by a vector of attribute values, belongs to.
7. **Overfitting** – This is a concept in decision tree learning that occurs when a statistical model describes random error or noise instead of the underlying relationship. It generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations.
8. **Entropy** – This is a measure of the impurity in a collection of training examples. Entropy is minimal (0) when all examples are positive or negative, maximal (1) when half are positive and half are negative.

$$\text{Entropy} = -pP * \log_2(pP) - pN * \log_2(pN)$$

Where pP is the proportion of positive (training) examples

And pN is the proportion of negative (training) examples



9. **Information Gain** – this is a measure of the effectiveness of an attribute in classifying the training data.

It is also a concept that measures the amount of information contained in a set of data. It gives the idea of importance of an attribute in a dataset. It is the difference between the entropy before and after a decision since it measures the expected reduction (in entropy) by partitioning the examples according to an attribute.

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_v (|S_v| / |S|) \text{Entropy}(S_v)$$

Where  $S$  – a collection of examples

$A$  – an attribute

$\text{Values}(A)$  – possible values of attribute  $A$

$S_v$  – the subset of  $S$  for which attribute  $A$  has value  $v$

10. **Classification** – This is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.
11. **Posterior Probability** - this is the conditional probability that is assigned after the relevant evidence or background of a random event is taken into account. It represents the probability that a hypothesis  $h$  holds given the observed training data  $D$ . It reflects confidence that  $h$  holds after we have seen the training data  $D$  (Tom, 1997). It is computed as  $P(h/D) = P(D/h)P(h)/P(D)$

# 1. INTRODUCTION

## 1.1. Background

Analytics is the future of big data because transforming data into information gives it value and can turn data in business into a competitive advantage (Ana-Ramona B. et al, 2013). Today, the insurance industry is totally dependent on the ability to convert raw data into intelligence that can support decisions on claims management, actuarial management, and customer relationship management. Over the years, data processing technology has progressed phenomenally and tools like data warehousing, OLAP and data mining, which constitute the cornerstone of effective business intelligence (BI) environments, have been widely accepted across industries. However, insurance companies have been relatively slow in adopting these tools, primarily because of lack of competition caused by protective regulations. But now, they can no longer afford to be complacent as the Internet, de-regulation, consolidation, and convergence of insurance with other financial services are fast changing the basic structure of the industry.

Insurance is quite diverse in terms of the portfolio of products provided. The products can be broadly classified into two product lines: general and life insurance. The life insurance product line can be further sub-divided into life insurance, health insurance and annuity products. Growing consolidation and change in the regulatory framework leads companies to add new products to their portfolio like Linda Jamii from BRITAM Investments Limited. This presents its own unique challenge to any insurer in leveraging its greatest asset – data while managing claims.

Business analytics solutions can help in addressing challenges and making decisions in four key areas:

- Retention and growth of a customer base by predicting the right offer for the right customer.
- Fast tracking claims and thus improve profitability by predicting claim complexity, severity and likelihood of fraud.
- Managing of risk across the enterprise and addressing regulatory compliance requirements.
- Creating and optimizing integrated distribution channels with sales performance management solutions.

A number of other trends in the insurance industry have also exponentially increased the importance of an effective business intelligence environment; at the same time, these trends are responsible for increasing the complexity of building such an environment. They include

- **Growing Consolidation:** Consolidation is a major force altering the structure of insurance, as insurers seek to create economies of scale and broaden their product portfolios.
- **Convergence of Financial Services:** Mergers and acquisition of insurance companies with other financial service providers like banks has led to the emergence of integrated financial services companies. E.g. BRITAM's latest acquisition of REAL insurance and its collaboration with SAFARICOM in their latest product Linda Jamii.

- **New Distribution Channels:** New distribution channels are fast catching up with the traditional insurance agent. Though these channels are not a major threat as yet, they are rapidly changing the way insurers and customers interact with each other.
- **Focus on Customer Relationship Management:** The only viable strategy for insurers today is to focus on the needs of the customers and strive to serve them better. Customers have extremely differentiated needs and, also, the profitability of individual customers differs significantly. Hence, an effective CRM strategy becomes the most vital component of an insurer's overall business strategy.

Below is an illustration of a typical claims process and its decision points.

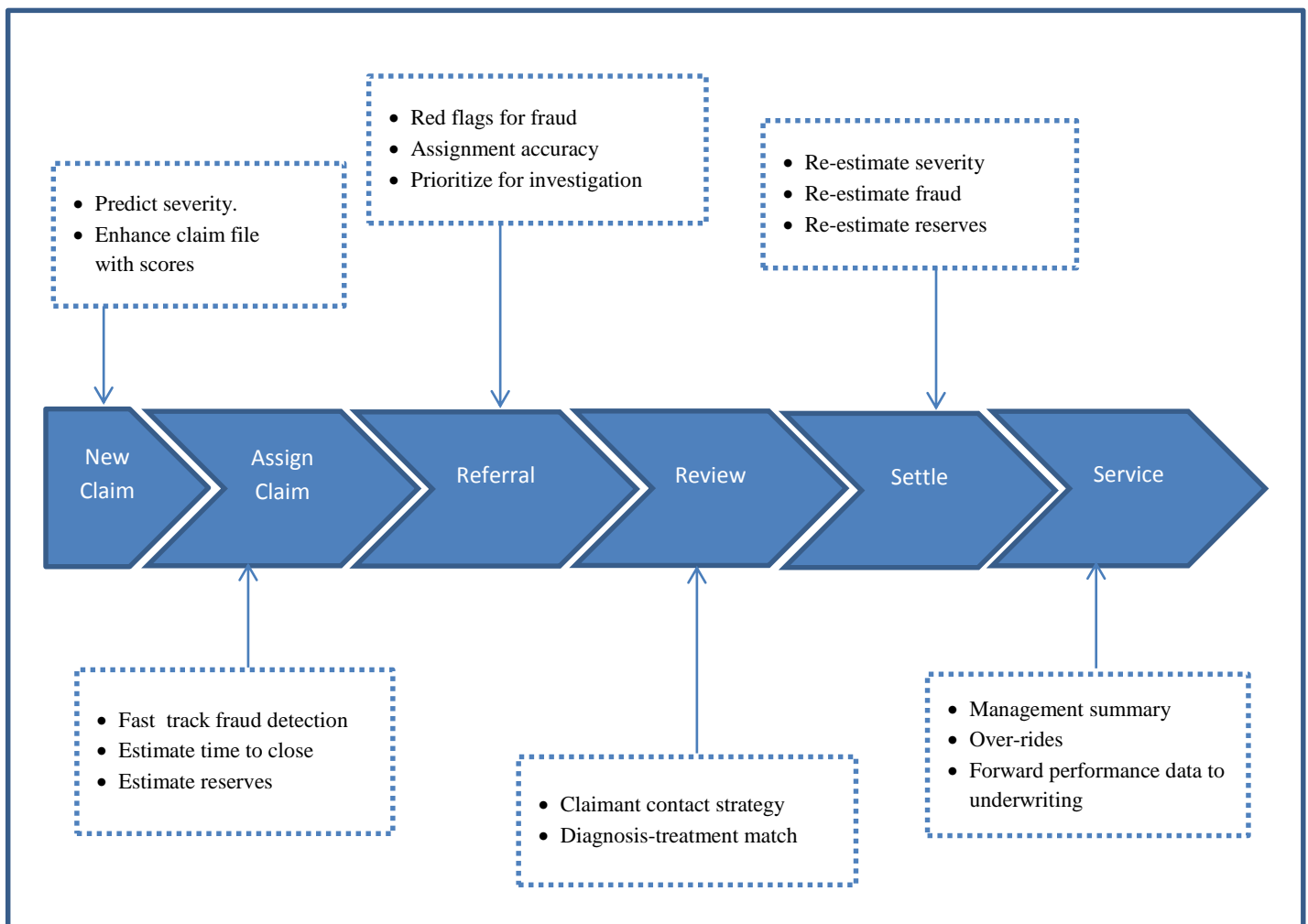


Figure 1: A typical claims process and its decision points.

## 1.2. Problem Statement

In the insurance industry, it is estimated that about 40-50 percent of claims contain some degree of suspicion concerning fraud (Derrig W., Chen X., 1994). There is also total reliance on the ability to convert raw data into

intelligence that can support decisions especially in claims analysis. The main issue lies in the fact that the collections of data are impossible to be processed by a human brain. Most insurance companies in the country employ spreadsheets as their primary tool for analytics. Whereas spreadsheets serve the required purpose, they require a lot of manual intervention, an understanding of the data formats of the set of data sources, data querying skills, and a good understanding of the software's inbuilt functions in order to produce any analytical report. This in turn leads to a slow reporting process. A delay in reporting on the other hand leads to inadequate claims management and therefore slow risk management and fraud management. Also, the kind of analytics produced by existing systems is *reactive*. Future plans that stem from the analytics reports are therefore largely based on experience and intuition. What this study is proposing is more of a *predictive* form of analytics where insurance companies would have data-supported facts for their future projections and can leverage on this insight to optimize every decision, transaction or process.

The presence of both legacy and more modern enterprise systems, multiple operating or administrative systems, custom rules, and so on, further complicate the process of effective claims analytics using spreadsheets that do not offer easy interfacing with other systems if any. This combined with the extra challenge of managing the growing speed, complexity and volumes of data together with growing product portfolios call for better solutions to analytics and BI in the Kenyan insurance industry.

Another problem is the high initial cost of purchasing, customizing and implementing existing Analytics and Business Intelligence tools. This is part of the reason why most industry players have chosen to work with spreadsheets.

Also, most of these solutions are technology driven, meaning that their sophisticated models are based on machine learning and other techniques which primarily rely on advancements in computing power rather than the business or user requirements. They are also sequential, meaning that since they are complete off-the shelf solutions, they might require well-defined data requirements up front in order to make an appropriate choice from amongst the many software vendors. And for such solutions, business value is delivered all at once at the end of the project. However, all these sequential steps would be appropriate if an analytics process is well-defined and predictable. Unfortunately, analytics is more of a discovery process where one rarely knows beforehand which paths to take and what would be required. Many times one rarely knows what they will be discovering. As a result, analytics and BI projects that take this approach often exceed timelines and run for years.

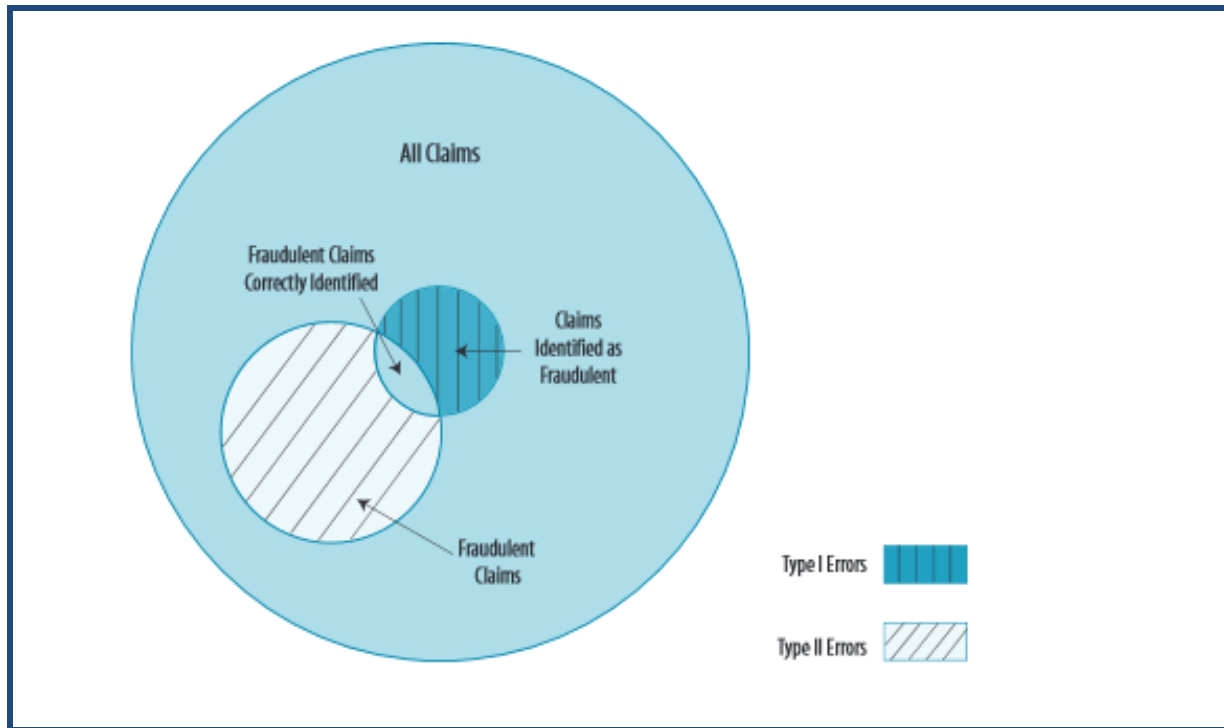


Figure 2: An illustration of the problem faced with little or no analytics in claims management

### 1.3. Research Objectives

The main objective of this research was to build a predictive model that can give useful insight from claims data and thus improve analytics and BI for claims management in the insurance industry. Other objectives were as follows:

- 1) To establish the level of use of analytics and BI in the insurance industry for claims management. This included determining the frequency of need and the urgency with which the output of analytics is usually required. This was to advise on the speed of response of the solution for every request.
- 2) To establish the value that can be possibly gained from predictive analytics and BI in claims management.
- 3) To establish current industry investments and acquisitions in analytics and BI for claims management, and any regulations that might pertain to analytics and BI directly or indirectly.
- 4) To make analytics and BI applicable at non-managerial levels in an insurance organization where staff are not necessarily actuaries.

### 1.4. Research Questions

This study focused on filling the gap left by traditional analytics that is done using spread sheets and by analytics done using off the shelf products. The key issue was to look for a cost-effective option that can be easily acquired, customized and implemented. Therefore the research intended to answer the following questions:

- 1) How much delay in analytical reporting can be tolerated while performing claims management?
  - a) How often are analytical reports required? How urgently are they usually needed?
  - b) How far in history should analytics and BI go while managing claims?

- c) What kind of essential analytics in claims cannot be generated from current tools especially spreadsheets?
- 2) What level of skill are companies willing to invest in to handle their analytics and BI in claims?
  - a) How much is spent on average on analytics and BI?
  - b) How can analytics and BI be used to detect fraudulent claims?
- 3) How does industry regulation affect the implementation of analytics and BI in managing claims?
  - a) What industry rules must any analytics and BI solution conform to?
- 4) What is the main target audience for the output of analytics and BI?

## 2. LITERATURE REVIEW

### 2.1. Previous Work

In a research done by Sharon T. et al (Sharon T et al, 2002), claims analytics techniques were found to highly rely on recorded statements. The most prevalent claims analysis technique was the recording of statements taken from a party to the claim incident (the claimant, the insured, or a witness). Another common technique was an independent examination of the claimant. Sworn statements, activity checks, medical audits, site investigations, and special investigative unit referrals were used much less frequently. In the same study, it was found that claims auditing patterns are consistent with the use of audits for both fraud detection and fraud deterrence. Consistent with a detection objective, subjective characteristics of claims that could only be determined on an individual claim basis were found to be significantly and positively related to the probability that a claim was investigated. The study, which aimed at determining how auditing was used primarily for detection or deterrence of future fraud, suggested that insurers pursue both detection and deterrent objectives in auditing.

In another research, Patrick and his co-researchers apply in their empirical study Kohonen's Self-organizing Feature Map to classify automobile bodily injury claims by the degree of fraud suspicion. They use feed forward neural networks and a back propagation algorithm to investigate the validity of the Feature Map approach (Patrick L. et al, 1998).

In a research by Patrick L. et al, they study a mathematical technique for an a priori classification of objects when no training sample exists for which the exact correct group membership is known. Using their technique, they attempt to reduce uncertainty and increase the chances of targeting the appropriate claims to uncover insurance fraud. Their technique also gave measures of the individual fraud indicator variables' worth and a measure of individual claim file suspicion level for the entire claim file (Patrick L. et al, 2002).

In another research, Richard and his co-researchers review fuzzy pattern recognition techniques and use them in clustering for risk and claims classification in automobile insurance data (Richard A., Ostaszewski K., 1995).

### 2.2. Use of Analytics in Insurance Today

Analytics has always existed in the insurance industry. However the main form of analytics that is currently being practiced is the reactive kind as opposed to the predictive one. While reactive analytics raises some benefits to a company, the predictive kind comes with extra benefits associated with the ability to read into the future with some data backing. These include effectively and proactively managing risks that come with claims management. Big Data technology and distributed processing power of big data cloud bring tasks like fraud detection in insurance to another level. Not long ago, insurance fraud detection was not considered cost-effective because the cost and duration of the investigations were too high, so many companies prefer to pay claims without investigation (Ana-Ramona B. et al, 2013). Applying Big Data analysis methods can lead to rapid detection of risks (Ana-Ramona B. et al, 2013), and then creates a new set of tests to automatically narrow the segment of potentially risky claim applications or to detect new patterns of fraud, previously unknown.

