BEHAVIOUR ANALYSIS OF CYBER ATTACKS USING KNIME

A Project

Report

Submitted in the partial fulfilment of the requirements for

the award of the degree of

Bachelor of Technology

in

Department of Computer Science and Engineering

by

N.Sri Vaishnavi 180031168

C.Gayathri 180030873

B.Sahithya 180031153

under the supervision of

Dr.Rajesh Pasupuleti

Associate Professor, Department of CSE

Department of Computer Science and Engineering

K L E F, Green Fields,

Vaddeswaram – 522502, Guntur(Dist), Andhra Pradesh, India.

November, 2021

Declaration

The Project Report entitled “BEHAVIOUR ANALYSIS OF CYBER ATTACKS USING KNIME” is a record of bonafide work of N.SRI VAISHNAVI-180031168, C.GAYATHRI-180030873, B.SAHITHYA-180031153 submitted in partial fulfillment for the award of B.Tech in Computer Science and Engineering to the K L University. The results embodied in this report have not been copied from any other departments/University/Institute.

<Signature of the Students >

Certificate

This is to certify that the Project Report entitled "BEHAVIOUR ANALYSIS OF CYBER ATTACKS USING KNIME " is being submitted by N.SRI VAISHNAVI-180031168, C.GAYATHRI-180030873, B.SAHITHYA-180031153 submitted in partial fulfillment for the award of B.Tech in Computer Science and Engineering to the K L University is a record of bonafide work carried out under our guidance and supervision.

The results embodied in this report have not been copied from any other departments/ University/Institute.

Signature of the Supervisor

Name and Designation

Signature of theHOD

Signature of the External Examiner

Acknowledgement

I would like to acknowledge and give my warmest thanks to my supervisor **Dr.Rajesh Pasupuleti** who made this work possible. Her guidance and advice carried me through all the stages of writing my project. I would also like to thank my committee members for letting my defence be an enjoyable moment, and for your brilliant comments and suggestions, thanks to you.

I would also like to give special thanks to our Head of the department of CSE, **Mr.Hari KiranVege** and my section incharges as a whole for their continuous support and understanding when undertaking my research and writing my project. Your prayer for me was what sustained me this far.

Finally, thank you for all my fellow friends who helped me out in all difficulties. I have experienced my supervisor guidance day by day. You are the one who let me finish my project with gaining lots of knowledge.

Date:

N.SRI VAISHNAVI 180031168

C.GAYATHRI 180030873

B.SAHITHYA 180031153

ABSTRACT

Stories of cyber attacks are becoming a routine in which cyber attackers show new levels of intention by sophisticated attacks on networks. Unfortunately, cybercriminals have figured out profitable business models and they take advantage of the online anonymity. A serious situation that needs to improve for networks' defenders.

In this project, we show that better results can be obtained by performing behavioural analysis on higher semantic level. We model this behaviour by creating customized normalcy profile of this system and evaluate how well does anomaly based detection work in this scenario.

This project is an effort to provide a review of relevant theories and principles, and gives insights including an interdisciplinary framework that combines behavioural cybersecurity, human factors, and modeling and simulation.

TABLE OF CONTENTS

INTRODUCTION

The asymmetric nature and ever-increasing degree of sophistication of cyber threats drive the need for assurance of critical infrastructure and systems. The conventional belief was that cyber attacks on critical infrastructures designed as embedded systems are of no concern because they were seldom connected to the network from which these attacks were enabled Behavioural analysis offers a more promising approach to malware detection since behavioural signatures are more obfuscation resilient than the binary ones. Indeed, changing behaviour while preserving the desired (malicious) functions of a program is much harder than changing only the binary structure. More importantly, to achieve its goal, malware usually has to perform some system operations.

Behavioural Analysis:

- Malicious attacks have one thing in common - they all behave differently than normal everyday behaviour within a system or network. Companies can often identify malicious behaviours through signatures that are directly related to certain types of well-known attacks. However, as attackers get more sophisticated, they continually develop new tactics, techniques, and procedures (TTPs) that allow them not only to enter vulnerable environments, but also to move laterally undetected.
- With the help of massive volumes of unfiltered endpoint data, security personnel can now use behavioural-based tools, algorithms, and machine learning to determine what the normal behaviour of everyday users is - and what it is not. Behavioural analysis can identify events, trends, and patterns - both current and historic - that are outside the parameters of everyday norms.
- By zeroing in on these anomalies, security teams can gain visibility and identify unexpected behavioural tactics of attackers early on, before they fully execute their plan of attack. Behavioural analysis can also help uncover root causes and provide insights for future identification and prediction of similar attacks.
- ABA therapy applies our understanding of how behaviour works to real situations. The goal is to increase behaviours that are helpful and decrease behaviours that are harmful or affect learning.

Machine Learning for Behavioural Analysis:

In the case of behaviour analysis and anomaly detection, a modern threat detection software may use a mix of ML techniques.

For example, a solution may use Classification in a Supervised ML algorithms to identify spam based on email content, Regression algorithms to dynamically identify risk levels while using the same software may use Unsupervised ML techniques to detect anomalies in data streams like network traffic.

Advantages of ML:

- Less supervision
- Scalability
- Establish correlation & regression
- Reduced number of false positives
- Faster detection and response time
- Continuous improvement

Importance of Behaviour Analysis:

Behavioral analytics is crucial in optimizing your company's conversion, engagement, and retention. With the right behavioral analytics tool, every member of your team should be able to gain the actionable insights they need to answer their own questions and leverage data in ways that didn't seem possible before.

LITERATURE SURVEY

[1]Architectural and Behavioral Analysis for Cyber Security:

In this we describe our tool for incorporating cyber security resiliency analysis and recommendations in the system design process that are automated, scalable, provide rich feedback, specify trade-offs and are easy to use by system architects. For this we abstract threat models in terms of an instrumentor that incorporates the effects of the threats. This allows us to aggregate classes of threats with the same effect so that they can be addressed at the effect.

[2]Review and insight on the behavioral aspects of cybersecurity:

In this paper they put effort to provide a review of relevant theories and principles, and gives insights including an interdisciplinary framework that combines behavioral cybersecurity, human factors, and modeling and simulation.

[3]Using Behavioral Modeling And Customized Normalcy Profiles As Protection Against Targeted Cyber-Attacks:

In this they observe that many critical computer systems serve a specific purpose and are expected to run strictly limited sets of software. We model this behavior by creating customized normalcy profile of this system and evaluate how well does anomaly based detection work in this scenario.

