Analyzing reasons for unanswered questions in Stack overflow

Name: Dissanayake B.A.K
Index No: 198748V

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of Degree of Master of Science in Information Technology.

July 2022

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student | Signature of Student

B.A.K. Dissanayake | ……

Date: ……..

Supervised by

Name of Supervisor | Signature of Supervisor

Dr. Chaman Wijesiriwardana | …….

Date: ……...

Acknowledgements

First and foremost, I would like to express my sincere gratitude towards my supervisor, Dr.Chaman Wijesiriwardana, Senior Lecturer, Faculty of Information Technology, University of Moratuwa for his guidance, supervision, advice and sparing valuable time thorough the research project. I have learned so much from our project discussions and his willingness to motivate me contributed tremendously to the study.

My big thank should goes to Dr.Saminda Premarathne for taught Data Mining subject and for all its efforts and facilities that it has contributed towards successful completion of this postgraduate programme.

Furthermore, my respect should go to all the lecturers in M.Sc in Information Technology degree program of Faculty of IT, who gave their hands to sharpen our knowledge and ideas throughout these two years as they were the illumination which lit up our path ways to success.

Also, I would like to thank my family for the greatest support they had provided me through my entire life by providing me with a good environment that influenced a lot to achieve this goal.

Abstract

In the digital world knowledge and learning depends on the internet, and it has many advantages to human. Stack overflow is significant site to software developers as well as all Information technology users. This research proposed analyzing reasons for unanswered questions in Stack overflow platform.

An end-user of stack overflow website must be analysis in various methods, the users don't have questions from one programming language and don't have equal knowledge, but their answer or feedback should be correct and must help to improve knowledge and resolved the issue. In this research primary dataset mainly gathered from stack overflow question database. It considers attributes of stack overflow dataset which are stored to analyze unanswered questions based on supervised learning, classification models.

In Stack Overflow more than 50% questions are not frequently use when compare with other sites, as a result of heigh filtering methods. Programmers usually update knowledge by reading and answering new problems. And share knowledge with other programmers frequently Stack overflow is a famous platform among IT users as well as programers to send question and a get answer. Most of problems immediately get an answer by other users and few questions remain without answers, and it is a problem for Stack overflow platform developers to keep these questions for long time as it take disk space only , because of platform developers remove these questions after some time. This is where this research problem starts. Find reason and help users to get answer is target of this project. Taken dataset with answered as well as unanswered questions with no upvoted or accepted answers. These unanswered are from android, localization, asp.net, JavaScript, java, xml, SharePoint, c, .net, mobil, sql, python, php, c#, html, jQuery, iOS, CSS …etc areas. At once it seems every question has an answer, when look at these counts it realized that there are lot of questions without answer. This project I want to find reason why questions posted on Stack Overflow has not answered. This reason analysis focuses 14 attributes of dataset questions. Day by day it rises questions and unanswered questions different areas of user knowledge so this will reflect developers how to get a proper answer quickly and it will make live changes in stack overflow site also. Based on dataset attribute values can not to find out proper path to analysis the unanswered questions at once. It needs to get more complicated datasets in different perspective of software development area.

Research has tended to focus for several clusters based on end user question, and it is easy to get efficient answer. This analysis method has used in my research work. It will be based on structured primary dataset, taken from previous research. By using this analysis users can easily predict, which questions will have answer efficiently or not answered reason.

Contents

Page

List of Figure

List of tables

Chapter 1

Introduction

1.1 Prolegomena

In the contemporary world many people depend on internet and their knowledge services because of new technology and fast changes in software industry every person can have a question at any moment. And because of expert knowledge availability ty there will be a possibility to get answer and gain required details. As the usage rate of stack overflow is becoming higher, the user also often take support to get lot of knowledge to improve their studies or software quality.

1.2 Background and Motivation

which helps to find website for professional and enthusiast programmers. It is the flagship site of the Stack Exchange Network, created in 2008 by Jeff Atwood and Joel Spolsky. “It features questions and answers on a wide range of topics in computer programming. It was created to be a more open alternative to earlier question and answer websites such as Experts-Exchange” [09].

A. Three primary reasons people hang around Stack Overflow.

First is the challenge. SO provides a unique environment where can come and browse through what are essentially challenges of my skills and abilities as a programmer. Can choose small challenge or a large one, and in most cases, it is a real-life problem. The experience of figuring these questions out is invaluable. Often there is a rush to be the first with the correct answer, and this adds another element of challenge to it.

Second, and together with the first, is the validation of skills. How to answer a question is met with feedback from peers in the industry, and it is very rewarding to see that solutions are correct.

Third is the learning. Just by reading questions, can get a sense of various technologies and how they are used. If see a question asked about our favorite language on how to accomplish a problem, might have never even considered tackling before, odds will learn a new way to use that language. can become a better programmer simply by answering questions here. It is a primary learning tool for every person in an industry

where things can change rapidly and training difficult to come by once in an established career.

The reputation and badges are the tangible rewards for the community and the learning that keeps users to use SO frequently.

B. People on Stack Overflow are always ready to help someone solve their problem.

What makes people want to help, want to share what they know to each other, and what makes them not. This research focus Analyzing reasons for unanswered questions.

C. Found stack overflow answers directly to the point. The site is carefully engineered to attract "people knowledgeable [and eager to share], or those eager to learn", and that is exactly what produces the "nice" behaviors.

A challenging area is user must wait for find the appropriate answer to their question. Some questions left without answer, user has no idea why it is or what should do next. This research output can help stack overflow users to know reason for not answering after analyzing the question. Few research has addressed questions categories and found datasets there are question and answer type, but unanswered questions left without any reason. This project shows some reason for them why get delay or not answer, user can fix the question and get support from stack overflow experts. Some research help to the software engineers' only, In this work general users have focused who cannot understand the detailed report. When give short problem to few main categories every person will understand clearly.

This research predict reason by analyzing and evaluating secondary dataset by using supervised learning technology. In this research mainly use primary data collected previous for questions analysis project, which had used text mining techniques. Here in this project, create some classifiers using data mining techniques. And second stage apply properly configured classification model, after testing for several algorithms it had decided most suitable algorithm to use with this dataset.

This research can aim is to find reason by using predictive analysis and conclude in various reasons of not answer usage of Stack overflow. Based on this outcome of the stack overflow could give reason if not answered for long time before remove question. In future idea, from this updated analysis link user can find better way to feed the question and get success answer fast.

1.3 Problem Statement

Who answers questions on Stack Overflow (SO)? Why answer? What they get for answer? what questions have answers? Why SO not answer our questions? How many unanswered questions remains in SO? whose questions are not answered?

When software developers having programming issues used take for help from expertise. As software development expands its subject area day by day stack overflow is very useful website to gather all expert's knowledge to one place. Some of question are not answering, wanted to find out what is are the reasons?

With so many users (particularly on stack overflow) there is some competition to answer questions first because earlier answers tend to be upvoted first. It takes time to accumulate reputation points, so keep reading questions, asking good questions and answer only when you are sure you know the answer. It is not clear whether those questions are always get an correct answer or supported by reliable sources for end users.

Many people depend on this platform and trust their can get proper answers in detail to achieve knowledge to solve the main issue. Some questions have minacious errors that affect answer, experts experience and their understanding about problem gives the correct path to user. Or must wait for long time without answer. This time can be minimized if this research could give reason for unanswered question to user. users can fix question and search for support if get answer.

People can add feedback in the stack overflow reviews. They are used to analyze the answering system. These reviews are necessary for users to trust the site.

In this research proposed to analyze the reasons for unanswered questions in stack overflow platform.

1.4 Aim and Objectives

The research aim and objectives are listed here.

1.4.1 Aim

The intention of this project is to analyze the reasons for unanswered questions in Stack Overflow by exploding supervised learning.

1.4.2 Objectives

- To review literature and find out research gap to find analyze reason for unanswered questions in stack overflow.
- To identify appropriate data analytics techniques and algorithms to predict the reasons.
- To prepare the data set for Model creation by preprocessing into a suitable format.
- To identify suitable features to predict the reason characteristics.
- To analyze metrics to evaluate the reason of the models build.
- To visualize and interpret research findings concisely.
- To recommend future works based on the research findings.

1.5 Proposed solution

By providing some clear reasons to SO user community, why questions left without answer It will be a helpful Developers and SO experts to expand the service among many Programmers. Before submitting question, user can edit question by avoiding reasons analyzed in this research.

Following steps will be taken to fulfill main aim.

1. Data sets will be extracted from the stack overflow site.
2. Identify the suitable methods for prediction.
3. Data preprocessing
4. Implement the classification models
5. Classifying appropriate technique
6. Expand clustering algorithm
7. Classification and regression method use to Analyze result and discovering reason patterns for unanswered accordingly.

This study will first explore the approaches used by SO to predict the answer characteristics. It will identify the shortcomings of the approaches and the obstacles faced by the specialist programmers when answering them. Because of the developers' questions and answer environment in the world has become complex after the rapid changes in software development sector. The study will analyze the reasons one by one of the unanswered questions. A model algorithm will be developed to predict the reason characteristics in the future by evaluating user request based on several aspects, such as demographic, geographical, and behavioral data of the user.

1.6 Resource requirements

Although the research has some analyzing had to use common data mining tool WEKA and resources like research papers, journals, books and magazines to do literature review. this research used secondary dataset to do prediction. Primary data has, reviews with non-technical attributes which are available from the stack overflow site. Other than these recourses used PC with High CPU and RAM helps to run application smoothly.

1.7 Summary

This chapter provide the introduction for analyzing reasons for unanswered questions in stack overflow platform using supervised learning approach. Chapter 2 presents research work in stack overflow and question analyzing methods in Data mining. Chapter 3 explain technology adapted. Chapter 4 describes the approach of the research and Chapter 5 about Preprocess analyzing and design. Chapter 6 Model creation and regression. Chapter 7 discuss about evaluation and findings. Chapter 8 Conclusion and list of references and appendixes last.

