National College of Ireland

Using IP address as a unique attribute for data deduplication in Network Devices

MSc Research Project
Cyber Security

Bhavana Bhavya
Student ID: 20128126

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Bhavana Bhavya		
Student ID:	20128126		
Programme:	Cyber Security	**Year:**	2021-2022
Module:	MSc Research Project		
Supervisor:	Vikas Sahni		
Submission Due Date:	07/01/22		
Project Title:	Using IP Address as a Unique Attribute for Data Deduplication in Network Devices		
Word Count:	5300	**Page Count:**	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Bhavana Bhavya

Date: 07/01/22

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Using IP address as a unique attribute for data deduplication in Network Devices

Bhavana Bhavya
x20128126@student.ncirl.ie

Abstract

This research paper contains descriptive techniques and methods followed for deduplicating record entries of large databases. It is useful in large organizations where duplication of records is a major concern as it affects the storage systems of these organizations significantly. In this paper, Logistic Regression machine learning technique has been introduced to deduplicate records and IP Address is used as major index. After implementation, the results indicate that IP Address can be considered a unique attribute for deduplicating data. Logistic Regression has accuracy of 95 percent in deduplicating records.

1 Introduction

Data Duplication and related issue is often ignored in many organizations. However, there are cases where it has led to major issues and setbacks causing harm to a company and its reputation in number of ways. Duplicate data causes poor hygiene of databases. In present world, database and hard drive space are crucial. Duplicate data may lead to loss of hard drive space resulting in less availability of space for important resources. Such cases also reduce database quality and hygiene significantly. This in turn also affects the budget cost of a company where company may end up spending more on purchasing cloud storage and hard drive storage to store other relevant data (QGate, n.d.).

Duplicate records also lead to confusion and employees end up spending more time on non-productive activities such as finding unique records among duplicate records. Due to duplicate records, employees need to pay more attention and invest more time in finding out any particular information which affects their overall productivity. Duplicate records often make employees clueless where one employee clarifies its doubt from another and involves multiple employees to finish a particular task. This in turn also interrupts the flow and interferes in the working pattern of many employees who directly or indirectly gets involved in finding relevant record among duplicate records (Ryan Bozeman, 2021).

Duplicate records can have serious threat on reputation of a company too. This seems irrelevant but if being a customer, we start receiving same mail from a company or agency multiple times in a day due to duplication of record in their database server resulting them in sending same mail multiple times to a single mail ID, it can be irritating and annoying for a customer which may lead to making the customer unsubscribe to the sender company or agency mail services and cause a negative impact of their brand name. Situation may get even worse if instead of mail service, a company or agency rely on call services to update their

https://www.qgate.co.uk/blog/crm/10-reasons-why-duplicate-data-is-harming-your-business/
https://www.impactplus.com/blog/reasons-duplicate-data-is-killing-your-marketing-and-sales-returns/

subscribed customers about their upcoming discounts and features. Receiving multiple calls from same company detailing same discounts and features can turn their customers furious and annoyed (QGate, n.d.).

Data duplication can be handled by multiple ways such as manually selecting and removing duplicate entries, writing a program to select and delete duplicate entries. However, using machine learning technique to get rid of duplicate data is most efficient method. Machine learning technique not only removes duplicate data but also prevents it from getting generated from the original source itself. Therefore, data training and testing technique used in machine learning helps in eradicating duplicate entries completely (Bettenburg, et al., 2008).
The research question "How can IP address be considered a unique attribute associated with assets based on the communication of the devices in network?" is interesting and worth implementing. There are many large organizations which deal with problems such as data duplication, asset details duplication where it becomes difficult to maintain unique record associated with each unique asset. Due to duplication of records, unique identities such as IP address, MAC address shows duplicity which pose question on basic networking infrastructure. It also becomes difficult to identify the inactive machine, unused machine which should be ideally deactivated and removed from company's network in order to provide better security and minimise cyber-attack risks such as Ransomware, DDoS attack which targets ideal machines connected to the company's internal network. This paper contributes towards bridging the gap by eliminating duplicate data and providing efficiency percentage of IP address which is be considered as unique attribute associated with assets connected to a company's network infrastructure. This process is generally done manually by employees of particular company. However, machine learning approach used in this paper can help in reducing human error and provide better productivity. Machine learning algorithms have high efficiency in solving such complex problems and it can be applied effectively on complicated network topologies belonging to large organisations. The research method proposed in this paper can help in solving the issue at the initial stage itself. This paper includes further sections such as background literature review of related works, basic methodology, design and implementation, main findings, conclusion and future scopes (Oleksii Tsymbal, 2020).

2 Related Work

There are many researches carried out in past to remove duplicate entries in data specific to many domains. However, there is similarity as well as differences in the type and field of work carried out in other research papers and this paper. Some of the prevailing research papers close to this research paper are discussed in this section.

In the research paper written by Souza, et al., the methodology consisted of three major steps: Indexing, Record Comparison and Classification. In the indexing step, a block key was assigned to each record. Each attribute value was encoded using Soundex algorithm where first letter of each attribute was kept as it is and subsequent letters were replaced by a digit

https://www.qgate.co.uk/blog/crm/10-reasons-why-duplicate-data-is-harming-your-business/
https://mobidev.biz/blog/5-essential-machine-learning-techniques/

from 0 to 6. In record comparison step, two algorithms were applied. Standard blocking algorithm was applied where similar records were grouped together in one block. Comparison of each record was done within the same block. Sorted neighbourhood algorithm used sorting key to combine records. In classification step, similarity algorithm ran through the records and returned 0,1 where 1 corresponded to perfect match. This paper used minimal process to evaluate and remove duplicate records. Therefore, it is similar in approach but different to the present research topic since machine learning approach of training and testing the dataset had not been applied in the previous paper (Souza, et al., 2018).

