THE ANALYSIS OF COVID-19 IMMUNIZATION DATA IN RAKHINE STATE USING KNN ALGORITHM

KHIN MYAT THU

M.C.Sc SEPTEMBER, 2022

The Analysis of COVID-19 Immunization Data in Rakhine State Using KNN Algorithm

By

Khin Myat Thu

B.C.Sc

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Computer Science

(M.C.Sc)

University of Computer Studies, Yangon

September 2022

ACKNOWLEDGEMENTS

Firstly, I would like to express my deepest appreciation to **Dr. Mie Mie Khin**, Rector of the University of Computer Studies, Yangon and **Dr. Zaw Tun**, Pro-rector of the University of Computer Studies (Sittway), for their kind permission to build up this thesis.

Secondly, I would like to extend my sincere thanks to **Dr. Si Si Mar Win, Dr. Tinzar Thaw**, Professor. University of Computer Studies, Yangon, for their warmhearted management throughout the completion of this thesis.

Thirdly, I would like to express my heartfelt thanks to my supervisor, **U Thaung Myint Htun**, Associate Professor, Head of Faculty of Information Science, University of Computer Studies (Sittway) for his encouragement, invaluable patient, and precious supervision of compiling the materials for my thesis.

Fourthly, I also deeply thank **Daw Aye Aye Khine,** Associate Professor, Head Department of English, University of Computer Studies, Yangon for editing my thesis from the language point of view.

Lastly, I would be remiss in not mentioning my family, especially my parents, my aunty, my helpmate, and all of my friends. Their suggestions, support and generous help rendered me during the development of thesis. Their love and concerned encouragement have strengthened me throughout the studies.

STATEMENT OF ORIGINALITY

By means of this, I officially state that the work represented in this thesis is the result of original study and has been presented for a higher degree to any other University or Institution.

…………………. ……………………..

Date Khin Myat Thu

Abstract

Data mining involves the searching of large information of the data or records to discover pattens and utilize these pattens in the prediction the future events. Classification is one of the methods in data mining for categorizing a particular group of items to targeted groups. Main goal of classification is to predict the nature of an items or data based on the available classes of items. Construction of the classification model always defined by the available training data set. In this system, an analysis of COVID-19 immunization results of Rakhine State was carried out using k-Nearest Neighbor (k NN) classification algorithm in data mining. The data set about COVID-19 immunization details are collected from General Administration Department, Rakhine State. The primary objective of this system is to evaluate algorithm in the prediction of COVID-19 immunization finishing rate and analysis result of Rakhine state. k-Nearest Neighbor algorithm is utilized to carry out for the prediction of COVID-19 immunization results.

CONTENSTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

CHAPTER 1

INTRODUCTION

COVID-19 is an infectious disease, a type of virus that can spread rapidly through the air. The infection is easy and fast and can be spread from person to person immediately. It spread to many countries worldwide, and the number of infected people increased rapidly daily.

The COVID-19 disease is caused by the coronavirus and has caused concern and fear in almost all countries, including Myanmar. It caused panic. The rate of spread was also speedy. The virus is not fixed, and the symptoms are similar, but there are differences. There were differences. Especially parts of the body related to the respiratory tract. In particular, the damage to the lung organ is very severe and may lead to loss of life.

Ministry of Health confirmed that the coronavirus was first detected in Myanmar on March 23, 2020. In the first wave, 374 cases were released in Myanmar. Six people lost their lives. It was announced that the infection was detected in Myanmar on July 16, 2020.

The infection prevention related to COVID-19; Control and safety measures have been implemented in all regions and states of Myanmar. Guidelines were issued for the public to follow. We should not go out of our homes unless necessary and wear a mask if we go outside. Thus, warnings begin with educating people not to live in densely populated areas. Vaccination of students, employees, people against COVID-19 is being carried out on a rotating basis, which are imported from abroad.

Making better plans, vaccination stations in towns and villages for the prevention of COVID-19 are organized and vaccinated in rural areas. Age-specific vaccination schedules; For the first time in a specified period, the second time, A third vaccination program is underway. In addition, children from primary schools, those working in government departments, residents in ward and village are also vaccinated against COVID-19 by setting up alternate programs.

On August 16, 2020, it was announced that the infection was detected in Rakhine State. In terms of the number of people affected by the disease, the second wave was more severe than the first wave. Relevant health workers and local authorities work together to administer vaccinations, but in Rakhine State, communication and transportation are limited. For the people who live in areas where communication is complicated, it is a rare opportunity to inject the COVID-19 vaccine into the country a specified number of times. Therefore, it can be said that it is necessary to do a specific business in combination with IT skills to collect accurate lists and meet the criteria for these immunization activities [6].

Data mining technology collects much information to obtain prediction results, anomalies patterns and relationships. It helps study in many areas, including health care and education. Therefore, data mining is a research activity. It can also be said that it is a functional modern and innovative technology that combines critical technologies to achieve local benefit projects [33].

1.1 Classification Rules

Classification is the set of modes that describes and distinguishes data concepts. A classification method means blocking a specific answer after checking against pre-constructed validation rules for a new record to be predicted. Records obtained from research. This is also called supervised learning. Easily adapted algorithms are used to get better data quality in this technique. It has many issues in classification. Some of them are accuracy, speed, comprehensibility and time to learn. In data mining, the classification model is used if the target variable is more "Qualitative". Even there have several different types of classification, each of these algorithms is applied to mine essential information from the dataset.

1.2 Objectives of the Thesis

The primary objectives of this system are:

- to analyze COVID-19 immunization results in Rakhine State.
- to know how to apply data with a data mining approach.
- to understand the evaluation of k-NN.
- to evaluate the need for COVID-19 Vaccines in Rakhine State.
- to provide health improvements by gathering valuable information and analysis.

1.3 System Overview

This system will analyze the COVID-19 vaccination data in Rakhine State. It uses the k-Nearest Neighbor classifier to classify the completion rate of the COVID-19 vaccine. It will work to categorize the completion rate of a given dose from the number of times that the COVID-19 vaccine has been administered.

The Euclidean distance method is used to search for the similarities between the record sets according to the input features. The prediction result can be identified by building the k-Nearest Neighbor (k-NN) model. The k-NN algorithm supposes similarity between the new case/data and the obtainable cases and sets the last case into the group most like the universal groups. All obtainable data are stored, and unique data points depending on similarity are classified by the k-NN algorithm.

Therefore, the system will use a k-NN classifier to estimate the completion rate and recommend reports for those regions. This work was undertaken to classify factors for vaccine completion rates using the k-NN algorithm.

1.4 Motivation of the Thesis

In the Rakhine state, the problematic transport congestion, unstable population, worst weather condition, low staff resources, and late vaccine coverage have led to a severe outbreak of COVID-19. The historical data and previous cases can consider for future work by using data mining technology for fulfilling those who are working to manage those areas. This study is focused on the COVID-19 immunization data with its related issues to those who are working to manage those areas.

1.5 Organization of the Thesis

This thesis implements an analysis of the COVID-19 immunization State using the k-NN classification. There have five sections in this thesis. First, Chapter 1 internalizes the data mining system, classification, objectives, and system overview. Second, Chapter 2 presents the theoretical background of the system. Third, Chapter 3 explains classification methods and k-Nearest Neighbor classification. Fourth, Chapter 4 illustrates the system design and implementation and analysis process. Finally, Chapter 5 shows the conclusion of the current work and future directions.

CHAPTER 2

BACKGROUND THEORY

This chapter comes up with the technical information for the tasks associated to the topic of data mining that will support to perspective throughout this study. The fundamental study of this work is the classification task based on supervised learning approach with k-NN methods for analyzing the finishing rate of COVID-19 immunization data in Rakhine state. Therefore, it is important to review the literature in classification areas. Before all, this chapter describes an explanation of the brief introduction of data mining covering its architectures and theoretical foundations. Then, the data mining functions, different kinds of classifications and types of data mining algorithms are described. Finally, this chapter presents data mining uses, machine learning and statistics, Data mining applications and related works respectively.

2.1 Data Mining

Data mining is the analysis study of large datasets and new theories. It is an emerging field of computer intelligence that provides new technologies and tools. It is a process that uses meaningful recording techniques, as well as numerical and mathematical techniques, to analyze large amounts of data in the storage and discover new patterns and correlations. Data mining refers to extracting knowledge from a large amount of data. It is an essential step in the knowledge discovery process in databases [7].

Finding Global Patterns and Relationships in Very Large Databases Data Mining These global patterns and relationships are hidden. These relationships represent valuable knowledge discovery about the database and the objects in the database. Data Mining is an interesting performance analysis without preconceived notions of what will happen. The success of data mining technique depends mainly on the amount of energy, knowledge, and innovation. In essence, data mining is similar to solving a puzzle. The individual pieces of the mystery are not complicated structures in and of themselves. However, taken as a united whole, they can construct very detailed systems. Data mining techniques can be concerned to problems of business process reengineering. The goal is to understand interactions and relationships within the business [13].

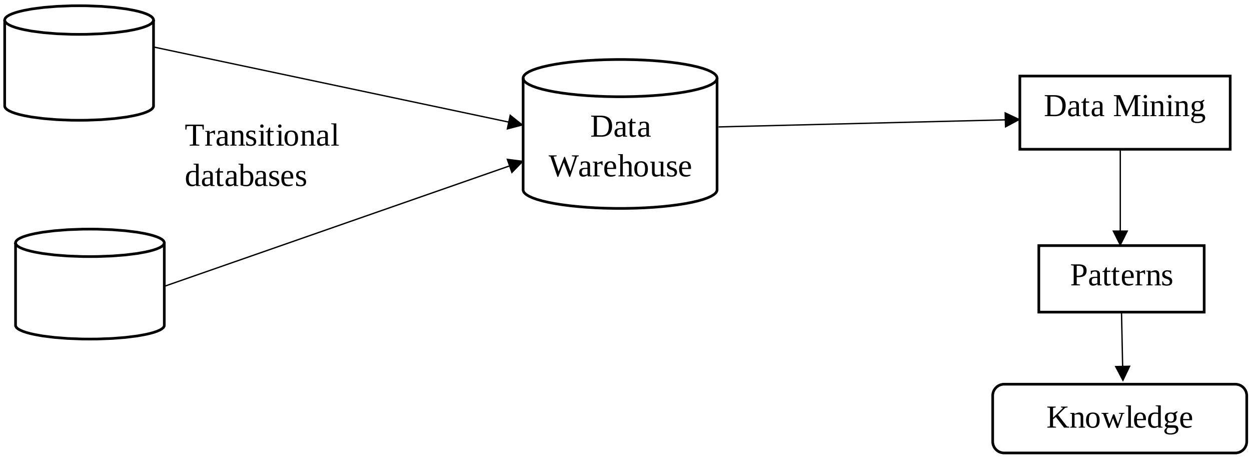


Figure: 2.1 Overview of Data Mining Concept

An essential part of the entire data analysis and a major discipline in data science that uses advanced analytical techniques to find useful information in data sets is called Data mining. It uses various data analysis tools to find patterns and correlations in data that can be used to make accurate predictions. It also handles various existing computational techniques from statistics, machine learning, and pattern recognition. In more detail, data mining is about the level of knowledge discovery in the databases (KDD) knowledge mining from data. Knowledge discovery is a process that contains a repetitive process [48] [7]:

1. Data cleaning (to carry out missing data, smooth out noise and correct deviations)
2. Data integration (merging of data from several data sources)
3. Data selection (Data can be retrieved from the database to be analyzed)
4. Data transformation (consolidating the data into suitable forms for mining tasks by attaining summary or aggregation operations)
5. Data mining (a vital process where imaginative methods are applied to clipping interesting patterns)
6. Pattern evaluation (to identify the truly fascinating patterns representing knowledge based on some allure measures)
7. Knowledge presentation (to give the mined knowledge to the user)

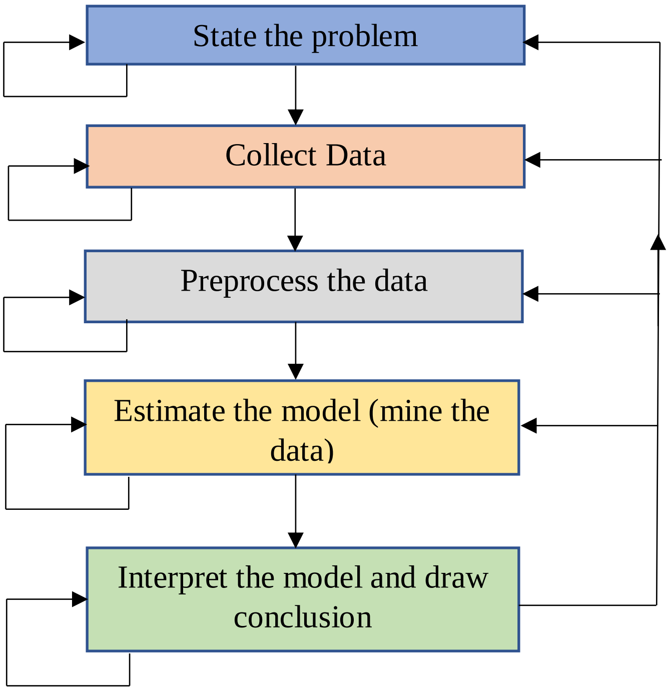

State the problem
Collect Data
Preprocess the data
Estimate the model (mine the data)
Interpret the model and draw conclusion

Figure:2.2 Process of Data Mining

2.2 The Architecture of a Data Mining

The techniques may have six major components:

1. **Database, data warehouse, or other information data center repositories:** This is one or a set of kinds of information repositories. Data preparation techniques may be performed.
2. **Database or data warehouse server:** It is responsible for taking back the relevant data depending on data mining request of the user [43].
3. **Knowledge base:** This is called the domain knowledge. It can be used to point the search or calculate the interestingness of the resulting patterns. Such knowledge can contain concept hierarchies used to organize attributes or attribute values into different abstraction levels Knowledge. That may be user beliefs, which can be used to assess a pattern's interestingness depending on its unexpectedness, may also be included. Other examples are additional interestingness constraints or thresholds and metadata.

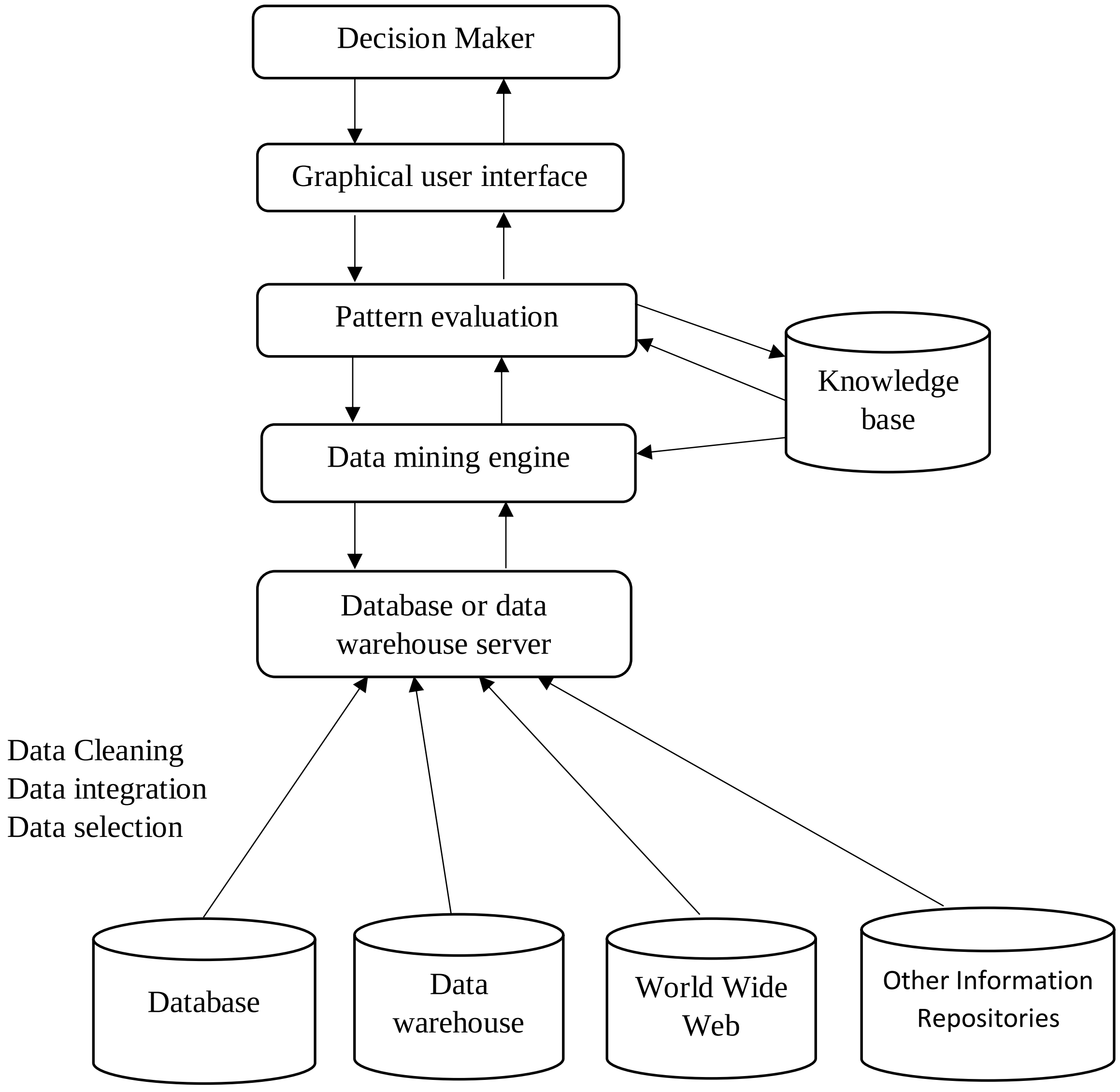


Figure:2.3 Architecture of a Typical Data Mining

4. **Data mining engine:** Data mining engines are important for a data mining system because they include a complete set of function modules such as characterization, association, classification, cluster analysis, and evolution and deviation.
5. **Pattern evaluation module:** This component typically engages interestingness measures and interacts with the data mining modules to force the search towards interesting patterns as deep as possible in the mining process, confining the search to only the interesting patterns.
6. **Graphical user interface:** The module connects between users and the data mining system. This module permits users to interact with the system by

determining a data mining query or task. It also provides information to help focus the search and perform exploratory data mining based on the intermediate data mining results. In addition, this component concedes the user to browse database and data warehouse schemas or data structures. It evaluates mined patterns and visualizes the patterns in various forms [7].

2.3 Theoretical Foundations of Data Mining

Data mining techniques come from extensive study and production development. This evolution was initiated when business data was first stored on computers. It carried on with progress in data access and, latest. This evolution also caused technologies that let users to cruise through their data in real-time. Data mining carries this development process beyond retrospective data access and find the way to potential and practical information distribution. Three technologies for business community like huge data collection, powerful multiprocessor computers, and data mining algorithms support it. There are several theories in the basis of data mining:

1. **Data reduction:** The foundation is to cut down the data representation. It deals with accuracy for speed in return for the need to get fast approximate answers to queries on enormous databases. Data reduction techniques contain singular value decomposition, wavelets, regression, log-linear models, histograms, clustering, sampling, and the construction of index trees.
2. **Data compression:** The primary of data mining is to wrap the given data by encoding it in terms of bits, association rules, decision trees, clusters and so on [32].
3. **Pattern discovery:** The basis of data mining is to find patterns in the database, Such as associations, classification models, and sequential pattern mining. Fields such as machine learning, neural networks, associated mining, sequential pattern mining, clustering, and several other subfields contribute to pattern discovery theory [51].
4. **Probability theory:** The base of data mining is to search joint probability distributions of random variables. This is based on statistical theory.

5. **Microeconomic view:** Small businesses be tasked with finding patterns of interest to the extent. The prospecting can be used in some business decision-making.
6. **Inductive database:** According to this theory, a database scheme consists of data patterns that are store database. Thus, the difficulty of performing induction on databases, where the task is to query the data and the theory of the database. Among many researchers in database systems, this view is popular.

2.4 Data Mining Functions

Functionalities of data mining are acclimated to determine the type of patterns to be discovered in data mining tasks. Tasks of data mining can be divided into two groups. They are descriptive and predictive. Descriptive mining can describe the properties of data in a database. The main job of predictive mining is to conclude to matrons on current data. Data mining functions and patterns are noted below [7].

2.4.1 Characterization and Discrimination

Data can be related with classes or concepts. It can be useful to specify individual classes and concepts in summarized, concise, and precise terms. Such explanation of a class or a concept are called class/concept descriptions [17]. These explanations can be derived in several ways:

1. **Data characterization:** A summarization of the general characteristics of features of a target class of data is called data characterization. The data corresponding to the user-specified class is typically collected by a database query [36] [29].
2. **Data discrimination:** Data discrimination compares the general features of target class data objects with the general features of objects from one or a set of contrasting classes. The target and contrasting classes can be specified by the user, and the corresponding data objects retrieved through database queries [18].

2.4.2 Classification and Prediction

A data analysis form used to extract models describing essential data classes or to predict future data trends with an unknown class label is called classification. The derived model is based on the analysis of a set of training data (i.e., data objects with a known class label). Classification can be used for predicting the class label of data objects. However, in many applications, users may wish to predict missing or unavailable data values rather than class labels. This is usually the case when the predicted values are numerical data, and it is often specifically referred to as prediction [30].

Classification and prediction may need to be proceeded by relevance analysis, which attempts to identify attributes that do not contribute to the classification or prediction process [40]. These attributes can then be excluded.

2.4.3 Association Analysis

One of the most vital and well-researched data mining techniques is called association rule mining [8]. It focuses to extract interesting correlations, frequent patterns, associations, or casual structures among sets of items in the transaction databases. It is the discovery of association rules showing attribute value conditions occurring frequently together in a given set of data [14]. And it is widely used for market basket or transaction data analysis.

More formally, association rules are of the form $X \Rightarrow Y$, that is, "$A_1 \wedge \ldots \wedge A_m \rightarrow B_1 \wedge \ldots \wedge B_n$", where A_i (for $i \varepsilon \{1, ..., m\}$) and B_j (for $j \varepsilon \{1, ..., n\}$) are attribute-value pairs. The association rule $X \Rightarrow Y$ is interpreted as "database tuples that satisfy the conditions in X are also likely to satisfy the conditions Y".

2.4.4 Cluster Analysis

Clustering analysis analyzes clusters embedded in the data. Clustering analyses data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. Each cluster that is formed can be viewed as a class of objects from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, organizing observations into a hierarchy of classes that group similar events together [31].

2.4.5 Outlier Analysis

A database may involve data objects that do not comply with the general behavior or model of the data. These data objects are outliers. The outlier data analysis is referred to as outlier mining. Outliers can be detected using distance measures where objects that are a substantial distance from any other clusters are considered outliers [7].

2.4.6 Evolution Analysis

Data evolution analysis expresses and models regularities or trends for objects whose behavior changes over time. Although this can contain characterization, discrimination, association, classification, or the clustering of time-related data, the distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis [7].

2.5 Classifications of Data Mining System

Data mining systems can be grouped according to various criteria, as follows:

1. **Classification according to the kinds of database mined:** A data mining system can be categorized according to the type of database mined [35]. Database systems themselves can be classified according to different criteria, each of which may require its own data mining technique.
2. **Classification according to the kinds of knowledge mined:** Data mining systems can be classified according to the kinds of knowledge they mine, depending on the data mining functionalities, such as characterization, discrimination, association, classification, clustering, outlier analysis and evaluation analysis.
3. **Classification according to the kinds of techniques utilized:** Data mining systems can be classified according to the underlying data mining techniques employed. These techniques can be expressed according to the degree of user interaction involved or the methods of data analysis employed.

4. **Classification according to the application adapted:** Data mining systems can be classified according to the applications they adapt. Different applications often need the integration of application-specific methods.

2.6 Types of Data Mining Algorithms

The services of analysis contain the following algorithm types:

1. Classification algorithms guess one or more discrete variables depending on the other attributes in the dataset [42] [7].
2. Regression algorithms guess one or more continuous variables, like profit or loss, based on other attributes in the dataset.
3. Segmentation algorithms split data into groups or clusters of items with similar properties.
4. Association algorithms search correlations between different attributes in a dataset. The most common application of this kind of algorithm is for creating association rules, which can be used in a market basket analysis.
5. Sequence analysis algorithms review frequent sequences or episodes in data, like a Web path flow.

2.7 Data Mining Uses

Data mining is applied for various purposes in the private and public sectors. Enterprise like banking, insurance, medicine, and retailing commonly apply data mining to reduce costs, enhance research, and increase sales. For instance, the insurance and banking industries apply data mining applications to discover fraud and assist in risk assessment (e.g., credit scoring). Applying customer data collected over several years, companies can build up models that predict whether a customer is a good credit risk or whether an accident claim may be fraudulent and requires closer investigation [42]. The medical community sometimes uses data mining to predict the effectiveness of a procedure or medicine. Pharmaceutical firms use data mining of chemical compounds and genetic material to help guide research on new treatments for diseases. Retailers can apply information collected through affinity programs (e.g., shoppers' club cards, frequent flyer points, and contests) to assess the effectiveness of coupon offers and decisions regarding product selection and placement. And that products are often purchased together. Companies like telephone service providers and music clubs can apply data mining. So that they can create a churn analysis to assess which customers are likely to remain subscribers and which are likely to switch to a competitor [54] [7].

In the public domain, data mining applications were initially applied to identify fraud and waste. However, they are now also applied to measure and improve program performance. It has been notified that data mining has assisted the federal government recover millions of dollars in fraudulent Medicare payments. The Justice Department has applied data mining to assess crime patterns and adjust resource allotments accordingly. Furthermore, the Department of Veterans Affairs has used data mining to make demographic changes in the constituency it serves to better estimate its budgetary needs [55]. One more example is the Federal Aviation Administration, which uses data mining to review plane crash data to recognize common defects and recommend preventive measures.

In recent times, data mining has been progressively cited as an essential tool in homeland security efforts. Some observers recommend that data mining should be used to analyze terrorist activities, such as money transfers and communications, and to analyze and track individual terrorists through travel and immigration records. Two initiatives that have attracted significant attention include the now-discontinued Terrorism Information Awareness (TIA) project conducted by the Defense Advanced Research Projects Agency (DARPA) and the now-cancelled Computer-Assisted Passenger Prescreening System II (CAPPS II) that was being developed by the Transportation Security Administration (TSA). CAPPS II is being replaced by a new program called Secure Flight [21] [12].

2.8 Data Mining, Machine Learning and Statistics

Data mining obtains benefit of advances in the artificial intelligence (AI) and statistics fields. Both regulations have been tracing problems of pattern recognition and classification [11]. Both societies have significantly supplied to the understanding and application of neural nets and decision trees [21].

Data mining does not take out traditional statistical techniques. Rather, it is an addition of statistical methods that stems partly from a major change in the statistics community. The development of most statistical techniques was, until recently, based on elegant theory and analytical methods that achieved success with the limited amounts of data being evaluated [47]. Computers’ increased power and lower cost, coupled with the need to analyses huge data sets with millions of rows, have permitted the development of new techniques based on a brute-force exploration of possible solutions.

New techniques contain relatively recent algorithms like neural nets and decision trees and new approaches to older algorithms like discriminate analysis. By employing the increased computer power on the massive volumes of available data, these techniques can estimate almost any functional form or interaction. Traditional statistical techniques based on the modeler to identify the functional form and interactions.

The point is that data mining is the applying of these and other AI and statistical techniques to common business problems whereby these techniques are available to the skilled knowledge worker and the trained statistics professional. A tool for increasing the productivity of people building predictive models is called data mining.

2.9 Data Mining Applications

Data mining technologies can be applied to diverse contexts in organizations. The significant payoffs areas are expected to involve the following:

Marketing: Applications contain analyzing consumer behavior depending on buying patterns; determining marketing strategies, involving advertising, keep location, and targeted mailing; partition of customers, stores, and products; and designing catalogues, keep layouts, and advertising campaigns.

Finance: Applications contain identifying the creditworthiness of clients' segmentation of account receivables; analyzing the performance of financial investments like stocks, bonds, and mutual funds; evaluating financing options; and detecting fraud [28].

Manufacturing: Applications include optimizing resources such as machines, manpower, and materials as well as optimizing the design involved in producing processes, shop floor layouts, and product design, such as for automobiles depending on customer requirements.

Health care: Applications contain identifying the effectiveness of certain treatments, optimizing processes within hospitals, relating patient wellness data to qualifications, and analyzing the side effects of drugs [7].

2.10 Related Works

Thirunavukkarasu K. et al. took the iris dataset and used the K-Nearest Neighbors (KNN) classification Algorithm. The model can recognize the iris species automatically. The dataset had 150 samples and three classes, each containing 50 samples— the choice of performance metrics as measured by the algorithms and compared [5].

Seda Çamalan and Gökhan Şengül classified the images using K-Nearest Neighbors (KNN) and Discriminant Analysis (DA)methods. Then, their performances according to the LBP parameters were compared. Also, classification methods' parameters were changed, and the comparison results were shown [3].

Siyabend Turgut et al. presented a classification of the patients who have breast cancer. The data they used was a microarray breast cancer dataset. The methods applied were SVM, KNN, MLP, Decision Trees, Random Forest, Logistic Regression, Adaboost, and Gradient Boosting Machines. The article states that feature selection methods have resulted in improved performance accuracy. They used two feature selection methods [4].

Prasannavenkatesan Theerthagiri et al. presented a predictive disease analysis. A study was conducted to achieve prediction on four types of classification algorithms. As a dataset, predictions were calculated using classification techniques on the dataset related to COVID-19. As a result of the analysis, the KNN algorithm produced the slightest error in accurate covid-19 disease prediction than other algorithms. Nevertheless, it is a good predictor of disease risk [2].

Hnin Yu Maw et al. took data sets related to heart disease diagnosis from the UCI machine learning repository and evaluated KNN, one of the most robust classification techniques. The similarity of the symptoms of patients suffering from heart disease was calculated by Euclidean distance. Performance results were also presented. The system is made by the person making the potential decision on behalf of the doctor [1].

CHAPTER 3

k-NEAREST NEIGHBOR (k-NN) CLASSIFICATION

The vital feature of this study is the illustration of the main methodology as the core purpose of this chapter. This chapter firstly presents classification and its methods. After that, the k-NN classification and an example calculation with that algorithm are described. Finally, this chapter offers the classifier accuracy that is essential to evaluate a great amount of data concerned with the accuracy performance.

3.1 Classification

To extract models which mark out vital data and to predict future data trends, two types of data analysis called classification and prediction are used [37]. During the time categorical labels are predicted by classification, continuous-valued functions are predicted by prediction models. For example, a classification model might be built to categorize bank loan applications as either safe or risky, while a prediction model might be built to predict the expenditures of potential customers on computer equipment given their incomes and occupations. Classifying a selection contains separating the items that make up the selection into groups or classes. Usage of a model which is built on historical data performs classification in the context of data mining. To precisely figure out the target class for each record is the purpose of predictive classification [15].

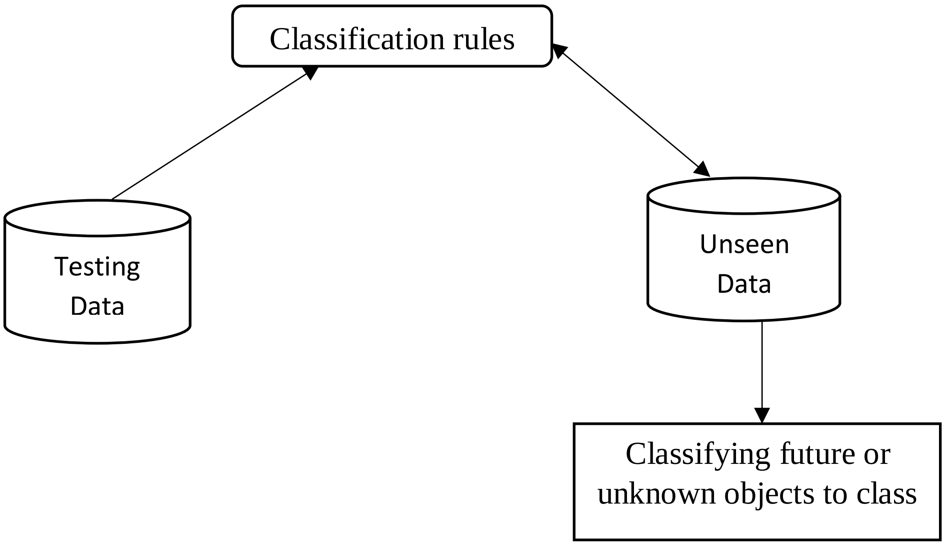


Figure 3.1 Classifier Architecture

A classification task initiates with building data (also known as training data) for which the target values (or class assignments) are known. Different techniques are used to search relationships between the predictor attributes' values and the target attributes' values in the building data by different classification algorithms. These relationships are summarized in a model; the model can then be applied to new cases with unknown target values to predict target values. A classification model can also be applied to data that was held aside from the training data to compare the predictions to the known target values; such data is known as test data or evaluation data. The comparison technique, called testing a model, measures a model's predictive accuracy.

Usage of classification is found in customer segmentation, business modelling, credit analysis, and many other applications. For instance, a credit card company may wish to predict which customers are likely to default on their payments. Customers are split into two classes: those who default and those who do not. Each customer corresponds to a case; data for each case might consist of attributes that describe the customer's spending habits, income, demographics, etc. These are the predictor attributes. The target attribute indicates whether or not the customer has defaulted. The building data is used to build a model that predicts whether new customers are likely to default.

Applying the model means the application of a classification model to new data, which is known as applying data or scoring data. Figure 3.1 explains the typical classifier process. A trained classifier is tested for accuracy using testing data with a known class. Then it can be tested on a data set with an unknown class. Either binary or multiclass targets can be identified as classification problems. Binary targets take on only two values, for example, good credit risk and poor credit risk.

3.2 Classification Methods

Many classification and prediction methods have been proposed by researchers in machine learning, expert systems, statistics, and neurobiology. The training data set is used in a classification algorithm. The classifier outputs are stored for later use; this stage is known as the LEARN model. Then test data (with known classes) is tested by the classifier (Apply Model or Test Model). If the classifier's output is the same as the known class, then the testing accuracy is good, which means this classifier algorithm and training sample has good performance. If poor accuracy results from the testing phase, the system has a bad classifier or bad training data set. The overview of classification is shown in Figure 3.2. There are training and testing processes in the classification. Training processes learn models using one of the learning algorithms (classification algorithm) from the training data set. The output model is then tested with the testing data set to deduct the testing samples. There are several classification methods, and the most widely used methods are:

1. Decision Tree Algorithm
2. Naïve Bayes Algorithm
3. *k*-Nearest Neighbors (k-NN) Algorithm
4. Neural Network
5. Support Vector Machine Algorithm

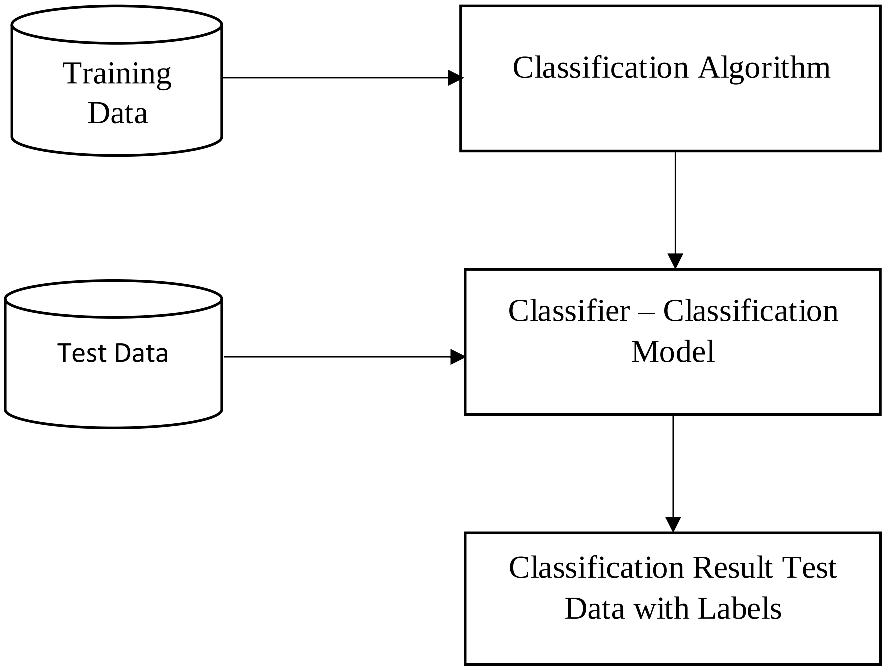

Training Data
Classification Algorithm
Test Data
Classifier – Classification Model
Classification Result Test Data with Labels

Figure 3.2 Classification Process

3.3 k-Nearest Neighbor (k-NN) Classification

In statistics, a non-parametric supervised learning method first developed by Evelyn Fix and Joseph Hodges in 1951 and later expanded by Thomas Cover, is the k-NN algorithm and is applied in classification and regression. The input contains the *k* closest training examples in a data set in each case [34].

k-NN is a kind of classification where the task is only estimated locally, and all computation is postponed until task evaluation. If the features express different physical units or come in massive different scales, then normalizing the training data can progress its accuracy dramatically as this algorithm depends on distance for classification.

An effective technique can be to assign weights to the supplement of the neighbors so that the nearer neighbors supply more to the average than the more distant ones for both classification and regression. As an example, a general weighting method involve in assigning each neighbor a weight of 1/d, where d is the distance to the neighbor.

The known class (for *k*-NN classification) or the object property value (for *k*-NN regression) of a set of objects are obtained as the neighbors. Despite no accurate training step is required, this can be consideration of as the training set for the algorithm. A peculiarity of the *k*-NN algorithm is that it is sensitive to the local structure of the data.

In this system, the data sets have numeric, binary, ordinal, and nominal data types. Data objects are contained in data sets. A data object stands for an entity. Customers, store items, and sales may be considered as the objects in a sales database. Patients could be considered as the objects in a medical database. Students, professors, and courses can be considered as the objects in a university database. Attributes express typically data objects. An attribute is a data field representing a characteristic or feature of a data object. The set of potential values (nominal, binary, ordinal, or numeric) that the attribute can have decide the kind of an attribute [46]. “Relating to names” can be denoted as nominal. A nominal attribute values are symbols or names of things. Each value expresses some type of category, code, or state, thus nominal attributes are also represented to as categorical. The values do not contain any relevant order. A nominal attribute with only two categories or states, 0 or 1, where 0 typically denotes that the attribute is absent and 1 denotes that it is present is called a binary attribute. If the two states correspond to true and false, Boolean is binary attributes. Quantitative attribute is a measurable quantity represented in integer or real values is called a numeric attribute. After data preparation, the system gets the data sets of all number types to make an evaluation.

The k-NN algorithm is one of the most widely used classification algorithms due to its simplicity and easy implementation. It is also used as the baseline classifier in many domain problems. The k-NN algorithm is a conventional non-parametric classifier for classification and regression problems. Based on the given problem or dataset, the learning and prediction analysis performs. Without any assumption on the dataset, the prediction bases on neighbor data values in the k-NN classification model. The k denotes the number of nearest neighbor data values in k-NN. It classifies the given dataset based on *k* (i.e., the number of nearest neighbors) and directly classifies the training dataset. It represents the prediction of a new instance is carried out by looking for similar "*k*" neighbor instances in the whole training set and classifying based on the class of highest instances [22]. The Euclidean distance formula is used to determine a similar instance. The Euclidean distance between two points or tuples, express, $\boldsymbol{X}_1 = (x_{11}, x_{12}, \ldots, x_{1n})$ and $\boldsymbol{X}_2 = (x_{21}, x_{22}, \ldots, x_{2n})$, is

$$dist(\boldsymbol{X}_1, \boldsymbol{X}_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}. \qquad (3.1)$$

That is, the difference between the interrelated values of the attribute in tuple $\boldsymbol{X_1}$ and in tuple $\boldsymbol{X_2}$ for each numeric attribute are calculated first. After that, that difference is squared and is accumulated. The calculation of the square root of the total accumulated distance count is done. To avoid attributes with initially huge ranges (e.g., income) from outbalancing attributes with initially smaller ranges (e.g., binary attributes), the values of each attribute are normalized generally before using Eq (3.1).

As an example, transforming a value *v* of a numeric attribute *A* to *v* 0 in the range [0, 1], that is called a Min-max normalization. That can be computed by the following equation:

$$v' = \frac{v - min_A}{max_A - min_A}, \qquad (3.2)$$

Where,

min_A = the minimum values of attribute *A*

$max_{A=}$ the maximum values of attribute A [50] [26].

Table 3.1 Variable Types with Examples.

Type	Description	Examples	Operations
Nominal	Uses a label or name to distinguish one object from another	Zip code, ID	= or not =
Binary	Uses a nominal attribute with only two categories or states	Medical test	0 or 1
Ordinal	Uses values to provide the ordering of objects	Opinion, grades	< or >

One of the simplest machine learning algorithms based on the supervised learning technique is also known as the k-nearest neighbor algorithm. It supposes a similarity between the new case/data and obtainable cases and sets the new case into the group that is most similar to the universal groups. All the obtainable data are stored, and a new data point based on the similarity are classified when new data appears. It can classify the new data into a relevant group and can be used for regression and classification. However, it is mostly used for classification problems [27]. Because of it does not make any presumption about fundamental data, it is also known as a nonparametric algorithm. Rather than storing the dataset at the time of classification, it performs an action on the dataset [44]. Throughout the training phrase, it just stores the dataset. When it gains new data, it classifies that data into a group that is nearly same as the new data [45]. Hence, it is also labeled as a lazy learner algorithm. It can be defined on the basis of the following:

- **First:** Choose the number (k) of neighbors.
- **Second:** Compute the Euclidean distance of the **number (k) of neighbors.**
- **Third:** Get the number (k) of the nearest neighbors per the computed Euclidean distance.
- **Fourth:** Count the number of data points in each category among these neighbors (k).

- **Fifth:** Select the new data points to that category for which the number of neighbors is the highest.
- **Sixth:** The model is standing by.

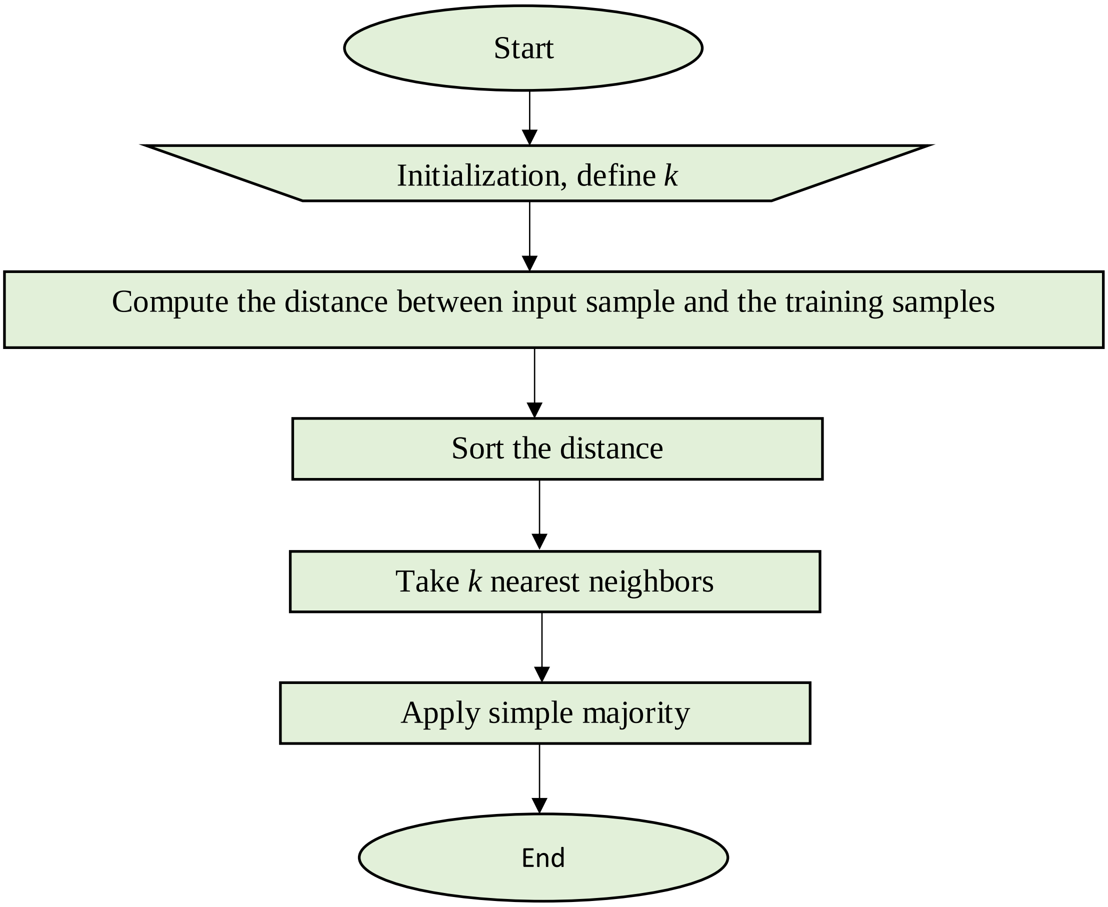

Start
Initialization, define *k*
Compute the distance between input sample and the training samples
Sort the distance
Take *k* nearest neighbors
Apply simple majority
End

Figure 3.3 Workflow diagram of k-Nearest Neighbor Algorithm

3.3.1 An Example Calculation

Let X be a data sample whose class label is unknown. Each data sample is represented by an n-dimensional feature vector, X = (X_1, X_2, …, X_n).

Table 3.2 Sample COVID-19 Immunization Training Data

ID	**Housing**	**House hold**	**L18_ M**	**L18_ F**	**O18_ M**	**O18_ F**	**Eligible**	**Quarter Village Tract or Small Village**	**…**	**Finishing Rate (Class)**
1	213	213	349	357	151	145	296	Yes	**…**	Yes
2	141	160	107	81	269	285	557	Yes	**…**	Yes

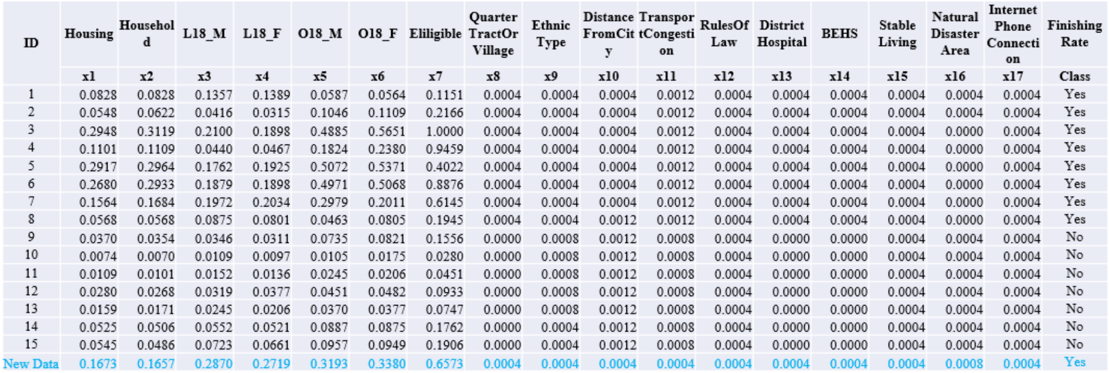
3	758	802	540	488	1256	1453	2571	Yes	...	Yes
4	283	285	113	120	469	612	2432	Yes	...	Yes
5	750	762	453	495	1304	1381	1034	Yes	...	Yes
6	689	754	483	488	1278	1303	2282	Yes	...	Yes
7	402	433	507	523	766	517	1580	Yes	...	Yes
8	146	146	225	206	119	207	500	Yes	...	Yes
9	95	91	89	80	189	211	400	No	...	No
10	19	18	28	25	27	45	72	No	...	No
11	28	26	39	35	63	53	116	No	...	No
12	72	69	82	97	116	124	240	No	...	No
13	41	44	63	53	95	97	192	No	...	No
14	135	130	142	134	228	225	453	No	...	No
15	140	125	186	170	246	244	490	No	...	No

To recognize the COVID-19 immunization finishing rate of one region in Rakhine state, the user fills the unknown sample (X) first that is unknown region in Rakhine state. Unknown sample (X) is shown in Table 3.3.

Table 3.3 Unknown Sample (X)

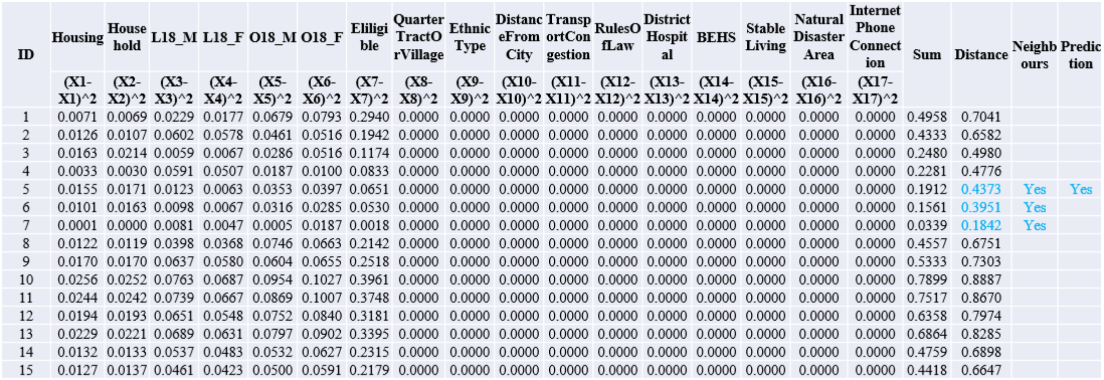
Housing	430
Household	426
L18_M	738
L18_F	699
O18_M	821
O18_F	869
Eligible	1690
QuarterTractOrVillage	Yes
...	...
Finishing Rate	Yes

Before calculation, the data are transformed by min-max normalization. After that, calculate the Euclidean distance of the new sample for each row of the training dataset. First, find the square difference for each attribute. Second, calculate the sum of the square differences. Finally, find the square root of the sum.



ID	Housing	Household	L18_M	L18_F	O18_M	O18_F	Eliligible	QuarterTractOrVillage	Ethnic Type	DistanceFromCity	TransportCongestion	RulesOf Law	District Hospital	BEHS	Stable Living	Natural Disaster Area	Internet Phone Connection	Finishing Rate
	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	Class
1	0.0828	0.0828	0.1357	0.1389	0.0587	0.0564	0.1151	0.0004	0.0004	0.0004	0.0012	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	Yes
2	0.0548	0.0622	0.0416	0.0315	0.1046	0.1109	0.2166	0.0004	0.0004	0.0004	0.0012	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	Yes
3	0.2948	0.3119	0.2100	0.1898	0.4885	0.5651	1.0000	0.0004	0.0004	0.0004	0.0012	0.0004	0.0004	0.0004	0.0004	0.0000	0.0004	Yes
4	0.1101	0.1109	0.0440	0.0467	0.1824	0.2380	0.9459	0.0004	0.0004	0.0004	0.0012	0.0004	0.0004	0.0004	0.0004	0.0000	0.0004	Yes
5	0.2917	0.2964	0.1762	0.1925	0.5072	0.5371	0.4022	0.0004	0.0004	0.0004	0.0012	0.0004	0.0004	0.0004	0.0004	0.0000	0.0004	Yes
6	0.2680	0.2933	0.1879	0.1898	0.4971	0.5068	0.8876	0.0004	0.0004	0.0004	0.0012	0.0004	0.0004	0.0004	0.0004	0.0000	0.0004	Yes
7	0.1564	0.1684	0.1972	0.2034	0.2979	0.2011	0.6145	0.0004	0.0004	0.0004	0.0012	0.0004	0.0004	0.0004	0.0004	0.0000	0.0004	Yes
8	0.0568	0.0568	0.0875	0.0801	0.0463	0.0805	0.1945	0.0004	0.0004	0.0012	0.0012	0.0004	0.0004	0.0004	0.0004	0.0000	0.0004	Yes
9	0.0370	0.0354	0.0346	0.0311	0.0735	0.0821	0.1556	0.0000	0.0008	0.0012	0.0008	0.0004	0.0000	0.0000	0.0004	0.0004	0.0004	No
10	0.0074	0.0070	0.0109	0.0097	0.0105	0.0175	0.0280	0.0000	0.0008	0.0012	0.0008	0.0004	0.0000	0.0000	0.0004	0.0004	0.0004	No
11	0.0109	0.0101	0.0152	0.0136	0.0245	0.0206	0.0451	0.0000	0.0008	0.0012	0.0008	0.0004	0.0000	0.0000	0.0004	0.0004	0.0004	No
12	0.0280	0.0268	0.0319	0.0377	0.0451	0.0482	0.0933	0.0000	0.0008	0.0012	0.0008	0.0004	0.0000	0.0000	0.0004	0.0004	0.0004	No
13	0.0159	0.0171	0.0245	0.0206	0.0370	0.0377	0.0747	0.0000	0.0008	0.0012	0.0008	0.0004	0.0000	0.0000	0.0004	0.0004	0.0004	No
14	0.0525	0.0506	0.0552	0.0521	0.0887	0.0875	0.1762	0.0000	0.0004	0.0012	0.0008	0.0004	0.0000	0.0000	0.0004	0.0004	0.0004	No
15	0.0545	0.0486	0.0723	0.0661	0.0957	0.0949	0.1906	0.0000	0.0004	0.0012	0.0008	0.0004	0.0000	0.0000	0.0004	0.0004	0.0004	No
New Data	0.1673	0.1657	0.2870	0.2719	0.3193	0.3380	0.6573	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0008	0.0004	Yes

Figure 3.4 Training Datasets and New Data for k-NN Evaluation



ID	Housing	Household	L18_M	L18_F	O18_M	O18_F	Eliligible	QuarterTractOrVillage	Ethnic Type	DistanceFromCity	TransportCongestion	RulesOfLaw	District Hospital	BEHS	Stable Living	Natural Disaster Area	Internet Phone Connection	Sum	Distance	Neighbours	Prediction
	(X1-X1)^2	(X2-X2)^2	(X3-X3)^2	(X4-X4)^2	(X5-X5)^2	(X6-X6)^2	(X7-X7)^2	(X8-X8)^2	(X9-X9)^2	(X10-X10)^2	(X11-X11)^2	(X12-X12)^2	(X13-X13)^2	(X14-X14)^2	(X15-X15)^2	(X16-X16)^2	(X17-X17)^2				
1	0.0071	0.0069	0.0229	0.0177	0.0679	0.0793	0.2940	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4958	0.7041		
2	0.0126	0.0107	0.0602	0.0578	0.0461	0.0516	0.1942	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4333	0.6582		
3	0.0163	0.0214	0.0059	0.0067	0.0286	0.0516	0.1174	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2480	0.4980		
4	0.0033	0.0030	0.0591	0.0507	0.0187	0.0100	0.0833	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2281	0.4776		
5	0.0155	0.0171	0.0123	0.0063	0.0353	0.0397	0.0651	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1912	0.4373	Yes	Yes
6	0.0101	0.0163	0.0098	0.0067	0.0316	0.0285	0.0530	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1561	0.3951	Yes	
7	0.0001	0.0000	0.0081	0.0047	0.0005	0.0187	0.0018	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0339	0.1842	Yes	
8	0.0122	0.0119	0.0398	0.0368	0.0746	0.0663	0.2142	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4557	0.6751		
9	0.0170	0.0170	0.0637	0.0580	0.0604	0.0655	0.2518	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5333	0.7303		
10	0.0256	0.0252	0.0763	0.0687	0.0954	0.1027	0.3961	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7899	0.8887		
11	0.0244	0.0242	0.0739	0.0667	0.0869	0.1007	0.3748	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7517	0.8670		
12	0.0194	0.0193	0.0651	0.0548	0.0752	0.0840	0.3181	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.6358	0.7974		
13	0.0229	0.0221	0.0689	0.0631	0.0797	0.0902	0.3395	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.6864	0.8285		
14	0.0132	0.0133	0.0537	0.0483	0.0532	0.0627	0.2315	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4759	0.6898		
15	0.0127	0.0137	0.0461	0.0423	0.0500	0.0591	0.2179	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4418	0.6647		

Figure 3.5 Sample Calculation with k-NN and Prediction

3.4 Classifier Accuracy

Estimating the classifier accuracy is important, as it permits one to calculate how precisely a given classifier will categorize future data or data on which the classifier has not been trained. It can get assistant from accuracy estimates in comparing different classifiers. Applying the training data to derive a classifier and guess its accuracy may lead to give the wrong impression and over-optimistic evaluation.

3.4.1 Cross-Validation

Each record is used the same number of times for training and once for testing in this approach. To demonstrate this method, assume the data can be split into two equal-sized subsets. Select one subset for training and one for testing first. After that, switch the tasks of the subsets so that the previous training set turns into the test set and in reverse. This approach is named twofold cross-validation [16]. The k-fold cross validation method derives this approach by dividing the data into k equal-sized partitions. The recuring processes of that procedure take k times hence each partition is used for testing precisely once. Moreover, the entire error is detected by summing up all errors for all k runs. A major case of the k-fold cross validation method puts k=N, the data set size [23]. Because of each test set involves only one record, it is called leave-one-out approach. This approach can use as much data as possible for training. Additionally, the test sets are mutually private, and they effectively coat the entire data set. The weakness of this approach is that it is computationally high-priced to recur the procedure N times [23] [9].

3.4.2 Holdout Method

The holdout method partitions randomly the given data are into two independent sets: a training set and a test set. Typically, two-thirds of the data are assigned to the training set, and the rest one-third is assigned to the test set. It uses the training set to derive the classifier and the test set estimates its accuracy [41].

In this system, about 900 COVID-19 immunization data points from Rakhine State's records are used as a dataset to calculate the classification system's performance [19]. The classifier performances in this system are estimated by the holdout method. It partitions randomly the given data into two independent sets a training set and a test set. Typically, two-thirds of the data are assigned to the training set, and the rest one-third is assigned to the test set [41].

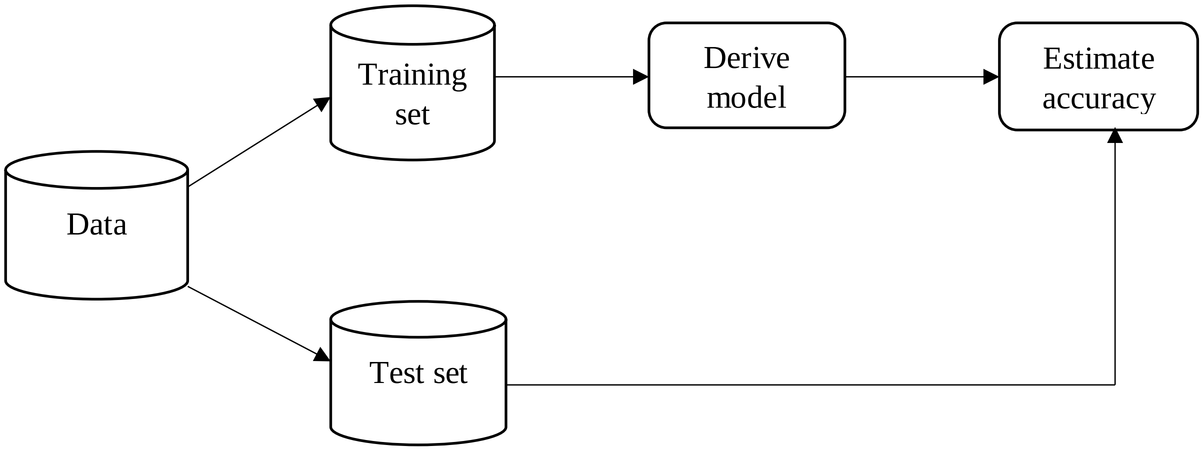


Figure 3.6 Estimating Classifier Accuracy with Holdout Method

The classifier evaluation measures include accuracy (also known as recognition rate), sensitivity (or recall), specificity, precision, and F-score. To evaluate the classification accuracy, two main matrices have been computed for the k-NN classifier in terms of the correct classification rate (%) in both the training and the testing phases. A 100% sensitivity represents that the test recognizes all 90% finishing rates that COVID-19 immunization requires. Therefore, the adequate finishing rate 90% of COVID-19 immunization in Rakhine State can be ruled out by a negative result in high-sensitivity. A 100% specificity represents that the test recognizes all 90% finishing rates of COVID-19 immunization meets as 90% finishing rate. Hence, a positive result in a high specificity test can be used to confirm the finishing rate. Precision may also be used to access the percentage of samples labelled, for instance, “Yes” of COVID-19 immunization 90% finishing rate that are actual “Yes” COVID-19 immunization 90% finishing rate samples.

Applying training data derived from the classifier or predictor to guess the accuracy of the resulting learned model can lead to misleading, over-optimistic estimation because of the overspecialization of the learning algorithm to the data. The accuracy of a classifier on a given test set is that the classifier classifies correctly the percentage of test set tuples [20].

Table 3.4 A Confusion Matrix for Positive and Negative Tuples

Predicted class

Actual

	Class 1	**Class 2**
Class 1	True positives	False negatives
Class 2	False positives	True negatives

class

$$\text{Sensitivity}= \frac{t_{pos}}{pos} \tag{3.3}$$

$$\text{Specificity}=\frac{t_{neg}}{neg} \tag{3.4}$$

$$\text{Precision}=\frac{t_{pos}}{t_{pos}+f_{pos}} \tag{3.5}$$

Where

t_{pos} = the number of true positives ("Yes" of finishing rate 90% samples that were correctly classified as such),

pos = the number of positives ("Yes" of finishing rate 90% samples),

t_{neg} = the number of true negatives ("No" of finishing rate 90% samples that were correctly classified as such),

neg = the number of negatives ("No" of finishing rate 90% samples),

f_{pos} = the number of false positives ("No" of finishing rate 90% samples that were incorrectly labelled as "Yes" of finishing rate 90%),

F-Measure= (2*Precision*Recall)/ (Precision + Recall) (3.6)

Explain that accuracy is a function of sensitivity and specificity:

$$\text{Accuracy} = \text{sensitivity}\ \frac{pos}{pos + neg} + \text{specificity}\ \frac{neg}{pos + neg} \tag{3.7}$$

The true positives, true negatives, and false positives are also useful in assessing the cost and benefits that have been computed and are associated with a classification model of this classification.

There are 900 records in this system. 600 datasets have been trained and 300 datasets are used as testing datasets. There are numerous methods to check accuracy which is used to evaluate the performance of k-nearest neighbor classification. Among them, hold-out method is used in this system. In accordance with table 3.5, testing accuracy has obtained 81%.

Table 3.5: Accuracy Result

True Positive	28
False Positive	5
True Negative	215
False Negative	52
Precision	0.8484849
Recall	0.35
F-Measure	0.4955752
Accuracy	0.81

CHAPTER 4

SYSTEM DESIGN AND IMPLEMENTATION

The overview design of the proposed system and attributes information and descriptions are firstly expressed in this chapter. And then, this chapter presents the demonstration of the system implementation and performance evaluation for k values. Finally, the attributes selection for analysis report is also described.

4.1 The Proposed System

The system is developed on Microsoft SQL Server 2017 Express Version for database and implemented using Microsoft Visual Studio 2017 Enterprise version. The system will give the prediction result for a specific region whether that region is required to the finishing rate of 90% immunization services or not by using k-NN classification and the analysis over the prediction result. The system computes Distance according to the Euclidean formula. The dataset consists of 900 regions in Rakhine State's records with 17 attributes. It makes the associated classification rules according to the k-NN method. There are two datasets: the training dataset and the testing dataset. Typically, two third of the data are assigned to the training dataset, and the rest one-third is assigned to the test set. The training dataset derives the classifier, whose accuracy is estimated with the testing data.

The user can input the attributes for the specific region in Rakhine State. Using this new region dataset, the user can know the prediction result of the finishing rate of 90% immunization services required or not, and the system gives the analysis result for that region over that prediction result.

4.2 Process Flow of the System

The system flow has described detailed specifications in the following Figure 4.1.

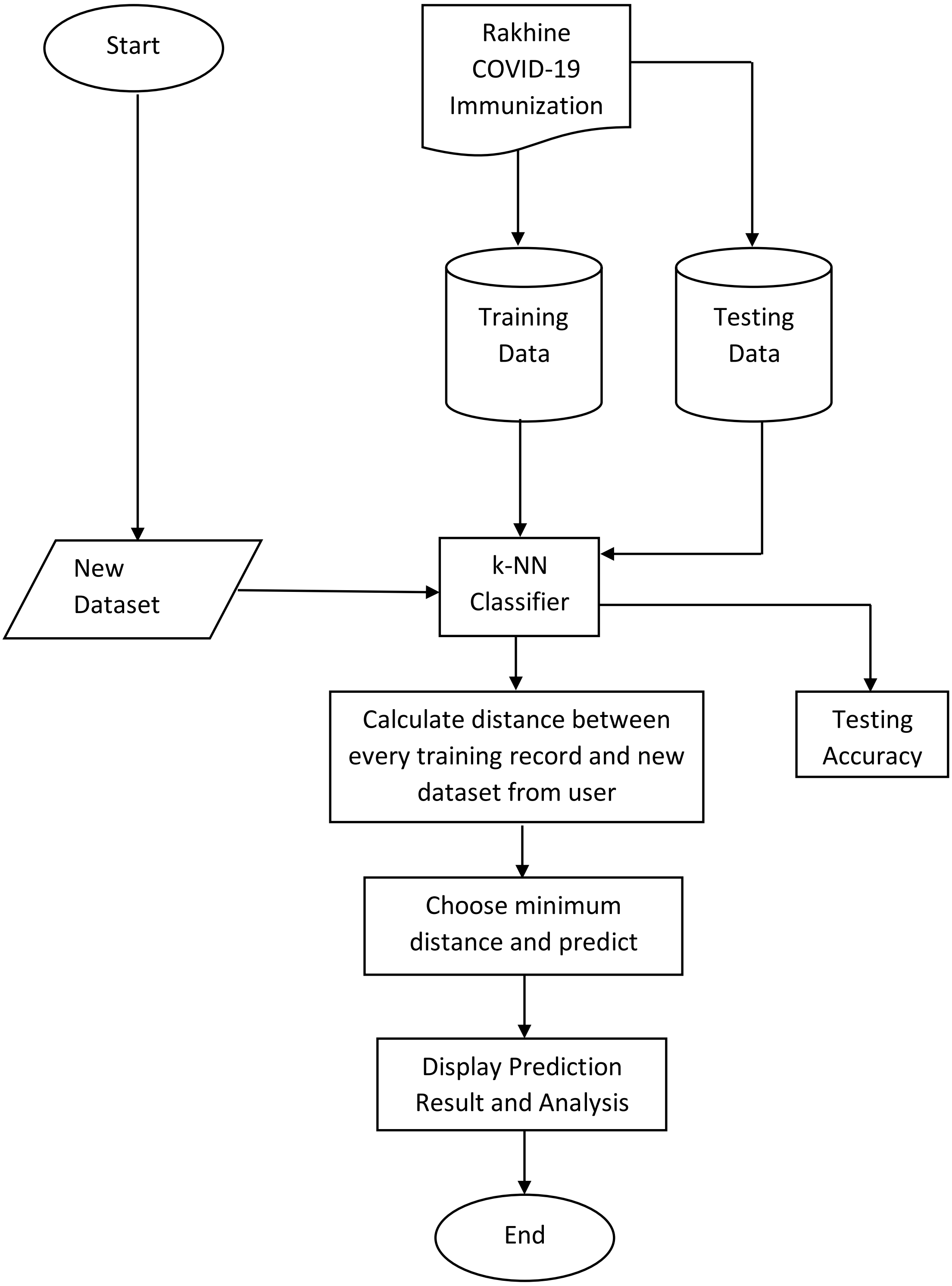

Start
Rakhine COVID-19 Immunization
Training Data
Testing Data
New Dataset
k-NN Classifier
Calculate distance between every training record and new dataset from user
Testing Accuracy
Choose minimum distance and predict
Display Prediction Result and Analysis
End

Figure: 4.1 System Flow Diagram of the System

4.3 Attributes Information and Descriptions

The attributes which are important for making a prediction whether finishing rate of 90% immunization services is required or not are shown in Table 4.1. Several pre-processing decisions had to be made before training and classifying. The system is used the dataset of COVID-19 immunization from General Administration Department, Rakhine State.

Table 4.1 Attributes Information and Descriptions

No.	Attribute Name	Description	Attribute Type	Attribute Values
1	Housing	အိမ်ခြေအရေအတွက်	Number	
2	Household	အိမ်ထောင်စုအရေအတွက်	Number	
3	L18M	၁၈ နှစ်အောက် ကျား အရေအတွက်	Number	
4	L18F	၁၈ နှစ်အောက် မ အရေအတွက်	Number	
5	O18M	၁၈ နှစ်အောက် ကျား အရေအတွက်	Number	
6	O18F	၁၈ နှစ်အောက် မ အရေအတွက်	Number	
7	Eligible	ဆေးထိုးနှံရန်လျာထားဦးရေ	Number	
8	Quarter, Village tract Or Small Village	ရပ်ကွက်၊ ကျေးရွာအုပ်စု (သို့) ကျေးရွာငယ် (ဟုတ်/မဟုတ်)	Yes	1
			No	0
9	Ethnic Type	ရပ်ကွက်/ရွာ ရှိ လူမျိုးစု (တိုင်းရင်းသား/ဘင်္ဂလီ/အရောအနှော)	Citizen	1
			Bengali	2
			Mix	3
10	Distance From City	မြို့မှ အကွာအဝေး (နီး/သင့်/ဝေး)	Near	1
			Fair	2
			Far	3

11	Transport Congestion	လမ်းပန်းဆက်သွယ်ရေး (လွယ်ကူ/ပုံမှန်/ခက်ခဲ)	Easy	1
			Normal	2
			Difficult	3
12	Rules of Law	ဒေသတွင်း တရားဥပဒေ စိုးမိုးမှု (ကောင်း/သင့်/ဆိုး)	Good	1
			Fair	2
			Bad	3
13	District Hospital	ကျန်းမာရေး တိုက်နယ်ဆေးရုံ (ရှိ/မရှိ)	Yes	1
			No	0
14	BES	အခြေခံပညာရေးဆိုင်ရာကျောင်း (ရှိ/မရှိ)	Yes	1
			No	0
15	Stable Living	ရပ်ကွက်/ရွာတွင် အတည်တကျ နေထိုင်ခြင်း (ဟုတ်/မဟုတ်)	Yes	1
			No	0
16	Natural Disaster Area	သဘာဝဘေးဒဏ်ခံရသောဒေသ (ဟုတ်/မဟုတ်)	Possible	1
			Rarely	0
17	Internet/Phone Connection	အင်တာနက်/ဖုန်း ဖြင့် ဆက်သွယ်နိုင်မှု (ပုံမှန်/ရံဖန်ရံခါ/လုံး၀)	Normal	1
			Often	2
			Never	3
18	Finishing Rate	ကာကွယ်ဆေးထိုးနှံခြင်း ၉၀% ပြည့်မှီခြင်း (ပြည့်မှီ/လိုအပ်)	Yes	1
			No	0

4.4 Detailed Implementation

Figure 4.2 shows the main form of the system. The sub-menus like ‘system’ and ‘COVID-19’ of the File menu explain the system and COVID-19. When the “Login” sub-menu is clicked from the File menu to management as admin role, the login form appears as shown in Figure 4.3. The ‘Exit’ sub-menu is to leave the system, and the “Prediction” menu is to check the prediction result and analysis from the user level.

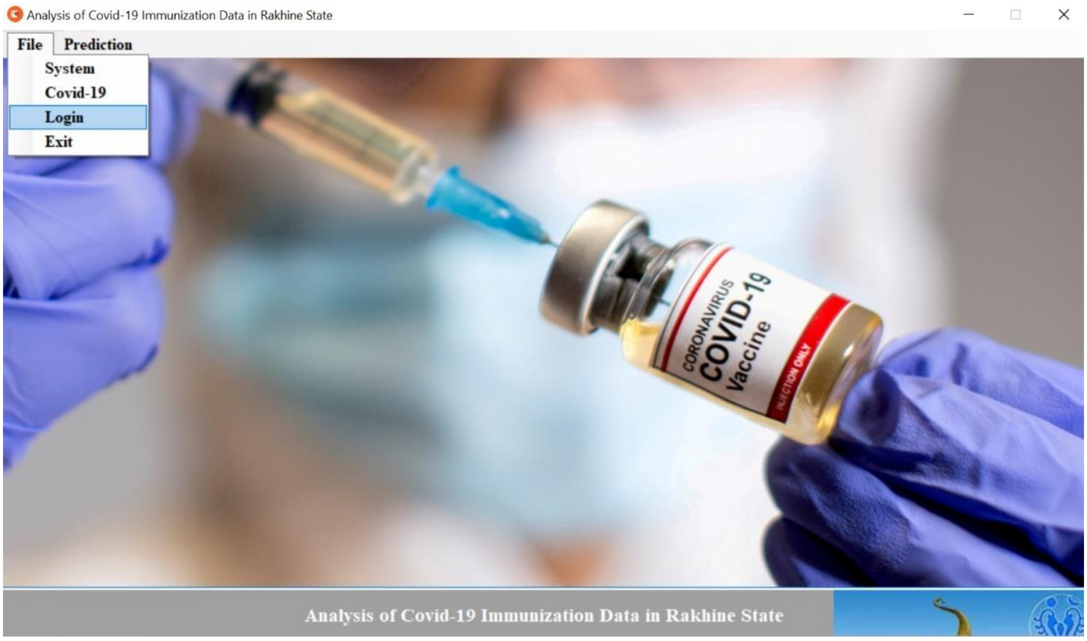

Analysis of Covid-19 Immunization Data in Rakhine State
File
Prediction
System
Covid-19
Login
Exit
CORONAVIRUS
COVID-19
Vaccine
INJECTION ONLY
Analysis of Covid-19 Immunization Data in Rakhine State

Figure 4.2 Main Form of the System

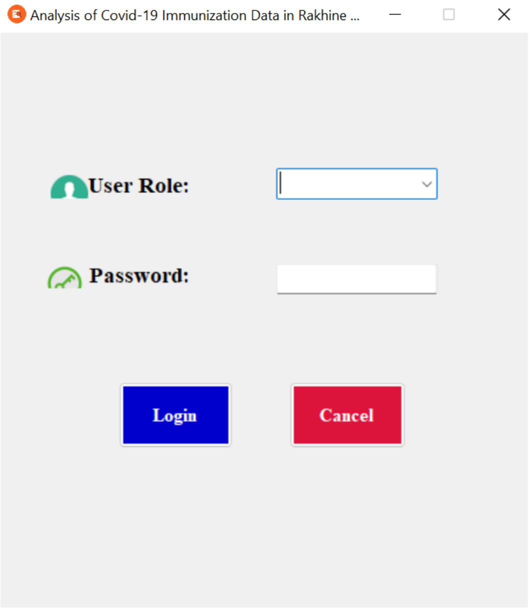


Figure 4.3 Login Form

After login as admin, the admin main form appears as shown in Figure 4.4. In this menu bar, there are three sub-menus, namely: "Training Dataset", "Testing Dataset", "Prediction", and the "logout" buttons. Training Dataset sub-menu is to build the model, the Testing Dataset sub-menu is to determine the model's accuracy, and the Prediction sub-menu is to make predictions and analysis results, as shown in Figure 4.8. The logout button is to log out from the admin role [49].

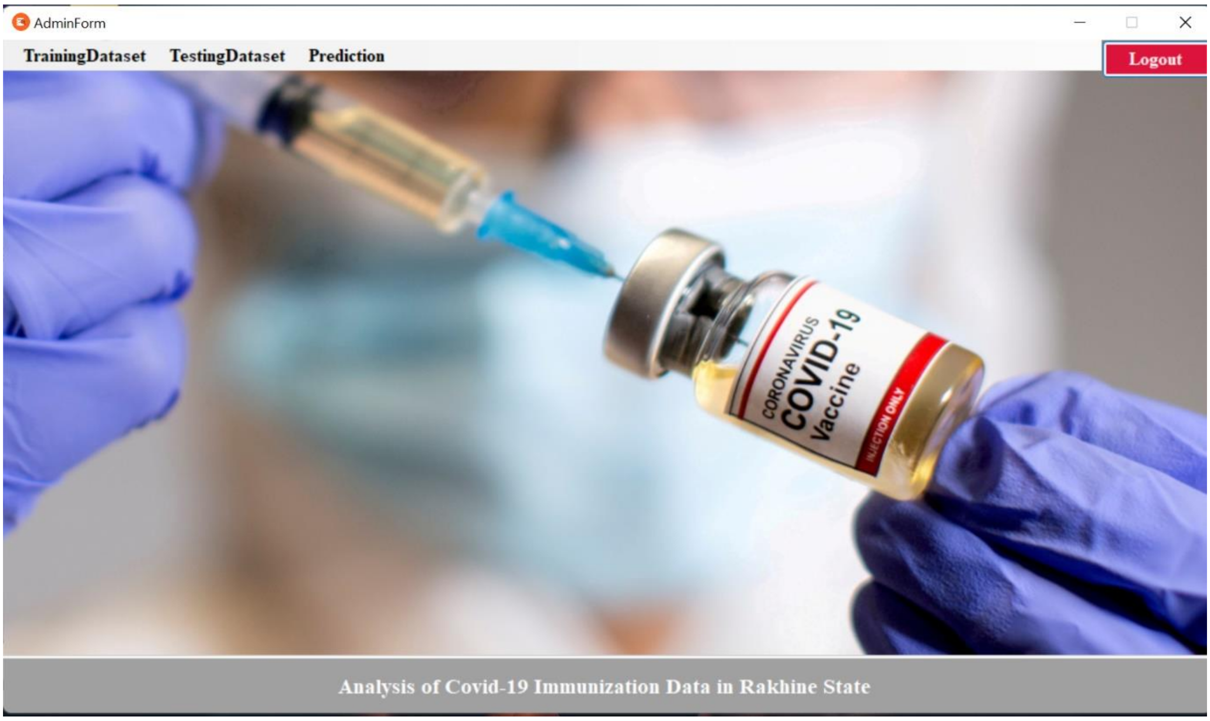
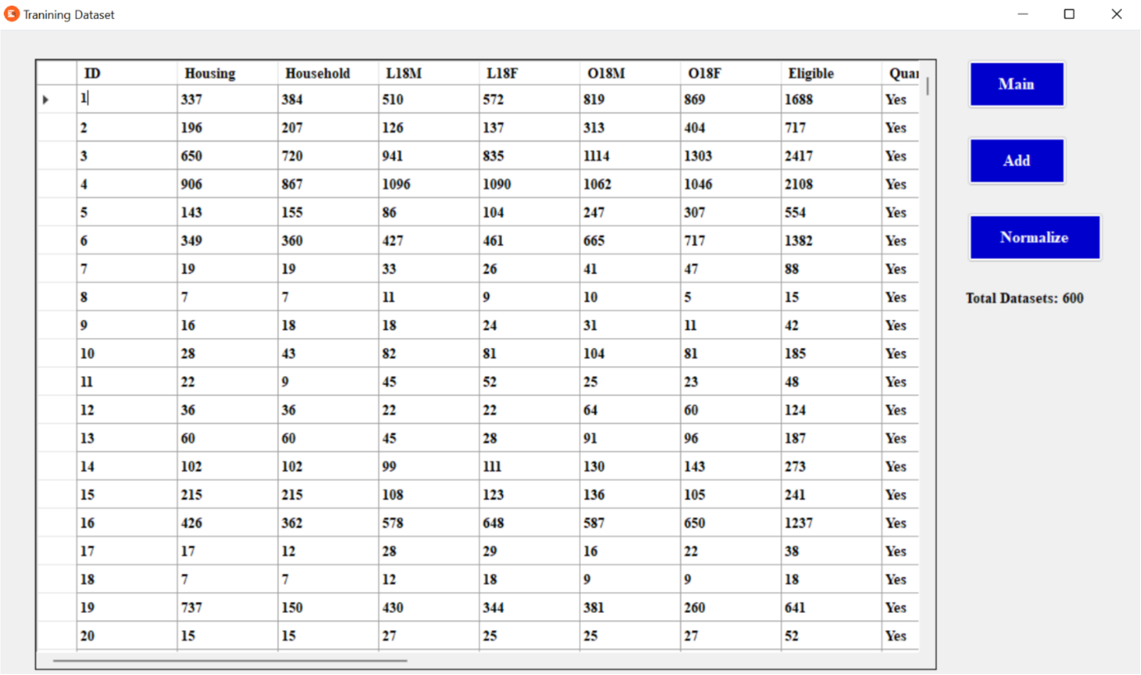


Figure 4.4 Admin Form

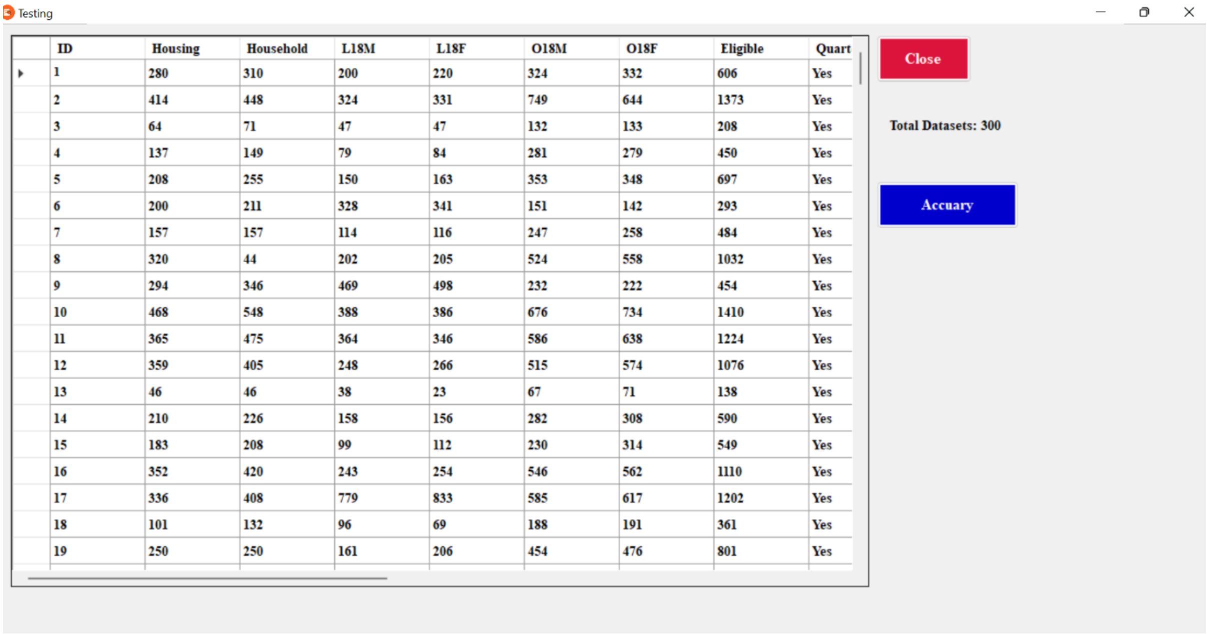
By clicking the training dataset sub-menu, training data set appears as shown in Figure 4.5. "Normalize" button is to show the normalized data of the training data set. The "Add" button is to save the new record for training, and the "Main" button is to go back Admin Form. After clicking the testing data set sub-menu, the testing data set appears as shown in Figure 4.6.



ID	Housing	Household	L18M	L18F	O18M	O18F	Eligible	Quai
1	337	384	510	572	819	869	1688	Yes
2	196	207	126	137	313	404	717	Yes
3	650	720	941	835	1114	1303	2417	Yes
4	906	867	1096	1090	1062	1046	2108	Yes
5	143	155	86	104	247	307	554	Yes
6	349	360	427	461	665	717	1382	Yes
7	19	19	33	26	41	47	88	Yes
8	7	7	11	9	10	5	15	Yes
9	16	18	18	24	31	11	42	Yes
10	28	43	82	81	104	81	185	Yes
11	22	9	45	52	25	23	48	Yes
12	36	36	22	22	64	60	124	Yes
13	60	60	45	28	91	96	187	Yes
14	102	102	99	111	130	143	273	Yes
15	215	215	108	123	136	105	241	Yes
16	426	362	578	648	587	650	1237	Yes
17	17	12	28	29	16	22	38	Yes
18	7	7	12	18	9	9	18	Yes
19	737	150	430	344	381	260	641	Yes
20	15	15	27	25	25	27	52	Yes

Total Datasets: 600

Figure 4.5 Training Dataset

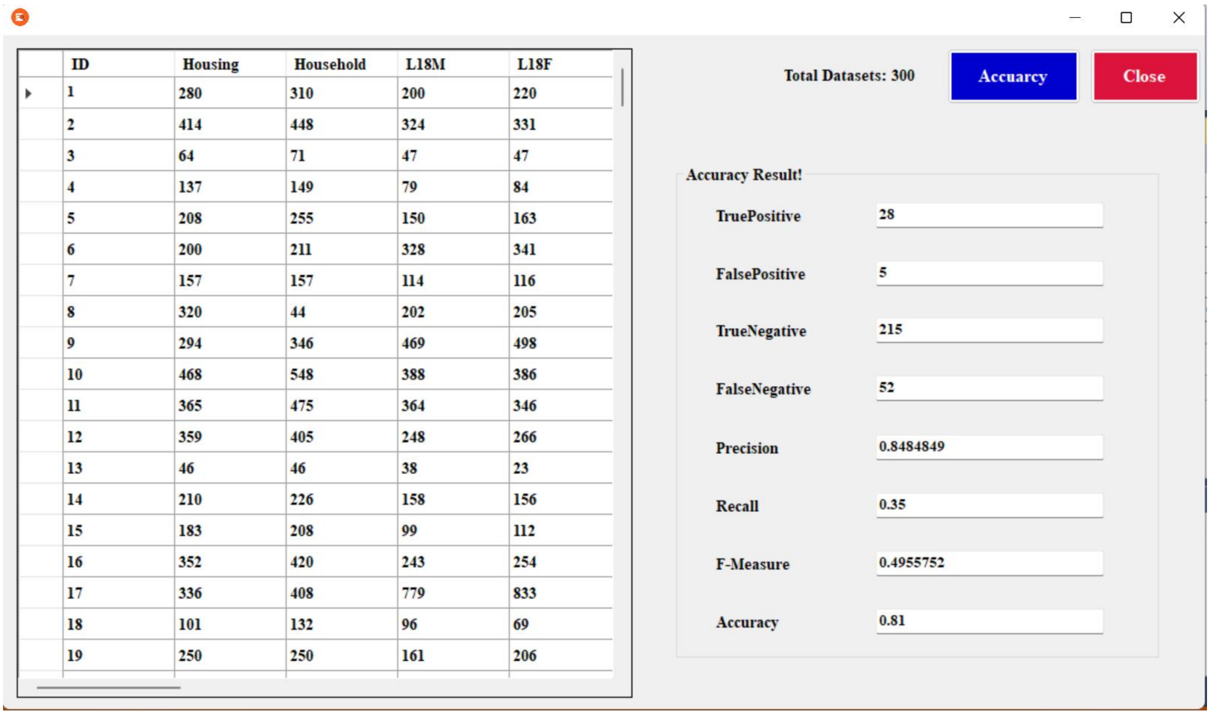


ID	Housing	Household	L18M	L18F	O18M	O18F	Eligible	Quart
1	280	310	200	220	324	332	606	Yes
2	414	448	324	331	749	644	1373	Yes
3	64	71	47	47	132	133	208	Yes
4	137	149	79	84	281	279	450	Yes
5	208	255	150	163	353	348	697	Yes
6	200	211	328	341	151	142	293	Yes
7	157	157	114	116	247	258	484	Yes
8	320	44	202	205	524	558	1032	Yes
9	294	346	469	498	232	222	454	Yes
10	468	548	388	386	676	734	1410	Yes
11	365	475	364	346	586	638	1224	Yes
12	359	405	248	266	515	574	1076	Yes
13	46	46	38	23	67	71	138	Yes
14	210	226	158	156	282	308	590	Yes
15	183	208	99	112	230	314	549	Yes
16	352	420	243	254	546	562	1110	Yes
17	336	408	779	833	585	617	1202	Yes
18	101	132	96	69	188	191	361	Yes
19	250	250	161	206	454	476	801	Yes

Total Datasets: 300

Figure 4.6 Testing Dataset

When "Accuracy" button is clicked, the accuracy result of this system appears as shown in Figure 4.7.



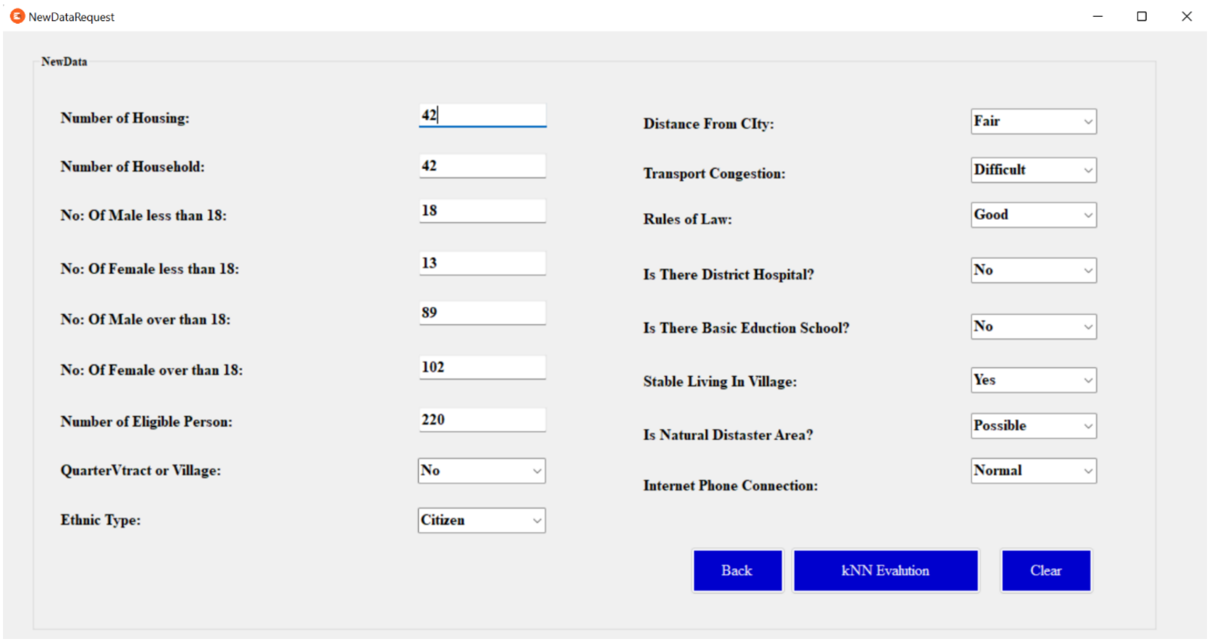
ID	Housing	Household	L18M	L18F
1	280	310	200	220
2	414	448	324	331
3	64	71	47	47
4	137	149	79	84
5	208	255	150	163
6	200	211	328	341
7	157	157	114	116
8	320	44	202	205
9	294	346	469	498
10	468	548	388	386
11	365	475	364	346
12	359	405	248	266
13	46	46	38	23
14	210	226	158	156
15	183	208	99	112
16	352	420	243	254
17	336	408	779	833
18	101	132	96	69
19	250	250	161	206

Total Datasets: 300

Accuracy Result!

TruePositive	28
FalsePositive	5
TrueNegative	215
FalseNegative	52
Precision	0.8484849
Recall	0.35
F-Measure	0.4955752
Accuracy	0.81

Figure 4.7 Accuracy Result

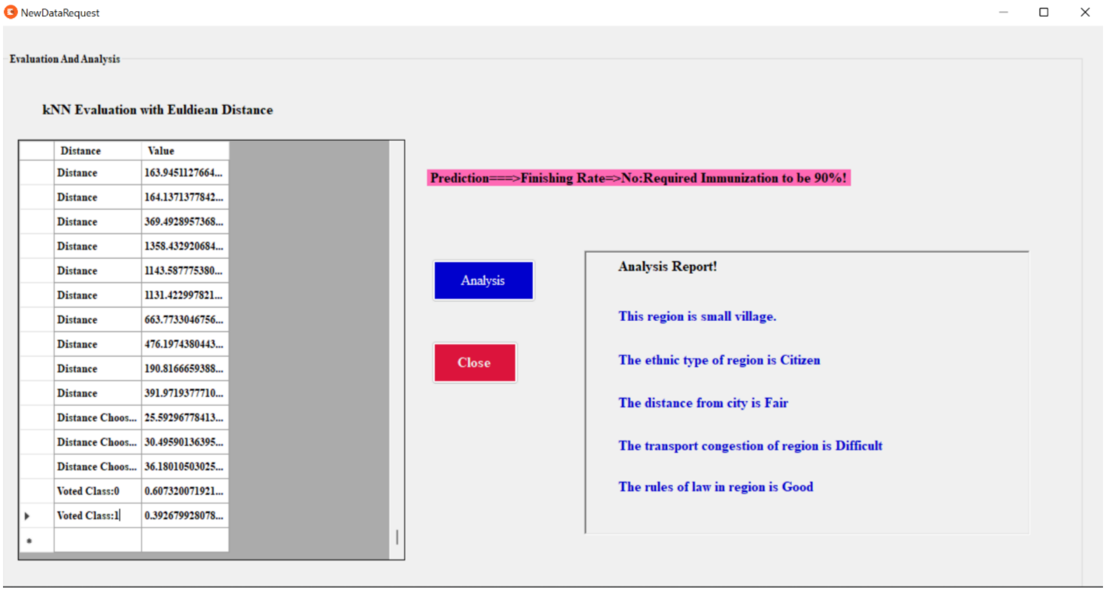


Number of Housing:	42	Distance From CIty:	Fair
Number of Household:	42	Transport Congestion:	Difficult
No: Of Male less than 18:	18	Rules of Law:	Good
No: Of Female less than 18:	13	Is There District Hospital?	No
No: Of Male over than 18:	89	Is There Basic Eduction School?	No
No: Of Female over than 18:	102	Stable Living In Village:	Yes
Number of Eligible Person:	220	Is Natural Distaster Area?	Possible
QuarterVtract or Village:	No	Internet Phone Connection:	Normal
Ethnic Type:	Citizen		

Figure 4.8 New Data Request Form

A k-NN classifier can be initialized by inputting a record set to calculate the prediction. New information can be entered by the user as shown in Figure 4.8.

When the data fill all inputs data and clicks “k-NN Evaluation” button, the result of Euclidean distance and prediction result appears as shown in Figure 4.9. “Analysis” button is clicked the analysis report appears.



kNN Evaluation with Euldiean Distance

Distance	Value
Distance	163.9451127664...
Distance	164.1371377842...
Distance	369.4928957368...
Distance	1358.432920684...
Distance	1143.587775380...
Distance	1131.422997821...
Distance	663.7733046756...
Distance	476.1974380443...
Distance	190.8166659388...
Distance	391.9719377710...
Distance Choos...	25.59296778413...
Distance Choos...	30.49590136395...
Distance Choos...	36.18010503025...
Voted Class:0	0.607320071921...
Voted Class:1	0.392679928078...

Prediction===>Finishing Rate=>No:Required Immunization to be 90%!

Analysis Report!

This region is small village.

The ethnic type of region is Citizen

The distance from city is Fair

The transport congestion of region is Difficult

The rules of law in region is Good

Figure 4.9 k-NN Evaluation and Analysis Result

4.5 Experimental Results

4.5.1 Performance Evaluation for k value

Before k-NN evaluation, the value of k must be chosen first. Thus, need to consider which value of k is the best suitable for calculation. The system is binary classification, the value of k should be used odd number not to confuse with the class. For instance, the minimum distance training examples output will be the predicted value for this testing examples for k=1. Consider the minimum distance training examples output and calculate the number of accept and reject for k=3. The largest one will be the predicted value for this testing example. The same procedure is repeated for k=5. Last of all for each k, the (%) of accuracy is calculated that is shown in Table 4.2. From Table 4.2, for this problem at k=3, get the maximum accuracy which is 81 (%).

Table 4.2 (%) of Accuracy for Different k Using k-NN

K	True Positive	True Negative	False Positive	False Negative	Accuracy
3	28	215	5	52	81%
5	8	202	72	18	70%

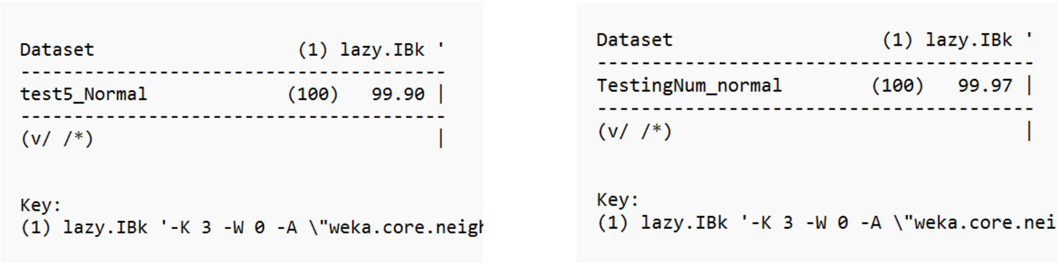
4.5.2 Attributes Selection for Analysis Report

According to the department, the 5 attributes that shown in Table 4.3 can test whether a region has reached 90% vaccination coverage or not. First, the 17 attributes are calculated in the system. Thus, calculation with those 5 attributes is done with weka tool. And it is found that the accuracy % of those 5 attributes is nearly the same as calculated with the previous 17 attributes as shown in Figure 4.10. Therefore, as per the instructions of the department and calculation of accuracy %, those 5 attributes are released for the analysis report.

Table 4.3 Attributes for Analysis Report

No	Attribute Name	Description	Attribute Type	Attribute Values
1	Quarter,		Yes	1

	Village tract Or Small Village	ရပ်ကွက်၊ ကျေးရွာအုပ်စု (သို့) ကျေးရွာငယ် (ဟုတ်/မဟုတ်)	No	0
2	Ethnic Type	ရပ်ကွက်/ရွာ ရှိ လူမျိုးစု (တိုင်းရင်းသား/ဘင်္ဂလီ/အရောအနှော)	Citizen	1
			Bengali	2
			Mix	3
3	Distance From City	မြို့မှ အကွာအဝေး (နီး/သင့်/ဝေး)	Near	1
			Fair	2
			Far	3
4	Transport Congestion	လမ်းပန်းဆက်သွယ်ရေး (လွယ်ကူ/ပုံမှန်/ခက်ခဲ)	Easy	1
			Normal	2
			Difficult	3
5	Rules of Law	ဒေသတွင်း တရားဥပဒေ စိုးမိုးမှု (ကောင်း/သင့်/ဆိုး)	Good	1
			Fair	2
			Bad	3



```
Dataset                       (1) lazy.IBk '
-----------------------------------------
test5_Normal                (100)    99.90 |
-----------------------------------------
(v/ /*)                                    |


Key:
(1) lazy.IBk '-K 3 -W 0 -A \"weka.core.neigh
```

```
Dataset                       (1) lazy.IBk '
-----------------------------------------
TestingNum_normal           (100)    99.97 |
-----------------------------------------
(v/ /*)                                    |


Key:
(1) lazy.IBk '-K 3 -W 0 -A \"weka.core.nei
```

Figure 4.10 Accuracy Result for 5 Attributes and 17 Attributes

CHAPTER 5

CONCLUSION AND FURTHER EXTENSION

A form of data analysis that can be used to extract models describing an important data class or to predict future data trends is called classification. This system is intended to develop an effective solution and suggestion for COVID-19 immunization results in Rakhine State. The finishing rate of the COVID-19 immunization results in Rakhine State can be produced as classification rules by using the k-NN algorithm and the classifier accuracy can be evaluated by using the holdout method. The performance and practical use of this system has been proved in testing the COVID-19 immunization data in Rakhine State.

One of the most efficient and effective predictive learning algorithms for data mining is k-NN. The system describes the analysis of the nonlinear algorithm for addressing the problem with COVID-19 immunization data in Rakhine State using the classification method.

5.1 Advantages of the System

This system can have some benefits for Rakhine State to make valuable predictions for COVID-19 immunization and analysis results. Thus, utilizing knowledge and experience of regions collected in databases is considered a valuable option. This study provides insights into using a classification model to predict COVID-19 immunization finishing rates in Rakhine State. The system can help the users to classify the finishing rate of immunization data for COVID-19 in Rakhine State. Therefore, it is hoped that the attempt based on the knowledge and experience, collection database of the regions can support the immunization process in Rakhine State.

5.2 Limitations of the System

There are some limitations to this system. This system was implemented by using k-NN classification. This system was tested in 900 regions of Rakhine State for COVID-19 immunization. The data was obtained from the General Administration Department of Rakhine State. Immunization of high school students was not included. If there have been more training datasets, it can increase the accuracy.

5.3 Further Extensions

The current work of this study analyzes the immunization data for people aged 60 and over in some regions of Rakhine state using k-NN supervised learning approach. Moreover., the performance of attributes selection for analysis report is also evaluated. As the future work, this system can be used to test all regions in Rakhine state, which support people who are working in COVID-19 defense areas.

REFERENCES

[1] Hnin Yu Maw et al., “Evaluation of Symptoms in Heart Disease Patients by using k - Nearest Neighbor Classification”

[2] Prasannavenkatesan Theerthagiri et al, “Prediction of COVID-19 Possibilities using KNearest Neighbour Classification Algorithm “, DOI:10.31782/IJCRR. 2021.SP173, Jan 2021

[3] Seda Çamalan et al., “Gender Prediction by Using Local Binary Pattern and K Nearest Neighbor and Discriminant Analysis Classifications”

[4] Siyabend Turgut et al., "Microarray breast cancer data classification using machine learning methods", April 2018, 2018 Electric Electronics, Computer Conference, DOI:10.1109/EBBT.2018.8391468

[5] Thirunavukkarasu K. et al., “Classification of Iris dataset using Classification based KNN Algorithms in Supervised Learning”, 2018 4th International Conference on Computing Communication and Automation (ICCCA) 2018 4th International Conference on Computing Communication and Automation (ICCCA)

[6] "Covid-19 and Escalating Conflict: Three Priorities for Rakhine State", Series No 3, Jan 2021, The Asia Foundation, Smart Peace

[7] Jiawei Han et al. "Data Mining Concepts and Techniques",Third Edition

[8] Hasegawa, Shiori, et al. “Thromboembolic Adverse Event Study of Combined Estrogen-Progestin Preparations Using Japanese Adverse Drug Event Report Database.” PLoS One, vol. 12, no. 7, Public Library of Science, July 2017, p. e0182045.

[9] https://18f.gsa.gov/2020/04/21/a-token-of-our-affection-uswds-2/

[10] https://ai.plainenglish.io/k-nearest-neighbour-as-a-feature-engineering-with-python -39b41f7a203b

[11] https://axon.cs.byu.edu/~martinez/classes/478/slides/dmintro.pdf

[12] https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.215.2703&rep=rep1 &type=pdf

[13] https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.677.3870&rep=rep1&type=pdf

[14] https://cyberleninka.org/article/n/188361

[15] https://docs.oracle.com/cd//B19306_01/datamine.102/b14339/3predictive.htm

[16] https://docs.tamr.com/new/v2021.3.0-2021.6.0/docs/glossary-1

[17] https://ebooks.ibsindia.org/business-intellegence/chapter/unit-2-preprocessing-the-data/

[18] https://geethacomputech.blogspot.com/2013/01/data-mining-unit-1-chapter-1.html

[19] https://github.com/CUAI/Intermediate-Level-Attack

[20] https://github.com/zainabmurtaza/Decision-Tree-Classifiers-and-Pipelining-in-Apache-Spark

[21] https://irp.fas.org/crs/RL31798.pdf

[22] https://ivypanda.com/essays/data-mining-essay/

[23] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0022626

[24] https://kpu.pressbooks.pub/psychmethods4e/chapter/finding-a-research-topic/

[25] https://learn.microsoft.com/en-us/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=asallproducts-allversions

[26] https://manualzz.com/doc/20847211/contents

[27] https://pdfs.semanticscholar.org/ad92/532208acaf70a6ed569f5ec68555f17b4e67.pdf

[28] https://remoteswap.club/retirement-planning-digital-nomads-us-citizen/

[29] https://socialanswerservice.com/qa/what-do-you-mean-by-technology-characterization.html

[30] https://studylib.net/doc/7206211/data-mining-functionalities-data-mining-functionalities-a...

[31] https://sungsoo.github.io/2015/04/30/cluster-analysis.html

[32] https://textbooks.elsevier.com/manualsprotectedtextbooks/9780123814791/Instructor's_manual.pdf

[33] https://thebrandboy.com/shoe-brand-names/

[34] https://theintactone.com/2021/11/28/k-nearest-neighbor/

[35] https://www.brainkart.com/article/Classification-of-Data-Mining-Systems_8309/

[36] https://www.brainkart.com/article/Data-Mining-Functionalities---What-Kinds-of-Patterns-Can-Be-Mined-_8307/

[37] https://www.coursehero.com/file/132438564/Classification-and-Prediction-of-Datadocx/

[38] https://www.coursehero.com/file/156713650/Classificationpdf/

[39] https://www.coursehero.com/file/pigk6g3/Data-mining-tasks-can-be-divided-into-two-categories-a-Descriptive-and/

[40] https://www.csithub.com/public/storage/file_downloads/GreatCompiledNotes DataMiningV1_1581660184.pdf

[41] https://www.datamining365.com/2020/04/classifier-accuracy.html

[42] https://www.everycrsreport.com/reports/RL31798.html

[43] https://www.gyanvihar.org/wp-content/uploads/2019/03/data-mining.pdf

[44] https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[45] https://www.kdnuggets.com/2021/05/essential-machine-learning-algorithms-beginners.html

[46] https://www.linkedin.com/pulse/attribute-data-types-asim-krticic

[47] https://www.linkedin.com/pulse/data-mining-machine-learning-statistics-hannington-wamala

[48] https://www.pba-analytics.com/

[49] https://www.privacyend.com/guides/tutanota-guide/

[50] https://www.sciencedirect.com/topics/computer-science/neighbor-classification

[51] https://www.slideserve.com/salvatore/applications-and-trends-in-data-mining

[52] https://www.slideshare.net/Niyitegekabilly/dam-techniques-and-dss

[53] iasir.net/IJSWSpapers/IJSWS16-201.pdf

[54] www.ijcee.org/papers/044.pdf

[55] www.ijcte.org/papers/024.pdf

[56] www.ijeir.org/Download/conference/15.pdf

AUTHOR'S PUBLICATION

[1] Khin Myat Thu, Thaung Myint Htun, "The Analysis of COVID-19 Immunization Data in Rakhine State Using KNN Algorithm", the Proceedings of the 10th Conference on Parallel and Soft Computing (PSC 2022), Yangon, Myanmar, 2022.