Chapter 2

Literature Review of datamining techniques

2.1 Introduction

This chapter describes about the currently existing research studies and methods of reason finding systems. And research carried out about machine learning and supervised learning methods, and data analyzing technologies [16]. Especially focused on stack overflow related data analyzing. Also, different methods used data classifying and regression. And data mining techniques and concept [3] most popular in classification and prediction [11] in question answering. Then recognized unanswered is not concerns in much research works in data mining prediction projects. This is how it recognized the requirement of this research. Summary of challenges and definitions taken to several tables after reading those papers.

2.2 Summary of research paper references

In research major task is to investigate the responses survey data. Though we consider sample survey or full survey, reason analyzing of unanswered questions is one use of data analysis. Handling this type of data set should be done with a special care to interpret the results accurately. Table 2.1 provide the details of an important papers found for this research data analysis.

Paper details	Supervised learning	Data analyzing	Clustering	Classification
a) Total number of papers returned.	35	61	45	37
b) Number of duplicate papers returned.	2	2	0	0
c) Number of papers selected for preliminary analysis.	35	38	27	21
d) Number of papers removed after preliminary analysis (reading the abstract and/or introduction).	4	11	18	16
e) Number of papers to be selected for analysis.	27	36	27	21

Table 2.1 Summary of research paper references

2.3 Supervised learning filtering techniques

After refence research papers relevant to this reason analysis on unanswered questions in Stack overflow research, it used that knowledge to understand and improve this

research work standard. Table 2.2 shows the supervised learning approaches of data filtering technics used for model creation in this research.

Year	Author	Approaches	Description
2018	Vapnik	SVM Approach	Binary representation is used for best result.
2019	Soonthornphisaj	"Centroid-Based approach" [69]	"The data items are represented using a Vector space model in Naïve Bayesian and KNN" [69]
2016, 2017	Graham	"Bayesian filter" [71]	"Caught 99.5% of spam with 0.03% false positives" [71]
2015	Woitaszek	Simple support vector machine with personalized dictionary	"It is an add-in for Microsoft Outlook XP providing sorting and grouping capabilities using Outlook's interface to the typical desktop E-mail user." [70]
2015	Clark	LINGER	"It is a Neural network-based system which uses a multi-layer perceptron" [73]
2016	Matsumoto	"Support Vector Machines (SVMs) and Naive Bayesian Classifier (NBC)" [72]	"Used both term frequency (TF) and term frequency with inverse document frequency (TF-IDF) for features vector construction" [72]
2018	Tavana	Artificial Neural network and Bayesian [74]	"Two ANN structures, single layer perceptron and multi-layer perceptron are considered and the inputs to the networks are determined using binary and probabilistic models" [74]

Table 2.2 Supervised learning filtering techniques

1. Stack overflow answer giving time prediction "system supporting uncertainty/ fuzziness This were implemented in Prolog or FuzzyCLIPS, in a particular area of Software Engineering" [49] support system stack overflow. The rules of the Data visualization contain also fuzzy grades to represent uncertainty. Skills and knowledge within cluster analysis and Tag has used there in revied project.

2. Deep learning to find bugs in software," diagnosing using Data Mining (containing information about previously diagnosed software bugs) may be mined to automatically extract rules that will indicate diagnoses based on symptoms. In other words, these rules may be used to automatically classify (find the Issues) of new (undiagnosed) bugs based on their known symptoms. Or they may be used to find hidden links between medical conditions and factors that can influence them, etc. The rules are the result of the application of data mining/machine learning algorithms, as those based on: - Decision trees (ID3, ID4.5, etc) - Simple rules (1R, etc) - K-nearest neighbor method - Bayesian inference/classification - Association rules - Genetic algorithms for supervised learning

- Neural Networks A few distinct projects may be tackled using a dataset containing information about software bugs (or, by choice, other datasets regarding other applications), which is to be mined by one algorithm based on one of the methods mentioned above. The algorithm is to be coded in your preferred programming language. The program will access a dataset stored in a text file or a database, based on choice" [43]

3. "Extensive mining of software issue data using machine learning techniques Alternatively, the theme may focus on the application of selected computing technologies in SE bug research. More precisely the subject concerns an advanced use of a software collection of machine learning and statistical techniques called Weka that may be used with a database in order to learn from such data. The results of learning may be used for diagnosing new bugs (classification in terms of data mining), and/or detecting unknown possible interactions between components of a solution. Most work will involve the in-depth study and application of various Data Mining algorithms and of the techniques of evaluation of the quality of the derived knowledge (models), collecting and understanding the bugs data and the problems that can be tackled in this context, and making an advanced use of the software Weka. This project is an application of all the phases of a KDD process model specification of the problem and collection of data, data pre-processing, data mining, interpretation and evaluation, deployment" [44]

2.4 Classify model creation techniques

After refence research papers relevant to this reason analysis on unanswered questions in Stack overflow research, it used that knowledge to understand and improve this research work standard. Table 2.3 shows the model creation approaches of data classify technics used for reason analysis in this research.

Year	**Author**	**Approaches**	**Description**
2017	Zhao and Zhang	Rough Set Based Model	Used classical rough set theory.
2016	Dong	"Bayesian spam filter" [7]	Based on cross N-gram
2017	Ichimura	Classification method based on the results of SpamAssassin	Proposed Self-organizing map (SOM) for spam classifications and

			Automatically defined group (ADG) to extract correct judgment rules.
2017	Pang Xiu-Li	"Anti-spam filter approach is based on support vector machine (SVM)" [61]	Tri-gram language model and discount smoothing algorithm was applied to perform word segmentation.
2017	Yang and Elfayoumy	Feed forward back propagation Neural network and Bayesian classifiers	Effectiveness of both classifiers were evaluated using accuracy and sensitivity metrics.
2018	"Lobato and Lobato"	Binary classification approach	Based on an extension of Bayes Point Machines
2019	Sun	"Spam filtering algorithm based on locality pursuit projection (LPP) and least square version of SVM" [7]	"The mail message features are first extracted by the LPP algorithm, then the LS-SVM classifier is used to classify mails into spam and legitimate" [42].
2019	Yong	Intelligent Spam filtering system based on fuzzy clustering	This System can work without training in advance.
2020	Na Songkhla and Piromsopa	Statistical rule-based spam detection system	Rules can be shared, but it must be updated frequently to cope with spammers' tactics.
2020	Qiu	Online linear classifiers: Perceptron Winnow and Naïve Bayesian	They have a state-of-the-art performance for filtering spam.

Table 2.3 Classify model creation techniques

4. This reason analysis could take "model-based clustering" [1]. This Clustering problems analysis can test with discriminative approaches. It is related to organizing pairwise similar points.

5. This dataset strong features are view count, score, answered or not, Tags, Tag counts, technology, are available with the primary dataset.

6. There are some projects experiments on predict model unanswered questions, but not about the stack overflow. These projects on some question analysis data bases. And those work shows a maximum accuracy of 79%. Comparison with these projects models that results taken from Naïve bayes and KNN are models are with statistically significant level.

7. It is possible to characterizing the Stack Overflow questions and investigated why questions remain unanswered. Is not the reason low interest, that's what understood by reading these papers?

8. People who had conducted research on stack overflow "Rahman & Roy", "Saha et al", "Lal et al", "Correa & Sureka". They presented "an in-depth characterization of migrated questions on five famous SE sites revealed that the suggested model is efficient by accomplishing a precision of 73% in predicting the questions migration" [35]

However, no systematic study has ever been done on the tags or comments.

2.5 Summary of cluster analysis approaches

Bellow mention research p apers had referred relevant to this reason analysis on unanswered questions in Stack overflow research, it used that knowledge to understand and improve this research work standard. But these researches are now about this topic. And knowledge gain by reading those articles were helpful to find idea how to do this research wok. Table 2.4 shows the existing cluster analysis approaches used for model creation in this research.

Author	**Algorithms**	**Corpus or Datasets**	**Accuracy / Performance**
Rathi et al.	Naive Bayes, Bayes Net, SVM, and Random Forest	Custom Collection	99.72% Accuracy Rate
Mohammed et al.	Word Filterization by Tokenization, Appling	Nielson Email-1431	Reported Satisfactory Accuracy for Proposed Method
Singh et al.	Naive Bayes, k-Nearest Neighbour, SVM, Artificial Neural Network.	Custom Collection	Reported Improvement of precision rate at least 2%
Abdulhamid et al.	Various Machine Learning Algorithms	UCI Machine Learning Repository	94.2% Accuracy Achieved
Sah et al.	Naïve Bayes, SVM	Custom Collection	Reported good Accuracy overall

Rusland et al.	Modified Naive Bayes with selective features	Spam Base, Spam Data	Spam Base get 88%Precision Rate and Spam Data get 83%
ksel et al.	Microsoft Azure platform defined decision tree and SVM	Custom Collection	SVM Accuracy 97.6% Decision Tree Accuracy 82.6%
Choudhary et al.	Feature Engineered Naive Bayes	The SMS Spam Corpus v.0.1	96.5% True Positive Rate Accuracy

Table 2.4 cluster analysis approaches

2.6 Summary

Many texts mining approach use keywords analysis techniques. And this method has used with pattern recognition or clustering methodology.

There is a considerable in the stack overflow scraping the data in a text format, so In this research use text mining approaches such as prediction model.

In this case when using Phrases based approach it carries more semantics like information, and it will create the large numbers of redundant noisy. Because some users type the question, they do not use the proper phrases. In this point can be applied the concept-based approach because this approach can improve prediction in overall analysis.

As this project take primary data after clean and systematically written user questions details. It can take as a best way to do a classification prediction model. also may be users ask several questions from different areas in one question before they get answer for one, then it is left as a duplicate question.

Here in this project discover several characteristics by analyzed more than 24 thousand questions.

Based on that assumption, these reasons find the keywords by Classification algorithms prediction methods. In this study needed to analysis the user questions of must be categories and can be divided into several subcategories there after applying the proper efficient algorithm to analysis the dataset via k-means methods. It is producing the reasonable analyze for prediction of reasons in a stake overflow.