In the research paper written by Canalle, et al., a systematic approach was used to identify significant attributes for carrying out research. This paper is important for review in order to make proper selection of attributes to be focused for deduplication activity. Each attribute has different significance and it is important to mark relevant attributes for removing duplicate entries. Therefore, IP address has been evaluated and marked as major attribute in removing duplicate entries. Entity resolution process explained in this paper consisted of three major steps: Blocking, Pair Comparison and Clustering. In blocking technique, instances belonging to same entity were placed in same block and partitioned using blocking key. The main motive of blocking technique was to decrease number of comparisons. Pair comparison step was used to evaluate two comparable instances and determine their similarity values. Clustering technique was used to group the attributes together to form a cluster. The discussed paper is important for significant attribute selection. However, it is not relevant in determining data deduplication strategy (Canalle, et al., 2017).

In the research paper written by Verschuuren, et al., invoice dataset was used to carry out study of data deduplication. There were many machine-learning specific algorithms that were applied to perform the experiment. Domain specific knowledge in combination with distance-based similarity function was used to compare different parameters of invoice dataset and identify duplicates using supervised learning tools. These tools were: Jaro-Winkler, N-gram, Smith-Waterman, Levenshtein and Damerau-Levenshtein, Longest Common Substring, Binary Comparison and Monge-Elkan algorithm. Dataset was categorized into separate groups by making use of hidden pattern identification and correlation matching technique performed using discussed tools. Basic steps were: data pre-processing, choosing best supervised machine learning technique and training the dataset and evaluating the results. The overall structure of this paper is similar to the present research paper. However, the selection of algorithms for carrying out experiment is different. Present paper makes use of only logistic regression model to carry out the experiment (Verschuuren, et al., 2020).

In the research paper written by Christen, et al., record linkage was done to remove duplicate entries from database. This paper is significant in understanding factors and methods important for linking records and removing duplicate records. In this research paper, dataset was collected from different tools. In order to remove duplicate data from combined dataset, it is highly important to determine common parameters and attributes that can be used to link data records collected from different tools. Record linkage process started with cleaning and standardizing dataset collected in two separate databases. Indexing of dataset was done and it

was preceded by record pair comparison. Later, similarity vector comparison method was applied over dataset to separate matching, non-matching and possible matchable records. Evaluation of each separated area was done to ensure accuracy. This paper is helpful for establishing importance and making strategies for considering significant parameters for record linkage process. Therefore, it can be considered as supporting paper to carry out intermediate process (Christen, 2011).

In the research paper written by Albalawi, et al., attempt was made to reduce side-channel attack caused by cloud-computing virtualisation feature or memory deduplication. Memory deduplication is process of minimising usage of memory by storing only single copy of code and data accessed by multiple VMs. Multiple storage can give opportunity to malicious VM to trigger an attack. The method proposed in this paper monitored sensitive data addresses and provided fake results to malicious VM therefore preventing potential attack. Memory deduplication was utilized for collecting reading of monitored functions and analyse it using logistic regression algorithm. This paper was helpful in providing in-depth application of logistic regression model for analysis and interpretation of data. It gave keen awareness of steps to be followed to implement this algorithm in present research work. However, the type of dataset and structural flow of the two research are completely different since memory deduplication of cloud platform was proposed in this paper which is contrast to eliminating duplicate entries (Albalawi, et al., 2021).

In the research paper written by Soru & Ngonga Ngomo, et al., claims were made that logistic regression is best suited for noisy dataset. Various supervised learning classifiers were kept under analysis for link discovery and the results were compared. Ten different approaches were adopted on three artificial and three real-world datasets. The result of each approach was compared to other approaches. In general, all algorithms performed well. However, multilayer perceptions performed best with logistic regression being best for noisy data. Since, the dataset used for this research also included duplicate values, unstructured values. Therefore, it can be concluded that logistic regression can be applied for noisy dataset. This paper is quite different from the present research topic however, it provides useful claim of logistic regression being one of the best algorithms for machine learning application with noisy dataset. Therefore, it is beneficial and significant paper to be reviewed. Also, it consists of comparison of ten different approaches which is significant to understand and kept into consideration while choosing an algorithm for implementation (Soru & Ngonga Ngomo, 2014).

In the research paper written by Christen, et al., combination of two algorithms had been done to get maximum output and better performance. Two-step approach was used where high-quality training examples was automatically selected from compared record pairs and used for training using Support Vector Machine (SVM) classifier. The results produced revealed that SVM technique is considered better for record linkage technique compared to other unsupervised approaches. Therefore, SVM is used as an alternative method while implementing this research project also (Christen, 2008).

Machine learning technique used to train and test data is highly efficient and is capable of permanent removal of duplicate records as done in this research project. Therefore, this method is extremely advanced and compatible with current technical scenario. Further, there is scope of advancement in the implemented technique by using same procedure and applying same technique on much larger dataset. Therefore, this work can be carried forward in other researches also. Due to the implementation of single technique, it is easy to understand the flow of algorithm and its associated application by anyone who has basic python and machine learning knowledge. Minimal memory and storage usage makes this project easy to integrate and implement. Since, data duplication is also generic and one of the most common issues faced by multiple companies and service providers, it can be easily adopted by many companies and will be helpful in database management. Simple and effective coding technique used for implementation also makes this project easy to understand and integrate (Fogel & Kvedar, 2017).