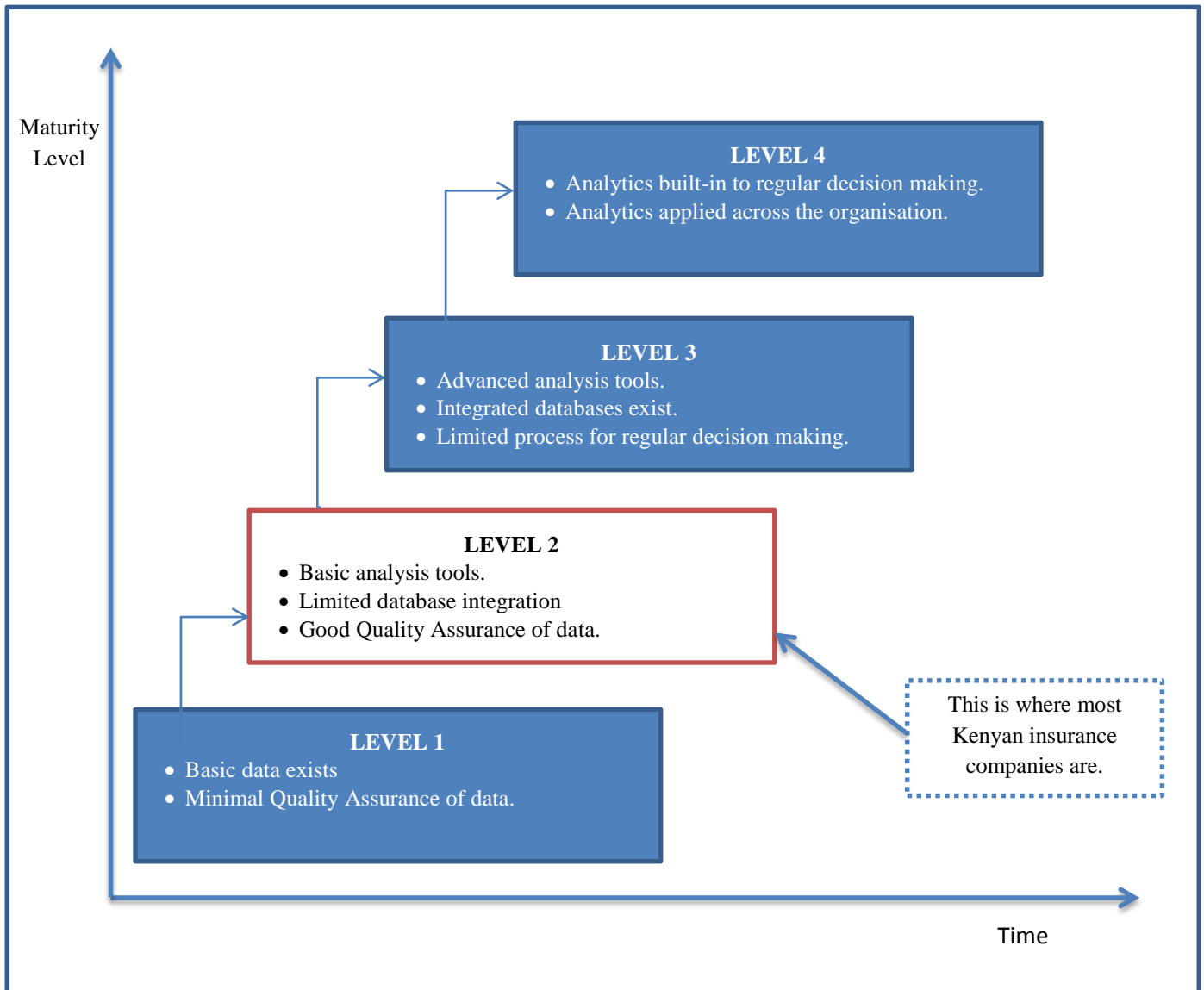


Figure 3: Analytics capability maturity against time in Kenya.

Currently, BRITAM, CIC and UAP use Microsoft Excel to support decisions in risk management, underwriting and policy management, claims management, actuarial management, customer relationship management, and sales agents' performance management. Spread sheets like Microsoft Excel do provide analytics capabilities but their main undoing is the inability to integrate various data sources and their limited query writing provisions.

As of 2011, analytics and BI was still in the 'emerging stage' and many insurance companies worldwide were proceeding cautiously towards its adoption (Steven 2012). After that, a research done by SAP in 2013 showed that its adoption had reached 10% and is expected to sky rocket to 75% by the year 2020. Most of the companies that had adopted analytics and BI by then were mostly concentrated in Europe, North America and in Asia Pacific. Africa



and Kenya in particular had not adopted it much as yet (SAP 2014). The two researches above however showed that analytics and BI were being looked to in solving big issues, with the primary focus being on improvement. The slow adoption has been attributed to IT not being agile enough, data quality, acquisition and integration, lack of proper analytical talent, and the ability to effectively analyze the growing volumes of data getting harder by the day. Culture was also determined to be a major barrier to the effective use of analytics (SAP, 2014).

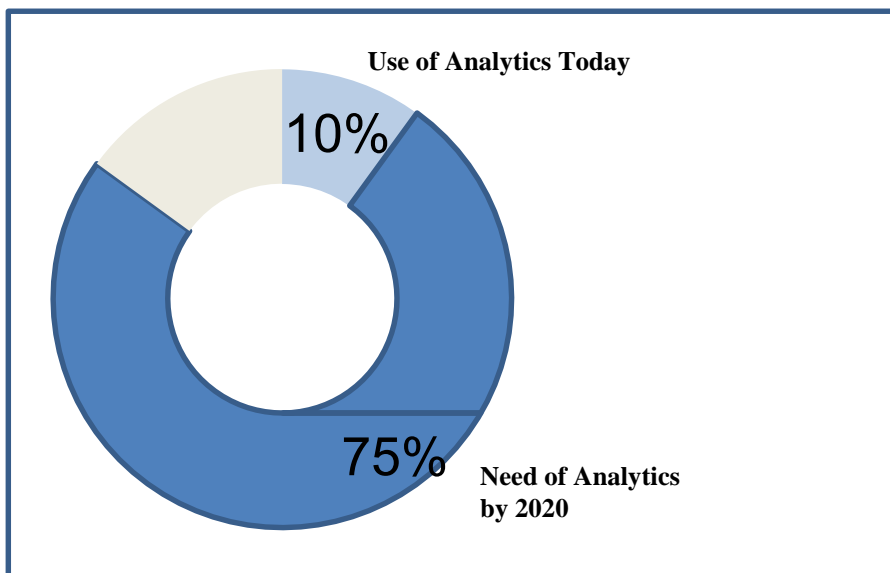


Figure 4: Adoption and Use of analytics as at 2013 (SAP 2014)

### 2.3. How This Study Was Different From What Has Been Done Before

There have been similar studies, for example one by Keith (2012). Most of them have however focused on the importance of claims analytics as opposed to reviewing the problem of its poor adoption. Others have done research on insurance analytics as a whole and thus generalize on their findings. In their research, Saama Executives (Saama 2014) have studied claims data analytics transformation using a commercial application called Guidewire Claim Center. This study however was more inclined towards organizations that already have implementations of claims analytics and so they use the new tool to build on to their existing analytics environments. This study also focused on those companies with little or no analytics implementation and which were (are) seeking a more cost-effective results-oriented approach. SaS has also done a study on predictive claims analytics that also gives great focus to SaS Company's claims analytics product as a solution (SaS 2013). Other studies on analytics but which are not necessarily focused on claims include Crain (2012) and Dan & Henry (2011).

Studies that have attempted the development of a BI claims fraud detection system or claim classification system include work done by the Automobile Insurance Bureau of Massachusetts (Derrig, R. A., Weisberg I, 1996), a study by Artis, Ayuso and Quillen to model the behavioural characteristics of claimants and insureds in the Spanish automobile insurance market (Artis, M, Ayuso M., Quillen M., 1997), and an expert system has been developed by Belhadji and Dionne (Belhadji, E. B., Dionne G., 1997) to aid insurance company adjusters in their decision making and to better equip them to fight fraud.

In this empirical study, we intended to apply a different approach to build a BI claim fraud detection or classification system. Specifically, we applied a decision tree approach, J48, to construct a claim classification system that uses similar collections of attributes in the classification.

### 2.4. Existing Solutions for Claims Analytics

Data mining techniques: these can be used for fraud detection for large sets of data from insurance system. These techniques detect behavior patterns in large datasets, so based on several cases considered fraudulent can calculate the probability that each record be fraudulent. There are two main criticisms of data-mining fraud detection tools: the dearth of publicly available data for analysis and the lack of published well-known methods and techniques that are specifically efficient for this field (Ana-Ramona B. et al, 2013).

Other examples of analytics and BI tools that are available in the market and that can be used to facilitate claims management include Business Analytics for Insurance Claims, SAS® Fraud Framework for Insurance, IBM Business Analytics, Microsoft Excel, Guidewire Claim Center by Saama, and IBM Business Analytics for Insurance

### 2.5. Barriers to Adoption of Analytics and BI

The chart below shows an analysis by percentage of respondents on some of the barriers to adopting analytics and BI by insurance companies.

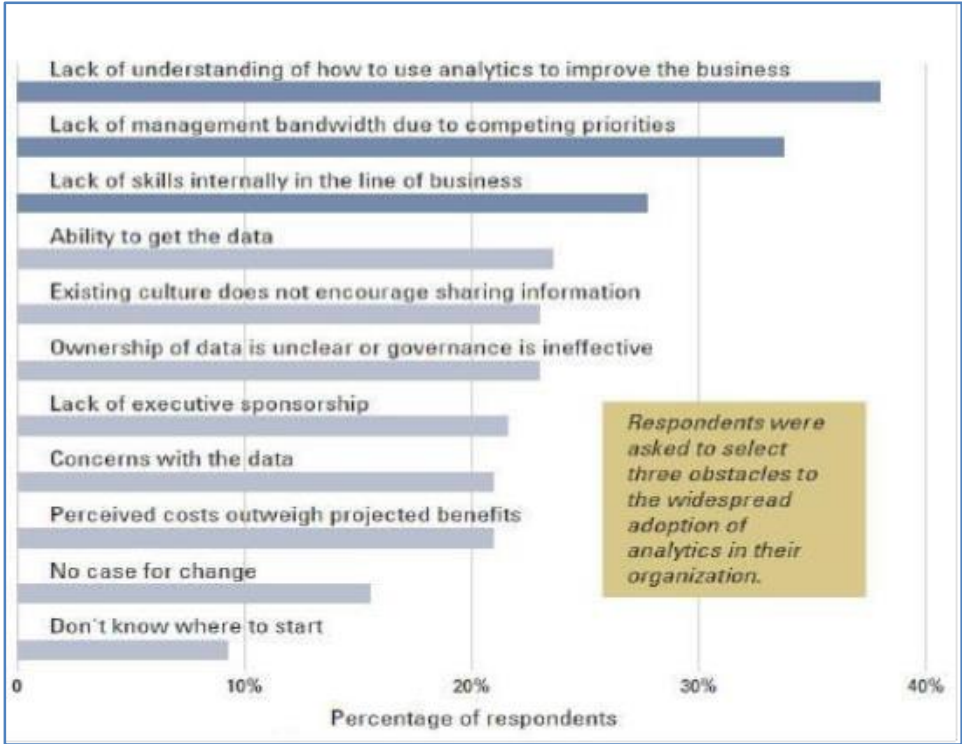


Figure 5: Barriers to analytics and BI adoption. Source: MIT Sloan Management Review 2012 (SAP 2014)

## 2.6. Models Categories

1. **Predictive models:** Predictive models analyze the past performance for future predictions. A predictive model provides a score to all claim applications based on the initial information provided by the applicant. It also classifies the claim application into a risk class category. The claims review applications used by insurance companies need to be optimized in order to incorporate the predictive model efficiently. If the information like score, risk class etc. generated by the predictive model can be presented to the claim administrator, it can augment the speed of claim processing for low risk claims (Akshay B., 2012). Also, by providing an indicator for the low risk cases, some of the requirements generated during the initial processing of a claim application can be skipped. But the final decision remains in the hands of the claim administrator.
2. **Descriptive models:** These models quantify the relationships in data in a way that is often used to classify datasets into groups.
3. **Decision models:** These models describe the relationship between all the elements of a decision involving many variables in order to predict the results.

## 2.7. Predictive Analytics

Predictive analytics is the process of using a predictive model to guess the probability of an outcome from a given set of input data. It's a powerful technique that converges technology, mathematical statistics, probability and other disciplines (George B et al, 2005). A predictive model solely corresponds to the requirements of a business. This model scrutinizes various data available for the customer through the existent customer logs, behavior and demographics (Akshay B., 2012). This information is encoded into a model, which, together with the business rules, calculates the risk factor for the customer. Predictive modeling is the process of creating, testing and validating a model to best predict the probability of an outcome. A number of modeling methods from machine learning, artificial intelligence, and statistics are used in predictive modeling.

In recent years, predictive modelling using general linear models (e.g. Poisson regression, logistic regression, log-linear analysis) have become immensely popular among actuaries and statisticians. Such modelling has the advantage of being more tractable and more amenable to meaningful interpretation (George B. et al, 2005) than results from common analysis tools.

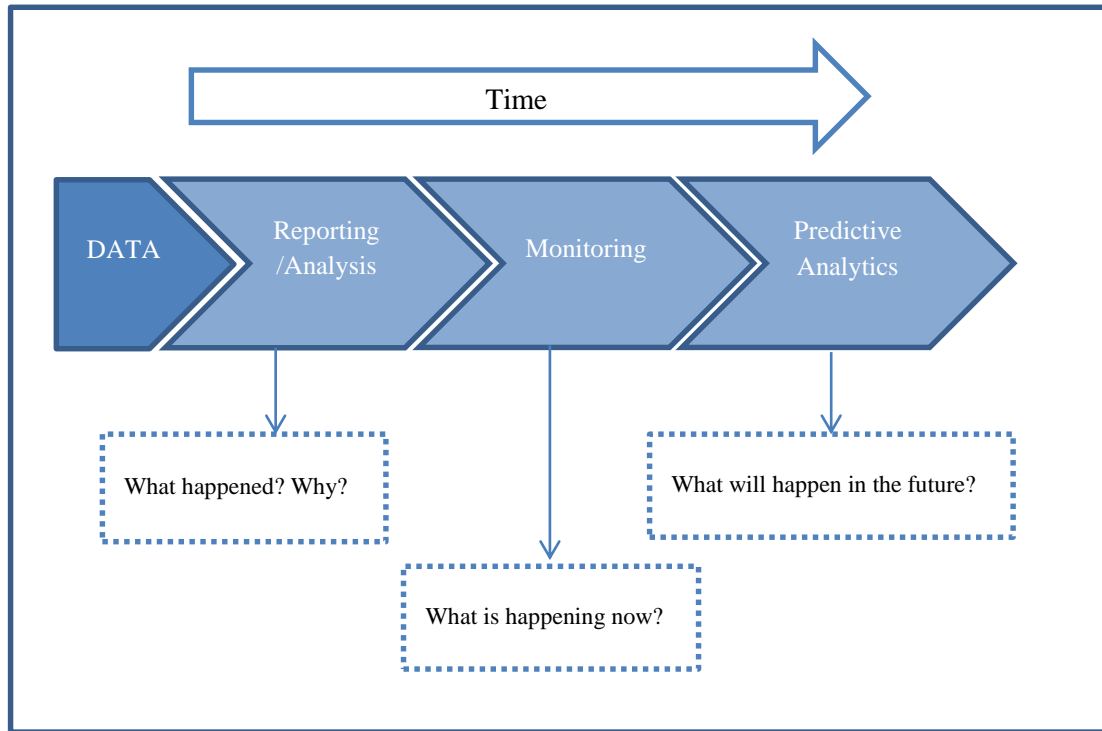


Figure 6: Predictive Analytics

## 2.8. Predictive Modeling Process

This is an iterative process that involves the following steps

1. Model Creation– here, a model based on one or more algorithms is created based on a given dataset.
2. Model Testing– here the model is tested on the dataset. The testing can also be done on past data to see how best the model predicts.
3. Model Validation – here the output of the model is then validated using visual tools and normal understanding of business data.
4. Model Evaluation– here, the model’s fit for data is determined.

## 2.9. Predictive Modeling Algorithms

The following are some of the algorithm categories that are used when performing data mining and statistical analysis in order to determine trends and patterns in data. It is however important to note that several binary classification techniques have been compared in auto-mobile insurance claims analytics to determine which approach is most effective, but no single method has been found to work best (Stijn V. et al, 2002).

1. **Time Series Algorithms:** these algorithms perform time based predictions. Examples are Single Exponential Smoothing, Double Exponential Smoothing and Triple Exponential Smoothing (Predictive Analytics Today, 2014).

2. **Regression Algorithms:** these algorithms predict continuous variables based on other variables in a dataset. Examples include Linear Regression, Exponential Regression, Geometric Regression, Logarithmic Regression and Multiple Linear Regression (Predictive Analytics Today, 2014).
3. **Association Algorithms:** these algorithms find frequent patterns in large transactional datasets to generate association rules. An example is the Apriori algorithm (Predictive Analytics Today, 2014).
4. **Clustering Algorithms:** these algorithms cluster observations into groups of similar properties. Examples of such algorithms are K-Means, Kohonen, and TwoStep (Predictive Analytics Today, 2014).
5. **Decision Tree Algorithms:** these algorithms classify and predict one or more discrete variables based on other variables in a dataset. The learning approach is to recursively divide the training data into buckets of homogeneous members through the most discriminative dividing criteria. The measurement of homogeneity is based on the output label; when it is a numeric value, the measurement will be the variance of the bucket; when it is a category, the measurement will be the entropy of the bucket. The training process stops when there is no significant gain in homogeneity by further splitting the tree. The members of the bucket represented at leaf node vote for the prediction; majority wins when the output is a category and member's average is taken when the output is numeric. Examples include C4.5 and CNR Tree (Predictive Analytics Today, 2014).
6. **Outlier Detection Algorithms:** these algorithms detect the outlying values in a dataset. Examples of such algorithms are Inter Quartile Range and Nearest Neighbour Outlier (Predictive Analytics Today, 2014).
7. **Neural Network Algorithms:** these algorithms do forecasting, classification, and statistical pattern recognition on datasets. Examples are NNet Neural Network and MONMLP Neural Network (Predictive Analytics Today, 2014).
8. **Ensemble Models:** these models use a form of Monte Carlo analysis where multiple numerical predictions are conducted using slightly different initial conditions (Predictive Analytics Today, 2014).
9. **Factor Analysis:** these deal with variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. An example of such an algorithm is the Maximum Likelihood Algorithm (Predictive Analytics Today, 2014).
10. **Naive Bayes:** this is a probabilistic classifier that is based on applying Bayes' Theorem with strong (naive) independence assumptions (Predictive Analytics Today, 2014).
11. **Support Vector Machines:** these are supervised learning models with associated learning algorithms that analyze data and recognize patterns. They are used for classification and regression analysis (Predictive Analytics Today, 2014).
12. **Uplift Modeling:** this models the incremental impact of a treatment on an individual's behavior (Predictive Analytics Today, 2014).
13. **Survival Analysis:** this refers to the analysis of time to events (Predictive Analytics Today, 2014).

## 2.10. Features of Predictive Modeling

Some of the features of predictive modeling are as follows

1. **Data analysis and manipulation:** This includes the creation of new datasets, modification, categorization, merging and filtering of datasets, and tools for data analysis.
2. **Visualization:** Visualization features include interactive graphics and reports.
3. **Statistics:** This includes tools to create and confirm the relationships between variables in data. Statistics from different statistical software can also be integrated into a solution.
4. **Hypothesis testing:** This involves the creation of models, evaluation and choosing of the right model.

### **2.11. Significance of the Study**

The research meant to show how analytics and BI in claims can be achieved cost effectively and meant to assist in managing losses incurred through fraudulent claims.

### **2.12. Beneficiaries of the Research**

1. Actuaries. These are professionals who deal with the financial impact of risk and uncertainty from business data including claims data.
2. Claims managers
3. Claims administrators
4. Company top-level management

### 3. METHODOLOGY

In this section, research methods for this study are described. The method of choice for the study was Action Research. The section is divided into four subsections. Sub-section one describes how Action Research was to be used i.e. describes the study strategy including the system development methodology chosen. Sub-section two gives reasons why Action Research was the preferred method. A test strategy for this study is highlighted briefly in sub-section three while sub-section four highlights expected outcomes of the study.

#### 3.1. Application of Action Research

The cyclic action research phases namely analysis (or fact finding), planning, acting (execution), observation and reflection (Reporting results) were followed through in this study.

- **Analysis:** here, an as-is analysis of the current claim analytics processes was done in depth to understand the basic processes, rules and reporting requirements from the process. This involved a review of existing tools and how they are linked with the mostly legacy systems in insurance.
- **Planning:** this was the second phase and involved a detailed plan definition of the activities that were to be carried out throughout this research from identification of data sources to the definition of the model/system requirements and the documentation of the final system and research outcomes. The details of this phase were mainly influenced by the outcomes of the first phase
- **Acting:** in this third phase, the activities in the plan were executed. This phase majorly involved the development of an analytics and BI prototype.
  - The chosen development methodology was **Extreme Programming Agile methodology (XP)**. XP was chosen because of its ability to deliver high-quality software quickly and continuously. It also promotes high end-user involvement, rapid feedback loops, continuous testing, and continuous planning to deliver a working software solution at very frequent intervals, typically every 1-3 weeks.
  - The chosen predictive modeling algorithms were **Decision Trees** and **Naïve Bayes**. These algorithms were chosen because of the following reasons.
    - They both have been widely and successfully applied in predictive analytics over time. This is because they are easy to interpret and explain to business users compared to the other methodologies (especially decision trees).
    - Decision trees implicitly perform variable screening or feature selection. This is a good thing because feature selection is very important in analytics. When one fits a decision tree to a training dataset, the top few nodes on which the tree is split are essentially the most important variables within the dataset and feature selection is completed automatically.
    - They both require relatively little effort from users for data preparation to overcome scale differences between parameters. For example, if we have a dataset which measures claim

amounts in millions and claimant ages in years, we will require some form of normalization or scaling before we can fit a regression model and interpret the coefficients. Such variable transformations are not required with decision trees for example because the tree structure will remain the same with or without the transformation.

- Nonlinear relationships between parameters do not affect a decision tree's performance. Highly nonlinear relationships between variables result in failing checks for simple regression models and thus make such models invalid. However, they (especially decision trees) do not require any assumptions of linearity in the data. Thus, we can use them in scenarios where we know the parameters are nonlinearly related.
- **Observation:** in this phase, the analytics and BI prototype was tested and its outputs reviewed. Its limitations and strengths were also evaluated and further action to possibly improve it proposed by potential users. This phase involved users who deal in claims analytics and actuaries where they took part in the model's testing.
- **Reporting/reflection:** In this phase, the research was documented with all its outcomes.

### 3.2. Why Action Research

Action research was chosen for this study because of the following reasons.

- a) Action Research accommodates user centered design and co-researching with target users. This study plans to involve target users and make them central to the study.
- b) Action Research is used in real situations as in this case of fraud in insurance claims, rather than in contrived, experimental studies, since its primary focus is on solving real problems.
- c) The development process for the study was expected to be cyclic. Action Research supports a cyclic process. It also easily accommodates the Extreme Programming system development methodology.
- d) Action Research has been successfully used in related areas of research e.g. in IBM (2011).

### 3.3. Testing Strategy

Users were to actively participate in the implementation and testing of the analytics and BI prototype output of the research. These tests involved functional tests and User Acceptance Tests where the users were given time to perform analytics functions using the model.

### 3.4. Anticipated Outcomes

The results of this study were, mainly, an analytics and BI prototype and a documentation of the study. The study was expected to be a major contribution in reducing the number of fraudulent claims that are processed by insurance companies and therefore reduce losses resulting from it. This would help the companies to also easily conform to industry standards that expect a certain limit of fraud in claims for every company year.

### 3.5. Expected Contribution

From the problem statement above, a more effective results-oriented approach or solution was required, and is what this research proposed. The approach was to be business-driven, iterative, incremental and continuous (Ravi 2012).



- **Business-driven:** Analytics is a bespoke craft that is used to gain a competitive advantage. Industry standard solutions or best practices borrowed from others would not provide that core advantage. A solution that is tailor-made for the specific needs of an organization was required. The proposed solution would provide a simple, easy to use interface that does not require its user to have query writing skills.
- **Iterative and incremental:** Analytics and BI implementation does not need to be front-loaded with high costs. The diagram below shows the different phases of an incremental and iterative analytics life cycle. The goal is to go through all phases of an analytics process within a reasonably short time, all the while preparing to iterate through it again. Each iteration would enable an implementation team to learn and improve. These low-risk small-budget iterations allow for the design of a long-term solution that is tailor-made to an organization's unique needs. Also, approaching a project in such an agile, iterative way produces results in a matter of weeks as opposed to a sequential approach which could take a year or more to begin to deliver value

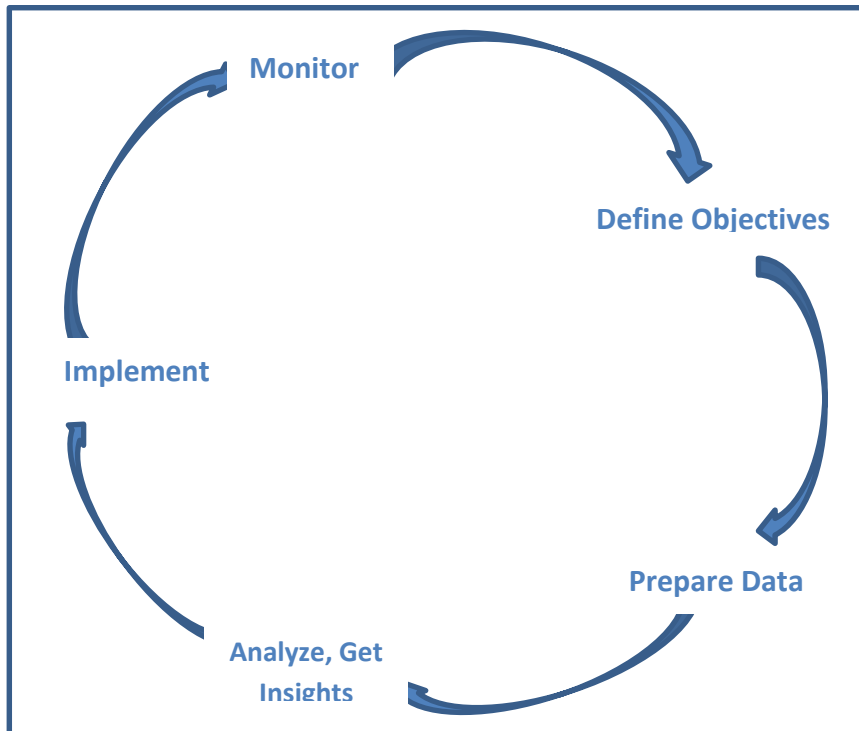


Figure 7: An Iterative approach to analytics and BI.

- **Continuous:** To be effective, analytics and BI should not be a one-time or once-every-few-years project. It requires constant calibration as adjusters, customers, and vendors react to changes already in place. In addition, analytics and BI must incorporate changing business needs on an ongoing basis.

### 3.6. Project Schedule

	Activity	Expected Duration (weeks)	Expected Duration (days)	Tentative Start Date
1.	Selection of participating companies	0.5	3	22-07-2014
2.	Sending of requests for participation and consent to selected companies	1	7	25-07-2014
3.	Identification of relevant data sources	1	7	01-08-2014
4.	As-is analysis	2	14	08-08-2014
5.	To-be analysis	2	14	22-08-2014
6.	Data collection and preparation	3	21	05-09-2014
7.	Data analysis and review	4	28	26-09-2014
8.	Development of the solution	8	56	24-10-2014
8.1	Iteration 1	2	14	07-11-2014
8.2	iteration 4	2	14	21-11-2014
8.3	Iteration 2	2	14	05-12-2014
8.4	Iteration 3	2	14	19-12-2014
9.	Solution testing	4	28	19-12-2014
10.	Research results interpretation and presentation	1	7	16-01-2015
11.	Research documentation	1	7	23-01-2015
12.	Expected end date			30-01-2015

Table 1: Project Schedule

### 3.7. Design and Implementation

In this section, the system (prototype) design and implementation process will be described. This predictive analytics research project was used to help identify potentially fraudulent claims. It was also used to predict the likelihood of a claimant to retain (or surrender) his policy after a claim, the likelihood that the claim amount would surpass the sum assured, chances that the claim is from a PSV owner, or chances that the claim is as a result of an accident.

### **3.7.1. Survey Setup**

Most insurance companies in Kenya have on average 7 claims administration and management staff for general insurance proper in the Kenyan offices. So from a sample of 6 companies, the entire population size is estimated to be about 42. This population includes claims administrators, claims managers, claims analysts and IT system administrators for general insurance proper. To facilitate the survey, employees of the companies were contacted and requested to participate in the research. A survey was set up on Google forms and formal participation requests sent out to them on e-mail (See appendix 1).

#### **3.7.1.1. Proposed Participants**

1. Geminia Insurance Kenya
2. Real Insurance Kenya
3. CIC insurance Kenya
4. BRITAM Insurance Kenya
5. UAP Insurance Kenya
6. ICEA Lion Group

### **3.7.2. Data Sources**

The data sources in use for this research are general insurance databases. The data was obtained from the internet (see reference 6) and was from outside the Kenyan industry. This was because of a major challenge faced obtaining real claims data from the proposed participants even with the data masked (see appendix 3). The main reason given the need to abide by contracts signed with clients on data privacy, and the data owners' lack of control of the research and how the data given would be used and managed.

### **3.7.3. As-Is Analysis**

The most common general insurance system is AIMs which can be safely categorized as a legacy system that is implemented using a command line language called CQSC (CyberQuery CyberScreen). This system does not use the currently common relational databases but instead stores its data in flat files. Occasionally this data is copied across to a relational database (e.g. MSSQL Server) and used for custom report generation and analytics by creating spreadsheet queries that are executed against the relational database.

Currently, spreadsheets are the most common analysis tools used for general insurance claims analytics. The figure below illustrates the as-is architecture.

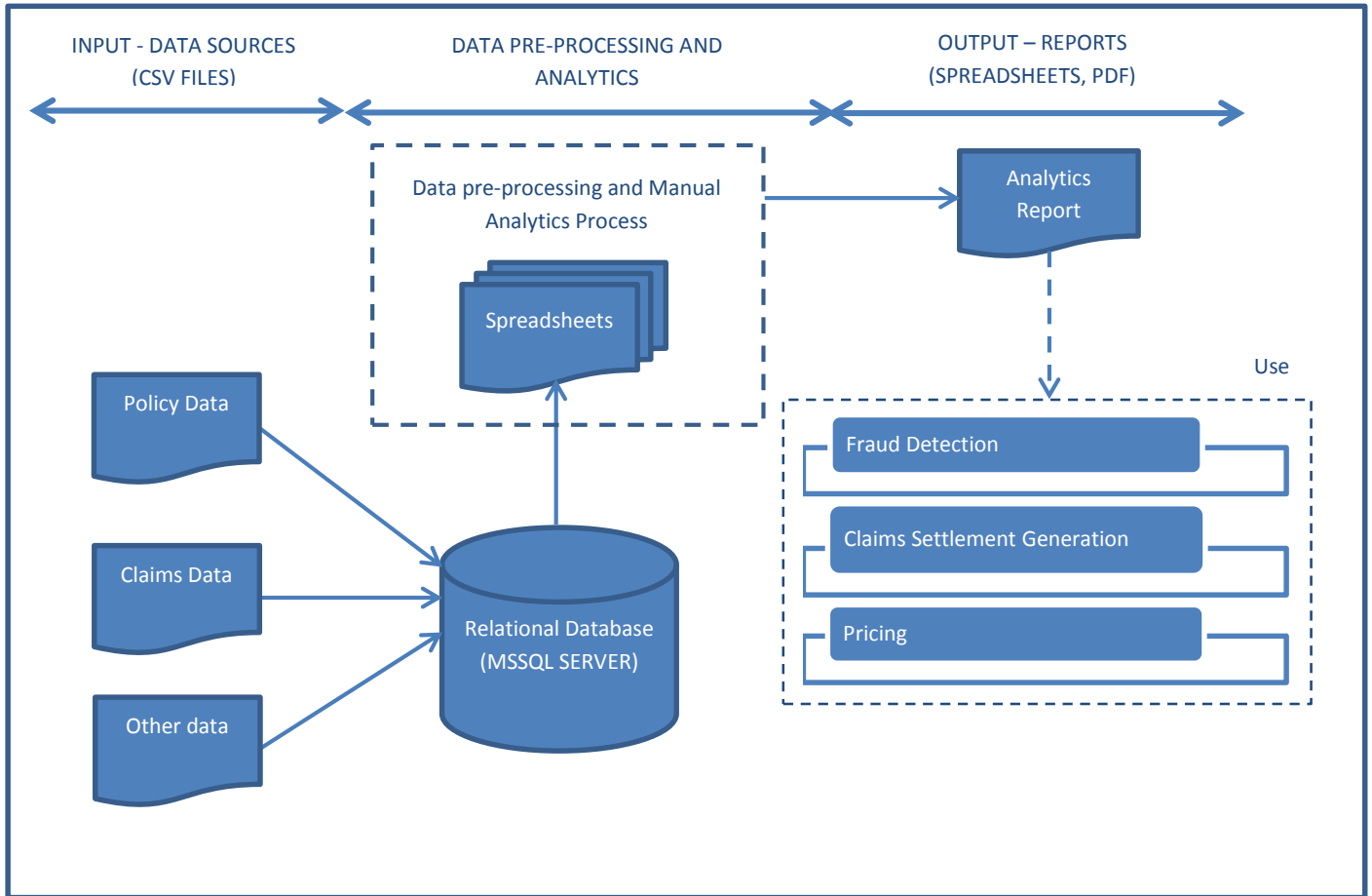


Figure 8: AS-IS Architectural Design

### 3.7.4. To-Be Analysis

Methods of identifying and preventing a risk like fraud must always be re-adjusted to rediscover the fraudulent actions, hence the need for a solution that provides for frequent re-modelling. A predictive analytics process that takes in comma separated files (CSV) as input and provides analytics as output was chosen as the to-be process. The process is as follows.

1. An initial set of claims data files are pre-processed into a single file of ARFF file format (University of Waikato, 2002) and used to build an initial model. ARFF file format is a CSV file with a header that describes the data variables in the file. This would constitute the training process of the model using company data.
2. The claims data flat files are then pre-processed into a single file of ARFF file format. Depending on restrictions placed by an organization on the use of third party applications within its production environment, the source flat files will be read directly from the production environment (flat files) or from a staging area built on MySQL Server. The process of building the ARFF file will be automated with no user intervention.

3. The ARFF formatted file is then fed into the model and executed against it for predictive analysis. This would also be automated with no user intervention.
4. The results (analytics) are formatted into a user-friendly format and presented to the system user.

To preserve accuracy, models must be constantly updated or new ones created to include new types of illegal events in modelling data. The prototype was based on a model built (trained) using existing data. The figure below illustrates the to-be architecture.

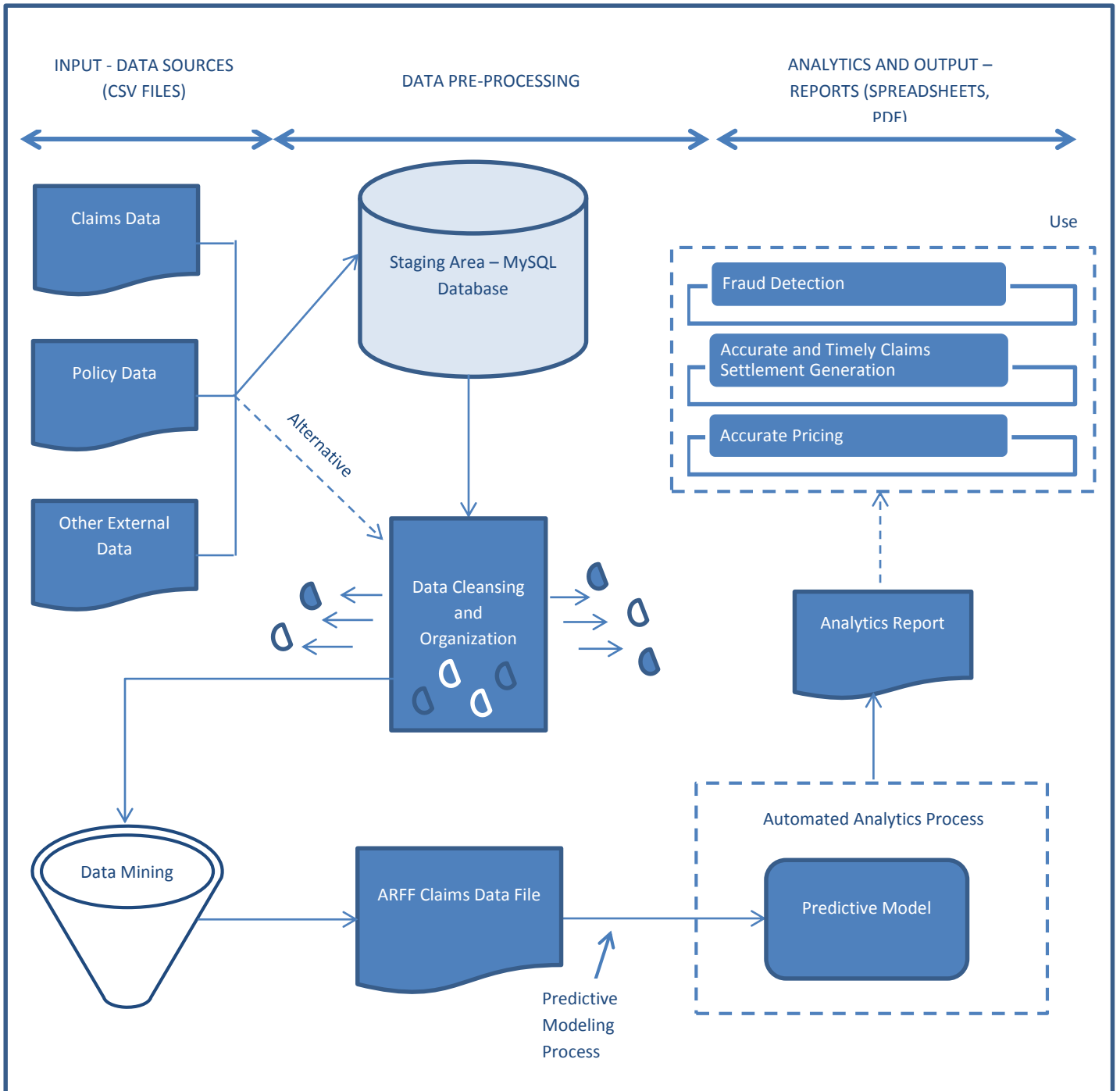


Figure 9: TO-BE High – Level Architectural Design

Other details of the implemented solution are as follows

Item	Description
1. Processing methodology	Predictive analysis
2. Analysis frequency	On demand
3. Analysis type	Real-time
4. Data type	Claims meta-data, claims historical data, claims transactional data
5. Content Format	Structured, text
6. Data Sources	Transactional data
7. Hardware	Commodity hardware

Table 2: Implementation Details

### 3.7.5. Data Collection and Preparation

For purposes of this research, the above mentioned data source was used. Based on the above architecture, data preparation would not be a one-off task and would have to occur in every analytics task. The frequent data preparation was necessitated by the need to have real-time analytics that includes any newly generated data. Data preparation would involve cleansing and organizing data ready for use in the prototype. This is be done by system users too and the process would be made easier by having the data in a staging area built from a relational database. This research project can be extended to have a friendly user interface that allows for data preparation for data sitting in the staging area.

### 3.7.6. Data Analysis and Review

Data from the above sources was analyzed in its original form. The analysis aimed at reviewing

1. Data complexity
2. Data volume, and
3. Data history

Based on the analysis, this research chose to focus on general insurance proper claims data which was available in higher volume than claims from other general insurance business categories, and was less complex that other claims data.

### 3.7.7. Design and Implementation

#### 3.7.7.1. Model Development Framework

WEKA (Waikato Environment for Knowledge Analysis), which is a collection of machine learning algorithms for data mining tasks, was selected as the model's development framework. The decision to use it was informed by the fact that WEKA is Java-based and therefore provided easy integration of models into this research's java-based prototype. Its algorithms can either be applied directly to a dataset or called from Java code. It also contains

tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Thus it was well-suited for developing the new machine learning models in this research project.

### 3.7.7.2. Algorithm – Decision Trees

Initial tests were done on WEKA to determine the level of accuracy of the previously selected predictive modeling algorithms - Decision Trees (implemented in WEKA as J48) and Naïve Bayes - before any further development can be done. After these initial tests, Decision Trees were found to have a lower error margin with 85.7 percent accuracy on training data, while Naïve Bayes was found to have an accuracy of 63.8 percent.

- **J48**

J48 is an open source Java implementation of the C4.5 pruned decision tree algorithm in WEKA (Wikipedia, 2014). C4.5 is a program that creates a decision tree based on a set of labeled input data. This algorithm was developed by Ross Quinlan. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier (Wikipedia, 2014). A C4.5 tree tries to recursively partition the dataset into subsets by evaluating the normalized information gain (difference in entropy) resulting from choosing a descriptor for splitting the data. The descriptor with the highest information gain is used on every step. The training process stops when the resulting nodes contain instances of single classes or if no descriptor can be found that would result to the information gain (Machine Learning Methods, 2012). The method is classification-only and was chosen for use in this research project to classify and predict on a general insurance claim's nominal attributes.

- **Naïve Bayes**

The idea behind Naïve Bayes algorithm is the posterior probability of a data instance  $t_i$  in a class  $c_j$  of the data model. The posterior probability  $P(t_i|c_j)$  is the possibility of that  $t_i$  can be labeled  $c_j$ .  $P(t_i|c_j)$  can be calculated by multiplying all probabilities of all attributes of the data instance in the data model:

$$P(t_i | c_j) = \prod_{k=1}^p P(x_{ik} | c_j) \quad \text{Equation 1}$$

with  $p$  denoted as the number of attributes in each data instance. The posterior probability is calculated for all classes, and the class with the highest probability will be the instance's label.

The two algorithms are classification-only and were chosen for use in this research project to classify and predict on a general insurance claim's nominal attributes.

### 3.7.7.3. Software Development Methodology

Extreme Programming Agile methodology (XP) was selected for the prototype's implementation. It is an agile software development methodology that improves software quality and responsiveness to changing business requirements. It advocates for frequent (iterative) releases in short development cycles, which is intended to improve productivity and introduce checkpoints at which new requirements can be adopted. Its core principles that were important to this research are

1. Iterative and incremental development.
2. Unit testing
3. Simplicity and clarity in design, and
4. Intensive user involvement

For every new functionality (requirement) to be deployed, the implementation process – based on XP – was as shown in the figure below.

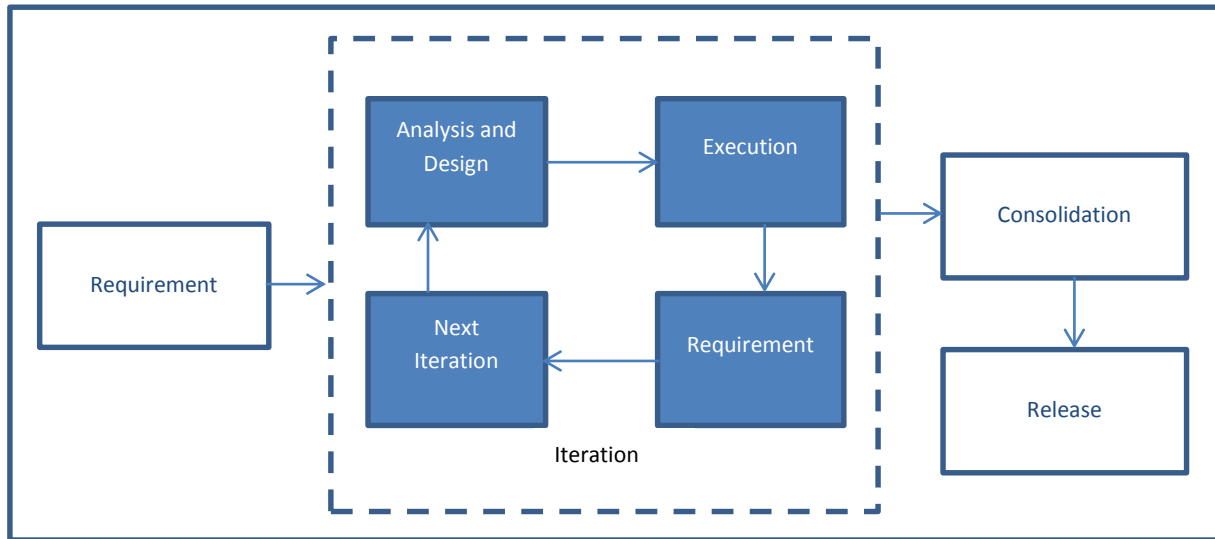


Figure 10: Extreme Programming Implementation Process

#### 3.7.7.4. Dataset Used

The following data set was used for training and testing and was applied for both J48 and Naïve Bayes algorithms in WEKA.

Training Set			Testing Set		
Claimant age	Count	%	Claimant age	Count	%
20-30	90	44.12	20-30	15	41.67
31-40	51	25	31-40	8	22.22
41-50	27	13.24	41-50	5	13.89
51-60	20	9.8	51-60	4	11.11
Above 60	16	7.84	Above 60	4	11.11

Table 3: Training Set Summary



### 3.7.7.5. Implementation Iterations

The following implementation iterations were adopted.

#### 1. Data Retrieval

This involved retrieval of claims data from flat files into a relational database. Data imported in this step is committed to a temporary data store (database table) and is also determined by a date range supplied by the user. This means that the history depth of analysis is left to the user's discretion.

#### 2. Data Organization And Cleansing

This involved removal of data rows unwanted in any analysis and any possible addition of external data to be considered for analysis.

#### 3. Data Modeling

This involved creation of a predictive model in WEKA using data from the relational database. The WEKA versions used were *weka.classifiers.j48.J48* for C4.5 and *weka.classifiers.NaiveBayes* for Naïve Bayes algorithm. A user supplies a date range that indicates the depth in history that the analysis should go, and also selects a list of attributes to be used in constructing the model. The model is built using either J48 or Naïve Bayes in WEKA and is trained on a given dataset and tested by using 10-split cross validation (i.e. using 10 cross validation folds). ). The modelling attributes selected were those that are typically available relatively early in the life of a claim. Another precondition for selecting the attributes was that an attribute would have at least ten data instances in our data set where the flag was set. A similar selection criteria was adopted by Stijn V. et al in a similar research (Stijn V. et al, 2002). One assumption made in the modelling process is that class distribution for an attribute is presumed to remain constant over time and relatively balanced. But even though classification based on accuracy alone may always predict the most prevalent class and thus yield very high performance (Stijn V. et al, 2002), such classifications always remain useful because they are indicative of a broader notion of good performance.

#### 1. Modeling Process

The modeling process is as follows

- a. **Data Selection:** A set of claim records are selected and submitted for use in the modeling process. The selection is made by submitting a date range, where all claims created within the supplied date range are selected for use in modeling.
- b. **Attribute Selection:** A modeling attribute is selected from a list as indicated in the modeling attributes listed in the table in the following section below.
- c. **Model Creation:** The selected claim records and the supplied attribute are used to create the model using either Naïve Bayes or J48 decision tree algorithm. The model is created, tested, and saved on disk. Details of the created model including the resulting statistics are saved and returned to the user for review. The statistics are described in the sections below.

#### 2. Modeling Output

- **Confusion Matrix**

This is a table layout that allows visualization of the performance of the J48 algorithm. It is also called a contingency table or an error matrix. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The matrix makes it easy to see if the prototype is confusing two classes (i.e. commonly mislabeling one as another). In this research project, the classification prototype trains to distinguish between various possible values of a selection of attributes. The attributes includes the following.

Attribute	Possible Values
1. IS_fraudulent_claim	{TRUE, FALSE}
2. IS_public_service_vehicle_claim	{TRUE, FALSE}
3. IS_caused_by_accident	{TRUE, FALSE}
4. HAS_costs_above_sum_assured	{TRUE, FALSE}
5. HAS_client_cover/policy_maintained	{TRUE, FALSE}
6. Plan (Product)	
7. Claim_Period	
8. Claimant_age	
9. Claim_amount	
10. Re-imburement_Cost	

Table 4: Modeling Attributes

For every model generated by the prototype, a confusion matrix is generated for further inspection. The matrix summarizes the results of the model generation process and testing using J48. A confusion matrix takes the format in the table below. Actual class refers to the true classification of an instance, while predicted class refers to a model’s classification of an instance. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. All correct guesses are located in the diagonal of the table, making it easy to visually inspect the table for errors, as they will be represented by values outside the diagonal. The matrix makes it easy to see if the prototype is confusing two classes (i.e. commonly mislabeling one as another). Performance measures that can be derived from a confusion matrix include sensitivity and specificity. They measure the proportion of +ve data instances that are predicted +ve and –ve data instances that are predicted –ve respectively (Stijn V. et al, 2002).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad \text{Equation 2}$$

$$\text{Specificity} = \frac{TN}{FP + TN} \quad \text{Equation 3}$$

In this research project, the classification prototype trains to distinguish between various possible values of a selection of attributes.

		Predicted Class	
		A	B
Actual Class	A	True Positives (items correctly classified as A)	False Negatives (items incorrectly classified as B)
	B	False Positives (items incorrectly classified as A)	True Negatives (items correctly classified as B)

Table 5: Confusion Matrix Structure

The tables below shows confusion matrices generated from models built from the prototype on attribute *IS\_fraudulent\_claim* using WEKA's J48 and Naïve Bayes classifiers. From these modelling results, the total number of correctly classified instances after training were 85 (50 + 35) by the model built using J48 algorithm, and 60 (32 + 28) by the model built using Naïve Bayes algorithm. The J48 model was therefore more accurate, although the Naïve Bayes model also gave a relatively good performance.

		Predicted Class	
		True	False
Actual Class	True	50	3
	False	12	35

Table 6: Confusion matrix using J48 algorithm

		Predicted Class	
		True	False
Actual Class	True	32	21
	False	19	28

Table 7: Confusion matrix using Naïve Bayes algorithm

▪ **Model Statistics**

These are statistics on the results of a modeling process. They include

- a) **Correctly Classified Instances** - This is expressed as a percentage. It shows the percentage of test instances that were correctly classified. The raw numbers are shown in the confusion matrix.
- b) **Incorrectly Classified Instances** - This is also expressed as a percentage. It shows the percentage of test instances that were incorrectly classified.

- c) **Kappa Statistic** - This is a chance-corrected measure of agreement between the classifications and the true classes. It is calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement. A value greater than 0 means that a model is doing better than chance, which was this research's objective. It is computed as follows

$$k = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)} \quad \text{Equation 4}$$

where  $\text{Pr}(a)$  is the relative observed agreement among raters, and  $\text{Pr}(e)$  is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then  $\kappa = 1$ . If there is no agreement among the raters other than what would be expected by chance (as defined by  $\text{Pr}(e)$ ),  $\kappa = 0$ .

The error rates below are provided by the model, but are used for numeric prediction rather than classification. In numeric prediction, predictions are not just right or wrong. Instead the error has a magnitude, and these measures reflect that. The research prototype, based on WEKA, computes these error measures by normalizing with respect to the performance obtained by predicting the classes' prior probabilities as estimated from the training data with a simple Laplace estimator. (This implies that a classifier like ZeroR for instance, always has a relative error of 100 %.) The error measures include

- d) **Mean Absolute Error** – This is a quantity used to measure how close forecasts or predictions are to the eventual outcomes (in this case the average error for testing cases). The mean absolute error is given by

$$MAE = \frac{\sum_{i=0}^n |p_i - a_i|}{n} \quad \text{Equation 5}$$

Where  $a_1 a_2 \dots a_n$  = actual target values

$p_1 p_2 \dots p_n$  = predicted target values

- e) **Root Mean Squared Error (RMSE)** – This is the sample standard deviation of the differences between predicted values and observed values. It measures the error rate of a regression model. It is given by

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (p_i - a_i)^2}{n}} \quad \text{Equation 6}$$

Where  $a_1 a_2 \dots a_n$  = actual target values

$p_1 p_2 \dots p_n$  = predicted target values

- f) **Root Relative Squared Error (Root RSE)** - Unlike RMSE, the relative squared error (RSE) can be compared between models whose errors are measured in the different units.

$$RSE = \frac{\sum_{i=1}^n (p_i - a_i)^2}{\sum_{i=1}^n (\bar{a} - a_i)^2} \quad \text{Equation 7}$$

Where  $a_1 a_2 \dots a_n$  = actual target values

$p_1 p_2 \dots p_n$  = predicted target values

- g) **Relative Absolute Error** – Like Root RSE, the relative absolute error (RAE) can be compared between models whose errors are measured in the different units.

$$RAE = \frac{\sum_{i=1}^n |p_i - a_i|}{\sum_{i=1}^n |\bar{a} - a_i|} \quad \text{Equation 8}$$

Where  $a_1 a_2 \dots a_n$  = actual target values

$p_1 p_2 \dots p_n$  = predicted target values

- h) **Total Number Of Instances** – This is the number of cases used in building and testing the model.

- **Decision Tree**

This is a visualization of the model from the prototype in tree form.

#### 4. Prediction And Reporting

This involved use of the model created above to perform predictive analysis using data imported into the relational database.

#### 3. Prediction Process

The modeling process is as follows:

- Data Selection:** A set of claim records are selected and submitted for prediction on a given attribute. The selection is made by submitting a date range, where all claims created within the supplied date range are selected for the prediction process.
- Attribute Selection:** The prediction attribute is selected from a list as indicated in the modeling attributes listed in table 3 above.
- Model Selection:** A previously created prediction model is selected from a list.

- d. **Model Creation:** The selected claim records and the supplied attribute are subjected to the model. The prediction process then runs and the output of the prediction returned to the user for review.

#### 4. Analysis Output

The output of the prediction process is as follows:

- **Predicted Attribute Values** – These are the predicted values of the prediction attribute for every claim subjected to the prediction process. E.g. TRUE or FALSE for the *IS\_fraudulent\_claim* attribute.
- **Predicted Probability** – This is the estimated membership probabilities of an instance in each class. It represents posterior probabilities that give some sort of confidence for the predicted class of a claim for a given attribute. For example \*0.70: 0.30 for a predicted value of TRUE on attribute *IS\_fraudulent\_claim* on a claim instance means that there is a 70% chance that the claim is truly fraudulent and a 30% chance that the claim is not fraudulent.

The figure below shows the prototype processes in all the implementation iterations described above.

XP ITERATIONS (IT)

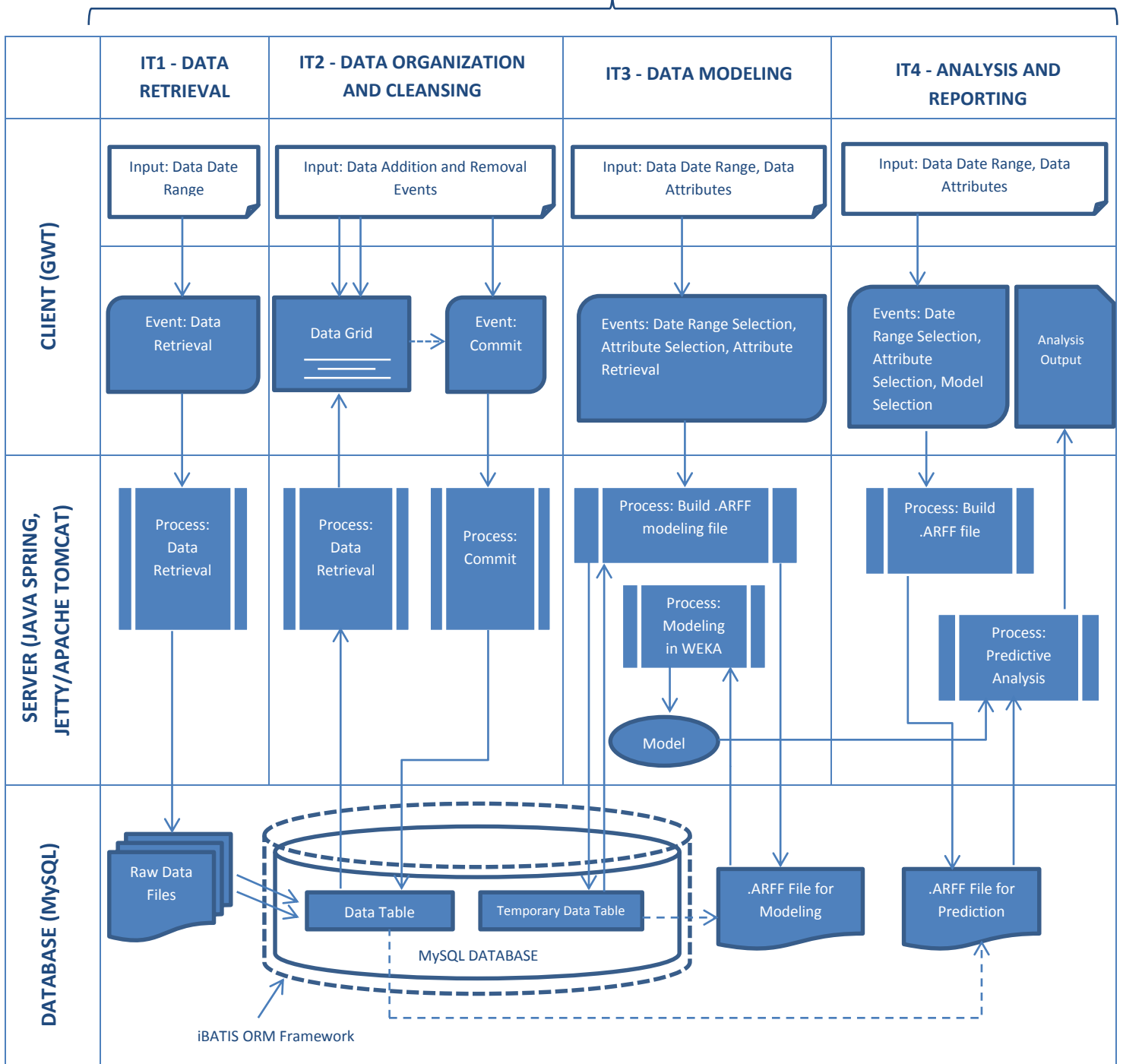


Figure 11: Process Flow and Implementation Iterations

### 3.7.7.6. Architectural Solution Design

The architectural design described in the to-be analysis was adopted. The high-level design details shown in the figure below were adopted for the solution’s implementation.

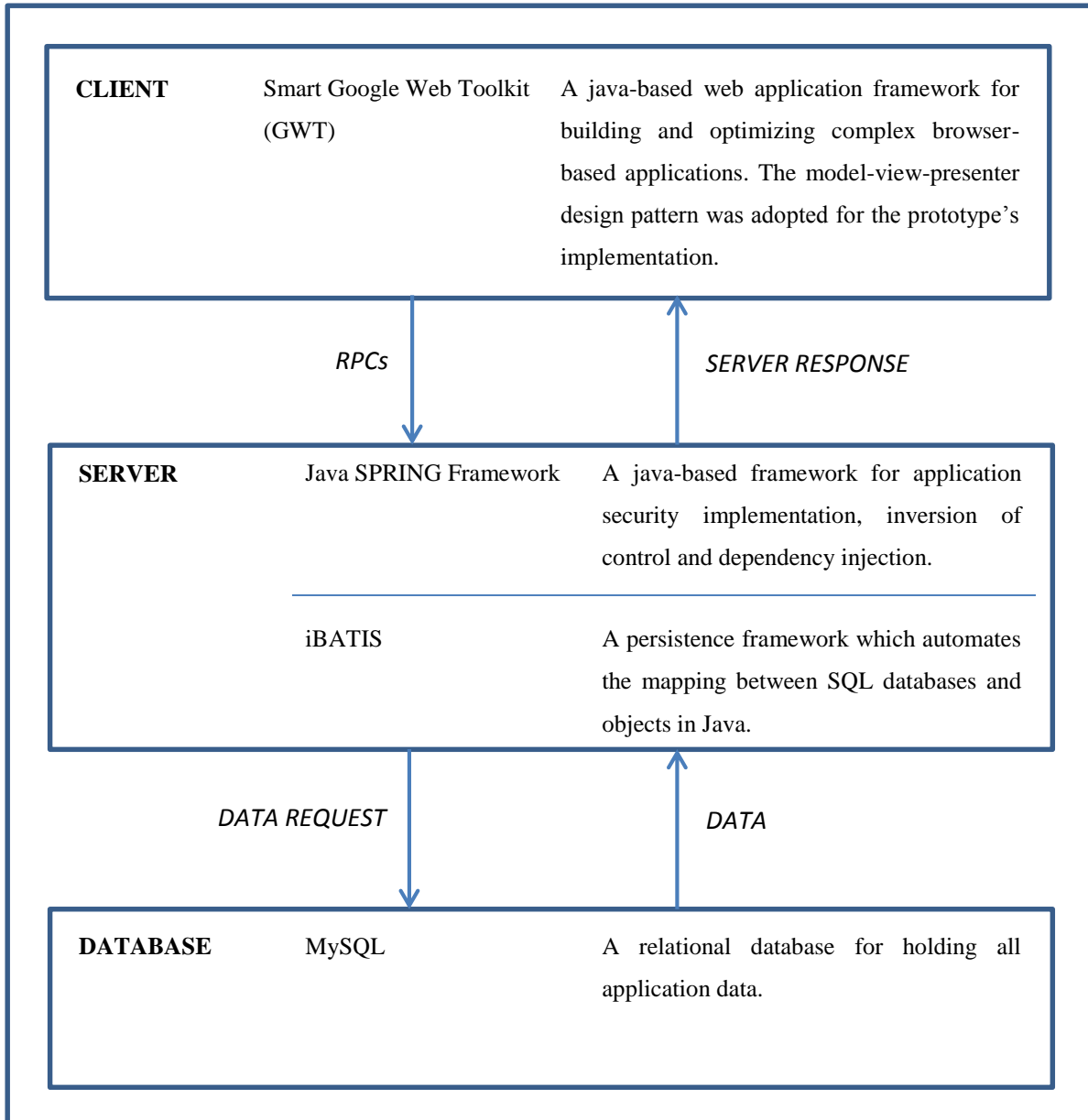


Figure 12: High Level Architectural Design Details

### 3.7.8. Testing

1. Testing of the prototype was performed after every implementation iteration. User involvement was key to the tests. The tests included unit tests, system tests and user acceptance tests.



## 4. RESULTS AND ANALYSIS

This chapter outlines the results of the mini-survey and the predictive analysis process as performed by the prototype.

### 4.1. Survey Results

#### 5. Responses Received

From the population of 42 as described under survey setup section, 21 survey responses were received. The responses to the survey are as follows.

##### 1. Organisation

As shown below, 62% of the respondents preferred not to reveal who their employers were.

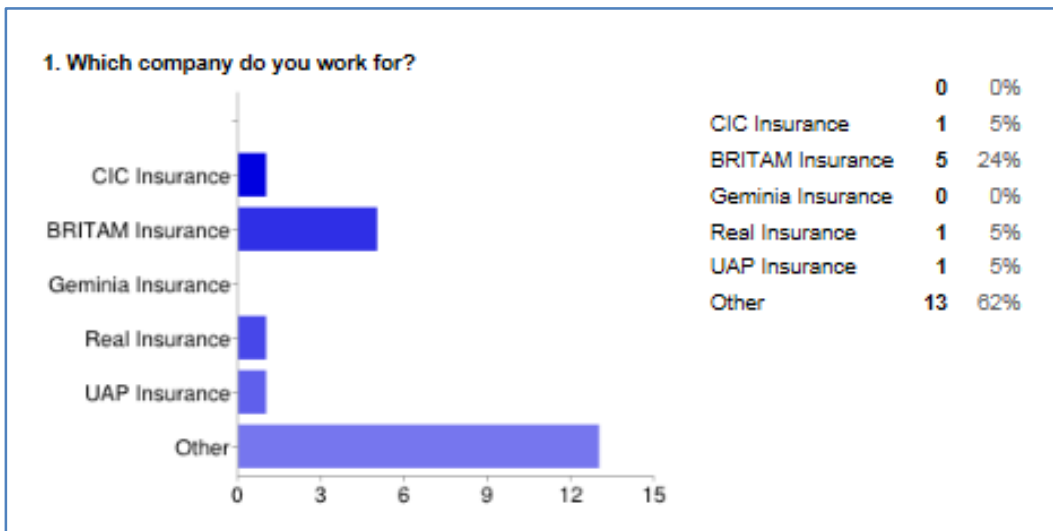


Figure 13: Survey respondents' organizations.

##### 2. Respondents' Roles

As shown below, majority of the respondents preferred not to reveal their roles in their respective organizations. A good percentage of them were claims administrators and IT system administrators for general insurance systems.

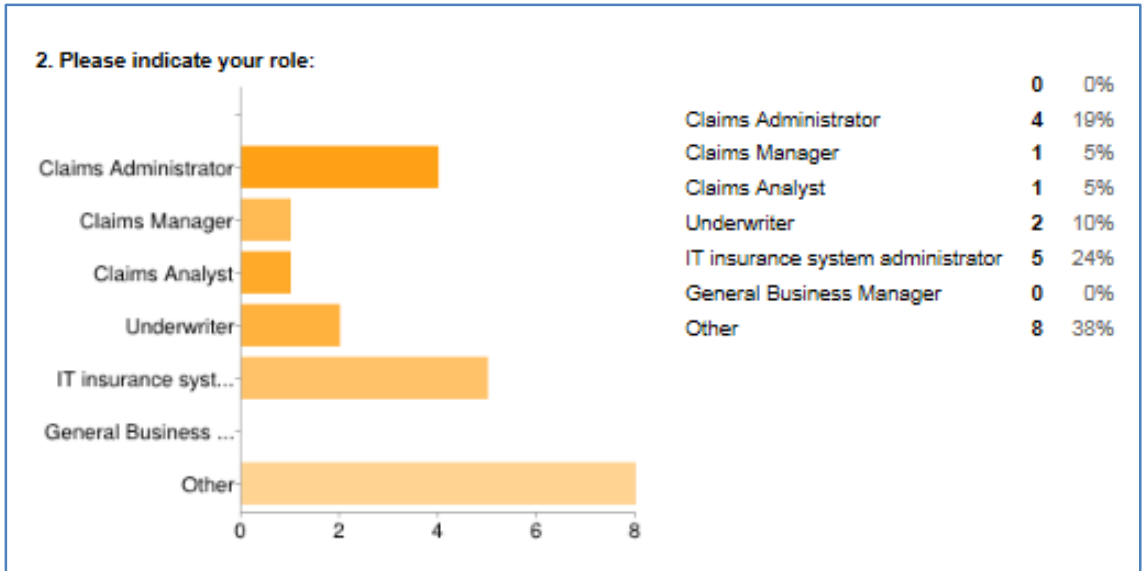


Figure 14: Survey respondents' roles

### 3. Analytics and BI Tools

On the question of tools used by the respondents in their organization for analytics and BI, the following responses were received.

- Crystal Reports
- System Reports (Line of Business reports)
- Microsoft Excel
- Oracle BI
- AIMS
- Oracle OBIEE
- MS Access
- BI360
- Sales Executive
- Media
- E - commerce
- Spreadsheets
- Sirius

From the responses above, spreadsheets and Microsoft Excel (which is also essentially a spreadsheet) had a combined frequency of 11 in total. This was the highest count over all the other tools in the responses, making spreadsheets the most common tools currently used in analytics and BI. Oracle BI and Crystal reports were the second most common having a frequency of 2 each. It was also clear that line of business

system reports were also used in analytics and BI with some respondents indicating that they are used in conjunction with spreadsheets.

The responses also show that the media is relied on for pre-compiled analytics data. This implies a lot of things as the media or research organizations that announce their results through the media might be having more advanced tools that enable them to produce statistics from the insurance industry which is readily consumed. All the responses show that there is still a lot of opportunity for implementation of advanced analytics and BI in the industry.

#### 4. Target Audience

On the question of who the target audience for analytics and BI is, majority of the respondents voted for management as the target audience. This shows that there is a need for high accuracy in the analytics and BI reports as they are likely to be used in decision making.

Top level management had the highest votes while claims analysts and claims administrators were also voted for as suitable audiences that can benefit from analytics and BI. This shows that analytics and BI is increasingly being used by non-managerial staff for their day to day work.

The responses also agree with the fourth objective of this research which was to make analytics and BI applicable at non-managerial levels in an insurance organization.

The chart below shows a summary of the responses.

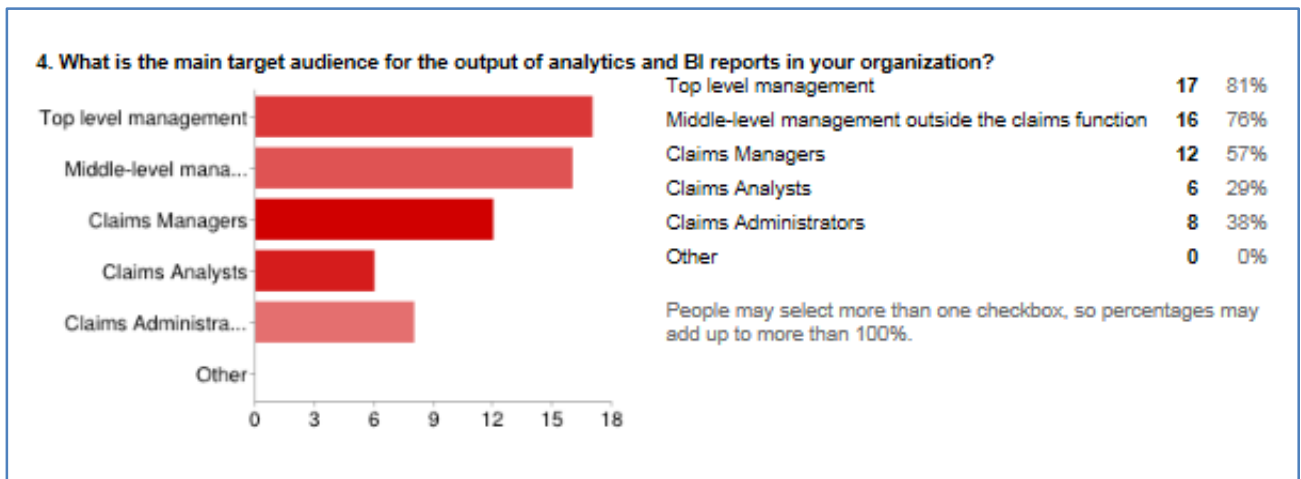


Figure 15: Survey responses on analytics and BI target audience.

#### 5. Level of urgency

On the question of how often analytical reports are normally required, the responses shown in the chart below were obtained.

A monthly frequency had the highest vote of 43percent but notably an on-demand frequency had the second highest votes with 24 percent. This implies that there is a need for tools that can generate analytics reports on demand. However there might be some level of tolerance for tools that are able to generate reports on a monthly basis.

The chart below shows a summary of the responses.

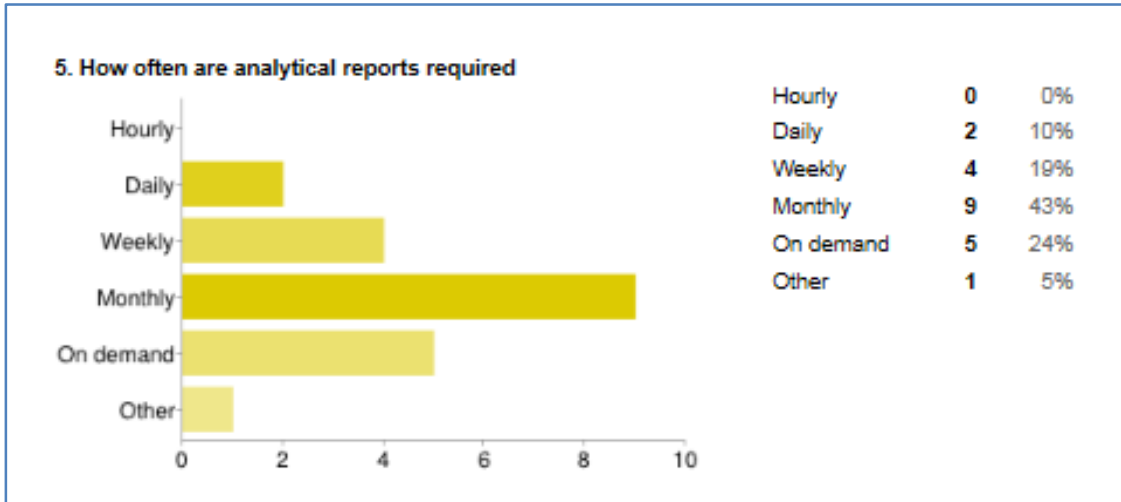


Figure 16: Survey responses on the frequency of need for analytics and BI reports.

### 6. Level of urgency

On the question of level of urgency, majority of the respondents voted for level 3 out of 5 with 5 being very urgent. However a combined percentage of the respondents (53 percent) indicated that the reports can be rated at levels 4 and 5 of urgency. This shows that there is a need for tools that can generate reports reliably and on-demand. This also shows that analytics and BI reports are most likely critical for day-to-day activities.

The chart below shows a summary of the responses.

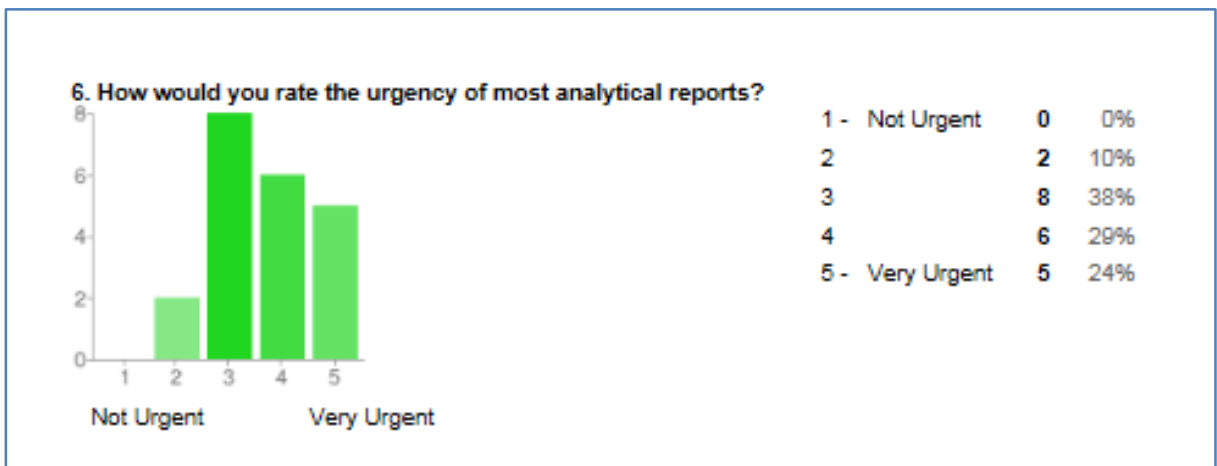


Figure 17: Survey responses on the level of urgency of analytics and BI reports.

### 7. Reporting history

On the question of how far in history analytical reports should go, 33 percent of the respondents indicated that they should be since the business began, and an equal 33 percent indicated that they should be as from the last 1 year.

From these statistics, there seems to be a consensus that the reports should preferably go far into history, but a divided opinion on exactly how far into history they should go. This might be dependent on a particular analysis being done. For this reason, the built prototype that accompanies this research leaves the decision on how far into history modeling and prediction should go to the user operating the system, by providing a date picker to select a start date.

The chart below shows a summary of the responses.

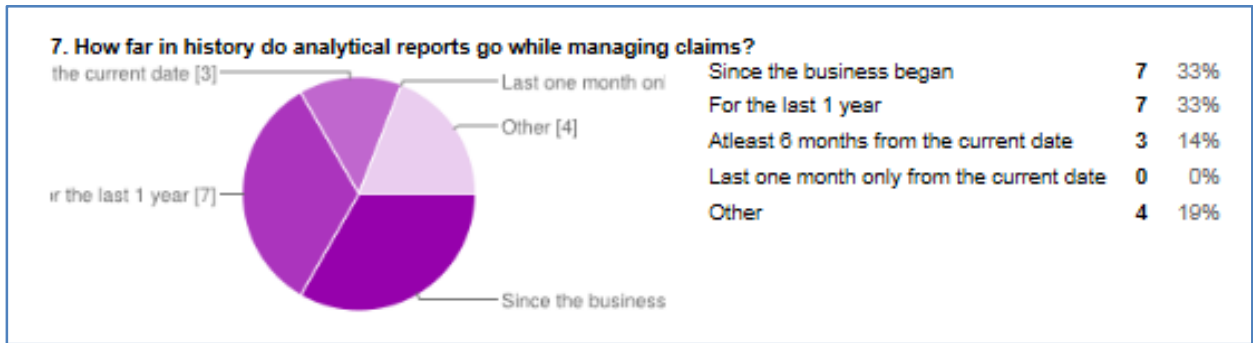


Figure 18: Survey responses on the reporting history for analytics and BI reports.

8. On the question of essential analytical information in claims that cannot be generated from currently available tools, the following responses were obtained.
  - Future performance projections
  - Possible future fraud count. Can be achieved but not efficiently.
  - Trends. This can be done but not easily achieved.
  - Possible payment failures
  - Agency future performance
  - Industry statistics and performance comparisons
  - Projected losses
  - Accurate loss ratios
  - Accuracy in reports due to data corruption.
  - Graphical dashboards
  - Claims triangulation: A table which charts that show the movement of total incurred losses from the original policy period, over several subsequent periods.
  - Reserve movements
  - Claims classification per motor body type
  - Customer claims experiences and loss ratios
  - Exposure information split by region
  - Rating factors
  - A one stop shop for information generation on customer reports.
  - None

From the responses above, it is clear that there is need to provide tools that can sufficiently furnish analytics needs. Also clear from the responses is the fact that most respondents are only familiar with reactive analytics and might need exposure to predictive analytics which might give them even further insights from their data. The key predictive analytics requirement which current tools are not able to generate is future performance projections from claims. The responses also show that the industry still has a long way to go when it comes to implementing analytics and BI, especially predictive analytics.

Other responses received that were not necessarily related to analytics and BI were

- I do not know.
- Determining claims movements.
- Determining client payment histories.
- Determining claims incidences.
- Determining the origin of claims.
- Locating of claims.
- Determining types of claims
- Determining clients to whom claims belong
- Servicing of providers' payment details
- Difficulty in linking a claim to the specific dependent for group covers.
- Treatment of invoices as claims thus rendering analysis of claim incidence rates and average amounts per claim erroneous.

These responses show that some of the respondents did not have a good understanding of what analytics and BI is and hence the misplaced answers. It can therefore be concluded that user education needs to be conducted in the industry to ensure maximum gain from the benefits of analytics and BI.

## **9. Value of Analytics and Business Intelligence**

On the question of how much is spent on average on analytics and BI, majority of the respondents indicated that they did not know. This could indicate that the respondents are not involved in planning for analytics and BI in their organizations. It is therefore difficult to conclusively determine how much investment is made on analytics and BI by an average insurance company in Kenya. However 19% of them agree that this figure is between 1 to 10 million.

The chart below shows a summary of the responses.

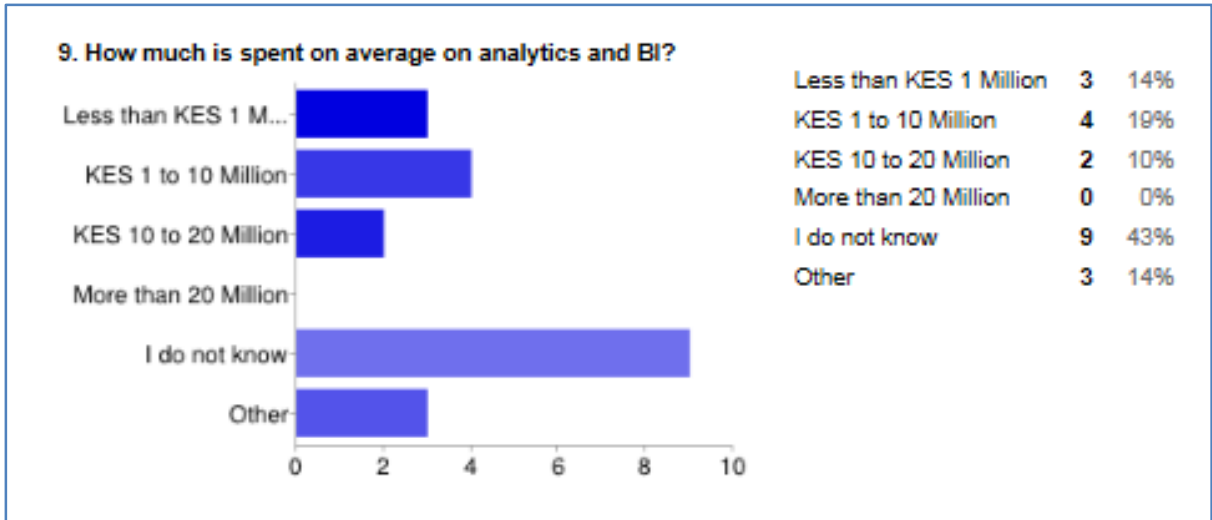


Figure 19: Survey responses on how much is spent on analytics and BI investments.

### 10. Value of Analytics and Business Intelligence

On the question of how analytics and BI reports are used by organizations, 52 percent of the respondents indicated that they are used for both product development and crafting strategies. 38 percent of them indicated that it is used for fraud detection. There was also an indication that analytics and BI is useful in other functions apart from the ones specified in the research as shown by the 19 percent figure below. The responses indicate that analytics and BI play a crucial role in the running and management of general insurance businesses.

The chart below shows a summary of the responses.



Figure 20: Survey responses on the use of analytics and BI reports.

11. On the question of how Analytics and Business Intelligence reports are used in the respondents' organizations, the following responses were obtained.

- To get fraudulent claims and using the fraudulent ones to check against new claims for similarity. This is done by IT support analysts.

- To show trends in performance.
- To service IRA reporting requirements on claims.
- To identify patterns in claims that can be of use in defining future strategies of profitability to the business.
- To perform market analysis that can lead to new product developments.
- To gain on the overall industry experience and rating against other insurers.
- To obtain projections on future product performances.
- To advise on options during reinsurance negotiations.
- To advise on capital conservation.
- They could be useful for previewing a business' growth.
- To provide mission dashboards where all analysis from different sections/departments are summarized and grouped together.
- To facilitate claims expense management.
- To assists in policy underwriting and product pricing.
- N/A. Meaning that analytics and BI was not applicable to the organization.

From the responses above, with the exception of the last response, it is clear that organizations reap a lot of benefits from analytics and BI and almost depend on it for their growth and survival. It is also clear that analytics and BI is essential to obtaining an aerial view of a business, making it easier for its owners to monitor it for failure or success.

### **Industry Regulation**

12. On the question of industry rules or regulations that Analytics and Business Intelligence functions should conform to, the responses obtained can be summarized as follows:

- Authorization and regulation by AKI regulatory board. This response had a frequency of 12 from amongst all the responses. Details of the authorization and/or regulation were however not provided.
- Unknown. Two of the respondents said they did not know of any regulations
- None. Three of the respondents said there was no industry regulation on analytics and BI.
- IRA and RBA regulations. This was provided by four of the respondents. Details of the regulations were however not provided.

Other responses that were not clear were

- i. Risk based management
- ii. Finance bill
- iii. Kenya Revenue Authority

From the responses above, it was clear that there was no outright stipulation of rules that govern analytics and BI in the industry, and if there are any at all, then the people in the industry are not



well aware of them. What was clear however was that if there was ever to be a regulation on the same it would come from AKI.

## 4.2. Prototype Results

### 6. Modeling

The prototype produced models to be used in prediction. Statistics from the modeling process were also produced and persisted alongside the model details and the model itself. The results included modeling statistics and a confusion matrix for both J48 and Naïve Bayes algorithms, and a tree visualization for J48 algorithm. Figure 22 below shows a sample server response from a modeling process. From the results, the model is shown to have an accuracy of 85 percent.

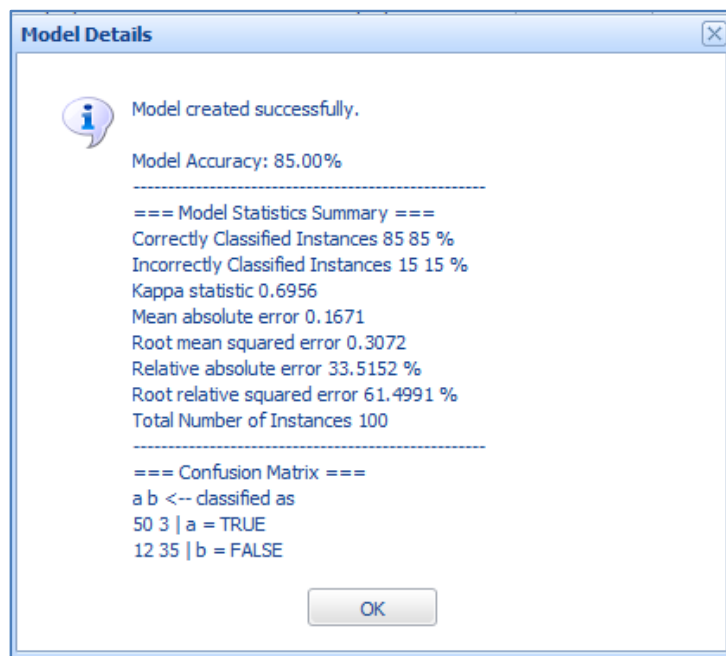


Figure 21: Sample modeling results on attribute *IS\_fraudulent\_claim* using J48.

#### ▪ J48 tree visualization

A visualization of a model created from the prototype is as shown in figure 23 below. Each path from the root node to a leaf node represents a rule in the model that leads to the classification in the leaf node. Alongside the classification on the leaf node, there are two numbers. The first number refers to the number of instances covered by the leaf, while the second number (after the slash) refers to the number of instances misclassified by the rule corresponding to the leaf. Misclassification mostly occurs as a result of pruning the tree, as achieving simplicity in the tree means a compromise on accuracy.

#### ▪ Accuracy

The accuracy is the proportion of the total number of predictions that were correct. The accuracy of models built using both algorithms is high as shown in table 6 below. C4.5 (J48) is however better than Naïve Bayes.

Criteria	J48	Naïve Bayes
Correct Classification	85	60
Incorrect Classification	15	40
Time to build model (seconds)	5.31	0.42
Accuracy (%)	85	60
Sensitivity	0.9434	0.6038
Specificity	0.7447	0.5957

Table 8: Model Results (based on models that resulted in the matrices on Tables 6 and 7).

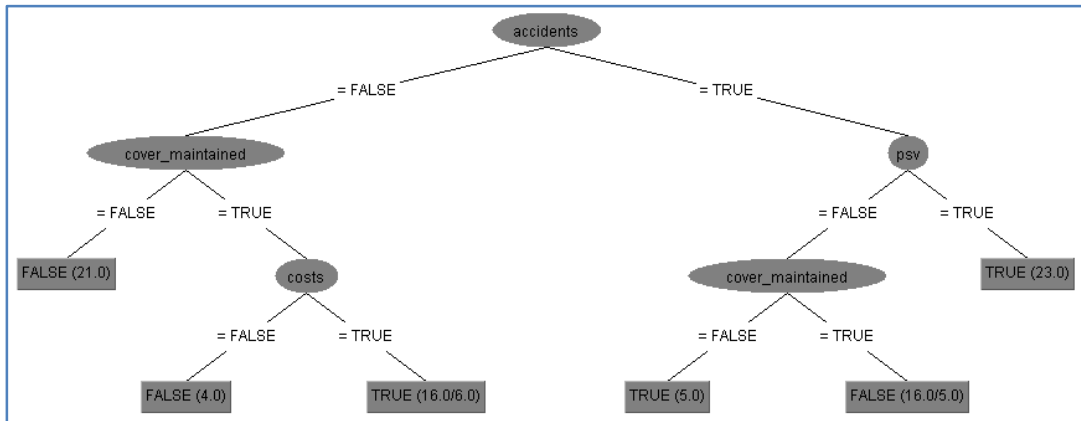


Figure 22: A model's tree visualization from the prototype.

## 7. Prediction

The prototype produced predicted attribute values for various claims records. The figure below shows sample results of a prediction run on a model built using Naïve Bayes algorithm. For the third claim in the figure, the result is a predicted probability of \*0.69 : 0.31 for a predicted value of TRUE on attribute *IS\_fraudulent\_claim*. This means that there is a 69 percent chance that the claim is truly fraudulent and a 31 percent chance that the claim is not fraudulent. This provides a starting point to any claims administrator to further investigate the claim for fraud.

Fraudulent?	From Accident?	PSV?	> Sum Assured	Cover Maintained?	Predicted Probability
TRUE	FALSE	FALSE	TRUE	FALSE	*1.0 : 0.0 :
TRUE	FALSE	TRUE	TRUE	TRUE	*1.0 : 0.0 :
TRUE	TRUE	FALSE	FALSE	TRUE	*0.69 : 0.31 :
TRUE	TRUE	FALSE	TRUE	TRUE	*0.69 : 0.31 :

Figure 23: Sample results of a prediction run on attribute *IS\_fraudulent\_claim*.

The figure below shows a chart of prediction results.

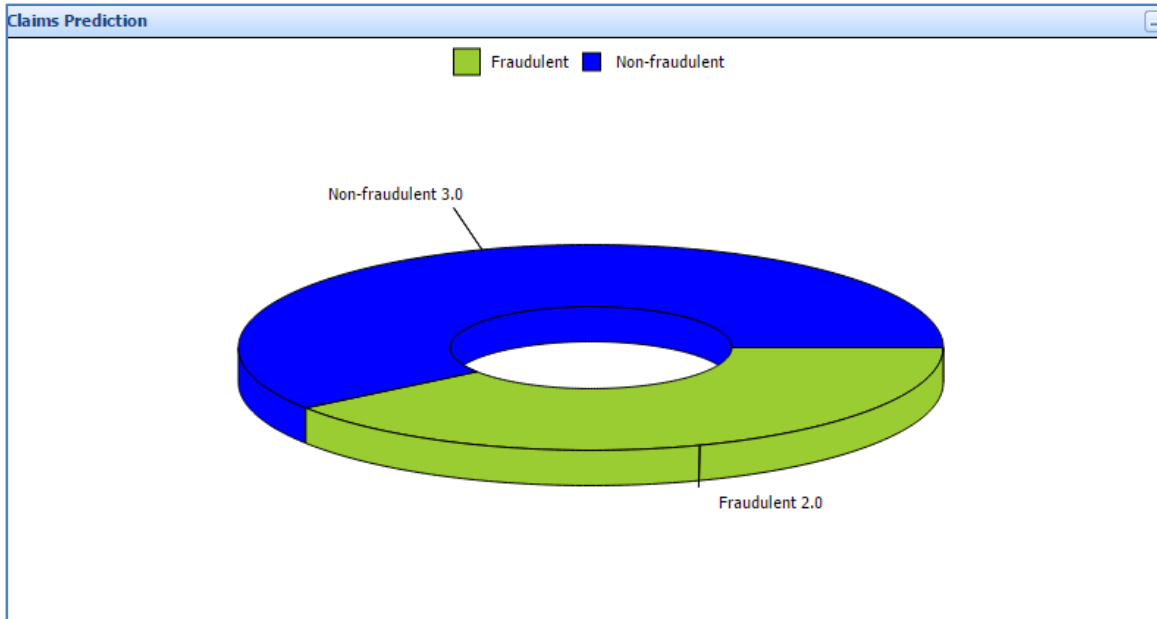


Figure 24: Sample results of a prediction run on attribute *IS\_fraudulent\_claim* in a chart.

Other prediction results from the prototype are as illustrated in the figures below.

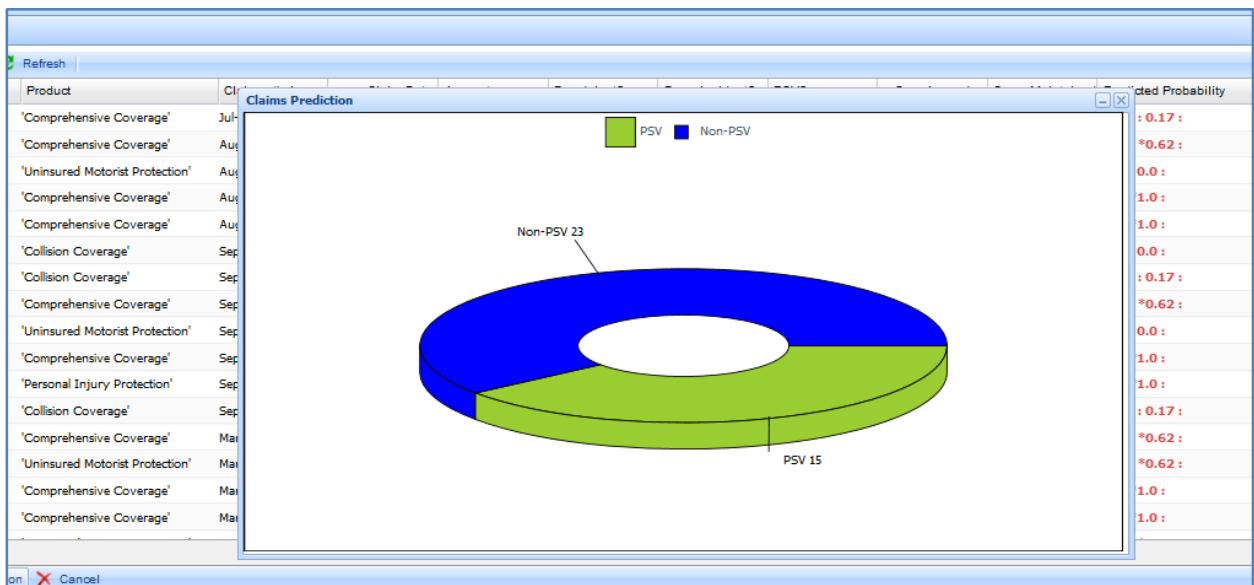


Figure 25: Results of a prediction run on attribute *IS\_PSV* in a chart.

Refresh									
Product	Claimant's Age	Claim Date	Amount	Fraudulent?	From Accident?	PSV?	> Sum Assured	Cover Maintained	Predicted Probability
'Comprehensive Coverage'	Jul-14	18/07/2014	47719	TRUE	TRUE	TRUE	FALSE	TRUE	*0.83 : 0.17 :
'Comprehensive Coverage'	Aug-14	19/08/2014	41039	FALSE	FALSE	FALSE	TRUE	FALSE	0.38 : *0.62 :
'Uninsured Motorist Protection'	Aug-14	22/08/2014	4668	TRUE	FALSE	TRUE	TRUE	TRUE	*1.0 : 0.0 :
'Comprehensive Coverage'	Aug-14	23/08/2014	32499	FALSE	TRUE	FALSE	FALSE	TRUE	0.0 : *1.0 :
'Comprehensive Coverage'	Aug-14	24/08/2014	11550	FALSE	TRUE	FALSE	TRUE	TRUE	0.0 : *1.0 :
'Collision Coverage'	Sep-14	01/09/2014	5233	FALSE	FALSE	TRUE	TRUE	TRUE	*1.0 : 0.0 :
'Collision Coverage'	Sep-14	01/09/2014	10217	TRUE	TRUE	TRUE	FALSE	TRUE	*0.83 : 0.17 :
'Comprehensive Coverage'	Sep-14	07/09/2014	24728	FALSE	FALSE	FALSE	TRUE	FALSE	0.38 : *0.62 :
'Uninsured Motorist Protection'	Sep-14	05/09/2014	34995	TRUE	FALSE	TRUE	TRUE	TRUE	*1.0 : 0.0 :
'Comprehensive Coverage'	Sep-14	06/09/2014	41916	FALSE	TRUE	FALSE	FALSE	TRUE	0.0 : *1.0 :
'Personal Injury Protection'	Sep-14	07/09/2014	22148	FALSE	TRUE	FALSE	TRUE	TRUE	0.0 : *1.0 :
'Collision Coverage'	Sep-14	07/09/2014	22781	TRUE	TRUE	TRUE	TRUE	TRUE	*0.83 : 0.17 :
'Comprehensive Coverage'	Mar-14	19/03/2014	15960	FALSE	FALSE	FALSE	TRUE	FALSE	0.38 : *0.62 :
'Uninsured Motorist Protection'	Mar-14	22/03/2014	21405	TRUE	FALSE	FALSE	TRUE	FALSE	0.38 : *0.62 :
'Comprehensive Coverage'	Mar-14	23/03/2014	48533	FALSE	TRUE	FALSE	FALSE	TRUE	0.0 : *1.0 :
'Comprehensive Coverage'	Mar-14	24/03/2014	26642	FALSE	TRUE	FALSE	TRUE	TRUE	0.0 : *1.0 :

ion

00024 Prediction Attribute\* PSV Claims From\*

03/06/2015 Created On Fri Mar 06 08:50:52 GMT+:

Figure 26: Results of a prediction run on attribute *IS\_PSV*.

Product	Claimant's Age	Claim Date	Amount	Fraudulent?	From Accident?	PSV?	> Sum Assured	Cover Maintained	Predicted Probability
'Comprehensive Coverage'	Aug-14	19/08/2014	41039	FALSE	FALSE	FALSE	TRUE	TRUE	*0.89 : 0.11 :
'Uninsured Motorist Protection'	Aug-14	22/08/2014	4668	TRUE	FALSE	TRUE	TRUE	TRUE	*1.0 : 0.0 :
'Comprehensive Coverage'	Aug-14	23/08/2014	32499	FALSE	TRUE	FALSE	FALSE	TRUE	*0.89 : 0.11 :
'Comprehensive Coverage'	Aug-14	24/08/2014	11550	FALSE	TRUE	FALSE	TRUE	TRUE	*0.89 : 0.11 :
'Collision Coverage'	Sep-14	01/09/2014	5233	FALSE	FALSE	TRUE	TRUE	TRUE	*0.89 : 0.11 :
'Collision Coverage'	Sep-14	01/09/2014	10217	TRUE	TRUE	TRUE	FALSE	TRUE	*1.0 : 0.0 :
'Comprehensive Coverage'	Sep-14	07/09/2014	24728	FALSE	FALSE	FALSE	TRUE	TRUE	*0.89 : 0.11 :
'Uninsured Motorist Protection'	Sep-14	05/09/2014	34995	TRUE	FALSE	TRUE	TRUE	TRUE	*1.0 : 0.0 :
'Comprehensive Coverage'	Sep-14	06/09/2014	41916	FALSE	TRUE	FALSE	FALSE	TRUE	*0.89 : 0.11 :
'Personal Injury Protection'	Sep-14	07/09/2014	22148	FALSE	TRUE	FALSE	TRUE	TRUE	*0.89 : 0.11 :
'Collision Coverage'	Sep-14	07/09/2014	22781	TRUE	TRUE	TRUE	TRUE	TRUE	*1.0 : 0.0 :

Figure 27: Results of a prediction run on attribute *Cover\_Maintained*.

#### 4.3. Research Evaluation

The following evaluation points were given on the prototype after testing and reviewing it.

- The research and its prototype could go a long way in assisting in claims administration and management.
- The prototype is simple and easy to use.
- A help menu should be available to explain the use of the prototype and the results that it gives.
- The prototype should be customized to use data specific to every company.
- The prototype should include a drag-drop report designer functionality.
- The prototype should be extended to show market dynamics from the claims data on charts and graphs.
- The prototype should be extended to handle medical insurance, and life insurance, and provide aggregated results for the three business categories.
- The prototype should show trends in performance.
- The research and the prototype should have been extended to potentially include sales, marketing and underwriting data in the modeling and prediction processes. This is because they play a big part in general insurance claims predictions.

The above evaluation points were reviewed. However, most of the items in the list were out of the scope of this research project. Also, some of the evaluation points were not feasible due the absence of real data and information that was not readily available (from the proposed research participants) to build a prototype that closely implements the users expectations.

The current implementation however provided a proof of concept on the possibility of predictive analytics on insurance claims.

#### **4.4. Challenges and Limitations**

The following challenges were faced in the research project:

1. Obtaining participation consent – Most companies were unwilling to provide data to be used in the research due to privacy concerns and company regulations. This made development of the prototype difficult as matching local scenarios was not easily achieved without real data (see appendix 3).
2. Data complexity – Use of data obtained online provided a great challenge as it included a lot of information including irrelevant columns, thus requiring time to cleanup and contextualize the data to the local scenarios.
3. Time - the time available to conduct this research was constrained by its due date. More time would have been needed to expand the prototype with additional functionality that would demonstrate the power of analytics and BI, allowing for the inclusion of some of the requirements noted in the survey.
4. Getting participants to the survey was a challenge due to company restrictions on employees to participate in such activities. . A guarantee had to be given to the participants that their identities would not be exposed for them to agree to participate. Access to the participants who hold managerial positions and getting them to participate was also a major challenge. Most of them referred the questions to their juniors and therefore their opinions could not be obtained.
5. Due to the above challenge, this researched was biased towards the opinion of non-managerial staff within the domain of general insurance proper.

#### **4.5. Recommendations and future work**

The following recommendations were made on the research:

1. The prototype can be enhanced to include a custom implementation of a data import functionality. This data import functionality will be dependent on the existing insurance system of a user/client and their claims data structure.
2. The prototype could be extended to predict on numeric attributes.
3. The prototype can be customized and extended appropriately to accommodate the evaluation remarks, and adopted for commercial use. It could also be extended to potentially include sales, marketing and underwriting data in the modeling and prediction processes.
4. This research could be extended to cover all of general insurance (including medical insurance).
5. A statistical regression analysis algorithm could be considered in extending the prototype to enable it to predict on numeric data attributes. This could be done by utilizing R libraries in the prototype. R is a free software environment for statistical computing and graphics, and is widely used for data analysis.

#### **4.6. Conclusion**

In today's economic climate where budget reductions are common, company executives are under continuous pressure to deliver profitable growth. They must therefore be able to identify and implement critical items that will facilitate growth and enable their companies to remain competitive. Predictive analytics and BI is one of those items. This work represents a start on the process of implementing predictive analytics and BI for determining the optimal strategy for investigating and managing claims.

## REFERENCES

- Accenture, 2012, Reaping the benefits of Analytics, viewed 18 March 2014, from <http://www.slideshare.net/AccentureInsurance/insurance-six-ways-to-improve-business-intelligence>
- Agile Process, Extreme Programming, Viewed 13-Sept-2014, from <http://www.extremeprogramming.org/>
- Akshay B., 2012, Enhancement in Predictive Model for Insurance Underwriting, International Journal of Computer Science & Engineering Technology.
- Ana-Ramona B. et al, 2013, Big Data and Specific Analysis Methods for, Insurance Fraud Detection, Database Systems Journal.
- Artis, M, Ayuso M., Quillen M., 1997, Modeling Different Types of Automobile Insurance Fraud Behavior in the Spanish Market, Paper presented at the First International Conference on Insurance: Mathematics and Economics.
- Attribute-Relation File Format (ARFF), University of WAIKATO, Viewed 10-Oct-2014, from <http://www.cs.waikato.ac.nz/ml/weka/arff.html>
- Belhadji, E. B., Dionne G., 1997, Development of an Expert System for the Automatic Detection of Automobile Insurance Fraud, Working Paper 97-06, Ecole des HEC.
- BRITAM, Viewed 12 July 2014, <http://www.britam.co.ke/site/index.php/who-we-are/2013-11-22-06-17-60/press-release>, BRITAM Limited website
- C4.5 Algorithm, Wikipedia, Viewed 08-Oct-2014, from [http://en.wikipedia.org/wiki/C4.5\\_algorithm](http://en.wikipedia.org/wiki/C4.5_algorithm)
- Charles N., 2007, Predictive Analytics, White Paper, Senior Director of Knowledge Resources, American Institute for CPCU/Insurance Institute of America
- Cohen's Kappa, Wikipedia, Viewed 10-Oct-2014, from [http://en.wikipedia.org/wiki/Cohen%27s\\_kappa](http://en.wikipedia.org/wiki/Cohen%27s_kappa)
- Confusion Matrix, Wikipedia, Viewed 09-Oct-2014, from [http://en.wikipedia.org/wiki/Confusion\\_matrix](http://en.wikipedia.org/wiki/Confusion_matrix)
- Crain Communications, 2012, The Toughest Workers Comp Claims: How to identify and resolve potentially catastrophic claims, White Paper
- Dan V., Henry.D. , 2011, The Business Value of Predictive Analytics, White Paper, IBM SPSS
- Derrig, R. A., Weisberg I, 1996, Coping with the Influx of Suspicious Strain and Sprain Claims, AIB 1997 Cost Containment and Fraudulent Claims Payments Filing. D.O.I., Docket R96-37.



Derrig, Weisberg I., Chen X. et al , 1994, Behavioral Factors and Lotteries under No-Fault with a Monetary Threshold: A Study of Massachusetts Automobile Claims, Journal of Risk and Insurance 61(2): 245-275.

George B., Michael K., Abraham P., 2005, Insurance Industry Decision Support: Data Marts, OLAP and Predictive Analytics, Casualty Actuarial Society Forum.

IBM, 2011, Three Ways to Improve Claims Management with Business Analytics, viewed 18 March 2014, from <http://www.influentialsoftware.com/assets/uploads/IBM%20Insurance%20-%20claims%20management%20with%20analytics%20with%20Influential.pdf>.

J48, Wikipedia, Viewed 08-Oct-2014, from <http://en.wikipedia.org/wiki/J48>

Keith P., 2012, Claims Analytics for Auto Casualty Insurers, White Paper, Vice President Advanced Analytics and Consulting at Mitchell Auto Casualty Solutions.

Linda Jamii, Viewed 12 July 2014, from <http://lindajamii.co.ke/>, Linda Jamii website.

Manish M., 2013, Decision Tree Analysis on J48 Algorithm for Data Mining, International Journal of Advanced Research in Computer Science and Software Engineering, Research Paper.

Matt F., 2011, Predicting Liability Insurance Claim Payments with Vehicle Data, <http://blog.nutonian.com/bid/348152/Predicting-Liability-Insurance-Claim-Payments-with-Vehicle-Data>

McLaughlin, 2014, Agile Methodologies for Software Development, viewed 04 Jul 2014, from <http://www.versionone.com/Agile101/Agile-Development-Methodologies-Scrum-Kanban-Lean-XP/>

Mean Absolute Error, Wikipedia, Viewed 10-Oct-2014, from [http://en.wikipedia.org/wiki/Mean\\_absolute\\_error](http://en.wikipedia.org/wiki/Mean_absolute_error)

MicroStrategy, 2012, Business Intelligence and Insurance, viewed 18 March 2014, from <http://www2.microstrategy.com/download/files/whitepapers/open/Business-Intelligence-and-Insurance.pdf>.

Model Evaluation – Regression, Viewed 10-Oct-2014, from [http://www.saedsayad.com/model\\_evaluation\\_r.htm](http://www.saedsayad.com/model_evaluation_r.htm)

Patrick L. et al, 1998, Using Kohonen's Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud, Journal of Risk and Insurance.

Patrick L. et al, 2002, Fraud Classification Using Principal Component Analysis of RIDITS, Journal of Risk and Insurance.

Posterior Probability, Wikipedia, Viewed 10-Oct-2014, from [http://en.wikipedia.org/wiki/Posterior\\_probability](http://en.wikipedia.org/wiki/Posterior_probability)

Predictive Modeling, Viewed 12 July 2014, from <http://www.predictiveanalyticstoday.com/predictive-modeling/>, Predictive Analytics Today Website

Prenhall, Analytical Model, Viewed 09 July 2014, from [http://wps.prenhall.com/bp\\_turban\\_dsbis\\_9/141/36108/9243675.cw/content/index.html](http://wps.prenhall.com/bp_turban_dsbis_9/141/36108/9243675.cw/content/index.html),

Ravi K., 2012, A cost-effective approach to casualty claims analytics, viewed 28 June 2014, from <http://www.milliman.com/insight/pc/A-cost-effective-approach-to-casualty-claims-analytics/>.

Richard A., Ostaszewski K., 1995, Fuzzy Techniques of Pattern Recognition in Risk and Claim Classification, Journal of Risk and Insurance.

Richard S., 2010, Action Research, viewed 02 July 2014, from <http://www.ascd.org/publications/books/100047/chapters/What-Is-Action-Research%C2%A2.aspx>

Ricky H. , Predictive Analytics: Decision Tree and Ensembles, Viewed 15 July 2014, from <http://horicky.blogspot.com/2012/06/predictive-analytics-decision-tree-and.html>

Root Mean Squared Error , Wikipedia, Viewed 10-Oct-2014, from [http://en.wikipedia.org/wiki/Root-mean-square\\_deviation](http://en.wikipedia.org/wiki/Root-mean-square_deviation)

Saama, 2014, Inject Speed, Confidence & Accuracy to Guidewire Claims Data Analytics Transformation, White Paper

Sam D., 2012, Decision Tree Analysis using WEKA, University of Miami.

SAP, 2014, Insurance & data analytics summit workforce science report, viewed 28 June 2014, from [http://www.slideshare.net/PatriciaSaporito/insurance-data-analytics-summit-workforce-science-final?qid=8f425f91-ae4d-4ee1-b6a3-33326be7b479&v=qf1&b=&from\\_search=6](http://www.slideshare.net/PatriciaSaporito/insurance-data-analytics-summit-workforce-science-final?qid=8f425f91-ae4d-4ee1-b6a3-33326be7b479&v=qf1&b=&from_search=6).

Sas, 2013, Predictive Claims Processing: Transforming the Insurance Claims Life Cycle Using Analytics, White Paper, Sas Institute Inc.

Sharon T., Tennyson, 2002, Claims Auditing in Automobile Insurance: Fraud Detection and Deterrence Objectives, Journal of Risk and Insurance.

Statistical Classification, Wikipedia, Viewed 10-Oct-2014, from [http://en.wikipedia.org/wiki/Statistical\\_classification](http://en.wikipedia.org/wiki/Statistical_classification)

Steven C., 2012, The Analytics “Gold Rush”: Mountains of Data, Hidden Profits, Robert E. Nolan, Presentation in conjunction with American Family on implementing and using analytics.

Stijn V. et al, 2002, A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection, Journal of Risk and Insurance".

The analytics compass blog, Viewed 15 July 2014, from <http://www.simafore.com/blog/bid/62333/4-key-advantages-of-using-decision-trees-for-predictive-analytics>

The Data Protection Bill, 2013

Tom M., 1997, Machine Learning, McGraw Hill

WEKA-J48, Machine Learning Methods, Viewed 10-Oct-2014, from <http://wiki.qspr-thesaurus.eu/w/WEKA-J48>

Wkato University, WEKA, viewed 13-Sept-2014, from <http://www.cs.waikato.ac.nz/ml/WEKA/>

## APPENDIX

### AP 1. Request for Participation E-mail (Sample)

**From:** Rosemary A. Onyango

**Sent:** 25 July, 2014

**To:** pwainaina@britam.co.ke

**Subject:** Request for your participation in insurance claims analytics survey.

Dear Pauline Wainaina,

I am writing to you to request your participation in the above survey. The aim of the survey is to develop a predictive analytics model that can be used to efficiently and effectively manage general insurance claims.

Your participation and responses to this survey will help in building a model that is custom to the Kenyan local market as it will be based on the input that you provide. To begin, kindly click the link below to go to a survey web site (or copy and paste the link into your Internet browser) and then enter the personal code to begin the survey. It is very brief and will only take about 2 minutes to complete.

**Survey link:**

<https://docs.google.com/spreadsheet/viewform?formkey=dEVrMEN0bmNXVW9HM1MzZka04zWVE6MA>

**Personal Access Code:** 2001

Your participation in the survey is completely voluntary and all of your responses will be kept confidential. The access code is used to uniquely identify you from the rest of the participants. No personally identifiable information will be associated with your responses to any reports of these data.

The research is done as a partial fulfillment of the requirements of the Degree of Master of Science in Computational Intelligence at the University of Nairobi by Rosemary Onyango.

Should you have any comments or questions, please feel free to contact me at [raonyango@gmail.com](mailto:raonyango@gmail.com) or +254 723 232 060.

Thank you very much for your time and co-operation. Your feedback is very important to this research.

Sincerely,

Rosemary A. Onyango

Student – University of Nairobi, Kenya

## AP 2. Online Questionnaire Used in the Research

The form can be found on

<https://docs.google.com/spreadsheets/viewform?formkey=dEVrMEN0bmNXVW9HM1MzbnZka04zWVE6MA>

# GENERAL INSURANCE CLAIMS ANALYTICS RESEARCH

This is a questionnaire on general insurance claims analytics that seeks to obtain a general opinion of your professional experience. Your opinion on the 12 questions shall be greatly valued. Kindly note that your e-mail not any information attached to your e-mail will NOT be recorded when you submit this form.

\* Required

### Personal Access Code

Key in the 4 digit number provided in your e-mail invite

## BASIC INFORMATION

1. Which company do you work for?

Optional

2. Please indicate your role:

Optional

3. What tools are in use by your organization to perform analytics and Business Intelligence? \*

Give a comma separated list

**4. What is the main target audience for the output of analytics and BI reports in your organization? \***

Optional

- Top level management
- Middle-level management outside the claims function
- Claims Managers
- Claims Analysts
- Claims Administrators
- Other:

**LEVEL OF URGENCY**

**5. How often are analytical reports required? \***

- Hourly
- Daily
- Weekly
- Monthly
- On demand
- Other:

**6. How would you rate the urgency of most analytical reports? \***

1 2 3 4 5

Not Urgent      Very Urgent

**7. How far in history do analytical reports go while managing claims? \***

- Since the business began
- For the last 1 year
- Atleast 6 months from the current date
- Last one month only from the current date
- Other:

**8. What kind of essential analytical information in claims cannot be generated from current tools especially spreadsheets? \***

Kindly list them down.

## VALUE OF ANALYTICS AND BUSINESS INTELLIGENCE

**9. How much is spent on average on analytics and BI? \***

- Less than KES 1 Million
- KES 1 to 10 Million
- KES 10 to 20 Million
- More than 20 Million
- I do not know
- Other:

**10. How are analytics and Business Intelligence reports used by your organization? \***

- For fraud detection
- For product development
- For crafting strategies
- For reporting
- Other:

**11. Give any further details if possible on the use of analytics and business intelligence reports in your organization**

Kindly provide a brief description.

## INDUSTRY REGULATION

**12. What industry rules or regulations must any of your analytics and Business Intelligence functions conform to?**

Kindly list down

**Submit**

Never submit passwords through Google Forms.



### AP 3. E-mail response to a request for general insurance claims data from BRITAM Insurance.

**Rosemary, Onyango**, <[ronyango@britam.co.ke](mailto:ronyango@britam.co.ke)> Sep 23 (3 days ago) ☆ ↶  
to Edward ▾

Morning Edward,

I have a request. I am doing a research project on general insurance claims analytics, and was asking if I can be allowed to use data from AIMS with client details masked for security purposes.

...

**Osiya, Osiya, Edward** Sep 24 (2 days ago) ☆ ↶  
to me ▾

Rosemary,

I have checked and confirmed that our Internal IT security policies do not allow use of company databases or proprietary information for research or academic purposes.

Regards,

Edward.

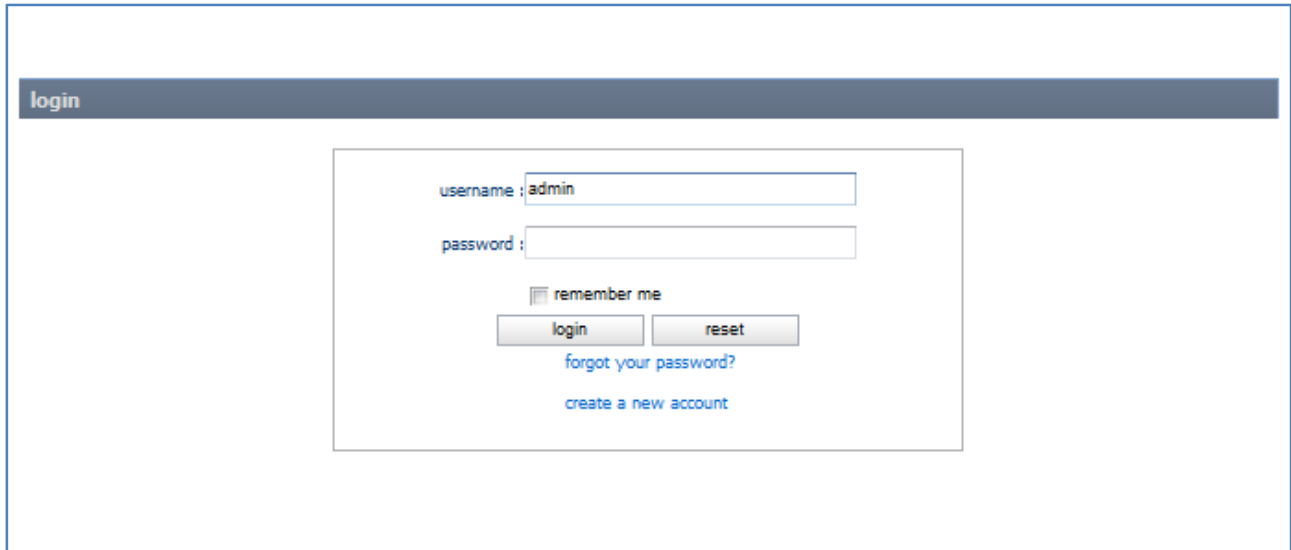
...

On Tue, Sep 23, 2014 at 3:34 PM, Rosemary, Onyango, <[ronyango@britam.co.ke](mailto:ronyango@britam.co.ke)> wrote:  
| Alright, thank you.

On Tue, Sep 23, 2014 at 3:32 PM, Osiya, Osiya, Edward <[eosiya@britam.co.ke](mailto:eosiya@britam.co.ke)> wrote:  
| Rosemary,

## AP 4. Research Prototype Screen Shots

1. The login screen. The application is secured by Spring Security.



login

username : admin

password :

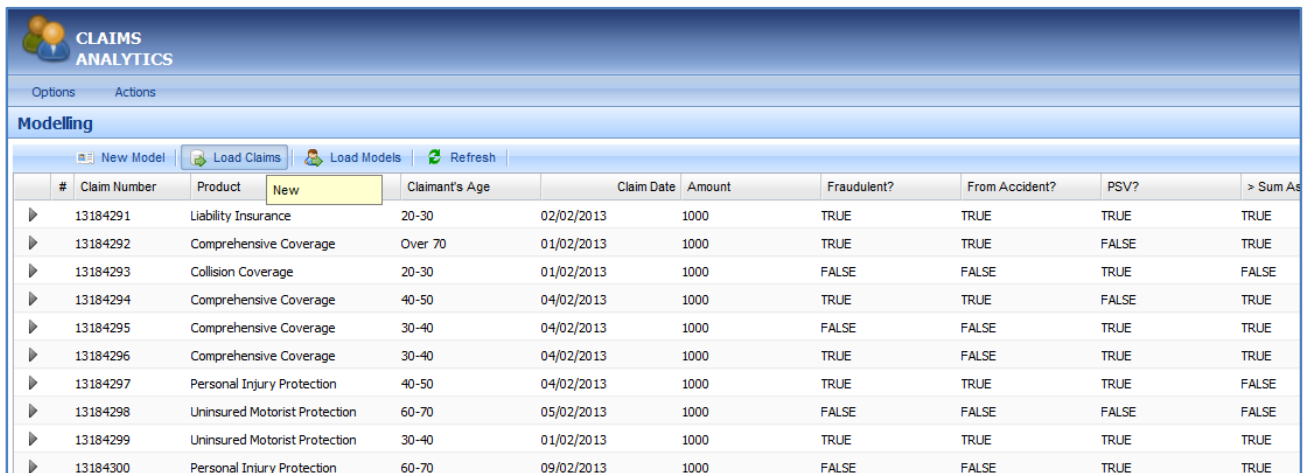
remember me

login reset

[forgot your password?](#)

[create a new account](#)

2. The modelling module with sample claims loaded.



CLAIMS ANALYTICS

Options Actions

Modelling

New Model Load Claims Load Models Refresh

#	Claim Number	Product	Claimant's Age	Claim Date	Amount	Fraudulent?	From Accident?	PSV?	> Sum As
▶	13184291	Liability Insurance	20-30	02/02/2013	1000	TRUE	TRUE	TRUE	TRUE
▶	13184292	Comprehensive Coverage	Over 70	01/02/2013	1000	TRUE	TRUE	FALSE	TRUE
▶	13184293	Collision Coverage	20-30	01/02/2013	1000	FALSE	FALSE	TRUE	FALSE
▶	13184294	Comprehensive Coverage	40-50	04/02/2013	1000	TRUE	TRUE	FALSE	TRUE
▶	13184295	Comprehensive Coverage	30-40	04/02/2013	1000	FALSE	FALSE	TRUE	TRUE
▶	13184296	Comprehensive Coverage	30-40	04/02/2013	1000	TRUE	FALSE	TRUE	TRUE
▶	13184297	Personal Injury Protection	40-50	04/02/2013	1000	TRUE	TRUE	TRUE	FALSE
▶	13184298	Uninsured Motorist Protection	60-70	05/02/2013	1000	FALSE	FALSE	FALSE	FALSE
▶	13184299	Uninsured Motorist Protection	30-40	01/02/2013	1000	TRUE	TRUE	TRUE	TRUE
▶	13184300	Personal Injury Protection	60-70	09/02/2013	1000	FALSE	FALSE	TRUE	TRUE

3. The modeling module with sample models loaded.

The screenshot shows the CLAIMS ANALYTICS Modelling interface. At the top, there are tabs for 'Options' and 'Actions'. Below that, the 'Modelling' section contains buttons for 'New Model', 'Load Claims', 'Load Models', and 'Refresh'. A table lists 10 models with columns for '#', 'Model Number', 'Attribute', 'Start Date', 'End Date', 'Model Details', and 'Visualize'. Each row includes a 'view stats' button and a 'view tree' button.

#	Model Number	Attribute	Start Date	End Date	Model Details	Visualize
▶	MDL000021	Cover Maintained	09/01/2013	08/10/2014	view stats	view tree
▶	MDL000020	Fraudulent	05/02/2013	07/10/2014	view stats	view tree
▶	MDL000015	Cover Maintained	01/01/2013	07/10/2014	view stats	view tree
▶	MDL000012	Cost (below sum assured)	11/02/2013	07/10/2014	view stats	view tree
▶	MDL000011	PSV	01/01/2013	07/10/2014	view stats	view tree
▶	MDL000010	PSV	04/03/2013	06/10/2014	view stats	view tree
▶	MDL000009	Fraudulent	06/02/2013	06/10/2014	view stats	view tree
▶	MDL000008	Fraudulent	01/02/2013	06/10/2014	view stats	view tree
▶	MDL000002	Cause Is Accidents	06/02/2013	02/10/2014	view stats	view tree
▶	MDL000001	Cause Is Accidents	28/09/2014	02/10/2014	view stats	view tree

4. Model creation: Sample form for creating a model.

The screenshot shows the CLAIMS ANALYTICS Modelling interface with the 'New Model' form open. The form has a 'Model' button and a 'Cancel' button. Below that, there is a 'Modelling Details' section with a 'Model and Save' button. The form contains the following fields:

- Model Attribute:** A dropdown menu with 'Fraudulent' selected.
- Claims From:** A date field with '01/01/2010' entered.
- (Claims) To:** A date field with '10/09/2014' entered.
- Created On:** A timestamp field with 'Thu Oct 09 09:49:20 GMT+300' entered.

At the bottom of the form, it indicates '0 of 10 selected'.

5. Model creation: Server response with results of a newly creating model.

The screenshot shows the CLAIMS ANALYTICS Modelling interface. A table lists models with their numbers and attributes. A 'Note' dialog box is open, displaying the following statistics:

Model Accuracy: 85.00%

=== Model Statistics Summary ===  
 Correctly Classified Instances 85 85 %  
 Incorrectly Classified Instances 15 15 %  
 Kappa statistic 0.632  
 Mean absolute error 0.1956  
 Root mean squared error 0.3356  
 Relative absolute error 47.2587 %  
 Root relative squared error 73.9215 %  
 Total Number of Instances 100

=== Confusion Matrix ===  
 a b <- classified as  
 64 7 | a = TRUE  
 8 21 | b = FALSE

6. Model creation: A tree visualization of a newly created model.

The screenshot shows the CLAIMS ANALYTICS Modelling interface with a 'Tree View' visualization of a decision tree. The tree structure is as follows:

```

  graph TD
    A(accidents) -- = FALSE --> B(cover_maintained)
    A -- = TRUE --> C(psv)
    B -- = FALSE --> D[FALSE (21.0)]
    B -- = TRUE --> E(costs)
    C -- = FALSE --> F(cover_maintained)
    C -- = TRUE --> G[TRUE (23.0)]
    E -- = FALSE --> H[FALSE (4.0)]
    E -- = TRUE --> I[TRUE (16.0/6.0)]
    F -- = FALSE --> J[TRUE (5.0)]
    F -- = TRUE --> K[FALSE (16.0/5.0)]
  
```



9. Running a Prediction: A prediction screen from the prototype.

CLAIMS ANALYTICS

Options Actions

Prediction

Predict Refresh

#	Claim Number	Product	Claimant's Age	Claim Date	Amount	Fraudulent?	From Accident?	PSV?	> Sum Assured	Cover Maintained?	Predicted Probability
No record found											

0 of 0 selected

Run Prediction Cancel

Prediction Details

Model# MDL000020 Prediction Attribute# Fraudulent Claims From# 08/01/2014

(Claims) To 10/09/2014 Created On Thu Oct 09 09:50:17 GMT+300

10. Running a Prediction: A screen with results of the above prediction.

CLAIMS ANALYTICS

Options Actions

Prediction

Predict Refresh

#	Claim Number	Product	Claimant's Age	Claim Date	Amount	Fraudulent?	From Accident?	PSV?	> Sum Assured	Cover Maintained?	Predicted Probability
▶	13184387	'Comprehensive Coverage'	30-40	19/08/2014	1000	TRUE	FALSE	FALSE	TRUE	FALSE	*1.0 : 0.0 :
▶	13184388	'Uninsured Motorist Protection'	20-30	22/08/2014	1000	TRUE	FALSE	TRUE	TRUE	TRUE	*1.0 : 0.0 :
▶	13184389	'Comprehensive Coverage'	20-30	23/08/2014	1000	TRUE	TRUE	FALSE	FALSE	TRUE	*0.69 : 0.31 :
▶	13184390	'Comprehensive Coverage'	30-40	24/08/2014	1000	TRUE	TRUE	FALSE	TRUE	TRUE	*0.69 : 0.31 :

**Prediction Results**

Prediction Process Complete.

OK

0 of 4 selected

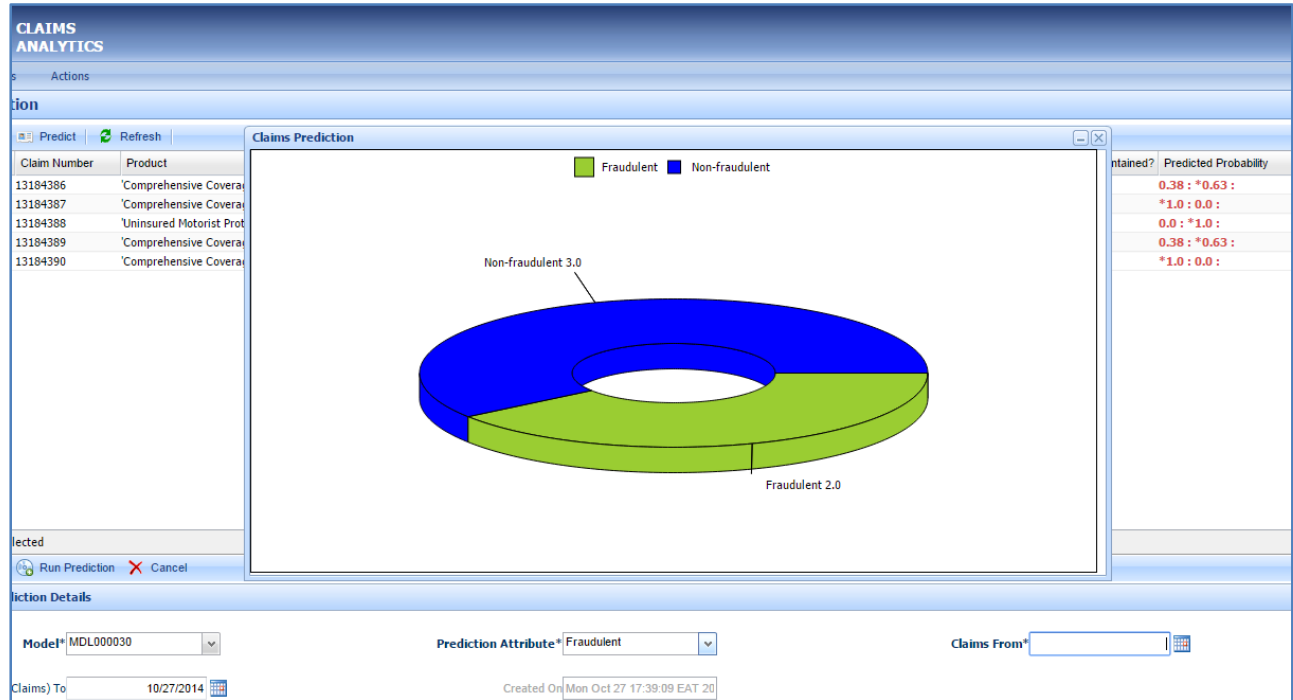
Run Prediction Cancel

Prediction Details

Model# MDL000020 Prediction Attribute# Fraudulent Claims From#

(Claims) To 10/09/2014 Created On Thu Oct 09 09:50:17 GMT+300

11. The predictions summary in a pie chart



12. Running a Prediction: A.ARFF unlabeled prediction file generated for the above prediction. The last two entries are dummy claims used to provide the model with a list of possible class values for the prediction attribute.

```

1 @relation PREDICT1149_4_14
2
3 @attribute product {'Comprehensive Coverage','Uninsured Motorist Protection'}
4 @attribute claim_period {Aug-13,Aug-14}
5 @attribute claimant_age {30-40,20-30}
6 @attribute amount numeric
7 @attribute psy {FALSE,TRUE}
8 @attribute accidents {FALSE,TRUE}
9 @attribute costs {TRUE,FALSE}
10 @attribute cover_maintained {FALSE,TRUE}
11 @attribute fraudulent {TRUE,FALSE}
12
13 @data
14 'Comprehensive Coverage',Aug-13,30-40,1000,FALSE,FALSE,TRUE,FALSE,?
15 'Uninsured Motorist Protection',Aug-14,20-30,1000,TRUE,FALSE,TRUE,TRUE,?
16 'Comprehensive Coverage',Aug-14,20-30,1000,FALSE,TRUE,FALSE,TRUE,?
17 'Comprehensive Coverage',Aug-14,30-40,1000,FALSE,TRUE,TRUE,TRUE,?
18 'Comprehensive Coverage',Aug-13,30-40,1000,FALSE,FALSE,TRUE,FALSE,TRUE
19 'Comprehensive Coverage',Aug-13,30-40,1000,FALSE,FALSE,TRUE,FALSE,FALSE
20

```