Chapter 3

Technology Adapted

3.1 Introduction

Last chapter shows different findings of supervised learning and data prediction research, but when looking at works carried out in the stack overflow Unanswered questions it is very limited. And this was a good opportunity to carried out this research work. Also identified classification supervised learning as the technology to address the reason finding for unanswered questions. This chapter gives the details about selected technologies used in the research.

3.2 What is supervised learning

Start this project by take used dataset for stack overflow questions as a primary data. In this project use secondary data. Classify used to make new model for finding reasons and regression use to evaluate the model. Predictive information is finding value of attribute using value of an attributes in information,

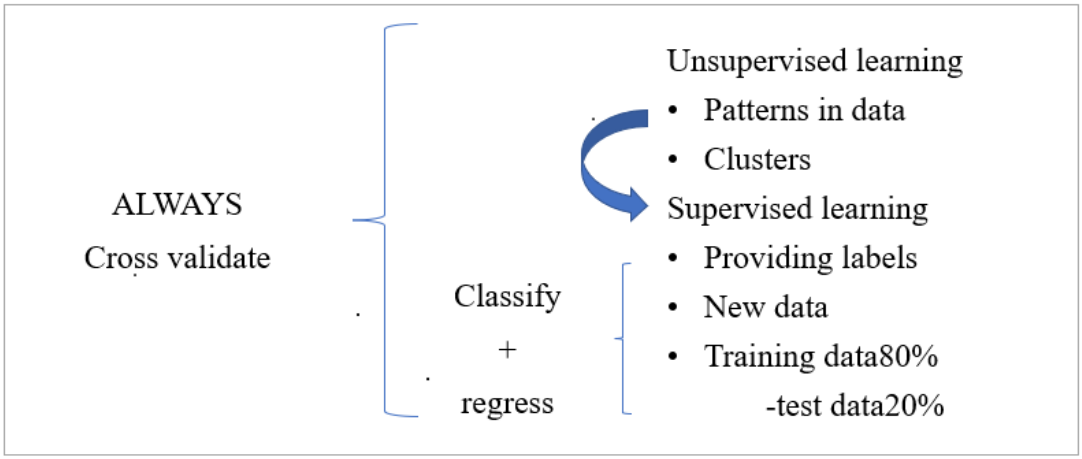


Figure 3.1 Supervised learning

Supervised learning process make data point categorizing. Data mining some important areas, Such as “association rule mining” [8], “classification” [8], “clustering”[8], “regression”[8] and “sequential patterns”[8]. Among them this research has considered Classifying data model to prediction and regression to evaluate dataset. And tested with clustering algorithms. Classification method can use to forecast or make decisions based on secondary dataset.

Create modal used to make local relationships between data points or samples. which can be applied to globally more efficiently and most appropriately.

Data pre-processing, selection before model creation and evaluation need after the process. Model can be used to make decisions based on accuracy, completeness, consistency, timeliness, believability, interpretability and accessibility after result analysis according to the generated Naïve Bayes algorithm and KNN train dataset.

3.3 The training algorithm.

The idea behind the training is to Use the fact that already know. train an algorithm to recognize what the key structures of the data are based upon expert loop pinion. And then can go test new data and see how well it work. want one idea to find when is clustering is happening here in the unsupervised learning or looking for patterns in the data. In contrast, here in the supervised learning, not necessarily looking for clusters it is providing the labels with the cluster areas. Here is do classification and regress. The idea behind regression is find curve or best least square fit type curve to the data doesn't have to be linear.

3.4 Weka data mining tool

Weka used for the preprocessing process of analyzing data and model creation in classify and clustering algorithms and evaluation by regression, performance testing and prediction which is retrieved at the beginning of the process. Can load dataset collected by different data sources, transform by filtering and do analysis.

Here in this research work analyzed secondary data to find suitable reason for unanswered question of stack overflow site, data collecting is not the main task of the project. Preprocessed data use to make train and test datasets which helped to make classify model and regressing evaluation.

3.5 Summary

This chapter presented supervised learning as technology proposed to analyze reasons for unanswered questions in stack overflow. According to the classification model. And it uses regression to evaluate results. And this preprocessing, model building and evaluation use WEKA API. Also, it shows how classification data mining algorithm use supervised learning to analyze reason for unanswered questions. Next chapter shows preprocess, model creation and evaluation process.

Chapter 4

Approach for analyzing reason in stack overflow unanswered questions

4.1 Introduction

Chapter three presented the tools and technology used to analyze these data and find reasons. In this chapter described the problem inadequate in evaluation reason for unanswered question in stack overflow. Find more accurate Classify model is the key module of this analysis. Will show here hypothesis, input, output, evaluation process and reasons recognized.

4.2 Hypothesis

Those unanswered questions waiting on Stack overflow can be solve by giving a reason for not answering to user. By using classifier analysis. This model will use various classify technologies and pickup highest accuracy model.

4.3 Input

Used the primary data set used for Stack overflow question analyzed dataset. As here in this project it used for classification building take secondary data with 10 attributes. And next must preprocessed and clean dataset and normalized dataset. And find most suitable attribute as class label. And make label as nominal. Next level selects all relevant attribute and make nominal. And secondly divide dataset in to Training and Testing parts as 80% to 20%.

4.4 Supervised learning

The design of the unanswered question analyzer working process has four stages. An overview of the various stages with respect predictive processes shown in Figure 4.1. Overview of the various stages with predictive analysis process.

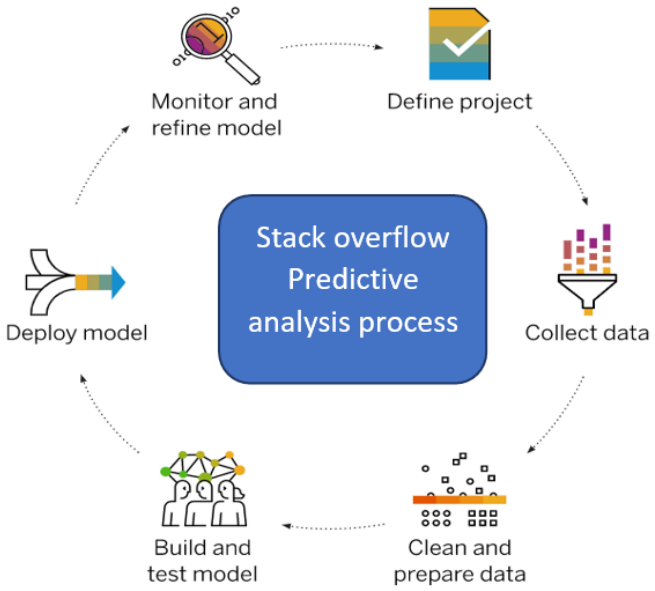
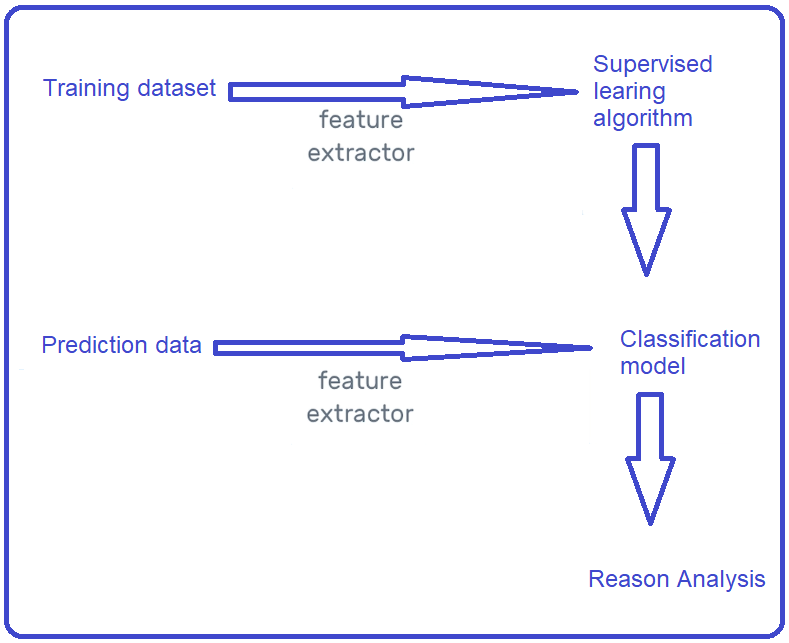


Figure 4.1. Framework of the processing data

This prediction will use attributes of secondary dataset and build test model by using the preprocessed data. To find most suitable classification model has to test dataset with all possible models. This has to consider normal PC with general configuration. Because this research focus on giving reason to stack overflow web users.

After Deploy the classification model it has to test with test data set with regression model. Has to use confusion matrix with suitable classification algorithm. As this project it focus on analysis reason, had to understand what data is going to use and what don't want for this work. Base on the decision removed attributes with date, index no, user id, user gradings, answer counts. Features have to select carefully after study about classification algorithm in this project gone studied about Naïve-Bayes, KNN, Decision Tree algorithm and found most useful features can take from the dataset. At this level it realized how important to have a good dataset without missing values and unwanted symbols. When put garbage value will get garbage outputs. Such situation results won't give a real situation to a un-answering or will fell several clusters irrelevant to main reasons giving by other prediction values. As here it use secondary dataset had to check it more carefully and data in attributes and content. Main feature of this dataset is all data we use is with numerical features. It shows in following chart How training data set combined with prediction reasons.



4.5 Summary

This chapter precented the supervised learning approach for secondary data analysis using classify and regression method to find reason for unanswered question in stack overflow. In preprocessing make dataset to use in classification algorithm model, in this work. According to maximize the speed and accuracy, clean and formatted is the main target of preprocessing. And have to focus on the output want to find from the dataset and do necessary steps for improve results. Develop process model helps to select algorithm use in prediction model with secondary results and find most relevant reason for not get an answer.

