INFOSYS 330

MACHINE LEARNING/ DATA MINING BUSINESS INTELLIGENCE

Shaquille Ali

2020

Table of Contents

[INTRODUCTION 2](#_Toc42020145)

[BUSINESS PROBLEM 2](#_Toc42020146)

[STAKEHOLDERS 3](#_Toc42020147)

[Stakeholder Analysis 3](#_Toc42020148)

[KPI for Anne Crowe: 4](#_Toc42020149)

[User Persona: 4](#_Toc42020150)

[DESIGN AND CHOICE 5](#_Toc42020151)

[Subproblem 1: 5](#_Toc42020152)

[Case based reasoning (CBR) 5](#_Toc42020153)

[Artificial Neural Network (ANN) 6](#_Toc42020154)

[Subproblem 2: 6](#_Toc42020155)

[Decision Trees 7](#_Toc42020156)

[Naive Bayes Algorithm 7](#_Toc42020157)

[IMPLEMENTATION 8](#_Toc42020158)

[Hardware requirements: 8](#_Toc42020159)

[Software requirements: 8](#_Toc42020160)

[System Visualisation: 8](#_Toc42020161)

[Conclusion: 9](#_Toc42020162)

[Limitations: 9](#_Toc42020163)

[System Infrastructure Diagram: 10](#_Toc42020164)

[Appendix 11](#_Toc42020166)

[Algorithm Diagrams 11](#_Toc42020167)

[CBR Figure 1 11](#_Toc42020168)

[ANN Figure 2 11](#_Toc42020170)

[Decision Tree Figure 3 12](#_Toc42020172)

[Naïve Bayes Figure 4 12](#_Toc42020174)

# Executive Summary

‘Kids Corp’ is an online store with the purpose of selling safe, quality products and being able to find information about various services provided. Currently the business has a very dated system, that requires plenty of oversight and tedious administrative tasks.

This system needs to be converted to new solutions by implementing business intelligence algorithms, that will help executives with the decision making process.

The biggest issue identified is time. Time management plays a huge role in any business as It impacts every part of a business, be it administrative work, sales, marketing, finance, delivery. My solution looks at how time can be more optimized and reduced in ‘Kids Corp’ daily/ administrative operations and in some of its implemented features. Quite a bit of time in ‘Kids Corp’ is being used up on the administrative side. A lot of parents who are using up their free time to work on maintaining the business are doing administrative tasks that can be automated or done faster using a business system.

The Proposed solution involves using multiple business intelligence algorithms to assist in fixing this problem. It involves using Case based reasoning (CBR), Artificial Neural Network (ANN) to solve with distinguishing the quality of goods supplied by analysing safety rating, quality of goods used to make the products and based on past reviews of products. Decision Trees and Naïve Bayes algorithm will pull data from reviews and recommend day-care centres and health-care providers.

With this integration of the solution this will enhance ‘Kids Corp’ in making better suited decisions.

# INTRODUCTION

‘Kids Corp’ is an ‘all about kids’ online platform that provides goods and services targeted towards helping parents make better ensured decisions for their kids. ‘Kids Corp’ is valued by parents because it provides them with reliable recommendations about goods and services for their kids. The business is quality driven, by ensuring the items being supplied and services being provided are of the best and safest quality. What first started as an online store has now also become a platform where parents can find various information about services provided that help them make informed decisions for their kids.

The biggest issue identified is time. Time management plays a huge role in any business as It impacts every part of a business, be it administrative work, sales, marketing, finance, delivery. My solution looks at how time can be more optimized and reduced in ‘Kids Corp’ daily/ administrative operations and in some of its implemented features. Quite a bit of time in ‘Kids Corp’ is being used up on the administrative side. A lot of parents who are using up their free time to work on maintaining the business are doing administrative tasks that can be automated or done faster using a business system.

## BUSINESS PROBLEM

Time management in the workplace. After analysing the case study, we can identify that a lot of the parents' time who offer to work for free for ‘Kids Corp’ is being used doing administrative tasks that can and should be using business systems.

Decision making is not carried out at a high level in ‘Kids Corp’ due to the uncertainty of business intelligence within the business and the inexperience of most parents with technology that work for ‘Kids Corp’. This makes it hard for ‘Kids Corp’ to know how to manage time effectively.

This problem is quite broad, so it has been broken down into smaller subproblems. These subproblems focus on specific areas/ operations and features of the business where time can be reduced and optimized for tasks. The sub problems are:

* The system should be able to automatically distinguish the quality of goods supplied by analysing safety rating, quality of goods used to make the products and based on past reviews of products. Currently each supplier seems to be reviewed by parents who work for ‘Kids Corp’ and then put into a recommendation system. By doing so they are spending too much time on something a business system can do quickly and efficiently.
* The system should automatically pull data from reviews and recommend day-care centres and health-care providers. Currently parents are reviewing day-care centres and health care providers in the system which in turn recommends them to parents. If we pull data from other sites, we will have a better understanding and a more unbiased review for the day-care centre and health-care provider as we will be working with a larger set of data.

Solving this problem is important to the organisation because time is money. It needs to be optimized and used efficiently to ensure growth and profitability of the business. The amount of time spent on these administrative tasks can be better spent on doing things like making new products or features to better enhance the business and attract more customers.

# STAKEHOLDERS

**Anne Crowe - Information Systems Manager**

As an Information Systems Manager, Anne Crowes’ role involves managing the system, managing the parents/ workers, allocating them tasks to complete within their skill level, ensure her teams KPIs are met and thinking of new ways to improve and or implement features into the system.

This problem affects her as she must use the system every day. She also spends a fair amount of time going over the parent’s technical knowledge so she can assign them work to be done for the day. This can be quite frustrating as she is spending more time managing parents than the system.

## Stakeholder Analysis

Challenges stakeholder currently face?

* Analysing quality of each product from suppliers
* Managing parents with various skill levels
* Working with lots of data

What improvements does the stakeholder wish to achieve with this project

* Time to focus on big picture projects.
* A lot less hands on with the raw data.
* Less managing parents, as processes will be more automated.
* Improve system efficiency.

Information expected to be given to stakeholders regarding the project.

* Grace Sparkle and Anne Crowe will be our high priority stakeholders so they will be briefed on the implementations via a weekly report and a bi-weekly in person meeting.
* Anne Crowe will be more interested in the implementation side of things and the steps required for the business to move to this new system.
* As an information systems manager she is very familiar with ‘Kids Corp’ current system. So, we will be consulting with her regarding the current system and improvements we are suggesting.

Is the stakeholder supportive, oppositional, or neutral towards the project? How much power does she have to sway others in the organisation towards their position?

* The stakeholder is supportive of this project as it will mean less time spent on training parents and less time spent on administrative tasks. Since our stakeholder is the IS manager of ‘Kids Corp’ she will be able to sway the parents who she manages and will likely be able to persuade Grace Sparkle (CEO) to support this project.

## KPI for Anne Crowe:

Number of total users

Number of new users per day/week/month/year

Daily/Weekly/Monthly website traffic

Number of products bought

Number of bought products reviewed

Number of products suppliers advertising per day/month/year

Minutes parents/ workers spent on a task

Employee satisfaction rating

## User Persona:

****

# DESIGN AND CHOICE

Data Warehouse

A data warehouse will be required for the Business Intelligence solution. As the data generated from this will be used to aid the decision making with the solution. Assuming that ‘Kids Corp’ currently has a data warehouse, this can be used to extract information and make informed decisions.

## Subproblem 1:

“The system should be able to automatically distinguish the quality of goods supplied by analysing safety rating, quality of goods used to make the products and based on past reviews of products. Currently each supplier seems to be reviewed by parents who work for ‘Kids Corp’ and then put into a recommendation system. By doing so they are spending too much time on something a business system can do quickly and efficiently.”

|  |  |  |  |
| --- | --- | --- | --- |
| Dimension | Target solution | CBR | ANN |
| Accuracy | High | Y | Y |
| Explainability | Moderate | Y |  |
| Response Speed | Moderate | Y | Y |
| Scalability | Moderate |  | Y |
| Flexibility | Moderate | Y | Y |
| Embeddability | High |  | Y |
| Tolerance for Complexity | High |  | Y |
| Tolerance for Noise in Data | High |  | Y |
| Independence | High | Y | Y |

### Case based reasoning (CBR)

Case based reasoning is a popular BI solution used in many businesses and will be used in our system to generate cases. It will first access the data warehouse and analyse the quantity supplied by supplier, product safety rating, material used to manufacture product and reviews of the product. CBR will take all of this into account and generate a collection of cases.

An example of a case created for ‘Kids Corp’ would be product supplied 20 times by Toys4Kids, product safety rating was 4 stars, materials used was hard plastic, reviews for this product are around an average of 4.5 stars.

To predict, we would filter and retrieve cases based on the probes we want. CBR will retrieve this and input this to ANN which would find an association or relationship and derive a prediction.

The benefits of CBR are that it solves problems without having to start from scratch each time and that it uses past knowledge and historical data to draw the solution. This will assist in the decision making of the subproblem. Some weaknesses for CBR are if attributes have subtle dependencies, it becomes difficult to compare cases and variables that encode soft knowledge can be difficult to deal with. Algorithm shown on Figure 1 of appendix.

### Artificial Neural Network (ANN)

Artificial Neural Network is a BI solution often used to find patterns within the business.  This occurs when ANN tries to guess the information provided to see if it is relevant or not, this can be predicted using historical data from CBR. The inputs for ANN will be the outputs from CBR and will be used to train data. We will be using supervised learning where ANN will get the sample data and the desired outputs. We will need to make sure that ANN is not over trained so that it can make predictions for the future.

Benefits of ANN are does not require a specification of knowledge or formulae and it can fill in missing or noisy data. Some weaknesses are that the process can be quite time consuming - this affects us as our main problem is time optimization and reduction. It also requires a lot of processing power. Artificial Neural Network can be used in our solution to fine-tune the data received from CBR and provide a more accurate and reliable solution. ANN works well side by side with CBR as they account for each other's weaknesses.

ANN will have 3 layers. Input layer obtained by CBR output. Hidden layer where the simulations occur. Lastly the output layer which will predict the results.

When a supplier adds a product to sell on the ‘Kids Corp’ web store, using an artificial neural network, the system will be able to take the data obtained from CBR about the supplier and their products and be able to determine whether the product will be listed or not.

Algorithm shown on Figure 2 of appendix.

## Subproblem 2:

“The system should automatically pull data from reviews and recommend day-care centres and health-care providers. Currently parents are reviewing day-care centres and health care providers in the system which in turn recommends them to parents. If we pull data from other sites, we will have a better understanding and a more unbiased review for the day-care centre and health-care provider as we will be working with a larger set of data.”

|  |  |  |  |
| --- | --- | --- | --- |
| Dimension | Target solution | NB | DT |
| Accuracy | High | Y | Y |
| Explainability | Moderate | Y | Y |
| Response Speed | Moderate | Y |  |
| Scalability | Moderate | Y | Y |
| Flexibility | Moderate | Y | Y |
| Embeddability | High |  | Y |
| Tolerance for Complexity | High |  |  |
| Tolerance for Noise in Data | High |  |  |
| Independence | High |  |  |

### Decision Trees

Decision Trees will be a very effective solution for this subproblem as it will work well with Naive Bayes as both algorithms use the concept of probability. Decision Trees are tree-based methods used to predict outcomes. ‘Kids Corp’ can use decision trees for web content mining to get reviews, prices, location, and services offered about health and day-care centres and return a prediction of the best overall health or day-care centre.

Decision Trees use an IF X Then Y algorithm. The algorithm starts off with a root node and then branches out to the child nodes. The child link may lead to another question being posed or returns a prediction. Decision Trees are used to create order via entropy and information gain. We want to avoid having an over-fitted model. To do so we need to use noise free data and not place too many details in the trees.

Benefits of decision trees include able to generate understandable rules, perform classification without too much computation and handle continuous or categorical attributes. Some weaknesses are it can be expensive to train, error prone, no sequence.

Algorithm shown on Figure 3 of appendix.

### Naive Bayes Algorithm

Naive Bayes algorithm can be used for this problem as it is simple and highly effective. It is an algorithm based on the concept of probability. The algorithm assumes that instances are independent of each other. The input will be the outputs from the decision trees. We will be taking advantage of the clustering and using deep knowledge of the health and day-care centres to solve our solution by taking in the reviews and outputting a cluster that will show the highest rated health and day-care centres with the parents current location.

By analysing the clusters around this location, we can recommend the highest rated health or day-care centre. Algorithm shown on Figure 4 of appendix.

# IMPLEMENTATION

## Hardware requirements:

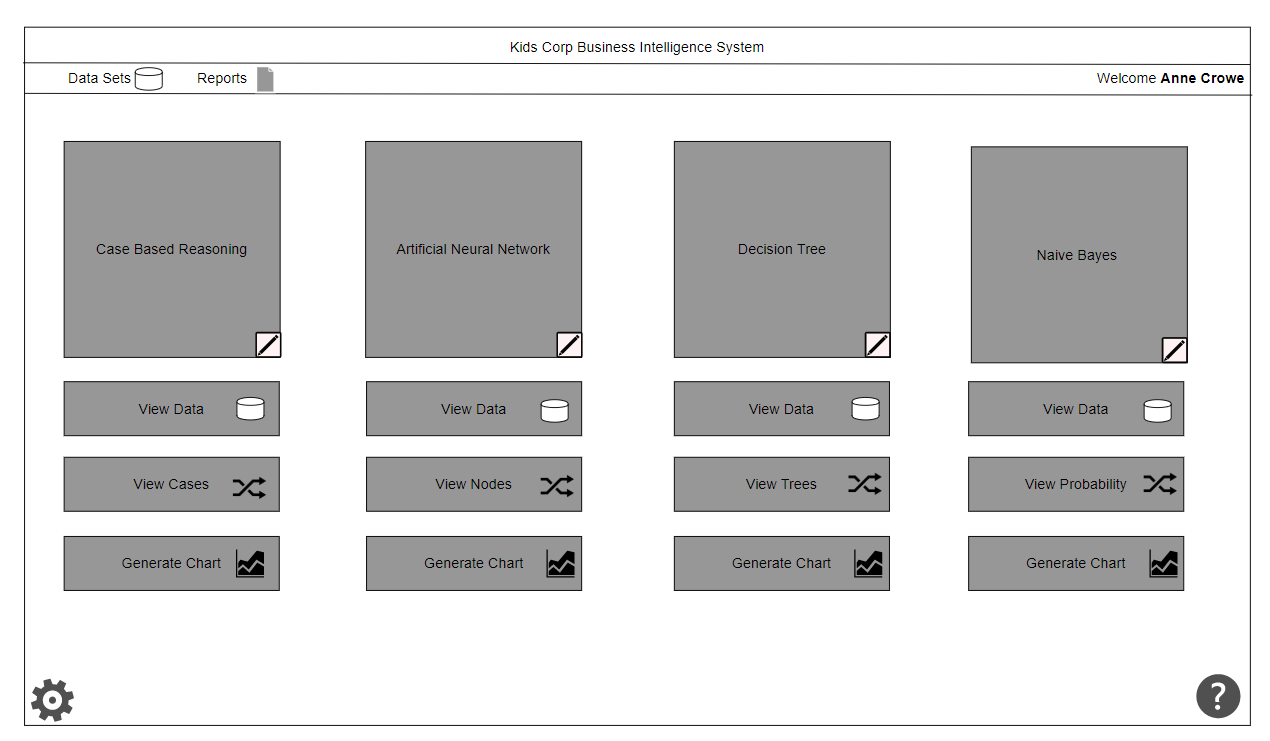
* Servers: needed to store data acquired from data warehouse and data mining.
* Large capacity RAM cards - needed for assisting the intensive computation from the algorithms
* Large capacity Hard drives - To store data/ backing up data.
* Advanced CPUs preferably with 8 cores and a high speed. This will be needed due to the high computation power of the algorithms.
* Strong Network Connection - Allow ‘Kids Corp’ servers to connect to the BI algorithms.
* Cloud Servers - backing up important data if the server crashes.

## Software requirements:

* CBR: jCOLIBRI - this software will enable users to find cases and match. It will allow the user to refine and analyse historical data. The structure of CBR would be a set of questions using texts and numbers. Example: Quantity Supplied: 20, Supplier: Toys4Kids, Safety rating: 4, Materials used: Hard plastic, Review: 4.5.
* ANN: Neural Designer - this software can help use identify relationships by profiling data sets that assist in the support of decision making. ANN will help with predicting whether an item will be listed based on the cases. Example: Product supplied by Toys4Kids is of high quality. Product supplied by ToysToysToys is of low quality.
* DT: AzureML - this software can assist with sort data and return a probability. Can be trained. Example of this can be: IF reviews < 3 stars THEN do not recommend. IF reviews > 3 stars THEN check prices etc.
* NB: AzureML - this software can assist with sort data and return a cluster prediction. Can be trained.

## System Visualisation:

This is a sketch of what the system may look like. It uses big clear buttons as it is easy to read and navigate. All the user will have to do is click on whatever algorithm they want to view, and the system will take care of the rest



# Conclusion:

The current system ‘Kids Corp’ are using is not quick and efficient which in turn does not assist with making better informed decisions. The proposed system will mitigate this problem as it will reduce time spent on administrative task by the workers and stakeholder.

This will allow ‘Kids Corp’ to make better business decisions as most of their time won’t be spent on administrative and menial tasks.

## Limitations:

The are a few limitations with this proposed solution. They are quite expensive due to the hardware needed to run and support these algorithms. Decision Tree can sometimes be inaccurate if the data is missing or corrupted. Initially the algorithms will be time consuming as the algorithms need to be trained until they start getting the desired outcomes. Employees will need to be trained on the system as well.

## System Infrastructure Diagram:

## 

# Appendix

## Algorithm Diagrams

### CBR Figure 1

### 

### ANN Figure 2

### 

### Decision Tree Figure 3

### 

### Naïve Bayes Figure 4

### 

\*Diagrams were drawn using draw.io