3 Research Methodology

3.1 Phase Description

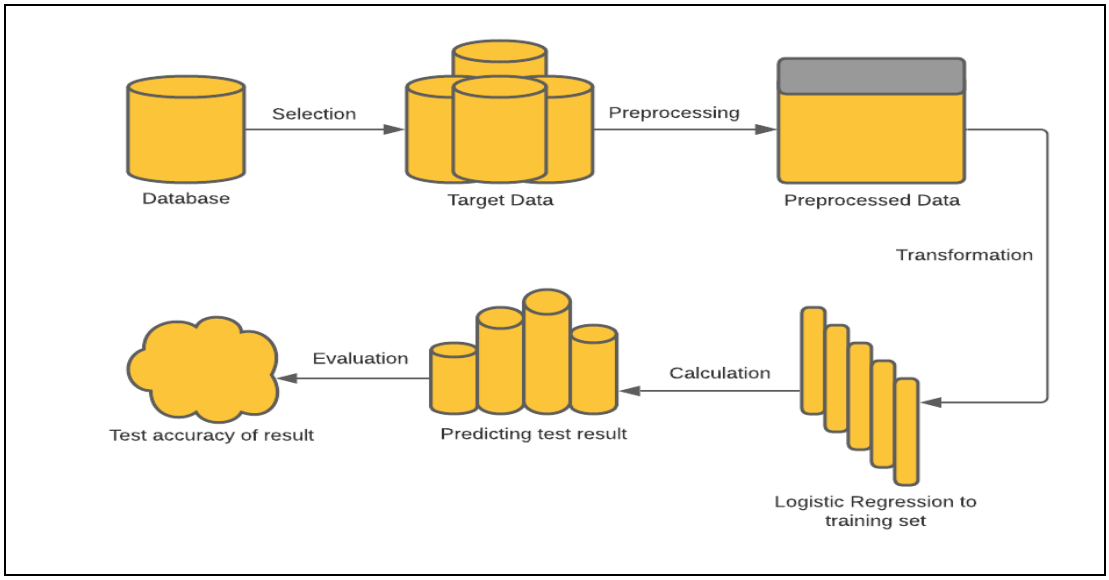

Figure 1: Phase Diagram

3.1.1 Database Selection

The following figure gives a complete network infrastructure where several devices and tools are interconnected to form a systematic, sectionalized and complex topology. Additional to all the devices mentioned below, Solarwinds Asset Inventory which forms an integral part of asset inventory management infrastructure is also taken into consideration.

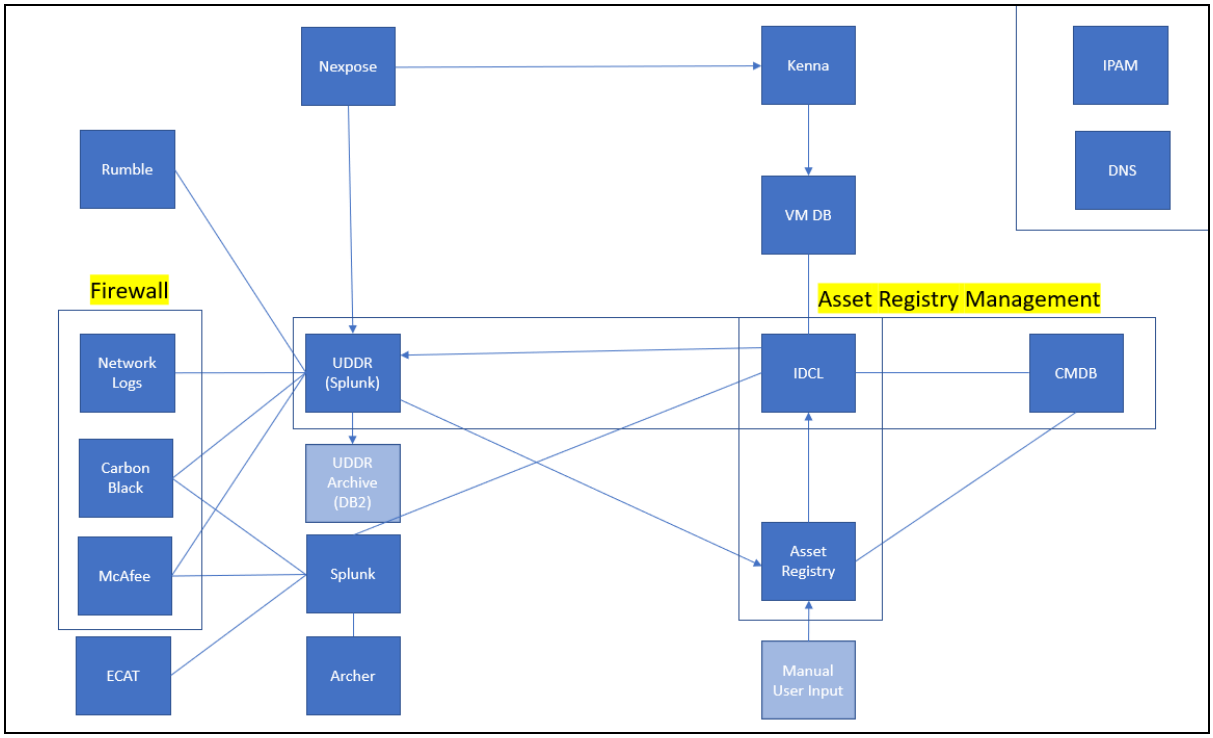

Figure 2: Network Infrastructure of an organization

3.1.2 Target Data Extraction

The above figure not only consists of asset related tools and devices used by company's employee but it also contains application responsible for firewall and other network related task. Therefore, selection of devices especially responsible for asset registry and the related stored database is crucial and one of the most important steps. In this research, UDDR and CMDB, which constitute major database systems have been considered for analysis. These databases also contain maximum duplicate entries of data that is removed and the database generated with unique entries is verified with database containing duplicates to check for difference and comparison.

3.1.3 Pre-processed Data

Next step is pre-processing the data which includes data hiding for better privacy and removing irrelevant columns. Character substitution and character shuffling process has been used to increase data security without disturbing the originality of database. Substitution of characters has been done in such a way that duplicity present in original dataset is not altered. This step has been applied on original dataset manually using Microsoft Excel application. Missing fields are filled and later empty rows are dropped for better data consistency. Also, all the entries are converted into uppercase to ensure data consistency (Jung & Yoo, 2009).