Chapter 5

Research Design for analyzing reason

5.1 Introduction

Chapter four presented the approach to analyses reason for unanswered questions in stack overflow platform. In this chapter elaborate the approach and focus to evaluation criteria to find more accurate Classification model is the key module of this analysis.

5.2 Research design

Those unanswered questions waiting on Stack overflow can be solve by giving a reason to user. Make a model in classify after preprocess by training dataset. Created classify models by Naïve Bayes, Decision Tree and KNN algorithms and pickup highest accuracy model. Evaluated each possible model separately with 20% test dataset.

Select the secondary dataset in csv format with 10 attributes. And preprocessed by remove unwanted attribute such as ID and date. then selected the label attribute from the dataset. Then changed class attribute to nominal. After this select all required attributes to nominal. And normalized relevant attribute from dataset. There after saved the new dataset with “arff” file format. Now secondary dataset is ready to make the classify model.

And secondly divide dataset in to Training 80% and Testing 20%. parts and save in “arff” format. After study about numerical data, it has decided what dataset tags it consider then features have to extract from original data. Next level has to find attribute selection, To select attribute missing value ratio, low variance filter, Heigh correlation filter, random forest, backward/ forward feature elimination has used to do feature extraction testing . Next must see what method to use for reduction, use the factor analysis, principle component analysis, t-SNE (distribute Stochastic Neighbor Embedding) , UMAP (uniform manifold approximation and project)- is a good for nonlinear and visual cluster groups. Then make suitable model for prediction. Here it has used classifier model by using Naïve bayes algorithm.

Figure 5.1 shows the theoretical framework of the design creation process from secondary dataset to the evaluation.

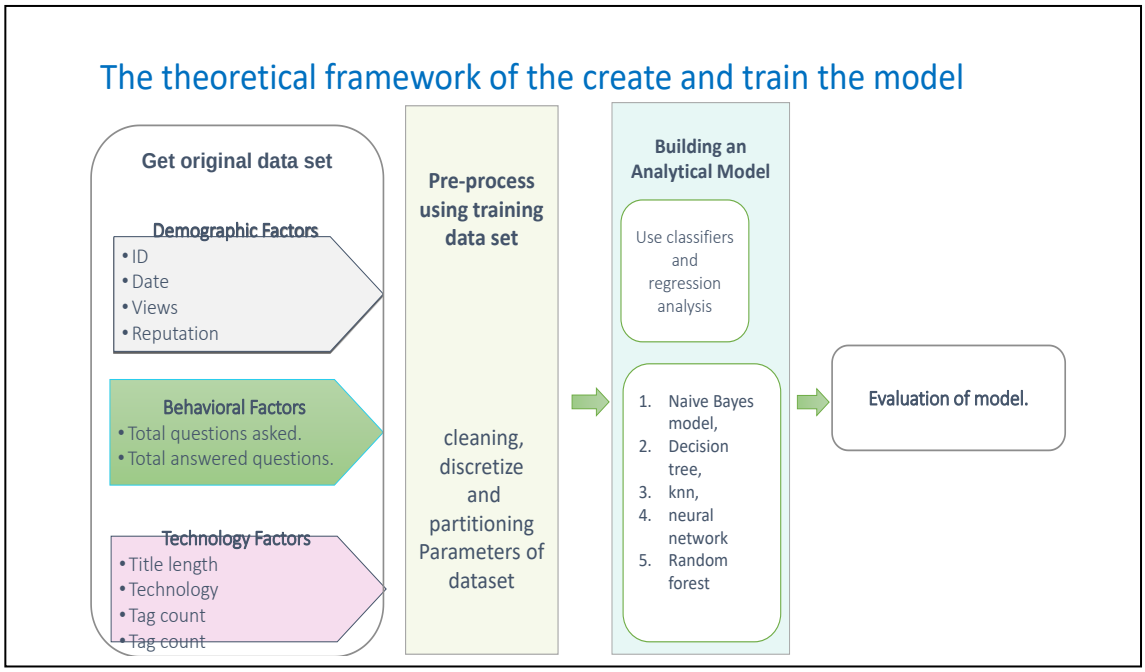


Figure 5.1 Framework of the preprocessing data

5.3 Top level Supervised learning design

The design of the unanswered question reason analyzer working process has four stages. Preprocess secondary data and select only suitable attributes for classification algorithm analyzing, find class attribute, normalize attributes and check with different classification model and find the accuracy and confusion matrix results. Figure 5.2 shows An overview of supervised learning processes

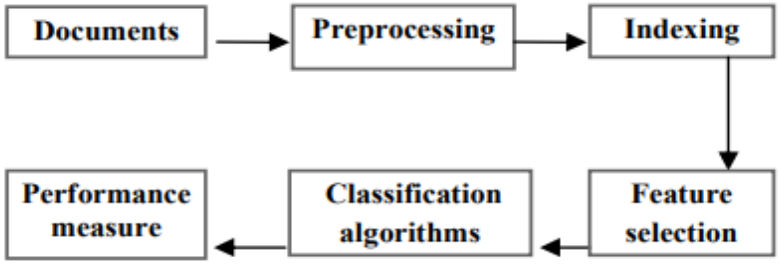


Figure 5.2 Supervised learning

According to the results of these classification models which captures best performance it can start data collection and results analysis process for classifiers.

5.4 Data collection and Preprocessing data

This analysis uses secondary data foe classification model creation. Primary data had collected from stack overflow knowledge platform by using R and Python programming datamining techniques. To use secondary data unwanted columns attributes have removed. To fix missing value it has used median and mean values. Then by using Weka data mining tool data has preprocessed and model has created. To find best algorithm for reason analysis it has run several classification algorithms for preprocessed data set separately. And have to finish one algorithm test and save results before go to next model. Then need to make regression model separately to confirm result by using the test data set.

5.5 Prediction modal

This is supervised learning process. This Use well completed labels with training data is used to make model. 5.3 Figure shows Prediction modal , those models use this data to make predictions and give the output. Labeled data some data is tagged with right output.

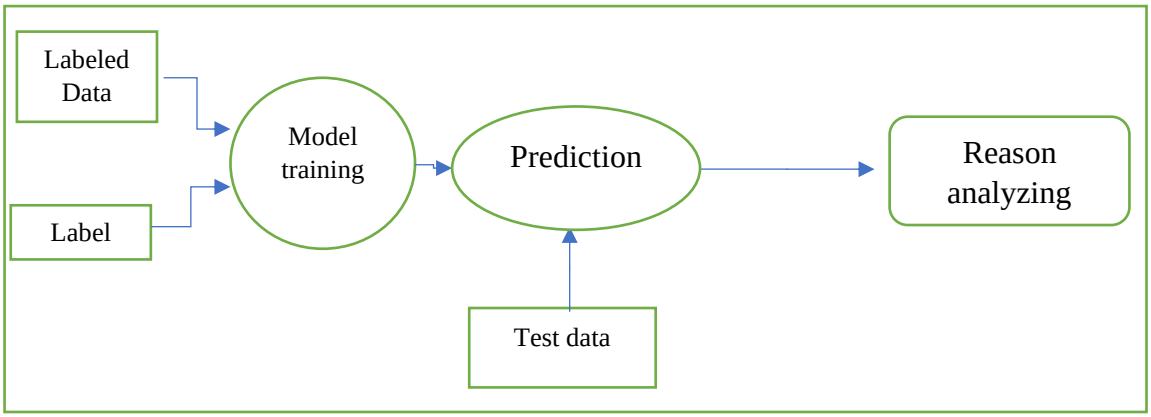


Figure 5.0.3 Prediction modal

Supervised learning takes training data (called data samples) and its associated outputs (also call labels or responses) use training data to process algorithm. Here it use supervised learning to find the association between input training data and their labels.

“KNN algorithm, Decision tree algorithm, Logic regression algorithm or Random forest algorithm” [25] use to do classification based task or Regression based task prediction.

5.6 Detail design of the research work

Make a classify model to analyzing stack overflow question details to find reasons for unanswered questions is identified as the research question. Finding accurate algorithm to do this analysis using secondary dataset is first important part and use testing dataset to evaluate results is other use of supervised learning taken in regression. “Linear regression” [26] based dependent and independent variables and error value. Which has calculated as zero.

5.7 Summary

This chapter precented design of the supervised learning approach for secondary data analysis using classify model and evaluation using regression method to analyzing reason for unanswered question in stack overflow.

Chapter 6

Implementation

6.1 Introduction

In previous chapter, it described the main method of analyze reason for unanswered question in stack overflow and the model creation. In this section will be discussed how to create the model in detail for all preprocessing and model creation. Nevertheless, this presents created process, classify model creation and evaluation and predictions with input dataset and outputs.

6.2 Overall implementation process

Here in this part, it describes all the steps taken in the research work. Initial data collection to the interpretation of reason and what else can propose to develop in next level.

Figure 6.1 shows the main steps for overall Implementation process

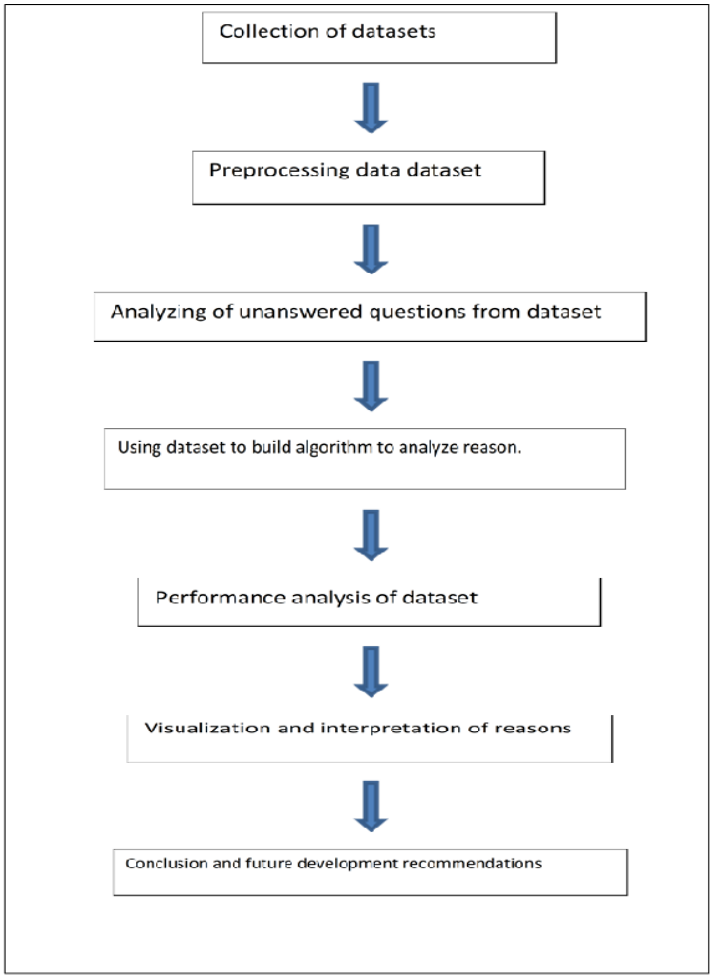


Figure 6.1 Implementation process

This reason analysis is an interesting research work. While model creation by using algorithm have to consider mainly on mathematical model of algorithm to understand dataset behavior. It is a big challenge in implementation part. After understand the mathematical model of algorithm we can understand how results depends on different algorithms. And run train dataset with good confusion matrix and good accuracy help to find suitable algorithm for the classify model creation. To find best algorithm with more accurate results had to test for all possible algorithms. Also, must consider for a WEKA application installed on normal PC. As this result aims to give fast feedback to Users to correct unanswered question and find answer from programmers in next step.

After satisfaction implementation Naïve Bayes, KNN and Decision Tree has given more trustful results with good accuracy among these datasets. Bellow figure 6.1 shows the KNN classify models analyzed results.

6.3 Preprocessing data

While preprocess secondary dataset, have removed some attributes such as Date, Index no. And partitioned as training and testing dataset in ratio of 80:20. Next found the suitable attribute for a for class attribute and converted it to nominal data type. In next level changed other attributes as numerical to nominal. And Normalized attribute for model creation.

6.4 Classifier model creation

Use classification model to predict accurate results by using target class. Prediction analysis used because of huge amount of data. There are two steps followed model construction and model usage. In main model it represented rules of model creation of Naïve-Bayes, KNN(K-Nearest Neighborhood) and Decision Tree

The method is tested with several other classification algorithms and selected for using in reason finding problem of data mining according to dataset possibility and hardware availability. According to reason finding requirements here use term based, selected after considering with Phrase based method, concept based and pattern based.

Got a requirement of analyzing a reason for left questions. Found already collected data set. It was a reason to select the supervised learning method. And mainly got possibility to use Support vector classification, KNN, Naïve bayes, support vector machine and

Neural Network to use in WEKA data mining tool. And finally found accurate results for Naïve Bayes algorithm.

“Data classification is the process of sorting and categorizing data into various types, forms or any other distinct class. Data classification enables the separation and classification of data according to data set requirements for various business or personal objectives. It is mainly a data management process” [44]

Figure 6.2 shows the classification model which used to developed different classifiers is from the begin, how and why here using classification and how to get data from secondary data set.

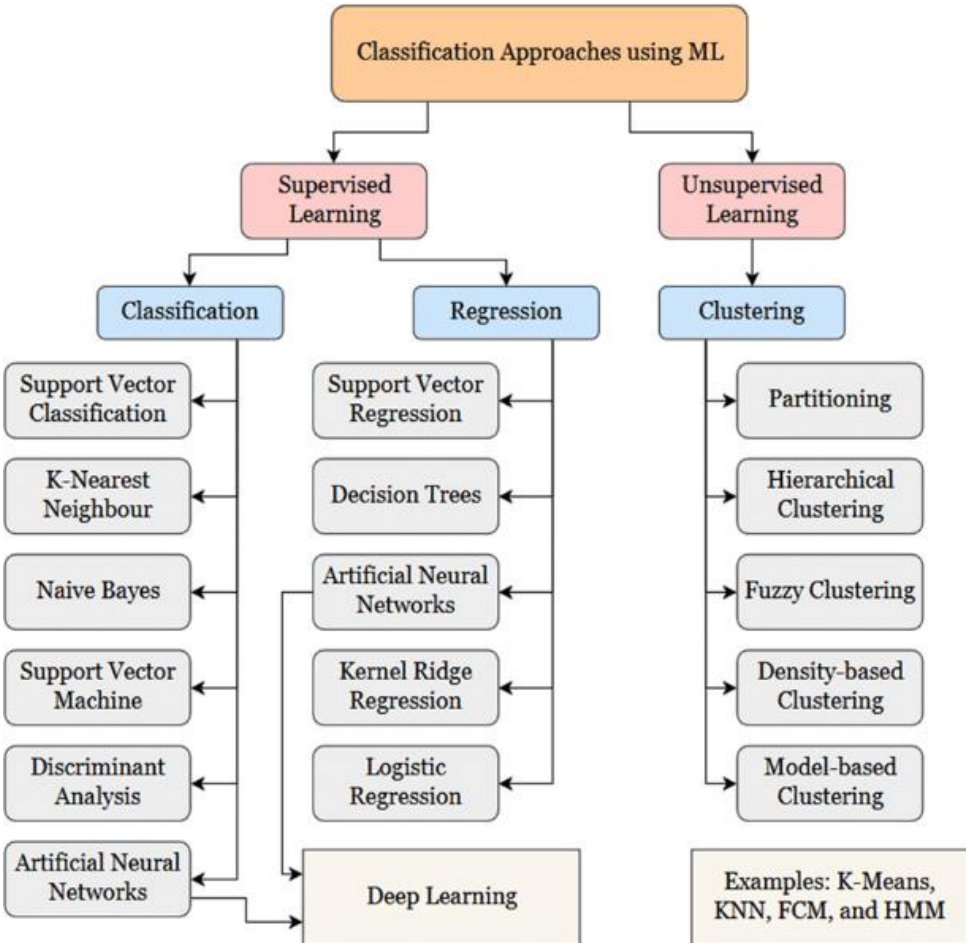


Figure 6.2 Supervised learning approach

After select the WEKA tool for data preprocessing and model creation it has uploaded training dataset to do analysis Figure 6.3 shows how classify model created for preprocessed data.

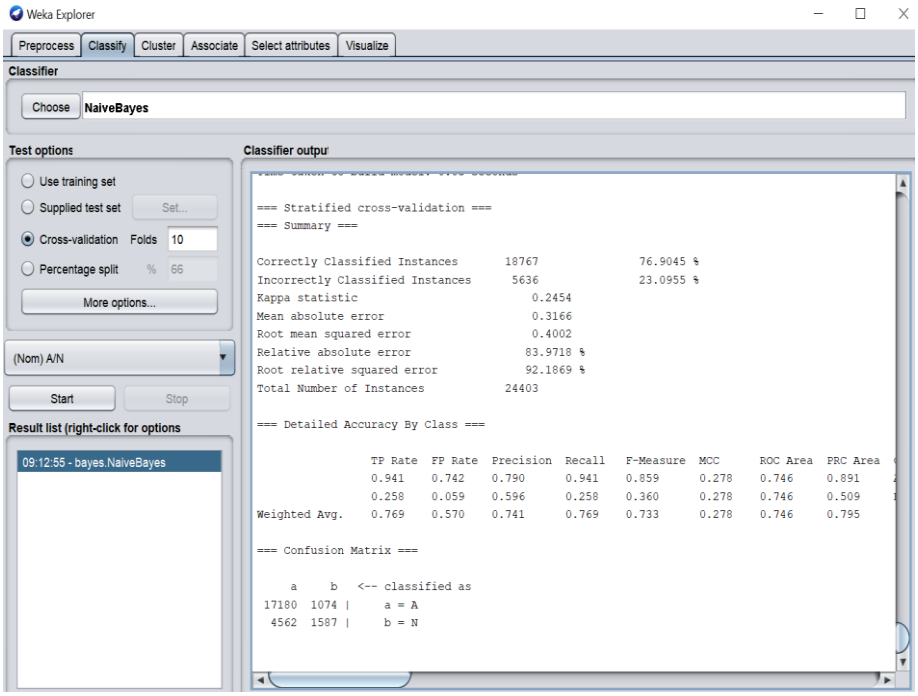


Figure 6.3 WEKA tool model construction

In classification model with KNN test results and basic configuration shows in figure 6.4 .

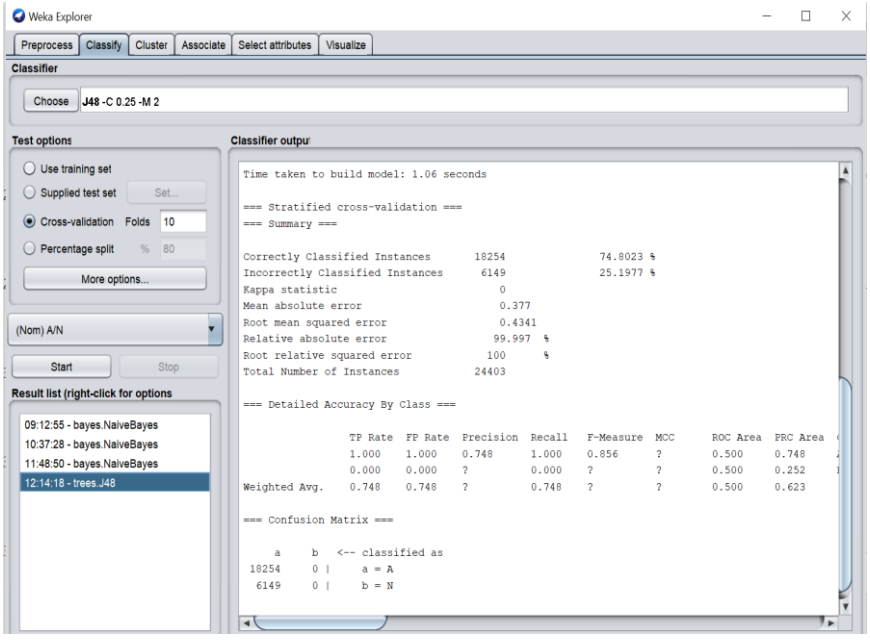
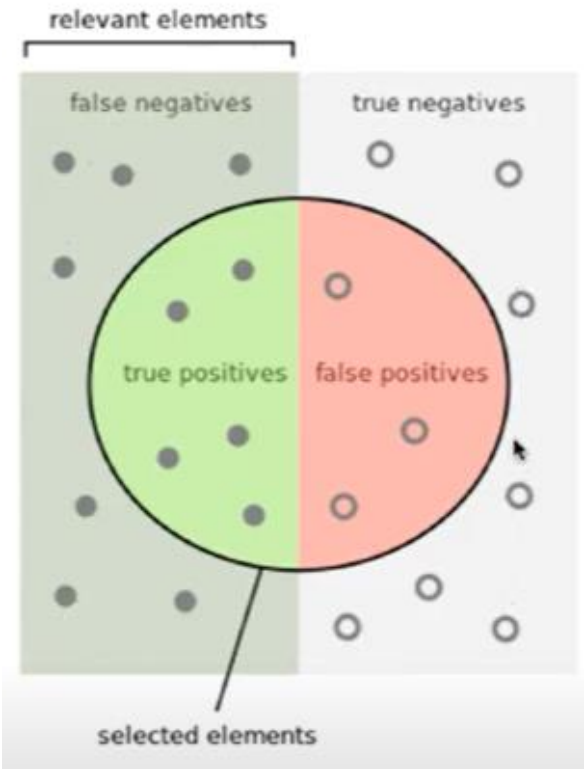


Figure 6.4 K-NN classify models results



By making classify model in Weka fond some interesting class with confusion matrix results True Positive(TP), False (FP), Ture negative(TN), False Negative(FN) areas. F-measure combine position, recall use for harmonic and precision ROC receiver operator characteristic. After compare results with Tags in original dataset found there having equal qualities between these groups. And evaluated by regression model to confirm results aby testing data set and all data.

Here it took the original and divided to training and testing dataset by80% and 20%, next created the regression model to evaluate supervised learning data set. And run on Support vector regression, Decision tree Artificial Neural network, Kernel Ridge regression after that evaluate by using test data found Linear regression was more fast and accurate while evaluating the test dataset.

Data Classification of Naïve Bayes allows to make predictions using the extracted knowledge from existing organized data efficiently.

6.5 Supervised Learning

Secondary data use in this research. Primary dataset had collected by UoM undergraduates. Primary data has taken by unsupervised learning method. After pre-processed and made clusters on this it had used for this research.

By using supervised learning method, found new data with labels for training classify and regress model. classification datapoints in quantitative data set. Training set use to generate model and test set use to find accuracy.

Found qualitative data there and can use several methods to analyze it. Firstly, divide to groups using WEKA tool classification.

6.6 Process of classification

This process depends on predictive analysis. It shows to what likely to happen next with new data. And can find what to do by using prescriptive analysis in next level.

1. predicts categorical class labels

2. classifies data models

Each time it is recommend doing cross validation to test the performance.

1. **Model construction stage**

It creates by using samples in dataset. Found most suitable attribute ‘Answer or not ‘ as the main class and named or labeled it as a class attribute. And change data type of all attributes to nominal to numeric at preprocess stage. Results given by Naïve bayes was more successful after looking at accuracy it confirms as well as in mathematical model. ,

2. Model usage stage

Naïve Bayes algorithm gave the most reliable details in all performance and taken which has highest accuracy and best confusion matrix as the suitable classification algorithm model for prediction. Take train dataset and use Naïve bayes algorithm and found 76% accuracy after using at testing datasets also gave same results with all models.

Following algorithms were tested during model creation for reason analyzing classify. “Decision Tree, Naïve Bayes, K Nearest Neighbor, Neural Networks, Support Vector Machines, Ensemble Methods” [56]

Selecting suitable model for reason analyzer by using data accuracy, confusion matrix and cross validation summary has used from classify models which developed by using stack overflow dataset. It is the hardest process in this research. it has algorithm with following category.

1. Complete algorithm.
2. Fastest algorithm in normal PC.
3. Tags separated to similar groups.
4. Find similarities of prediction model.

Characteristics of prediction clusters it has evaluated main 4 reasons.

Moreover, the analysis of prediction algorithm help to understand and can suggest reasons to gather as separate classifiers.

6.7 Applied to algorithm

Supervise learning in machine learning has the computer learn and analyze after understanding the developed model base on the algorithm. At this situation have applied the data to the Naïve bayes mathematical function where it shows p(class evidence)=p(evidance class)*p(class)/p(evidence) . As here selected a single most likely class, and that based on the main requirement of this project which need to find reason for unanswered questions. That shows only main one class is answered or not. Then it only need to find maximum p(e| class)*p(class). And have to estimate the p(e class) where this data evidence e has set of observation as True positive, False positive, False negative and false positive points. Then E=(e1,e2,e3,…en).

$$P(e|C) \geq \prod_{i=1}^{n} P(e_i|C) \qquad P(C|e) = \frac{P(C)\prod_{i=1}^{n} P(e_i|C)}{P(e)}$$

----------[25]

Regression model “evaluated the predicted values given by algorithm. Also used to predicted **numerical values by** Linear regression as this **supervised learning** algorithm in a linear model.

6.8 Data analyzing

Will do the analysis according to the prediction approach as

Original dataset csv file taken into preprocess and store as arff for model creation

- First create Naïvebayse models with 60% second 80% third decision tree J48 forth-KNN [iBk]Models
- Can use it for prediction.
- Confusion matrix use for evaluation purpose.

- Do the same process to testing dataset as well

Figure 6.5 shows classify model result for naïve bayes algorithm. And for full dataset also gives accuracy 76% Appendix A shows classify output with comparison matrix.

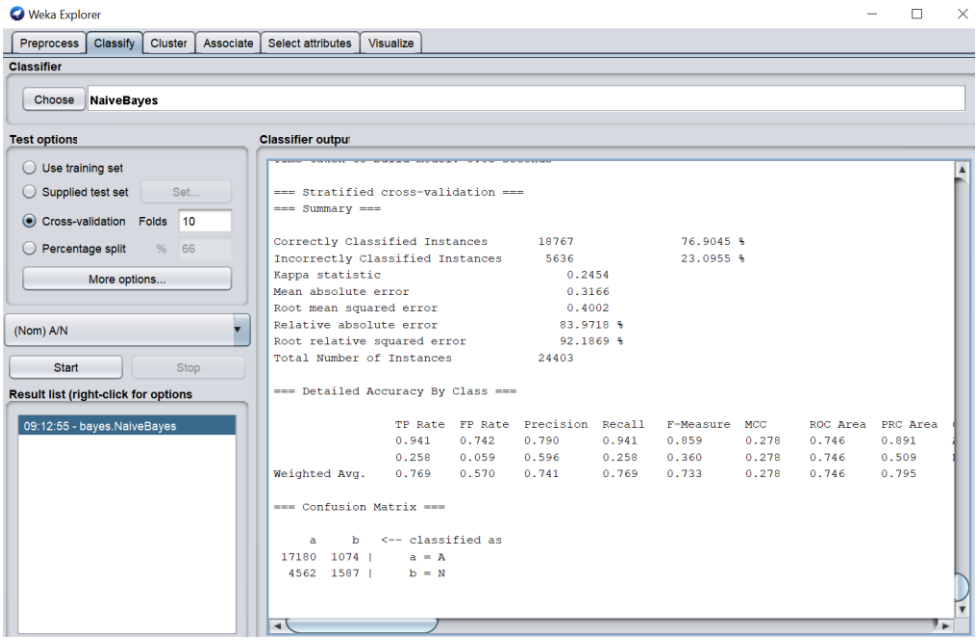


Figure 6.5 classify model naïve bayes

6.9 Summary

This chapter provide full design of implementation of each modal creation with training dataset. It shows data process and model creation as in main design. Next chapter gives the evaluation details of tested modules.

Chapter 7

Evaluation

7.1 Introduction

Previous chapter shows how model creation works and usage works in modules which has mention in design level. In this chapter going to talk about why prediction is correct for supervised learning classification model evaluation and regression models with testing results. And give predictions of results.

7.2 Cross-validation Summary for selected algorithm.

In following table 7.1 shows the results considered in the taking model in

Correctly Classified Instances	18767 ----76.9045 %
Incorrectly Classified Instances	5636------23.0955 %
Kappa statistic	0.2454
Mean absolute error	0.3166
Root mean squared error	0.4002
Relative absolute error	83.9718 %
Root relative squared error	92.1869 %
Total Number of Instances	24403

Table 7.1 Cross-validation Summary

7.3 Detail accuracy by class

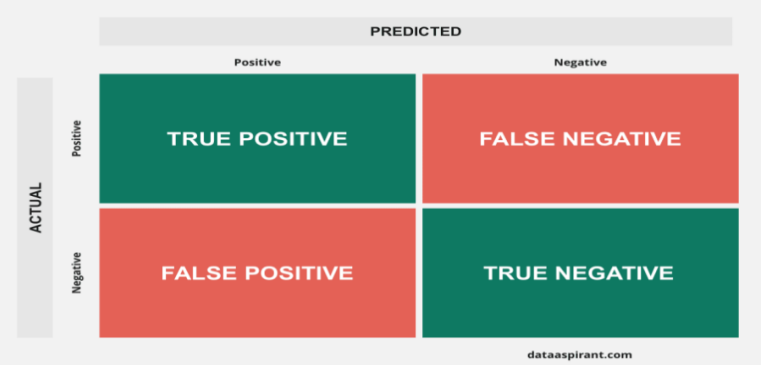
Classify model Naïve Bayes results for accuracy 76% details gives in table 7.2

TP Rate	FP Rate	Precision	Recall		F-Measure	MCC	ROC	Area	PRC Area
0.941	0.742	0.790	0.941		0.859	0.278	0.746	0.891	A
0.258	0.059	0.596	0.258		0.360	0.278	0.746	0.509	N
Weighted Avg.	0.769	0.570	0.741		0.769	0.733	0.278	0.746	0.795

Table 7.2 Accuracy by class

7.4 Confusion matrix

This confusion matrix must give a value equal to answered for Ture positive vales and true negative values. Should not have zero value or low values. Must have a good distribution on prediction value.



=== Confusion Matrix ===

a b <-- classified as

17180 1074 | a = A

456287 | b = N

---

a. Evaluation of classification techniques

In this research evaluated dataset using different algorithms namely Naïve Bayes, Decision tree and KNN model by using WEKA tool. Expectation show is the most input for expository motor. This assessing a classifier quality utilized perplexity lattice, which can be assessed different estimations. such as cross-validation, accurately classified occasions, erroneously classified occasion, cruel supreme blunder, root cruel square blunder, point by point precision by course and Perplexity lattice [52]. (Appendix A,B.C)

Using different algorithms created several classify models with training dataset and evaluated with test data. Comparison of classify model technique Naïve Bayes, Decision Tree, KNN. Table 7.3 shows the comparison on supervise learning classify models which taken to evaluate suitable technique.

Technique	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC
Naïve Bayes	0.946	0.082	0.982	0.982	0.981	0.978
Decision Tree	0.941	0.092	0.975	0.975	0.974	0.998
KNN	0.947	0.746	0.691	0.772	0.704	0.897

Table 7.3 Classify results between completed models

b. Evaluation of prediction

Using Linear regression mathematical model "Y=a=bX1+cX2+dX3+ ϵ" [63] where "Y-dependent variable, X-independent(explanatory) variables, a intercept, b,c,d-slopes., **ϵ**-Residual(error)"[63]. This regression follows the linear analysis. Other mandatory condition is non-collinearity, which shows minimum correlation with each other. These groups reasons identified as **"**duplicate of another question", "off topic", "asking debugging", "typographical error". Table 7.4 shows the Predicted reason, Formula and Description which has applied in evaluation.

Predicted Reasons	Formula	TAG Description
Duplicate of another question	"TP/(TP + FP)"[56]	The rate of positive expectations those are correct.
Off topic	"TP / (TP + FN)" [56]	"The rate of positive labeled occurrences that were anticipated as positive" [63]
Asking debugging	"TN / (TN + FP)" [56]	The rate of "negative labeled occurrences that were anticipated as negative" [63]
Typographical error	"(TP + TN) / (TP + TN + FP + FN)" [56]	The rate of forecasts those are redress

Table 7.4 Evaluation details

And using different results derived TP rate equal to off topic, FP rate gives debugging, f-measure give duplicate and accuracy value near 0 shows typographical error.

All Naive Bayes, Decision tree and KNN models were tested and found accuracy level high in KNN and Naïve Bayes models and tested again with test results and used for regression. All results are attached in Appendix (D, E, F,G,H,I,J)

c. Reasons analysis by cluster

According to the divided clusters analysis done by using in training and testing dataset. there was main 4 categories with countable percentage out of 6 clusters Taken all attributes as numerical and 2 independents nominal (Appendix K,L)

.	Clustered	Instances
Asking debugging	0	860 (14%)
Duplicate of another question	3	2597 (42%)
Off topic	4	1428 (23%)
Typographical error	5	10781(18%)

Table 7.5 Reasons analysis by cluster

Reason	Probability
Duplicate of another question	42%
Off topic	26%
Asking debugging	14%
Typographical error	18%

Table 7.6 Evaluated reasons analysis

Reasons for unanswered questions as results of tags details in dataset

```
kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 93036.0

Initial starting points (random):

Cluster 0: <c#><.net><refactoring><constructor>,1221,10,0,2,4809,4,A
Cluster 1: <capture><avi>,227,2,0,1,5396,2,A
Cluster 2: <sql><sqlite>,6974,3,0,1,151061,2,A
Cluster 3: <c#><winapi><windows-7><autohotkey><keyboard-hook>,6451,3,1,1,1130,5,A
Cluster 4: <bash><cygwin>,1365,3,0,1,49543,2,A

Missing values globally replaced with mean/mode


Time taken to build model (full training data) : 0.41 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      8791 ( 45%)
1      5510 ( 28%)
2      2067 ( 11%)
3      1548 (  8%)
4      1606 (  8%)
```

- The first option is it a duplicate of another question which has already been asked. It asks us for the link outside of rule, so if it's already been asked. Then it will be marked as duplicate, and not giving answer.
- Then, apart from that, another reason which we have is off topic. When can it be off topic? Hardware or software kind of question which is not related about programming.
- Another option is asking debugging help. For example, some users put 200 lines of code and ask people, why is it not working? in those cases the question would be not answered.
- Another reason to unanswered is when it cannot be reproduced or a simple typographical error. this happens a lot. People post a question on stack overflow and then they realize they have just missed comma or semi colon, something like that. Very simple issue in that case it is unanswered as a typographical error or something which that cannot be reproduced.

d. Summary

This chapter evaluated the methodologies and the discussed results. Classifier comes as category and regression comes as number. Next chapter discuss limitations and future improvements to reason analyzing for stack overflow.

Chapter 8

Discussion

8.1 Introduction

The various approaches have been tested in this research study for the classifying model creation. Have tried to find reason for not getting answer to stack overflow questions posting to site. can use the K-nearest neighbor algorithm is best because of dataset will be used the several type of category and when we can use these methods easily organized the user tags and numerical attributes of the secondary dataset. The outcome of this research will be a produce the several models of classifications based on the user TAGS. “The dependable methods which are specified the over investigate technique will be deliver the expected result and based on that result it predicted the reason. But at some point, calculated result may not create proficient forecast” [56]

When get new dataset can easily preprocess and create prediction for new values and reproduce the result again and send feedback to user, immediately by referencing User ID. Correspondingly, can acknowledge new users about these reasons why people not getting answers from Stack Overflow before submitting the question.

8.2 Limitations

The various approaches have been tested in this research study for the classifying model can use the K-nearest neighbor algorithm is giving best results because of dataset will be used the several type of categories. Had to test several times to find best model of classify because of large dataset. Also had to face application termination because this research conducted on normal Laptop computer.

8.3 Future Developments

Need to test on prediction algorithms to find most reliable prediction engine for online datamining if this research can update user withing question uploading time. Looking to do improvement for correlated features. Where it is possible to remove the highly correlated ones using Naive Bayes is based on the assumption. By improving feature extraction of prediction model try to get advantage to improve classification model performances.

8.4 Summary

Here in this chapter conclude research work after describing the reasons given with supervised learning to analyze reason to unanswered questions in stack overflow platform. It will help to the improve user questions and probability of answering to question could heigh.

This research studies about reasons for Stack Overflow unanswered question. by using supervised learning technology. Here first it categorizes by to classification method and did predictions by regression model using secondary data.

- First preprocessed dataset
- After the preprocessed and normalization, the dataset has divided to trained and tested data set by using, instance randomize method. And saved dataset.
- Taken labeled training dataset. Secondly created classification model and Naive Bayes algorithm KNN and Decision tree classification models and cross validated results. Using tested dataset, it evaluates the results.
- Found accuracy of 83.14% when use Naïve Bayes and KNN model.
- The outcome of this entire research is a find reasons by evaluating standard classification algorithm predictions results. Also, by Analyzing k-mean clustering

To giving a reason used Confusion matrix to evaluate matrixes. algorithm methods which are said the over evaluated the reasons for unanswered questions. Based on the result it can make good question which will be answered by system level.

Reference

[1] Sonali Vijay Gaikwad, Archana Chaugule, Pramod Patil, *Text Mining Methods and Techniques*

[2] Min-Yuh Day, Yue-Da Lin (2017), *Deep Learning for Sentiment Analysis on Google Play Consumer Review*, IEEE International Conference on Information Reuse and Integration in pp1,2

[3] Phong Minh Vu, Tam The Nguyen, Hung Viet Pham, Tung Thanh Nguyen, *Mining User Opinions in Mobile App Reviews: A Keyword-based Approach,* in ppr 1, 3, 4.

[4] A. Shabtai, Y. Fledel, and Y. Elovici (2010), *Automated static code analysis for classifying Android applications using machine learning,* in Proc. Int. Conf. Comput. Intell. Secur., pp. 329-333.

[5] A.Arleo, W. Didimo, G. Liotta et al. (2021), *Large graph visualizations using a distributed computing platform,* Information Sciences, vol. 381, pp. 124-141.

[6] X. Chen, W. Liu, H. Qiu et al. (2019), *APSCAN: A parameter free algorithm for clustering*, Pattern Recognition Letters, vol. 32, no. 7, pp. 973-986.

[7] H. Guo, Y. Yu, and M. Skitmore (2020), *Visualization technology-based construction safety management: A review*, Automation in Construction, vol. 73, pp. 135-144.

[8] T. İnkaya, (2019), *A density and connectivity-based decision rule for pattern classification*, Expert Systems with Applications, vol. 42, no. 2, pp. 906- 912.

[9] Iyigun, and A. Ben-Israel (2021), *Probabilistic distance clustering adjusted for cluster size*, Experts-Exchange Probability in the Engineering and Informational Sciences, vol. 22, no. 4, pp. 603-621.

[10] A.K. Jain, (2020), *Data clustering: 50 years beyond Kmeans*, Pattern recognition letters, vol. 31, no. 8, pp. 651-666,.

[11] M. Maier, U. Von Luxburg, and M. Hein (2018), *How the result of graph clustering methods depends on the construction of the graph*, ESAIM: Probability and Statistics, vol. 17, pp. 370-418.

[12] S. Pourbahrami, L. M. Khanli, and S. Azimpour (2019), *A novel and efficient data point neighborhood construction algorithm based on Apollonius circle*, Expert Systems with Applications, vol. 115, pp. 57-67.

[13] J. Sander, M. Ester, H.-P. Kriegel et al. (2017), *Densitybased clustering in spatial databases: The algorithm gdbscan and its applications*, Data mining and knowledge discovery, vol. 2, no. 2, pp. 169-194.

[14] B. K. Singh, K. Verma, and A. Thoke (2016), *Fuzzy cluster based neural network classifier for classifying breast tumors in ultrasound images*," Expert Systems with Applications, vol. 66, pp. 114- 123.

[15] U. Von Luxburg, (2017), A *tutorial on spectral clustering*, Statistics and computing, vol. 17, no. 4, pp. 395-416.

[16] G. Wang, and Q. Song (2016), *Automatic clustering via outward statistical testing on density metrics*, IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 8, pp. 1971-1985.

[17] T.-S. Wang, H.-T. Lin, and P. Wang (2017), *Weighted spectral clustering algorithm for detecting community structures in complex networks*, Artificial Intelligence Review, vol. 47, no. 4, pp. 463- 483.

[18] H. Zhou, J. Li, J. Li et al. (2017), *A graph clustering method for community detection in complex networks*, Physica A: Statistical Mechanics and Its Applications, vol. 469, pp. 551-562.

[19] T. İnkaya, (2015), *A parameter-free similarity graph for spectral clustering*, Expert Systems with Applications, vol. 42, no. 24, pp. 9489-9498.

[20] J. F. O'Callaghan (2003), *An Alternative Definition for Neighborhood of a Point,* IEEE Transactions on Computers, vol. 100, no. 11, pp. 1121-1125, www.ngweb.gre.ac.uk

[21] R. K. Ahuja, Ö. Ergun, J. B. Orlin et al. (2012), *A survey of very large-scale neighborhood search techniques,* preparing and analyzing texts Discrete Applied Mathematics, vol. 123, no. 1-3, pp. 75-102.

[22] Rodriguez, and A. Laio (2014), *Clustering by fast search and find of density peaks*, Science, vol. 344, no. 6191, pp. 1492-1496.

[23] W. Pedrycz (2019), *The design of cognitive maps: A study in synergy of granular computing and evolutionary optimization*, Expert Systems with Applications, vol. 37, no. 10, pp. 7288-7294.

[24] T. İnkaya, S. Kayalıgil, and N. E. Özdemirel (2015), *An adaptive neighbourhood construction algorithm based on density and connectivity*, Pattern Recognition Letters, vol. 52, pp. 17-24.

[25] M. Du, S. Ding, and H. Jia (2016), *Study on density peaks clustering based on k-nearest neighbors and principal component analysis*, Knowledge Based Systems, Random forest algorithm vol. 99, pp. 135-145.

[26] J. Xie, H. Gao, W. Xie et al. (2016), *Robust clustering by detecting density peaks and assigning points based on fuzzy weighted K-nearest neighbors*, Linear regression , Information Sciences, vol. 354, pp. 19-40.

[27] L. Yaohui, M. Zhengming, and Y. Fang (2017), *Adaptive density peak clustering based on Knearest neighbors with aggregating strategy*, Knowledge-Based Systems, vol. 133, pp. 208-220.

[28] R. Mehmood, G. Zhang, R. Bie et al. (2016), *Clustering by fast search and find of density peaks via heat diffusion*, Neurocomputing, vol. 208, pp. 210-217.

[29] Q. Zhang, C. Zhu, L. T. Yang et al. (2017), *An incremental CFS algorithm for clustering large data in industrial Internet of Things*, IEEE Transactions on Industrial Informatics, vol. 13, no. 3, pp. 1193-1201.

[30] B. Wu, and B. M. Wilamowski (2016), *A fast density and grid-based clustering method for data with arbitrary shapes and noise,* IEEE Transactions on Industrial Informatics, vol. 13, no. 4, pp. 1620-1628.

[31] B. Shi, L. Han, and H. Yan (2018), *Adaptive clustering algorithm based on KNN and density*, Pattern Recognition Letters, vol. 104, pp. 37-44.

[32] Z. Liang, and P. Chen (2016), *Delta-density based clustering with a divide-and-conquer strategy: 3DC clustering*, Pattern Recognition Letters, vol. 73, pp. 52-59.

[33] Y. Zhang, S. Chen, and G. Yu (2016), *Efficient distributed density peaks for clustering large data sets in map reduce*, IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 12, pp. 3218-3230.

[34] J. Xu, G. Wang, and W. Deng (2016), *DenPEHC: Density peak based efficient hierarchical clustering*, Information Sciences, vol. 373, pp. 200- 218.

[35] G. P. Guedes, E. Ogasawara, E. Bezerra et al. (2016), *questions migration Discovering top-k non-redundant clustering in attributed graphs*, Neurocomputing, vol. 210, pp. 45-54.

[36] A. T. Hashem, I. Yaqoob, N. B. Anuar et al. (2015), *The rise of big data on cloud computing: Review and open research issues*, Information systems, vol. 47, pp. 98-115.

[37] W. L. G. Koontz, P. M. Narendra, and K. Fukunaga (2012), *A graph-theoretic approach to nonparametric cluster analysis,* IEEE Transactions on Computers, no. 9, pp. 936-944.

[38] R. Urquhart, (1982) *Graph theoretical clustering based on limited neighborhood sets*, Pattern recognition, vol. 15, no. 3, pp. 173-187.

[39] C. T. Zahn (2003), *Graph theoretical methods for detecting and describing gestalt clusters*, IEEE Trans. Comput., vol. 20, no. SLAC-PUB-0672-REV, pp. 68.

[40] J. C. S. C. S. Langerman (2009), *Empty Region Graphs*, Comput. Geom.: Theory Appl., vol. 42, pp. 183-195.

[41] Y. Yang, D.-Q. Han, and J. Dezert (2016), *An angle-based neighborhood graph classifier with evidential reasoning*, Pattern Recognition Letters, vol. 71, pp. 78-85.

[42] Liu, G. V. Nosovskiy, and O. Sourina (2008), *Effective clustering and boundary detection algorithm based on Delaunay triangulation*, Pattern Recognition Letters, vol. 29, no. 9, pp. 1261- 1273.

[43] A. Esmin, R. A. Coelho, and S. Matwin (2015), *A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data*, Artificial Intelligence Review, vol. 44, no. 1, pp. 23-45.

[44] Huang, T. Liu, Y. Yang et al. (2015), *Discriminative clustering via extreme learning machine*, Neural Networks, vol. 70, pp. 1-8.

[45] J. Luo, L. Jiao, R. Shang et al. (2016), *Learning simultaneous adaptive clustering and classification via MOEA*, Pattern Recognition, vol. 60, pp. 37-50.

[46] Y. Peng, W.-L. Zheng, and B.-L. Lu (2016), *An unsupervised discriminative extreme learning machine and its applications to data clustering*, Neurocomputing, vol. 174, pp. 250-264.

[47] S. Rana, S. Jasola, and R. Kumar (2011), *A review on particle swarm optimization algorithms and their applications to data clustering*, Artificial Intelligence Review, vol. 35, no. 3, pp. 211-222.

[48] Q. Wang, Y. Dou, X. Liu et al. (2016), *Multi-view clustering with extreme learning machine*, Neurocomputing, vol. 214, pp. 483-494.

[49] Shen, and O. Hasegawa (2008), "*A fast nearest neighbor classifier based on self-organizing incremental neural network*," Neural Networks, vol. 21, no. 10, pp. 1537-1547.

[50] S. Güney, and A. Atasoy (2012), *Multiclass classification of n-butanol concentrations with knearest neighbor algorithm and support vector machine in an electronic nose*," Sensors and Actuators B: Chemical, vol. 166, pp. 721-725.

[51] X. Pan, Y. Luo, and Y. Xu (2015), *K-nearest neighbor based structural twin support vector machine,* Knowledge-Based Systems, vol. 88, pp. 34-44.

[52] Chatterjee, and P. Raghavan (2012), *Similarity graph neighborhoods for enhanced supervised classification,* Procedia Computer Science, vol. 9, pp. 577-586.

[53] Hajjar, and H. Hamdan (2013), *Interval data clustering using self-organizing maps based on adaptive Mahalanobis distances,* Neural Networks, vol. 46, pp. 124-132.

[54] M. Oja, S. Kaski, and T. Kohonen (2013), *Bibliography of self-organizing map (SOM) papers: 1998-2001 addendum*, Neural computing surveys, vol. 3, no. 1, pp. 1-156.

[55] M. Roigé, M. Parry, C. Phillips et al. (2016), *Selforganizing maps for analysing pest profiles: Sensitivity analysis of weights and ranks,* Ecological Modelling, vol. 342, pp. 113-122.

[56] J. Vesanto, and E. Alhoniemi (2010), *Clustering of the self-organizing map*, IEEE Transactions on neural networks, vol. 11, no. 3, pp. 586-600.

[57] Y. Qin, Z. L. Yu, C.-D. Wang et al. (2018), *A Novel clustering method based on hybrid K-nearest neighbor graph,* Pattern Recognition, vol. 74, pp. 1-14.

[58] Y. Tao, M. L. Yiu, and N. Mamoulis (2006), *Reverse nearest neighbor search in metric spaces,*" IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 9, pp. 1239-1252.

[59] J. P. Papa, S. E. N. Fernandes, and A. X. Falcao (2017), *Optimum-path forest based on k-connectivity: Theory and applications*, Pattern Recognition Letters, vol. 87, pp. 117-126.

[60] C.-D. Wang, J.-H. Lai, and J.-Y. Zhu (2011), *Graph based multi prototype competitive learning and its applications,* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 6, pp. 934-946.

[61] J. Chen, H.-r. Fang, and Y. Saad (2009), *Fast approximate kNN graph construction for high dimensional data via recursive Lanczos bisection*, Journal of Machine Learning Research, vol. 10, no. Sep, pp. 1989-2012.

[62] P. Franti, O. Virmajoki, and V. