Child Maltreatment Forecast using Bigdata Intelligent Approaches

Abdurazzag A Aburas*[1], Mohammad Hassan[2], Hilary Lin [3], Shreshtha Batshu[4]
[1]Computer Engineering Dept. University of KwaZulu-Natal, Durban, South Africa
[2]Faculty of Engineering, Al-Ahliyya Amman University, Jordan, Amman,
[3]Computer Engineering Dept. University of KwaZulu-Natal, Durban, South Africa
[4]Computer Engineering Dept. University of KwaZulu-Natal, Durban, South Africa
1: aburasa@ukzn.ac.za, 2: mhassan@ammanu.edu.jo 3: 215040470@ukzn.ac.za, 4: 215012460@ukzn.ac.za

***Abstract*-Child Welfare associations collect large datasets that they are required to process to assess the risk posed to children within their living environment. Current methods for dealing with these large datasets reduce the time caseworkers are able to spend with the children assigned to their care. Within the following research work, methods for obtaining trends in child abuse and neglect datasets are outlined using self-populated datasets. The potential of Machine Learning algorithms in supporting child welfare associations is illustrated through the use of two unsupervised machine learning algorithms. The simple unsupervised learning cycle of the C4.5 algorithm together with the Apriori algorithm assist in ensuring changes in trends can be identified. The results from using Big Data intelligent algorithms indicate that child maltreatment cases can be more efficiently prioritized and handled through Big Data methods. The datasets and machine learning algorithms were run and stored on the Hippo Cluster.**

Keywords: child welfare, risk, abuse, neglect, datasets, machine learning

1 Introduction

Child welfare is a group of services developed to improve quality of life and promote the mental, physical, and emotional well-being of children. These services were established in an attempt to reduce child abuse and neglect by families and/or parents and guardians of children. Many families become involved with the child welfare due to reports of suspected abuse or neglect, considered in this research work under the term of 'Child Maltreatment'.

There are two principal methods of identifying child abuse such as spotting physical injuries, and detecting enough information to warrant an investigation. When reports are received and logged by the Child Protective Services [27], they may fall under two broad categories, namely screened-in or screened-out. Screened-in cases are classified by having enough information to warrant an investigation. Screened-out cases are classified by lack of information or not meeting the legal definition of abuse or neglect in the jurisdiction [1]. Many reports go unnoticed due to lack of evidence, and often these victims are left for dead. Early identification of child maltreatment, reduces later negative effects on the mental and psychological health of children. Within South Africa, child welfare groups arrange foster families and maintain orphanages for children who are not safe in their homes and have been removed from their parent or guardian's care. However, the children's situations may not improve by being adopted into a foster home or orphanage. In these cases, the caseworkers may be blamed for the child maltreatment that occurs in foster homes and orphanages [1].

2 Research related works

Child Welfare associations collect large amounts of datasets that they are required to process to assess the risk posed to the child from their current environment. Current methods for dealing with these large datasets are cumbersome and inefficient [30, 31]. This leaves a wide scope for a Bigdata Analytics datasets to identify child abuse and neglect with a reasonable accuracy.

Lery et al [34] suggests four principles for the successful use of Big Data strategies in child welfare decision making, with an emphasis on the consumer's role in the process and the need for a continuous improvement cycle to the solution. A preliminary study presented a proposal of a framework to obtain a solution for decision making in the forensic field centred on sexual abuse in children through an intelligent decision support system [28]. Some research work deals with child information such as Electronic Child Records (ECRs) [29] with improved design and integrated approach with the intention to help create the right balance between personal interests, system performance, and social values within the records.

Hadoop is a set of open source programs or procedures that can be used for storage or analysis of large datasets that allow the addition or modification of data systems as is required by the user [2]. Since Hadoop is open source, it can be modified into more specific software for the intended usage. It consists of Distributed File System [3, 4], MapReduce which carries out two operations: reading from the database, arranging the database into a suitable format for analysis, performing mathematical operations and MapReduce uses datasets locality [2]. Data locality minimizes network congestion by moving the

978-1-5386-9588-3/18/$31.00 ©2018 IEEE

computation close to where the data resides within the cluster. Hadoop needs to know where the datasets are located as well as the topology of the nodes where the tasks are executed [4]. Hadoop Common, also known as Hadoop Core, the module provides the Java tools required to read the data from the Hadoop file system between different operating systems [2][5]. Finally, YARN was designed to solve the scalability problem faced by MapReduce. Through the usage of the three managers and distributed application, YARN creates a generic approach to running tasks that allows the cluster to run various workloads, and more efficient solution to data processing [6].

3 Machine Learning C4.5 Algorithm

The child maltreatment project uses two Big Data intelligent algorithms: the C4.5 and Apriori algorithms. Both algorithms are unsupervised learners. Unsupervised learning is a type of machine learning algorithm where sequence and patterns are identified within datasets that do not have obvious relations [16]. The C4.5 Algorithm [19, 25] is a collection of algorithms that perform classifications in data mining and machine learning. This C4.5 algorithm consists of two phases for decision tree construction: growing tree phase and pruning tree phase. The growing tree phase is a top-down approach which repeatedly builds the tree. The pruning-tree phase is a bottom-up approach which removes trees by replacing them with leaves (decisions).

Decision trees are an effective method of unsupervised learning. The aim is to partition the dataset into groups that contain the most similarities in terms of the variable to be predicted. It takes data as an input, and outputs a tree where each end node (leaf) is a decision and each non- final node represents a test [17, 18]. Figure 1 illustrates the splitting of the dataset into the three levels based on the maltreatment type. A larger number of tests may further increase the number of levels of maltreatment for more precise subdivisions.

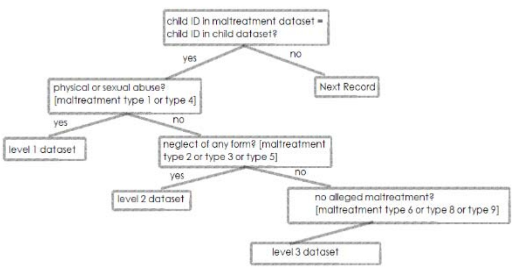


Figure 1: Decision making tree based on dataset

Figure 2 shows the system's flowchart based on the maltreatment type (decision tree). The decisions follow a simple algorithm: first, check if there is a record to analyze. Check if child is under 18. Check if child ID in maltreatment dataset matches the child ID in child dataset. Check if type 1 or 4, split into new dataset. Check if type 2 or 3 or 5, split into new dataset. Check if type 6 or 8 or 9, split into new dataset. Finally, repeat with remaining dataset records until there are no remaining uncatalogued records.

3.1 Machine Learning Apriori Algorithm

The Apriori Algorithm [22, 23, 24, 26] is a bottom-up method to understand how things are related to each other. It is used for mining frequent item sets and association rules associated with the item sets. A usage of the Apriori algorithm in social welfare applications is due to the quick accumulation of data. In addition, the Apriori algorithm is designed to operate on a database with a lot of transactions [20]. The input of the Apriori algorithm is a set of transactions called a transaction datasets, where a transaction is a set of items [21]. The output is a frequent item set that is dependent on the user defined parameter. The results are interpreted in terms of the support of an item set, or the number of times the item set appears in the transaction database.

Within the scope of child maltreatment, the transaction datasets are the children who have been categorized by type by the C4.5 algorithm. The resultant support items are the similarities between the datasets for each type. This enables improved tests for each of the categories in the C4.5 algorithm.

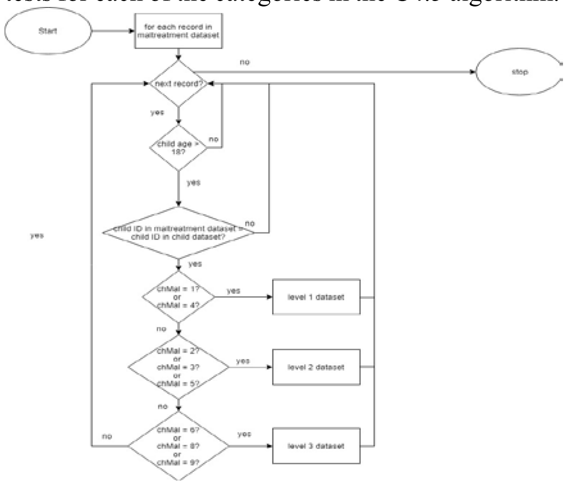


Figure 2: flowchart for C4.5 Algorithm [19]

4 Child welfare Dataset Design

For this project, the datasets have been designed using child datasets as well as perpetrator datasets. In application, the datasets focused on would be the reports that are received and the frequency in which these reports are received based on the individual child.

4.1 Dataset Identification

Within a child welfare organization, the data acquisition will come from internal datasets for children that are under the care of the organization. For the project undertaken, self-populated datasets were used. The datasets used for this project were designed around the exhaustive codes taken from the National Child Abuse and Neglect Data System (NCANDS) Child File code book [8]. The datasets recorded and processed to acquire the relevant data needed for this project. All unnecessary data may be discarded as it would be irrelevant. With the datasets being self-generated, there was little need for filtering of datasets. However, real datasets would require a level of filtering and sorting that was not considered within the context of this project. The data would then be processed according to childID and be extracted in terms of different categories that involve varying levels of attention needed in the child's case.

Within the context of this project, the scope was narrowed. Within a large child welfare organization, a wider scope will yield more accurate results.

4.1.1 Child Dataset

In Figure 3, contains all information necessary about the child such as name, surname, gender, area.

1	chID	chName	chSurnam	chAge	chSex	chRace	chLvng	chMil	chPrior	area	province
2	1	Dominic	White	2	1	American	4	1	2	Masibekela	Mpumalanga
3	2	Pippa	Wright	1	2	Asian	11	1	2	KwaNgongoma	KwaZulu Natal
4	5	Lisa	Rutherfor	7	2	Hispanic	7	1	2	Makgoba	Limpopo
5	6	Benjamin	Mills	14	1	Asian	7	1	9	Mamosweu	Limpopo
6	7	Jan	Arnold	12	2	White	10	1	2	Tyusha	Eastern Cape

Figure 3: Child Dataset

However, the most important fields are the child ID (chID), child age (chAge) and the child gender (chSex). Report Dataset, which logs all reports with date of report, state/country in which the report has come in from, source of report and ends with disposition of report (rptID).

4.1.2 Child Maltreatment Dataset

Records a maltreatment cases for a child and the disposition for each as shown in Figure 4. These maltreatment records are acquired from the reports of child maltreatment initially received by the organization. Each code represents a different category of maltreatment, likely to be specified by the child welfare organization's own standard.

1	chMal1	mal1Lev	chMal2	mal2Lev	chMal3	mal3Lev	chMal4	mal4Lev	malDeath	childID	dateReported
2	6	2	3	7	8	88	4	1	2	1148	2004/01/28
3	3	1	6	88	5	4	5	5	2	1109	2016/06/16
4	9	4	9	4	3	8	6	99	1	833	2012/09/15
5	2	8	2	3	9	2	1	4	2	928	2010/05/22
6	2	6	1	88	9	3	9	7	2	307	2014/05/16

Figure 4: Child Maltreatment Dataset

4.1.3 Child Risk Factors

Figure 5, shows child risk factors which include information such as the child's physical and mental well-being, any substance abuse that the child may influenced by and any behavioural or medical issues the child may have. These risk factors may be identified as a potential effect of child maltreatment.

1	crfID	cdAlc	cdDrug	cdRtrd	cdEmotnl	cdVisual	cdLearn	cdPhys	cdBehav	cdMedicl	chID
2	1	1	1	2	2	2	1	1	1	1	83
3	2	1	2	1	1	2	2	2	2	2	44
4	3	9	2	9	9	1	2	2	9	2	42
5	4	2	2	2	2	2	1	2	1	1	91
6	5	1	2	1	2	2	2	2	2	1	55

Figure 5: Child Risk Factors

4.1.4 Caretaker Risk Factors

Figure 6 shows information containing the caretaker's physical and mental well-being, any substance abuse that the caretaker may be involved in and any behavioural or medical issues the caretaker may have. Resources the caretaker may/may not have access to, to take care of the child may also be included as a caretaker risk factor.

1	carfID	fcAlc	fcDrug	fcRtrd	fcEmotnl	fcVisual	fcLearn	fcPhys	fcMed	fcViol	fcHouse	fcMoney	fcPublic	chID
2	1	1	1	1	1	1	2	1	2	2	2	1	1	2
3	2	1	9	2	2	1	2	2	9	2	2	1	9	66
4	3	2	2	2	1	2	2	1	2	1	2	2	1	39
5	4	2	1	9	2	2	2	9	2	1	2	1	1	23
6	5	2	2	1	9	2	1	9	2	9	2	9	9	11

Figure 6: Caretaker Risk Factors

4.2 Datasets hardware implementation

The Hippo CMS is a java-based open source web content management platform [9]. The Hippo Cluster at UKZN is a 1000 core Ivy-Bridge cluster that consists of 50 nodes, each with 20 cores. Each node contains 64GB of RAM, and 1TB of local disk space. Storage can be read from simultaneously at a 1GB bandwidth.

5 Services provided for perpetrators datasets

The proposed system is using Structured Query Language (SQL) which is a standardized programming language. The main uses for SQL is adding, updating, deleting rows of dataset, and retrieving data for analytical purposes [10]. The database should contain information of the perpetrators involved in the logged report which is compared against the Child Maltreatment Dataset records [8]. The lists and all services that may be provided to the families are as follows:

5.1 Reduced child maltreatment dataset

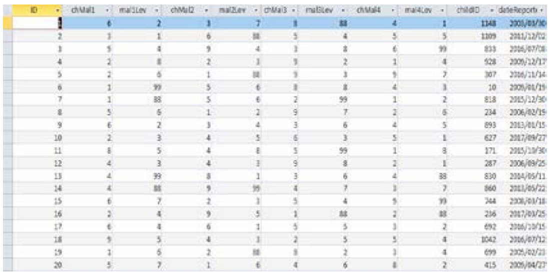
From an excerpt of our child dataset, which has a total of 1200 records, it is observable that the dataset has unnecessary information such as the child's living arrangements (chLvng), if the child has a family member in the military (chMil), or if the child is a prior victim to any maltreatment (chPrior). This would be vital information however, a child maltreatment dataset has already been identified, which makes the information redundant, and thus useless in this child dataset. It was noticed that certain children in the dataset such as the age of 18, which is the legal age of adulthood according to the "Children's Act" [11]. Therefore, these records must be deleted as they are no longer a part of the Child Welfare System. The SQL query used as follows:

```
DELETE childDB.chAge FROM childDB WHERE (((childDB.chAge)>= 18));
```

Deleting all records from the child dataset with a condition of age being over 18. This resulted in 108 records being deleted, leaving us with 1092 records to work with. The next reduction was to remove the unnecessary fields and only keep child ID (chID), the child's age (chAge) and the child's gender (chSex).

6.2 Child maltreatment dataset

In Figure 7, an excerpt of the child maltreatment dataset is shown.



ID	chMal1	mal1Lev	chMal2	mal2Lev	chMal3	mal3Lev	chMal4	mal4Lev	childID	dateReporte
1	6	2	3	7	8	88	4	1	1148	2005/03/30
2	3	1	6	88	5	4	5	5	1109	2011/12/02
3	5	4	9	4	3	8	6	99	833	2016/07/08
4	2	8	2	3	9	2	1	4	528	2009/12/17
5	2	6	1	88	9	3	9	7	307	2016/11/14
6	3	99	5	6	8	8	4	3	10	2009/01/19
7	1	88	5	6	2	99	1	2	818	2015/12/30
8	5	6	1	2	9	7	2	6	234	2006/02/15
9	6	2	3	4	3	6	4	5	893	2013/01/15
10	2	3	4	5	6	3	5	1	627	2017/09/27
11	8	5	4	8	5	99	1	8	171	2015/10/30
12	4	3	4	3	9	8	2	1	287	2006/09/25
13	4	99	8	1	3	6	4	88	830	2014/05/11
14	4	88	9	99	4	7	3	7	860	2013/05/22
15	6	7	2	3	5	4	9	99	344	2008/03/18
16	2	4	9	5	1	88	2	88	256	2017/03/25
17	6	4	6	3	5	5	3	2	692	2016/10/15
18	9	5	4	3	2	5	5	4	1042	2016/07/12
19	1	6	2	88	8	2	3	4	699	2005/02/21
20	5	7	1	6	4	6	8	2	415	2009/04/27

Figure 7: an excerpt of data from the maltreatment dataset

The above dataset correlates with the child dataset, where it is linked with the child ID. This maltreatment dataset considers the last 4 reported instances and combines these 4 instances of abuse to make a record in this dataset. Each instance of abuse is important as, together, they were used to classify into the levels of abuse that was used for graphical purposes. The dataset above is used to report instances of child abuse and is therefore linked to the child database by the child's unique identifier (childID). Each maltreatment field (mal1, mal2, mal3, mal4) must correlate to integers used to classify the type of abuse, such as physical abuse, neglect oo deprivation of necessities, medical neglect, sexual abuse, psychological or emotional maltreatment. Also, 'No alleged maltreatment', 'Other', 'Unknown' or 'Missing' are also classifiers.

Each maltreatment instance (mal1, mal2, mal3, mal4) must have a substantiation level (mal1lev, mal2lev, mal3lev, mal4lev). This must correlate to an integer used to show how well the reported maltreatment is substantiated. Date recorded field is the date whereby the maltreatments were grouped for the child identified by unique child identifier. Sub-index dataset, used for child's private datasets, uses the child's unique identifier key. With the concept of the C4.5 Algorithm [19, 25], the maltreatment dataset was narrowed to just 3 levels that are based on the maltreatment instances such as

- *Level-1*: includes all instances of physical (1) and sexual abuse (4).
- *Level-2*: includes all instances of neglect (2), medical neglect (3) and emotional maltreatment (2)
- *Level-3*: all instances that include no alleged maltreatment (6), other (8) and unknown (9)

These levels would represent the importance of the cases as level-1 should be investigated within 24 hours, level-2 within 48 hours and level-3 within 3-5 working days. The following SQL Query is used to create the level 1 dataset (malDBlev1) as follows:

```
SELECT chMal.ID, chMal.childID, childDB.province INTO malDBlev1 FROM chMal, childDB WHERE
((([chMal].[childID]=[childDB].[chID]) AND ((chMal.chMal1)=1 Or (chMal.chMal1)=4)) Or
(([chMal].[childID]=[childDB].[chID]) AND ((chMal.chMal2)=1 Or (chMal.chMal2)=4));
```

Create a new dataset, to group level 1 maltreatment and include unique identifier, maltreatment ID, child ID. Select first two instances where the maltreatment is 1 (physical abuse) or 4 (Sexual abuse) and the child ID in the maltreatment data set matches the child ID in the child dataset to obtain the province (graphical purpose). The level 1 dataset created from the physical and sexual abuse is then tallied and grouped by province which is shown in the Figure 8.

Create a new dataset, to group level 2 maltreatment and include unique identifier, maltreatment ID, child ID. Select first two instances where the maltreatment is 2 (neglect), 3 (medical neglect) or 5 (emotional maltreatment) and the child ID in the maltreatment data set matches the child ID in the child dataset to obtain the province (for display graphical purpose). The level 2 dataset was tallied and grouped by province. Create a new dataset, to group level 3 maltreatment and include unique identifier, maltreatment ID, child ID. Select first two instances where the maltreatment is is 6 (no alleged maltreatment), 8 (other) or 9 (unknown) and the child ID in the maltreatment data set matches the child ID in the child dataset to obtain the province.

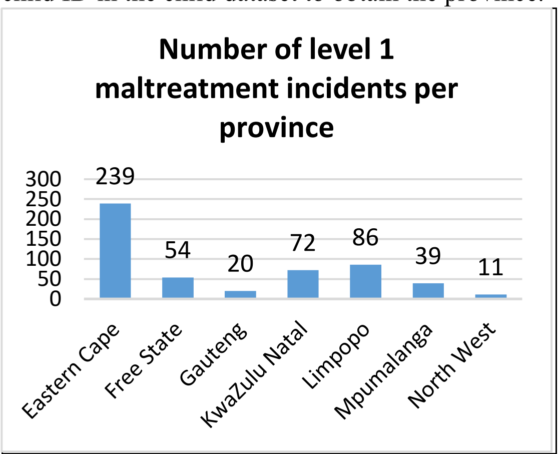


Figure 8: level 1 counts in each province

As illustrated in Figure 8, the Eastern Cape Province in South Africa is recorded as the highest in child abuse for level 1 incidents.

6 Display datasets methods

The visual methods decided upon to display the data were a bar graph as well as a line graph.

6.1 Bar Graph Method

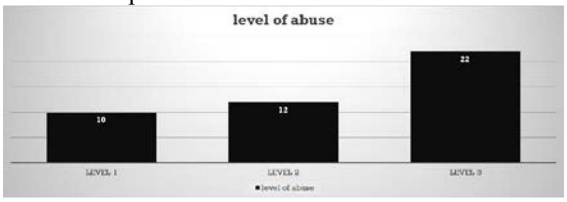


Figure 9: example of a bar graph depicting levels of abuse [12]

The bar graph as illustrated in Figure 9, depicts the percentage of each level of physical abuse, with level 1 is respond within 24 hours, level 2 is respond within 48 hours and level 3 is respond within 2-5 working days respectively [12]. After grouping the maltreatment reports into the 3 levels, each group was further analysed into total number of reports based on the province the report came from, with number of reports under each level grouped by province. Within an organisation, these groupings may be distinguished by the neighborhoods, towns, or cities, and not necessarily provinces or states.

6.2. Line Graph Method

Figure 10 breaks down the counts of maltreatment grouped by province into years spanning from 2010 to 2017. The intention is to make trends more visible so that improvement in investigations by province is more easily identified. Figure 25 shows provinces' yearly total using Datawrapper [32] to visualize the datasets.

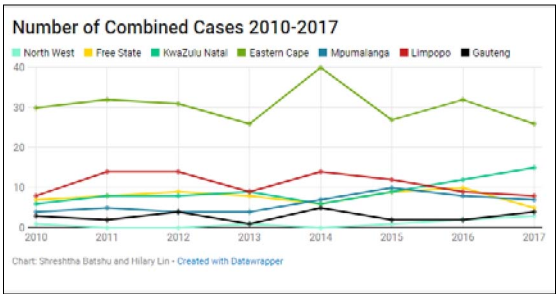


Figure 10: 2010 - 2017 data visualized [14]

From the data presented in Figure 10, it is clear that Eastern Cape has the highest percentage overall. This would indicate that investigations need to either be done more thoroughly or be taken more seriously as there is a growing problem of maltreatment in this province.

Using the averaging method to forecast number of cases that will be reported in the years 2018 – 2020 as shown in Figure 11 below:

province	Limpopo	Forecast(Limpopo)	North W	Foreca	Eastern	Foreca	Free St	Foreca	Gauten	Foreca	KwaZu	Foreca	Mpum	Foreca
2010	8		1		30		7		3		6		4	
2011	14		0		32		8		2		8		5	
2012	14		0		31		9		4		8		4	
2013	9		1		26		8		1		9		4	
2014	14		0		40		6		5		6		7	
2015	12		1		27		9		2		9		10	
2016	9		2		32		10		2		12		8	
2017	8	8	3	3	26	26	5	5	4	4	15	15	7	7
2018		8,695443365		1,232523		30,32268		7,052388		3,459056		13,86749		9,415754
2019		8,339214531		1,541886		27,27655		7,026006		3,515957		14,87294		10,09236
2020		8,088847255		1,85115		30,00317		6,959623		3,572858		15,8584		10,75897

Figure 11: Forecasted dataset

Figure 11 displays the datasets for the years 2010 to 2017 in the *solid* colors according to the keys are set by each dataset. The forecasts for each province is depicted by the *dotted* lines for the years 2018 to 2020. This will leads to descriptive analytics that creates a summary of historical data. This is useful to identify patterns and possibly assist in future data analysis. A typical question for descriptive analytics is: "What happened?" [14].

'Historical data' in terms of descriptive analysis refers to anything before the occurrence of an event, no matter how short or long the time between them. Descriptive analytics can provide historical insights that are needed to understand current events at an aggregate level [15]. Attempting to predict the future is never absolute, but predictive analysis helps draw a likely conclusion based on the current information to assist in adjusting efforts so that goals are met. This assists in being adequately prepared for possible increases or decreases in child maltreatment cases.

7 Conclusion and future research work

The use of machine learning tools and algorithms can significantly increase the efficiency of storing and processing data. This is especially important as an increase in further data, especially unstructured data, may lead to more accurate predictive analytics in child welfare cases. The machine learning algorithms, C4.5 decision tree and Apriori Algorithm, were chosen due to their obvious relation to the difficulties faced within child welfare organizations, namely the identification of trends and risk factors over large datasets. However, it is not impossible for alternate machine learning algorithms to more efficiently solve current and future problems facing child welfare organizations.

The scope of the project was limited to the cases identified to be 'Level 1' in the dataset containing the characteristic to be analyzed with the chosen algorithms and/or other algorithms in the future. Future research work is suggested to further identify and analyze 'Level 2' and 'Level 3' characteristics.

References

[1] J. McLaughlin, "The Child Welfare System: Kids Falling Through the Cracks," 06 June

2015. [Online]. Available: https://lawstreetmedia.com/issues/law-and-politics/child-welfare-systems-falling-cracks/. [Accessed on March2018]
[2] B. Marr, "Big Data: What Is Hadoop - An Easy Explanation For Absolutely Anyone," [Online]. Available: https://www.bernardmarr.com/default.asp?contentID=1080. [Accessed on March2018]
[3] "HDFS Apache Hadoop Distributed File System," [Online]. Available: https://www.ibm.com/analytics/hadoop/hdfs . [Accessed March2018].
[4] S. P, "Data locality in Hadoop: The Most Comprehensive Guide," 22 April 2017. [Online]. Available: https://data-flair.training/blogs/data-locality-in-hadoop-mapreduce/. [Accessed March2018].
[5] "techopedia," [Online]. Available: https://www.techopedia.com/definition/30427/hadoop-common. [Accessed 05 March 2018].
[6] A. Kawa, "IBM," 01 April 2014. [Online]. Available: https://www.ibm.com/developerworks/library/bd-yarn-intro/. [Accessed March 2018].
[7] K. Krishnan, "Building the Business Case for Big Data," tdwi, 12 November 2013. [Online]. Available: https://tdwi.org/Articles/2013/11/12/BI-Business-Case.aspx?Page=2. [Accessed 27 March 2018].
[8] NDACAN, "NDACAN," 09 September 2016. [Online]. Available: https://www.ndacan.cornell.edu/datasets/pdfs_user_guides/NCANDSChildFileCodebook.pdf. [Accessed 03 March 2018].
[9] HIPPO, "HIPPO CMS APPLICATION ARCHITECTURE," [Online]. Available: https://www.onehippo.org/library/architecture/hippo-cms-architecture.html. [Accessed April 2018].
[10] M. Rouse, "SQL (Structured Query Language)," September 2016. [Online]. Available: https://searchsqlserver.techtarget.com/definition/SQL. [Accessed April 2018].
[11] Brand South Africa, "Government explains new Children's Act," 2007. [Online]. Available: https://www.brandsouthafrica.com/governance/services/government-explains-new-childrens-act. [Accessed April 2018].
[12] S. Batshu and H. Lin, Artists, *Bar graph Depicting statistics of Abuse*. [Art]. UKZN, 2018.
[13] E. Putnam-Hornstein. [Online]. Available: http://www.chhs.ca.gov/Child%20Welfare/Predictive%20Risk%20Modeling.pdf. [Accessed 05 March 2018].
[14] H. Lin and S. Batshu, Artists, *Reports 2010 - 2017 grouped by province*. [Art]. UKZN, 2018.
[15] H. Lin and S. Batshu, Artists, *Forecast 2018 - 2020*. [Art]. UKZN, 2018.
[16] M. Rouse, "Descriptive Analytics," December 2015. [Online]. Available: http://whatis.techtarget.com/definition/descriptive-analytics. [Accessed March 2018].
[17] "Descriptive, Predictive, and Prescriptive Analytics Explained," [Online]. Available: https://halobi.com/blog/descriptive-predictive-and-prescriptive-analytics-explained/. [Accessed March 2018].
[18] MathWorks, "Machine learning technique for finding hidden patterns or intrinsic structures in data," [Online]. Available: https://www.mathworks.com/discovery/unsupervised-learning.html. [Accessed March 2018].
[19] A. M. E. E. Badr HSSINA, "A comparative study of decision tree ID3 and C4.5," [Online]. http://saiconference.com/Downloads/SpecialIssueNo10/Paper_3-A_comparative_study_of_decision_tree_ID3_and_C4.5.pdf. [Accessed March 2018].
[20] *Response Time*. [Art]. 2018
[21] R. M. R. Masud Karim, "Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing," *Journal of Software Engineering and Applications*, vol. VI, no. 4, p. 11, 2013.
[22] R. Jain, "A beginner's tutorial on the apriori algorithm in data mining with R implementation," 24 March 2017. [Online]. Available: https://www.hackerearth.com/blog/machine-learning/beginners-tutorial-apriori-algorithm-data-mining-r-implementation/. [Accessed March 2018].
[23] R. A. a. R. Srikant, "Mining Frequent Itemsets using the Apriori Algorithm," 1994. [Online]. Available: https://www.philippe-fournier-viger.com/spmf/Apriori.php. [Accessed March 2018].
[24] R. Jain, "A beginner's tutorial on the apriori algorithm in data mining with R implementation," 24 March 2017. [Online]. Available: https://www.hackerearth.com/blog/machine-learning/beginners-tutorial-apriori-algorithm-data-mining-r-implementation/. [Accessed 9 March 2018].
[25] "C4.5 Decision Tree Implementation," [Online]. Available: http://www.otnira.com/2013/03/25/c4-5/. [Accessed March 2018]

[26] L.-Y. C. E. H. Y.-D. L. C.-H. H. H.-W. C. Jen-Yang Tang, “Identifying the Association Rules between Clinicopathologic Factors and Higher Survival Performance in Operation-Centric Oral Cancer Patients Using the Apriori Algorithm,” Hindawi Publishing Corporation, 2013

[27] http://www.dsd.gov.za/index.php?option=com_content&task=view&id=89 . [Accessed on June2018]

[28] Noor Maizura Mohamad Noor ; Salwana Mohamad @ Asmara, Intelligent Interpretation and Analysis of Child Sexual Abuse Forensic Evidence: A Preliminary Study, 2010 International Symposium on Information Technology, 02 September 2010, Kuala Lumpur, Malaysia, DOI**:** 10.1109/ITSIM.2010.5561610.

[29] Guido van Heck, Itamar Sharon, Paulus Kampert and Jan van den Berg, Introducing Electronic Child Records: Balancing Personal Interests, System Performance, and Social Values, 2009 International Conference on Computational Science and Engineering, Vancouver, Canada 29-31 August 2009, PP: 568 – 575, DOI 10.1109/CSE.2009.42

[30] Y. Jewkes, C. Andrews, “Policing the filth: the problems of investigating online child pornography in England and Wales”, *Policing and Society*,vol. 15, no. 1, pp. 42-62, 2005.

[31] F. Sebastiani, “Machine learning in automated text categorization”, *ACM Computing Surveys*, vol. 34, pp. 147, 2002

[32] https://www.datawrapper.de/ . [Accessed on july2018].

[33] https://www.ndacan.cornell.edu/. [Accessed on july2018].

[34] B. Lery, JM. Haight, L.Alpert, “Four Principles of Big Data Practice for Effective Child Welfare Decision Making”, *Journal of Public Child Welfare*, vol. 10, no.4, pp. 466-474, 2006