3.1.4 Algorithm Application

Logistic Regression is one of the best algorithms for removing duplicate data entries on the basis of literature review carried out in above section. Here, IP address is considered as the basis to prove uniqueness of database with negligible duplicate entries. Therefore, IP address is considered for indexing and duplicate entries detection.

Duplicate entries are categorized in pairs where first entry is kept undisturbed and second entry is dropped. Later, Jaro-Winkler method is used to compare the unique records and verify if there are any duplicate entries present. At the end, training and testing of dataset is done using logistic regression algorithm (Sarvagya Agrawal, 2021).

3.1.5 Accuracy Calculation

Any algorithm or process is only considered complete and efficient if it provides better performance and accuracy percentage. Therefore, this step is most crucial which decides the performance of an algorithm. Logistic Regression works among the best while dealing with duplicate entries, its removal and later training and testing the dataset. This is proven by accuracy percentage which is 95 percent for Logistic Regression algorithm in this research project. One more algorithm, Support Vector Machine is used for performing same process to check for comparison. However, Logistic Regression proved to be more efficient in deduplicating records and offered higher accuracy percentage (GeeksforGeeks, 2021).

4 Design Specification

4.1 Machine learning Technique

Machine learning is a research technique that provides cognitive functionality to machines and helps in making them perform similar to human mind. In this technique a set of data is used to make the machine learn which is called as training the machine and later it is tested by method called testing the data. For example, to develop automatic systems such as autonomous vehicle, first a set of image data is marked manually with road side signals, lane marking, cyclist, cars, trucks and other vehicles and scenarios. Later, these image data are used to train the system and test it in next stage of vehicle production. Similarly, machine learning technique can be used in all fields of technology with much greater ease. Machine learning methods have been categorised into two separate fields, supervised learning and unsupervised learning. Supervised learning method is used when the steps to be followed is already laid out and a set of data is used to teach the machine a specific procedure where as unsupervised learning method is used when we want the machine to explore by itself and discover hidden patterns. Classification and Regression technique is used on supervised learning and Clustering technique is used in unsupervised learning (Meyfroidt, et al., 2009).

4.2 Record Linkage Library

It is a record linkage library as the name suggests, which is used for linking records between data sources. It is a toolkit that provides plenty of tools for data deduplication and linking records. It also contains packages which includes indexing methods, record comparison functions which proves to be helpful during implementation of Logistic Regression algorithm for removing duplicate. Data manipulation tools such as Pandas and Numpy are used extensively by record linkage library. Pandas can be used easily for data manipulation and integration of record linkage to existing data manipulation projects. Some of the features of record linkage toolkit are multiple easy to use tools, efficient indexing methods which makes pairing of records easy such as blocking, sorted neighbourhood indexing. Huge

https://www.analyticsvidhya.com/blog/2021/05/logistic-regression-supervised-learning-algorithm-for-classification/
https://www.geeksforgeeks.org/understanding-logistic-regression/

records can be compared easily using plenty of comparison and similarity measures which includes multiple variables types such as strings, numbers, dates. The basic steps to link records are importing ‘recordlinkage’ library at the starting of source code. Linking two or more datasets using ‘pandas.DataFrame’. Using built-in indexing technique ‘recordlinkage.Index()’. Comparing the records using ‘recordlinkage.Compare()’ (Jonathan de Bruin, 2019).

4.3 Sklearn framework

Sklearn also known as Scikit-learn framework is a user-friendly framework that consists of multiple efficient tools such as classification, regression and clustering, pre-processing models and evaluation tools. This is an open-source framework which is popular and easy to integrate. It helps in crunching structured data, for example, it can be used for predicting price of upcoming sports game. It is easy to learn and implement as simple model can be built and trained using simple three lines of code, for example:

```
from sklearn.svm import SVC
model = SVC()
model.fit(X, y)
```

Matrix function, which is generally used to evaluate performance of an algorithm or technique can be easily and effectively implemented using metric function of scikit-learn library. It consists of multiple numbers of functions that can be used for implementing machine learning technique efficiently. Some of the examples are Pandas, Xarray, Auto-Sklearn, Tpot, Featuretools, Neptune, Scikit-Optimize, Sklearn-eap, Sklearn-Onnx, Treelite etc (Ugo Loobuyck, 2020).

4.4 SVM Architecture and Technique

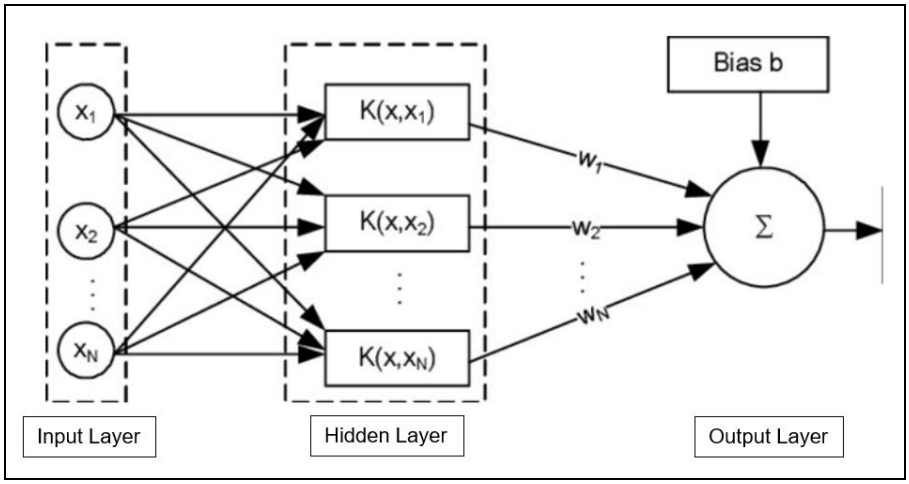

Figure 3: SVM Algorithm generic architecture

SVM is considered both supervised and unsupervised learning algorithm depending upon the case. It is used for both linear and non-linear classification and regression problems. SVM is used as unsupervised learning algorithm when data is unlabelled. It consists of three layers: Input layer, Hidden layer and Output layer. Input layer is used for providing input, hidden layer is where all the processing takes place and output layer is used for displaying output (Mayank Tripathi, 2020).

https://recordlinkage.readthedocs.io/en/latest/about.html/
https://towardsdatascience.com/scikit-learn-tensorflow-pytorch-keras-but-where-to-begin-9b499e2547d0/
https://datascience.foundation/datatalk/basic-overview-of-svm-algorithm/

4.5 Logistic Regression Technique

Logistic Regression technique is one among the probabilistic method used to understand relationship between various attributes and parameters. One of the main features of Logistic Regression is, it gives the result in binary, i.e., 0 and 1. We can also find percentage value using Logistic Regression algorithm. For example, we can experiment a set of data and conclude that smoking increases lung cancer risk by 20% where a person may or may not have lung cancer which is a binary variable. They either have it or not. Formula is:

$$P(Y = 1) = \frac{1}{1+e^{-(b_0+B*X)}}$$

Where P (Y = 1) represents probability of Y being 1, while b0 is a parameter not linked to X and B is vector of coefficients representing relationship between Y and each one of X1, X2 and so on. Below is the graph generated for logistic regression

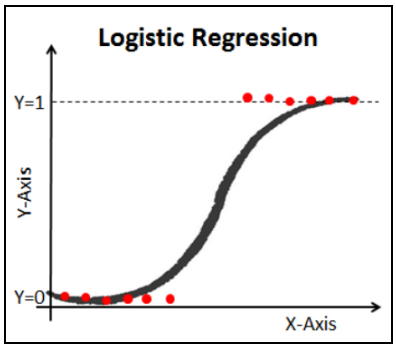

Figure 4: Logistic Regression Curve

The generic steps for logistic regression are same as other algorithms. Data cleaning followed by training and testing data and data modelling. These processes have been already explained in the above section (Ayush Pant, 2019).

4.6 Logistic Regression Architecture and Flow Diagram

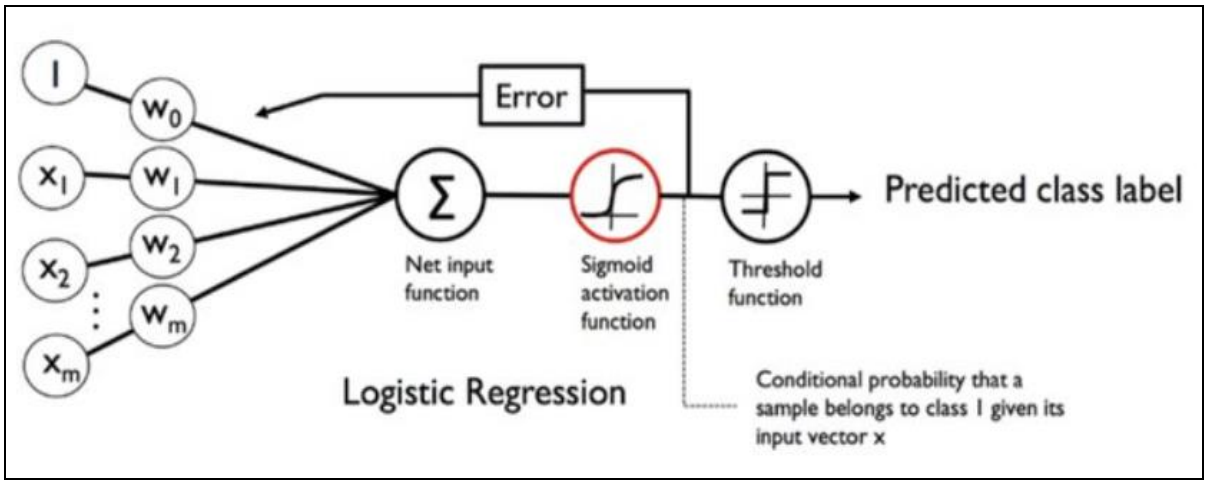

Figure 5: Logistic Regression generic architecture

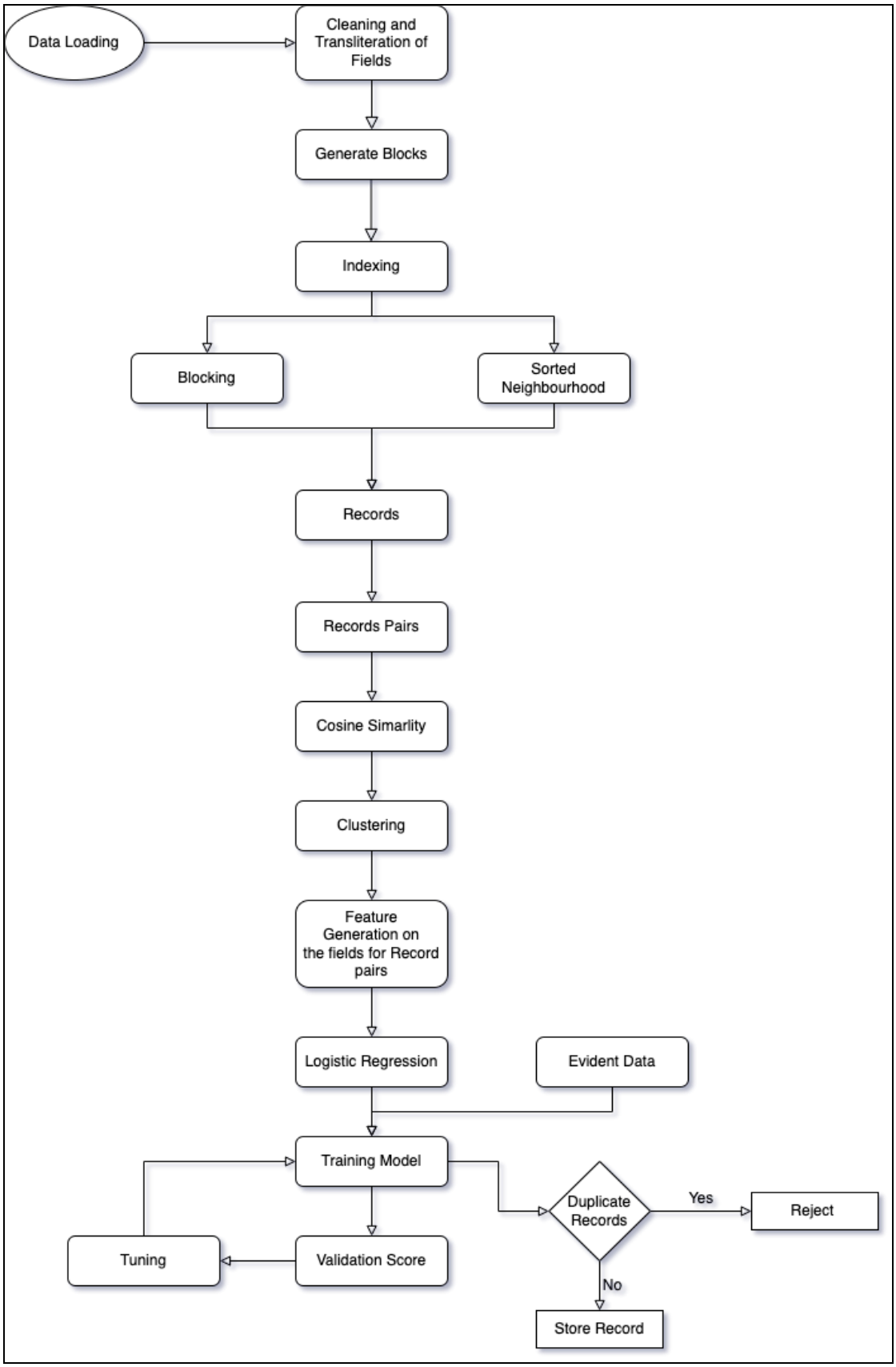

Figure 5: Logistic Regression flow-diagram

5 Implementation

Below is the end-to-end algorithmic step followed while implementing Logistic Regression algorithm:

Step 1: **START**

Step 2: Data loading as a data frame

Step 3: Generate the block of IP Address data

Step 4: Create indexing based on IP Address

Step 5: Create index based on block method and sort neighborhood method

Step 6: Combine the records

Step 7: Create the record pairs for clustering

Step 8: Using Cosine Similarity identify duplicate pairs

Step 9: Group the pairs based on the Euclidean distance

Step 10: Feature Generation on the field of IP Address from record pairs

Step 11: Create a logistic regression model as a classifier

Step 12: Train the model with the record pairs

Step 13: Check the validation metrics for performance

Step 14: Tune the model parameters for the better performance

Step 15: Get new evident data

Step 16: Identify whether it is duplicate records or nonduplicate records

Step 17: **IF YES**

Step 18: Reject the record

Step 19: **ELSE**

Step 20: Store the record

Step 21: **STOP**

During implementation the following figure shows the row numbers which are duplicates where first-row number is kept in dataset and the second-row number is dropped.

```
Out[425]: MultiIndex([(   2,    0),
                      (  35,   33),
                      ( 101,   48),
                      (  84,   79),
                      ( 120,  118),
                      ( 378,  119),
                      ( 305,  181),
                      ( 368,  194),
                      ( 322,  204),
                      ( 371,  210),
                      ...
```

Figure 6: Intermediate output showing duplicate records row number

Final output is the accuracy percentage obtained after implementing the research project completely. This percentage indicates the percentage of accuracy of Logistic Regression machine learning algorithm in deduplicating dataset, training and testing dataset with considering IP Address as index. Accuracy percentage obtained is **95 percent** which is considered as a good value.

Following table includes all the tools, language and libraries used for implementing the algorithm.

Tools, Language, Libraries Used	Functionality
Anaconda Navigator (tool)	A desktop graphical user interface (GUI) that allows easy launch and management of libraries and packages without using command-line (Rolon-Mérette, et al., 2016)
Jupyter Notebook (tool)	A server-client application that allows editing and running codes and documents via web browser without making use of internet (Kluyver, et al., 2016)
Python3 (language)	An interpreted open-source, high-level programming language used for writing code. It is widely used for machine learning techniques as it contains huge variety of libraries and functions for such implementation (Rolon-Mérette, et al., 2016)
pandas (library)	A python library used extensively for manipulation and analysis of data (Millman & Aivazis, 2011)
numpy (library)	A python library used extensively for working with arrays (Millman & Aivazis, 2011)

recordlinkage (library)	A python library used extensively for linking record and data deduplication in or between data sources (Millman & Aivazis, 2011)
sklearn (library)	A machine learning library used extensively for implementing various algorithms such as svm, random forest, knn, logistic regression etc (Millman & Aivazis, 2011)
matplotlib (library)	A visualization python library for plotting 2D arrays (Millman & Aivazis, 2011)

Figure 7: Tools, Technologies and Libraries used

6 Evaluation

This section is one of the critical sections where comparisons of all the methods and techniques applied on dataset is evaluated and the method followed is proved and justified for being used for implementation. In this research project, two techniques have been used and the one providing better results has been selected as final method. The reason to choose two algorithms is to compare the results and establish the conclusion. Since, literature review has been done thoroughly which implies that even if only two separate methods have been used for analysis, proper reasoning can be provided regarding selection of algorithm and implementation process using appropriate confusion matrix, graph and figures.

6.1 Experiment 1 – Support Vector Machine Algorithm Implementation and Analysis

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	0
1.0	1.00	0.91	0.95	85
accuracy			0.91	85
macro avg	0.50	0.45	0.48	85
weighted avg	1.00	0.91	0.95	85

Figure 8: SVM Algorithm calculated output

The above figure indicates that values of accuracy, precision, recall, f1-score are 0.91, 1.00, 0.91 and 0.95 respectively. This shows that performance of SVM is good in deduplicating the record entries.

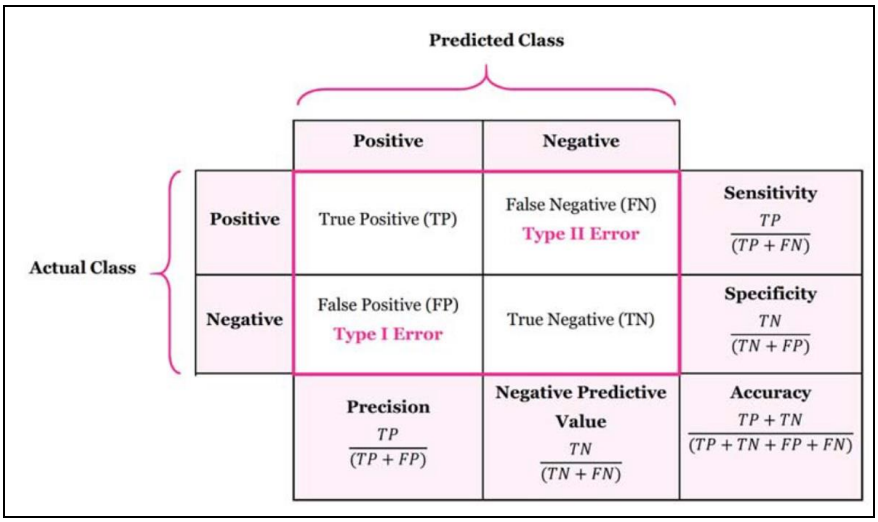

Figure 9: Confusion Matrix

The above figure indicates confusion matrix which is used for denoting parameters such as Sensitivity (recall), Specificity, Accuracy, Negative Predictive Value and Precision. Therefore, calculating these parameters on the basis of below generated confusion matrix will be:

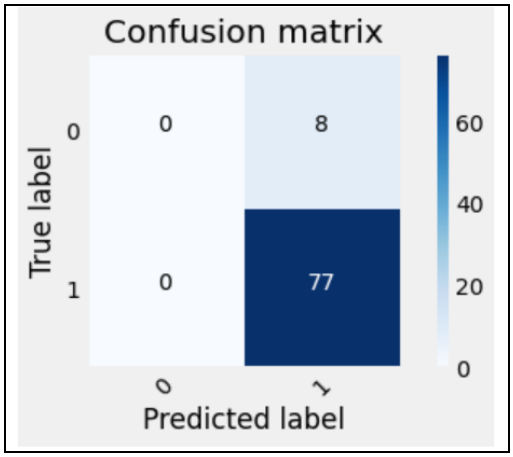

Figure 10: Confusion Matrix for SVM Algorithm

TP = 0 FP = 0 FN = 8 TN = 77
Sensitivity = 0.91 Specificity = 1 Accuracy = 0.90 Precision = 1
Therefore, performance of SVM can be considered good.

6.2 Experiment 2 – Logistic Regression Algorithm Implementation and Analysis

	precision	recall	f1-score	support
0	0.00	0.00	0.00	0
1	1.00	0.95	0.98	85
accuracy			0.95	85
macro avg	0.50	0.48	0.49	85
weighted avg	1.00	0.95	0.98	85

Figure 11: Logistic Regression calculated output

The above figure indicates that values of accuracy, precision, recall, f1-score are 0.95, 1.00, 0.95 and 0.98 respectively. This shows that performance of Logistic Regression is best in deduplicating the record entries.

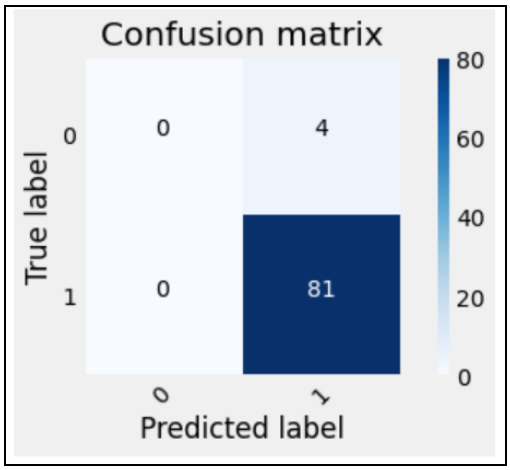

Figure 12: Confusion Matrix for Logistic Regression

On the basis of confusion matrix generated for Logistic Regression Model, calculation on parameters such as Sensitivity (recall), Specificity, Accuracy, Negative Predictive Value and Precision using formula can be done in following way:
TP = 0 FP = 0 FN = 4 TN = 81
Sensitivity = 0.95 Specificity = 1 Accuracy = 0.95 Precision = 1
Therefore, performance of Logistic Regression can be considered better than SVM.

6.3 Discussion

Algorithm	Accuracy	Precision	Recall	F1-score
Support Vector Machine	91	1	91	95
Logistic Regression	95	1	95	98

Figure 13: Comparison of results

By the above table, it is proved that Logistic Regression performs comparatively well. However, precision is same for both the algorithms, sensitivity (recall) and f1-score are better

for Logistic Regression. To answer the research question 'How can IP address be considered a unique attribute associated with assets based on the communication of the devices in network?', Logistic Regression Algorithm is implemented which proves that IP address can be considered as a unique attribute associated with assets by **95 percent**. SVM Algorithm also performs well however, Logistic Regression performance is outstanding.

7 Conclusion and Future Work

Finally, it can be concluded that IP address can be considered as a unique attribute associated with assets based on the communication of the devices in network by 95 percent using Logistic Regression model of machine learning technique. This is very satisfactory number and provides solution to many data duplication problem existing in many organizations dealing with network related devices and their storage details. Future work of this research project could be generating output with no duplicate entries in excel file format. Also, using the technique demonstrated in this research project as part of asset management infrastructure in various organizations so that duplicate data can be removed within the process flow itself without taking much efforts later by manually deduplicating it.

8 References

[1] Albalawi, A., Vassilakis, V. & Calinescu, R., 2021. *Memory Deduplication as a Protective Factor in Virtualized Systems.* s.l., International Conference on Applied Cryptography and Network Security, Springer.

[2] Arthur Mello, 2020. *towards data science.* [Online] Available at: https://towardsdatascience.com/logistic-regression-the-basics-b1716661c71b [Accessed 07 12 2021].

[3] Ayush Pant, 2019. *towards data science.* [Online] Available at: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148 [Accessed 27 11 2021].

[4] Bettenburg, N., Premraj, R., Zimmermann, T. & Kim, S., 2008. *Duplicate bug reports considered harmful… really?.* s.l., IEEE International Conference on Software Maintenance.

[5] Canalle, G., Lóscio, B. & Salgado, A., 2017. *A Strategy for Selecting Relevant Attributes for Entity Resolution in Data Integration Systems.* s.l., ICEIS.

[6] Christen, P., 2008. *Automatic record linkage using seeded nearest neighbour and support vector machine classification.* s.l., Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.

[7] Christen, P., 2011. *A survey of indexing techniques for scalable record linkage and deduplication.* s.l., IEEE transactions on knowledge and data engineering.

[8] Derrick Mwiti, 2021. *neptuneblog.* [Online] Available at: https://neptune.ai/blog/the-best-ml-framework-extensions-for-scikit-learn [Accessed 17 12 2021].

[9] Fogel, A. & Kvedar, J., 2017. *Benefits and risks of machine learning decision support systems.* s.l., s.n.

[10] GeeksforGeeks, 2021. *GeeksforGeeks.* [Online] Available at: https://www.geeksforgeeks.org/understanding-logistic-regression/ [Accessed 26 11 2021].
[11] Jonathan de Bruin, 2019. *Python Record Linkage Toolkit.* [Online] Available at: https://recordlinkage.readthedocs.io/en/latest/about.html [Accessed 09 12 2021].
[12] Jung, K. & Yoo, K., 2009. *Data hiding method using image interpolation.* s.l., Computer Standards & Interfaces.
[13] Kluyver, T. et al., 2016. *Jupyter Notebooks-a publishing format for reproducible computational workflows.* s.l., s.n.
[14] Mayank Tripathi, 2020. *DataScience foundation.* [Online] Available at: https://datascience.foundation/datatalk/basic-overview-of-svm-algorithm [Accessed 19 12 2021].
[15] Meyfroidt, G., Güiza, F., Ramon, . J. & Bruynooghe, M., 2009. *Machine learning techniques to examine large patient databases.* s.l., Best Practice & Research Clinical Anaesthesiology.
[16] Millman, K. & Aivazis, M., 2011. *Python for scientists and engineers. Computing in Science & Engineering.* s.l., s.n.
[17] Oleksii Tsymbal, 2020. *mobidev.* [Online] Available at: https://mobidev.biz/blog/5-essential-machine-learning-techniques [Accessed 05 12 2021].
[18] Peter Christen, 2008. *ACM Digital Library.* [Online] Available at: https://dl.acm.org/doi/abs/10.1145/1401890.1401913 [Accessed 18 12 2021].
[19] QGate, n.d. *QGate.* [Online] Available at: https://www.qgate.co.uk/blog/crm/10-reasons-why-duplicate-data-is-harming-your-business/ [Accessed 02 12 2021].
[20] Rolon-Mérette, T., Rolon-Mérette, D., Ross, M. & Church, . K., 2016. *Introduction to Anaconda and Python: Installation and setup.* s.l., Python for research in psychology.
[21] Ryan Bozeman, 2021. *IMPACT.* [Online] Available at: https://www.impactplus.com/blog/reasons-duplicate-data-is-killing-your-marketing-and-sales-returns [Accessed 15 12 2021].
[22] Sarvagya Agrawal, 2021. *Analytics Vidhya.* [Online] Available at: https://www.analyticsvidhya.com/blog/2021/05/logistic-regression-supervised-learning-algorithm-for-classification/ [Accessed 01 12 2021].
[23] Soru, T. & Ngonga Ngomo, A.-C., 2014. *A comparison of supervised learning classifiers for link discovery.* s.l., Proceedings of the 10th international conference on semantic systems.
[24] Souza, L., Murai, F., da Silva, A. & Moro, M., 2018. *Automatic identification of best attributes for indexing in data deduplication.* Cali, Colombia, In Proceedings of the 12th Alberto Mendelzon International Workshop on Foundations of Data Management.
[25] Ugo Loobuyck, 2020. *towards data science.* [Online] Available at: https://towardsdatascience.com/scikit-learn-tensorflow-pytorch-keras-but-where-to-begin-9b499e2547d0 [Accessed 15 12 2021].

[26] Verschuuren, P. et al., 2020. *Supervised machine learning techniques for data matching based on similarity metrics.* s.l., s.n.