[4]Behavioral Analysis of Insider Threat: A Survey :

This paper starts by presenting a broad, multidisciplinary survey of insider threat capturing contributions from computer scientists, psychologists, criminologists, and security practitioners.they develop bootstrapping algorithms that learn from highly imbalanced data, mostly unlabeled, and almost no history of user behavior from an insider threat perspective.

THEORITICAL ANALYSIS

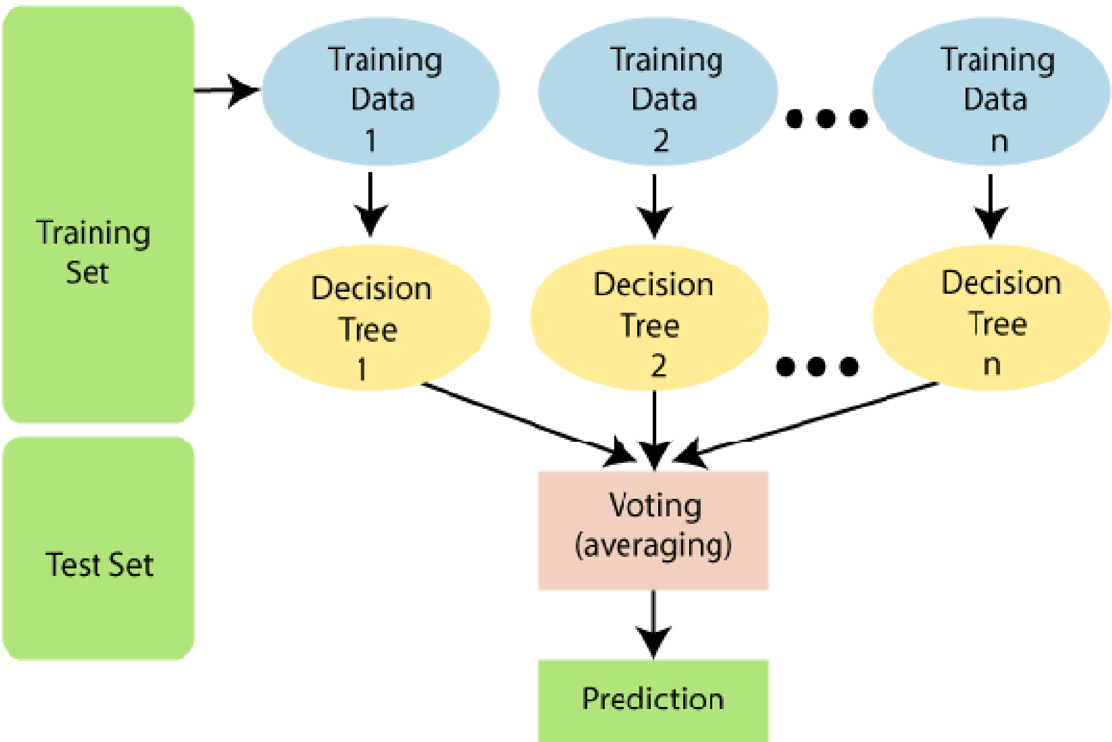
RANDOM FOREST:

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification.

Advantages:

- It performs better results for classification problems.
- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.
- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the over fitting issue.



TREE ENSEMBLE:

Ensemble methods, which combines several decision trees to produce better predictive performance than utilizing a single decision tree. The main principle behind the ensemble model is that a group of weak learners come together to form a strong learner.

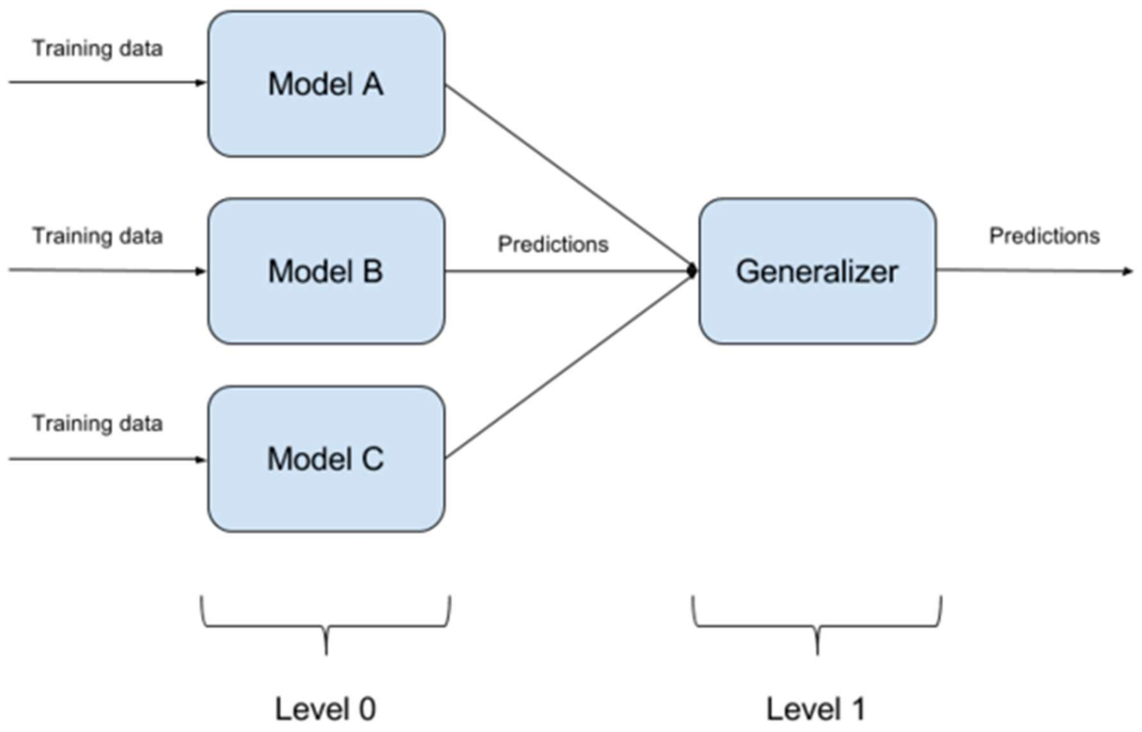
Let's talk about few techniques to perform ensemble decision trees:

- Bagging
- Boosting

Bagging (Bootstrap Aggregation) is used when our goal is to reduce the variance of a decision tree. Here idea is to create several subsets of data from training sample chosen randomly with replacement. Now, each collection of subset data is used to train their decision trees. As a result, we end up with an ensemble of different models.

Boosting is another ensemble technique to create a collection of predictors. In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analyzing data for errors. In other words, we fit consecutive trees (random sample) and at every step, the goal is to solve for net error from the prior tree.

An ensemble of trees are built one by one and individual trees are summed sequentially. Next tree tries to recover the loss (difference between actual and predicted values).



K-MEANS:

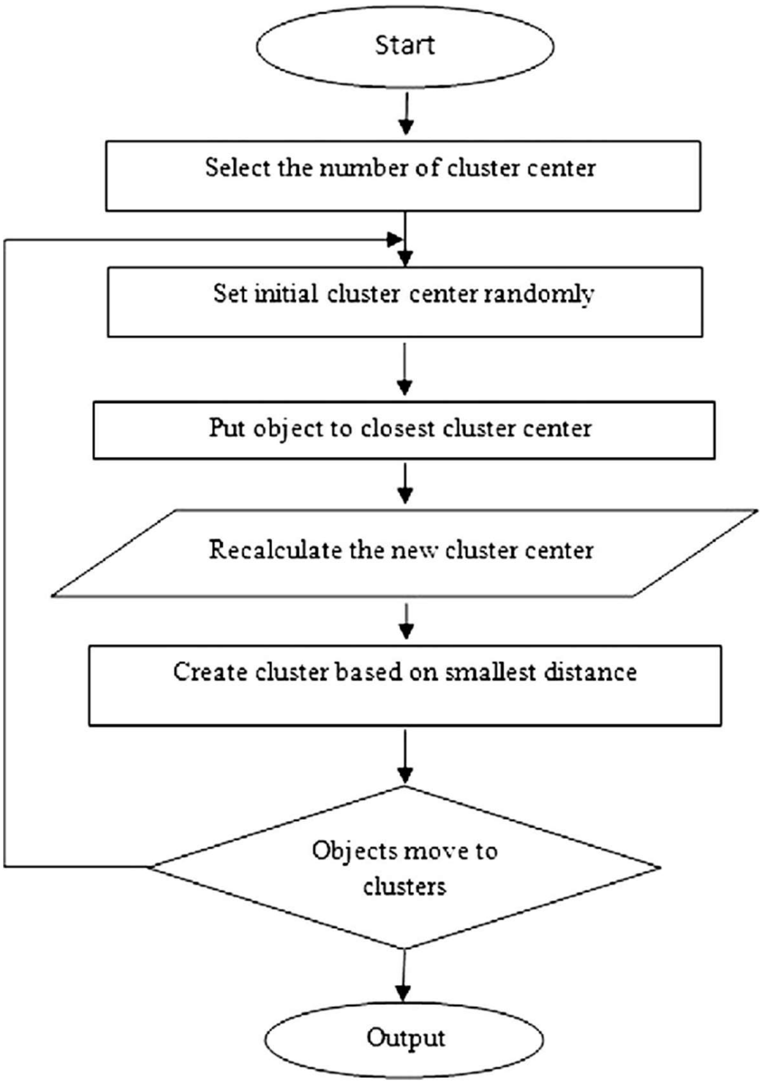
K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away from other clusters.



HIERARCHICAL CLUSTERING:

Hierarchical Clustering Algorithm also called Hierarchical cluster analysis or HCA is an unsupervised clustering algorithm which involves creating clusters that have predominant ordering from top to bottom.

For e.g: All files and folders on our hard disk are organized in a hierarchy.

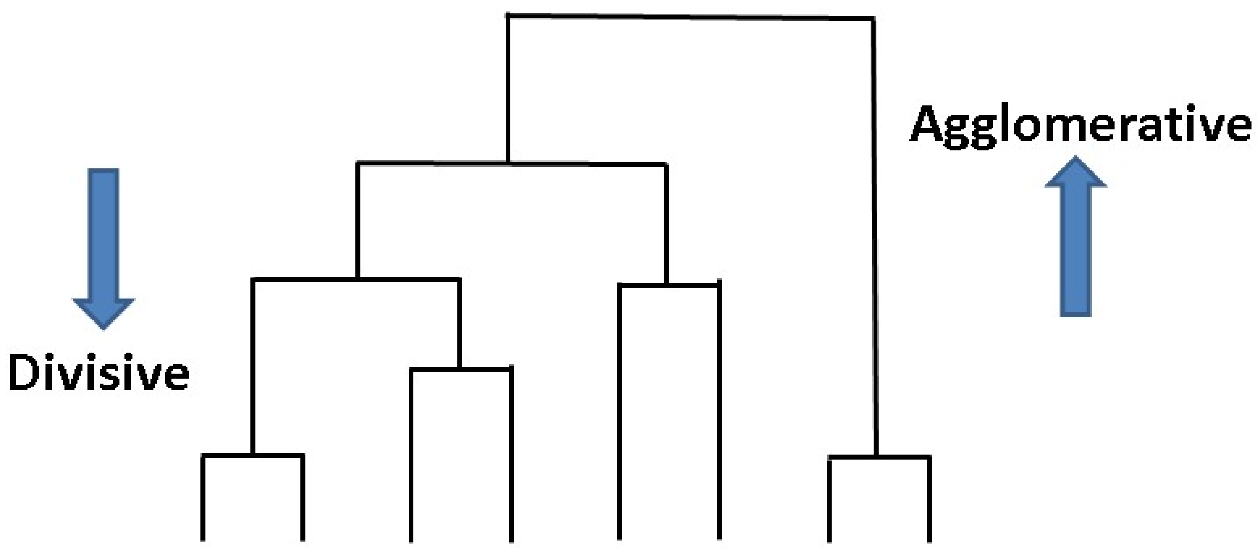
The algorithm groups similar objects into groups called clusters. The endpoint is a set of clusters or groups, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

This clustering technique is divided into two types:

- Agglomerative Hierarchical Clustering
- Divisive Hierarchical Clustering

Agglomerative: Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.

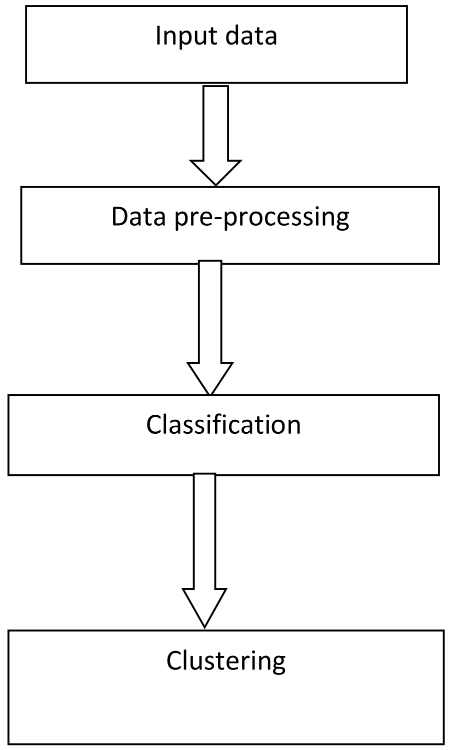
Divisive: Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.



PROCEDURE

Applying Data Pre Processing techniques and understanding the data and then performing data visualisation. Finally, classification of data and further clustering it and drawing valuable insights.

DIAGRAM



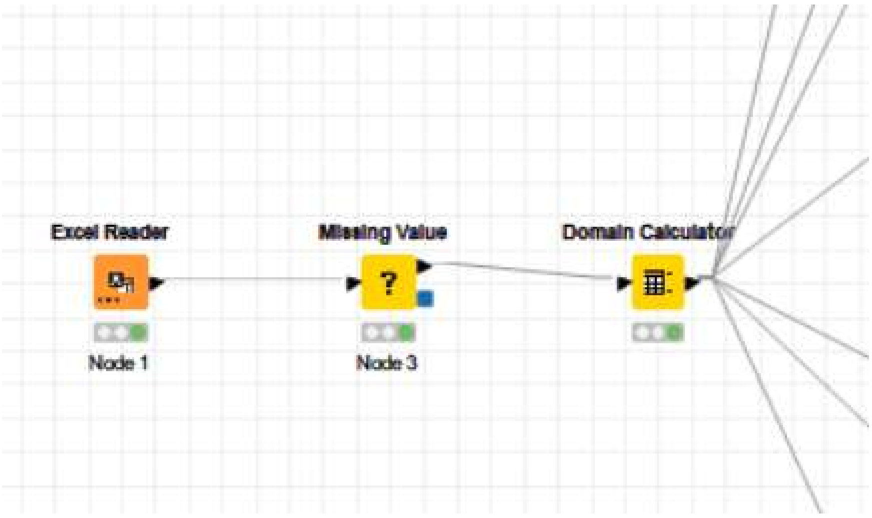
IMPLEMENTATION

We have done this project with the help of a tool called KNIME.

KNIME Analytics Platform is the open source software for creating data science. Intuitive, open, and continuously integrating new developments, KNIME makes understanding data and designing data science workflows and reusable components accessible to everyone.

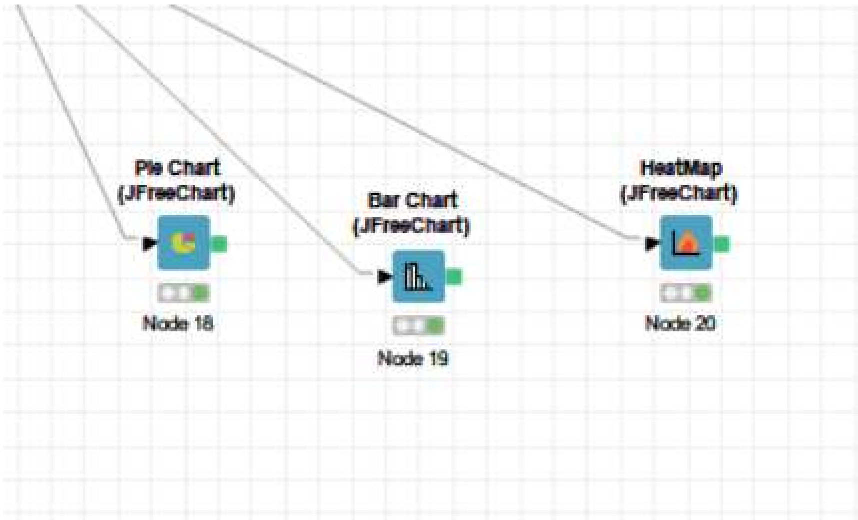
1. **Data loading and pre-processing**

   In this step the data set is being loaded and cleaned and processed for further analysis to be done over it.



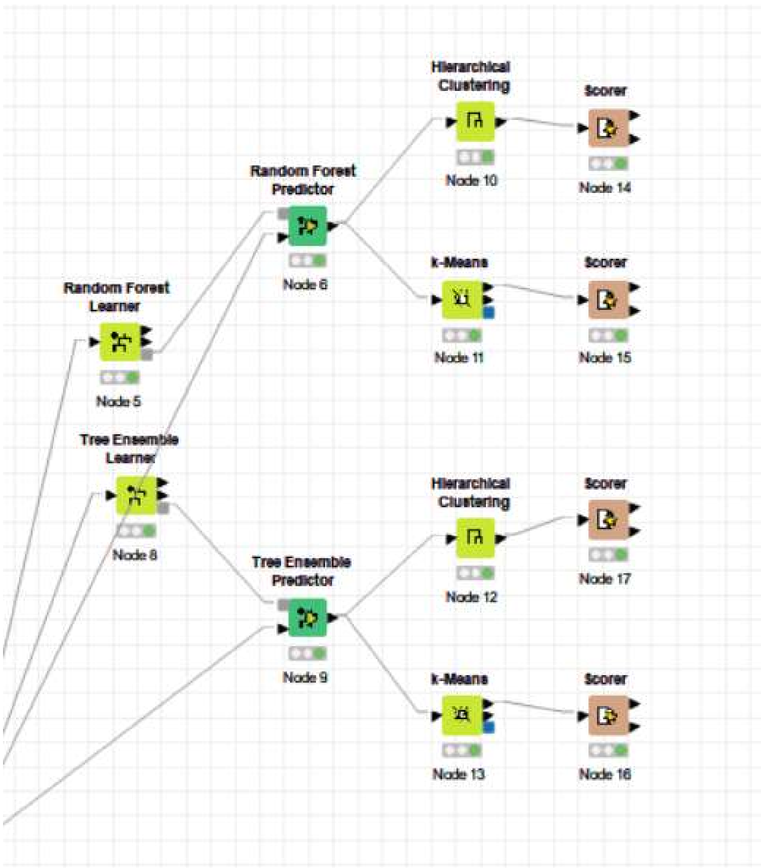
2. **Data Visualization and Understanding of Data**

   In this step we are using Data Visualisation techniques and representing the data as several charts and studying the data for better further analysis. We used Pie Chart, Bar Chart and Heat Map for understanding of the various columns present in the data set (Method, Entity, Records, Year, Organisation Type).



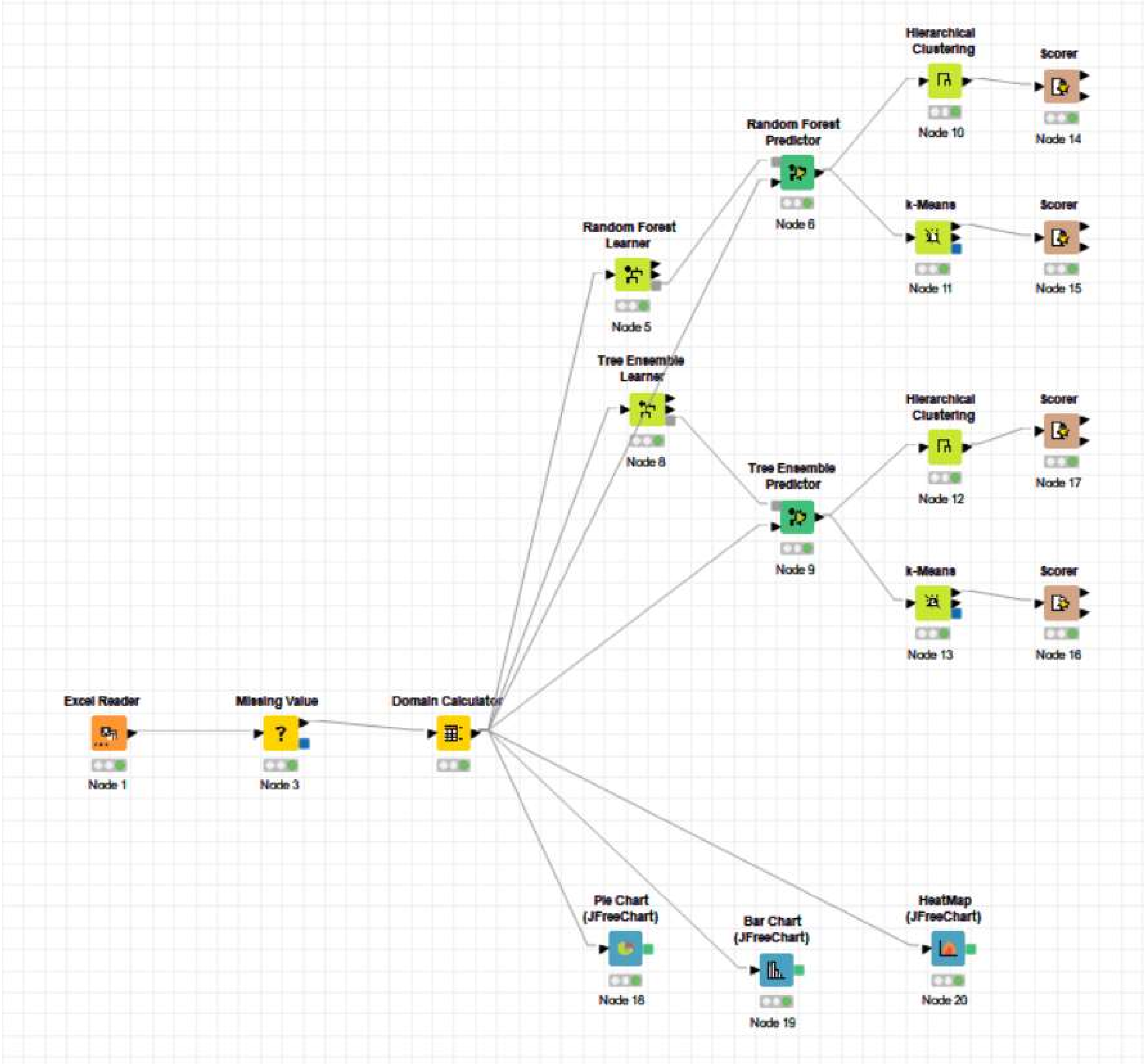
3. Classification and Clustering of Data

Here we are Classifying using Random Forest and Tree Ensemble algorithms and further Clustering the data using k-means and Hierarchical Clustering techniques and scoring our models and observing the accuracy.



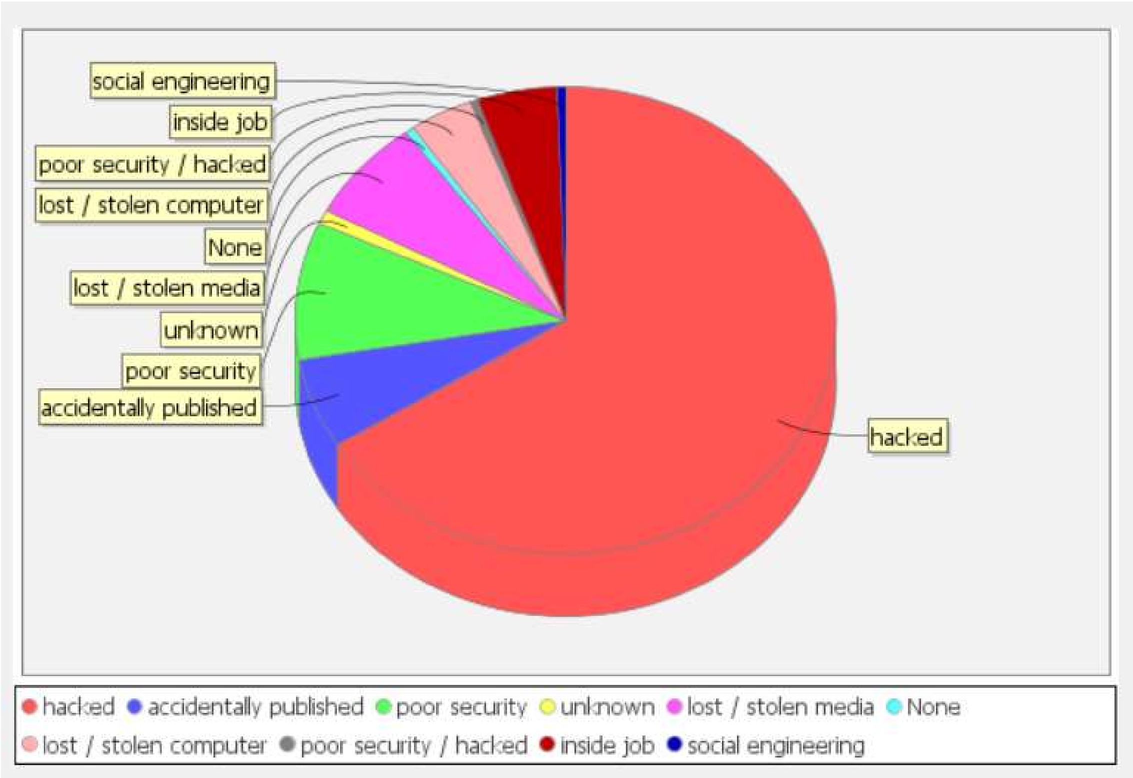
Complete Workflow

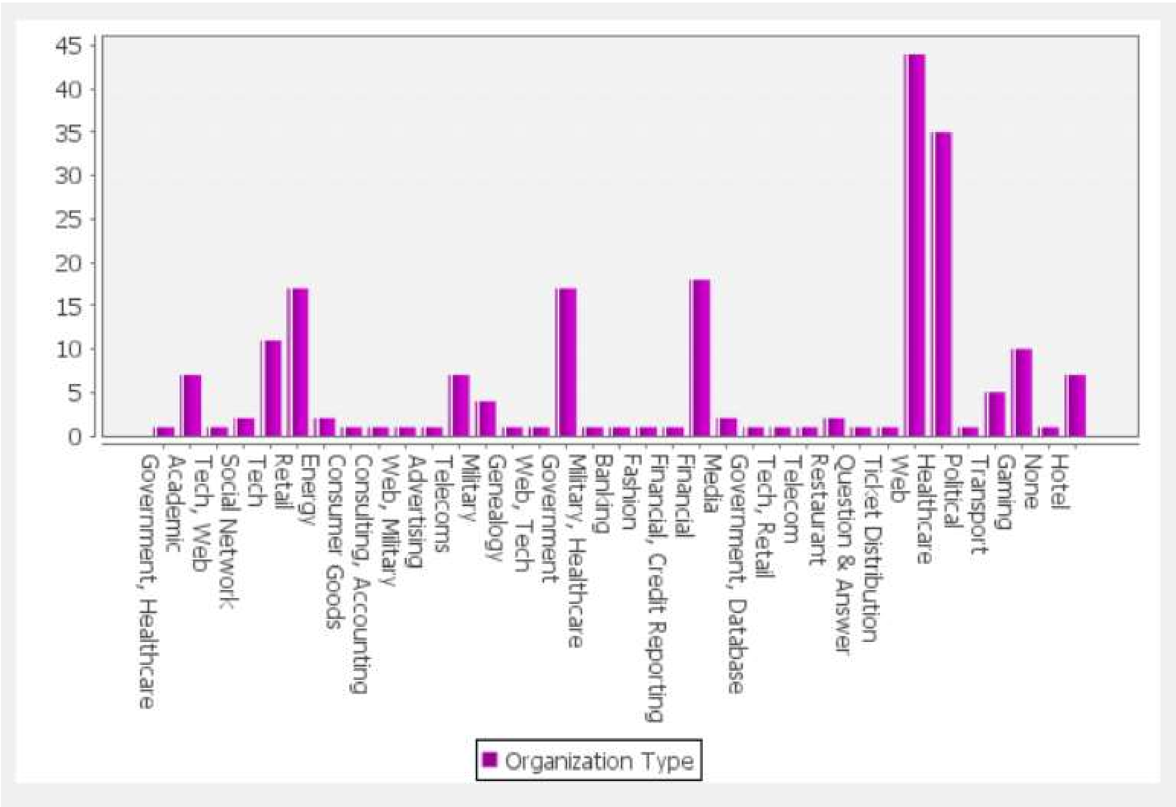
Below is the Complete Workflow build using KNIME for the analysis .

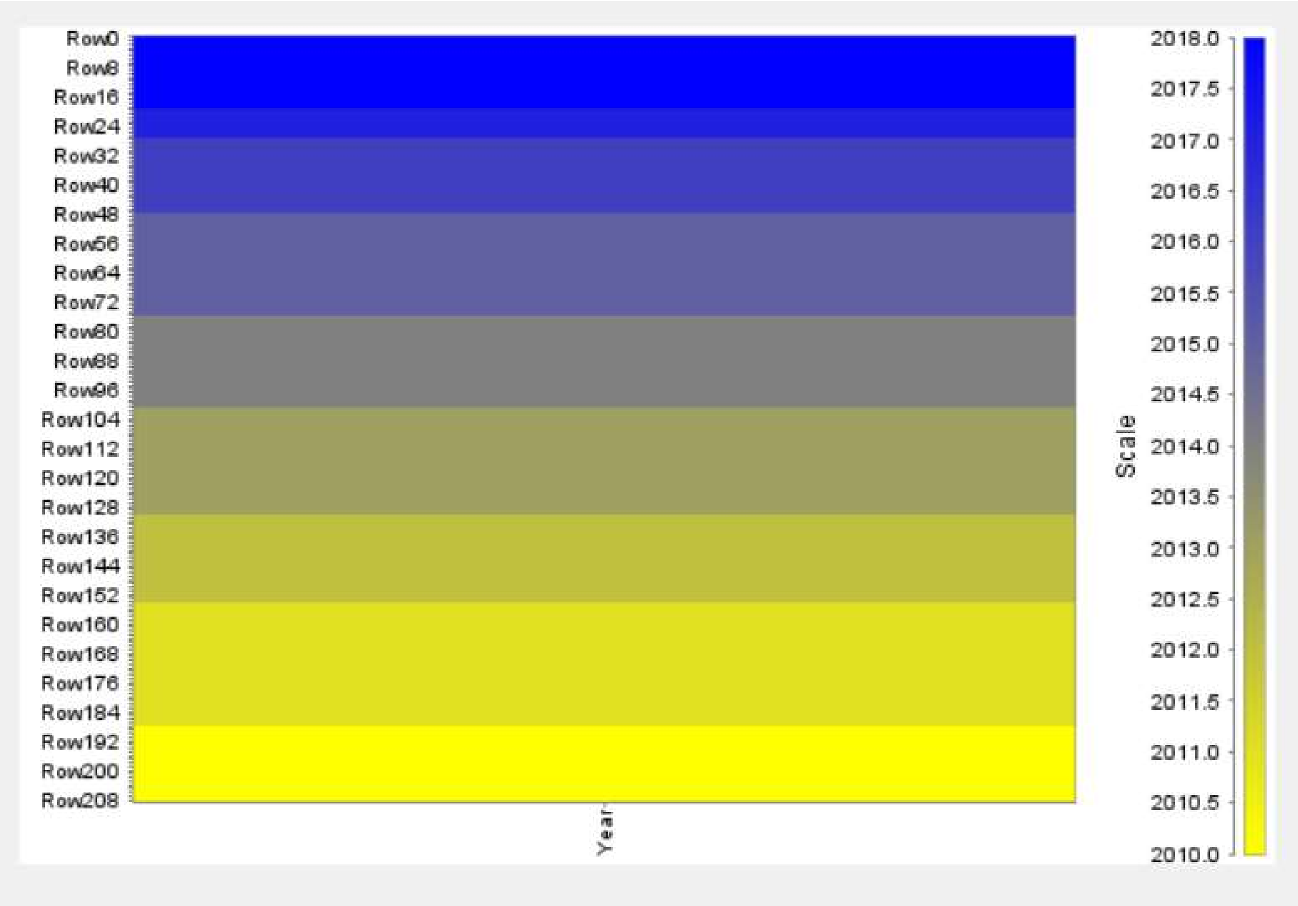


RESULTS

Data Visualisation:






Row0
Row8
Row16
Row24
Row32
Row40
Row48
Row56
Row64
Row72
Row80
Row88
Row96
Row104
Row112
Row120
Row128
Row136
Row144
Row152
Row160
Row168
Row176
Row184
Row192
Row200
Row208
Year
Scale
2018.0
2017.5
2017.0
2016.5
2016.0
2015.5
2015.0
2014.5
2014.0
2013.5
2013.0
2012.5
2012.0
2011.5
2011.0
2010.5
2010.0

Algorithms Analysis :

Confusion Matrix

Table "spec_name" - Rows: 10 | Spec - Columns: 10 | Properties | Flow Variables

Row ID	hacked	acciden…	poor se…	lost / st…	lost / st…	inside job	unknown	None	poor se…	social e…
hacked	134	0	0	4	0	0	0	0	0	0
accidentally p…	9	4	0	0	0	0	0	0	0	0
poor security	13	0	6	1	0	0	0	0	0	0
lost / stolen m…	6	0	0	9	0	0	0	0	0	0
lost / stolen c…	3	0	0	4	1	0	0	0	0	0
inside job	6	0	0	0	0	4	0	0	0	0
unknown	2	0	0	0	0	0	0	0	0	0
None	1	0	0	0	0	0	0	0	0	0
poor security …	1	0	0	0	0	0	0	0	0	0
social enginee…	1	0	0	0	0	0	0	0	0	0

Accuracy Statistics

Table "default" - Rows: 11 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo…	FalsePo…	TrueNe…	FalseN…	Recall	Precision	Sensitivity	Specificity	F-meas…	Accuracy	Cohen'…
hacked	133	41	30	5	0.964	0.764	0.964	0.423	0.853	?	?
accidentally p…	4	0	196	9	0.308	1	0.308	1	0.471	?	?
poor security	6	0	189	14	0.3	1	0.3	1	0.462	?	?
lost / stolen m…	11	10	184	4	0.733	0.524	0.733	0.948	0.611	?	?
None	1	0	208	0	1	1	1	1	1	?	?
inside job	3	0	199	7	0.3	1	0.3	1	0.462	?	?
unknown	0	0	207	2	0	?	0	1	?	?	?
lost / stolen c…	0	0	201	8	0	?	0	1	?	?	?
poor security …	0	0	208	1	0	?	0	1	?	?	?
social enginee…	0	0	208	1	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.756	0.443

Random Forest

Row ID	S Entity	I Year	S Records	S Organiz...	S Method	S Predicti...	D Predicti...
Row0	AerServ (subsidiary of InMobi)	2018	75000	Advertising	hacked	hacked	1
Row1	Bethesda Game Studios	2018	0	Gaming	accidentally ...	hacked	0.8
Row2	BlankMediaGames	2018	7633234	Gaming	hacked	hacked	0.8
Row3	BMO and Simplii	2018	90000	Banking	poor security	poor security	0.6
Row4	British Airways	2018	380000	Transport	hacked	hacked	0.95
Row5	Cathay Pacific Airways	2018	9400000	Transport	hacked	hacked	0.95
Row6	Centers for Medicare & Medi...	2018	75000	Healthcare	hacked	hacked	0.81
Row7	Facebook	2018	50000000	Social Network	poor security	poor security	0.84
Row8	Google Plus	2018	500000	Social Network	poor security	poor security	0.84
Row9	Marriott International	2018	500000000	Hotel	hacked	hacked	0.99
Row10	MyHeritage	2018	92283889	Genealogy	unknown	hacked	0.44
Row11	Orbitz	2018	880000	Web	hacked	hacked	0.98
Row12	Popsugar	2018	123857	Fashion	hacked	hacked	0.99
Row13	Quora	2018	100000000	Question & ...	hacked	hacked	1
Row14	Reddit	2018	unknown	Web	hacked	hacked	0.98
Row15	SingHealth	2018	1500000	Government...	hacked	hacked	0.99
Row16	Ticketfly (subsidiary of Even...	2018	26151608	Ticket Distri...	hacked	hacked	1
Row17	Typeform	2018	unknown	Tech	poor security	poor security	0.53
Row18	Under Armour	2018	150000000	Consumer G...	hacked	hacked	0.99
Row19	Wordpress	2018	0	None	hacked	hacked	1
Row20	Defense Integrated Data Ce...	2017	235 GB	Military	hacked	hacked	0.8
Row21	Deloitte	2017	0	Consulting, ...	poor security	poor security	0.62
Row22	Erie County Medical Center	2017	unknown	Healthcare	poor security	hacked	0.71
Row23	Equifax	2017	143000000	Financial, Cr...	poor security	poor security	0.63
Row24	Grozio Chirurgija	2017	25000	Healthcare	hacked	hacked	0.71
Row25	Heathrow Airport	2017	2.5GB	Transport	lost / stolen ...	hacked	0.66
Row26	Taringa!	2017	28722877	Web	hacked	hacked	0.92
Row27	Uber	2017	57000000	Transport	hacked	hacked	0.66
Row28	21st Century Oncology	2016	2200000	Healthcare	hacked	hacked	0.77
Row29	Apple Health Medicaid	2016	91000	Healthcare	poor security	hacked	0.77
Row30	Central Coast Credit Union	2016	60000	Financial	hacked	hacked	0.96
Row31	Philippines Commission on El...	2016	55000000	Government	hacked	hacked	0.79
Row32	Cox Communications	2016	40000	Telecoms	hacked	hacked	0.98
Row33	Democratic National Committee	2016	19252	Political	None	hacked	0.53
Row34	US Department of Homeland...	2016	30000	Government	poor security	hacked	0.79
Row35	EyeWire	2016	unknown	Tech	lost / stolen ...	hacked	0.54
Row36	Friend Finder Networks	2016	412214295	Web	poor securit...	hacked	0.97
Row37	Gyft	2016	unknown	Web	hacked	hacked	0.97
Row38	Inuvik hospital	2016	6700	Healthcare	inside job	hacked	0.77
Row39	KM.RU	2016	1500000	Web	hacked	hacked	0.97
Row40	Nival Networks	2016	1500000	Gaming	hacked	hacked	0.99
Row41	Ofcom	2016	unknown	Telecom	inside job	inside job	0.48
Row42	Rosen Hotels	2016	unknown	Hotel	hacked	hacked	1
Row43	Taobao	2016	20000000	Retail	hacked	hacked	1
Row44	TaxSlayer.com	2016	unknown	Web	hacked	hacked	0.97
Row45	University of California, Berk...	2016	80000	Academic	hacked	hacked	0.99

Tree Ensemble

Row ID	S Entity	I Year	S Records	S Organiz...	S Method	S Predicti...	D Predicti...
Row0	AerServ (subsidiary of InMobi)	2018	75000	Advertising	hacked	hacked	0.99
Row1	Bethesda Game Studios	2018	0	Gaming	accidentally ...	hacked	0.84
Row2	BlankMediaGames	2018	7633234	Gaming	hacked	hacked	0.84
Row3	BMO and Simplii	2018	90000	Banking	poor security	poor security	0.58
Row4	British Airways	2018	380000	Transport	hacked	hacked	0.92
Row5	Cathay Pacific Airways	2018	9400000	Transport	hacked	hacked	0.92
Row6	Centers for Medicare & Medi...	2018	75000	Healthcare	hacked	hacked	0.88
Row7	Facebook	2018	50000000	Social Network	poor security	poor security	0.71
Row8	Google Plus	2018	500000	Social Network	poor security	poor security	0.71
Row9	Marriott International	2018	500000000	Hotel	hacked	hacked	1
Row10	MyHeritage	2018	92283889	Genealogy	unknown	hacked	0.46
Row11	Orbitz	2018	880000	Web	hacked	hacked	0.98
Row12	Popsugar	2018	123857	Fashion	hacked	hacked	0.99
Row13	Quora	2018	100000000	Question & ...	hacked	hacked	1
Row14	Reddit	2018	unknown	Web	hacked	hacked	0.98
Row15	SingHealth	2018	1500000	Government...	hacked	hacked	1
Row16	Ticketfly (subsidiary of Even...	2018	26151608	Ticket Distri...	hacked	hacked	1
Row17	Typeform	2018	unknown	Tech	poor security	poor security	0.52
Row18	Under Armour	2018	150000000	Consumer G...	hacked	hacked	0.99
Row19	Wordpress	2018	0	None	hacked	hacked	1
Row20	Defense Integrated Data Ce...	2017	235 GB	Military	hacked	hacked	0.77
Row21	Deloitte	2017	0	Consulting, ...	poor security	poor security	0.56
Row22	Erie County Medical Center	2017	unknown	Healthcare	poor security	hacked	0.78
Row23	Equifax	2017	143000000	Financial, Cr...	poor security	poor security	0.67
Row24	Grozio Chirurgija	2017	25000	Healthcare	hacked	hacked	0.78
Row25	Heathrow Airport	2017	2.5GB	Transport	lost / stolen ...	hacked	0.77
Row26	Taringa!	2017	28722877	Web	hacked	hacked	0.97
Row27	Uber	2017	57000000	Transport	hacked	hacked	0.77
Row28	21st Century Oncology	2016	2200000	Healthcare	hacked	hacked	0.81
Row29	Apple Health Medicaid	2016	91000	Healthcare	poor security	hacked	0.81
Row30	Central Coast Credit Union	2016	60000	Financial	hacked	hacked	0.96
Row31	Philippines Commission on El...	2016	55000000	Government	hacked	hacked	0.74
Row32	Cox Communications	2016	40000	Telecoms	hacked	hacked	1
Row33	Democratic National Committee	2016	19252	Political	None	None	0.47
Row34	US Department of Homeland...	2016	30000	Government	poor security	hacked	0.74
Row35	EyeWire	2016	unknown	Tech	lost / stolen ...	hacked	0.52
Row36	Friend Finder Networks	2016	412214295	Web	poor securit...	hacked	0.97
Row37	Gyft	2016	unknown	Web	hacked	hacked	0.97
Row38	Inuvik hospital	2016	6700	Healthcare	inside job	hacked	0.81
Row39	KM.RU	2016	1500000	Web	hacked	hacked	0.97
Row40	Nival Networks	2016	1500000	Gaming	hacked	hacked	0.96
Row41	Ofcom	2016	unknown	Telecom	inside job	hacked	0.46
Row42	Rosen Hotels	2016	unknown	Hotel	hacked	hacked	1
Row43	Taobao	2016	20000000	Retail	hacked	hacked	1
Row44	TaxSlayer.com	2016	unknown	Web	hacked	hacked	0.97
Row45	University of California, Berk...	2016	80000	Academic	hacked	hacked	1

CONCLUSION

Behavioural aspects of cyber security are becoming a vital area to research. The unpredictable nature of human behaviour and actions make Human an important element and enabler of the level of cyber security. The goal from discussing reviewed theories is to underscore importance of social behaviour, environment, biases, perceptions, deterrence, intent, attitude, norms, alternatives, sanctions, decision making, etc; The implementation of the described approach includes the development and periodic updating of the normalcy profile, and the on-going tasks of the functionality extraction, detection of known malicious functionalities, and the anomaly detection in network operation.

FUTURE SCOPE

These algorithms are flexible and can easily adapt to any changes in the environment and can solve a wide range of complex problems in easy way. Behavioural models are generally used for analysis and study without complexity.

REFERENCES:

[1] Behavioural analysis of insider threat: A survey and bootstrapped prediction in imbalanced data:

Azaria A, Richardson A, Kraus S, Subrahmanian VS

https://ieeexplore.ieee.org/document/7010900

[2] Using Behavioural Modeling And Customized Normalcy Profiles As Protection Against Targeted Cyber-Attacks :

Andrey Dolgikh, Tomas Nykodym, Victor Skormin, and Zachary Birnbaum Binghamton University, Binghamton, NY, USA

https://ia.binghamton.edu/publication/SkorminPDF/StPetersburg2012_03.pdf

[3] Review and insight on the behavioural aspects of cybersecurity :

Rachid Ait Maalem Lahcen1*, Bruce Caulkins2, Ram Mohapatra1 and Manish Kumar3

https://cybersecurity.springeropen.com/articles/10.1186/s42400-020-00050-w

[4] Architectural and Behavioural Analysis for Cyber Security:

Kit Siu; Abha Moitra; Meng Li; Michael Durling; Heber Herencia-Zapana; John Interrante; Baoluo Meng; Cesare Tinelli

https://ieeexplore.ieee.org/abstract/document/9081652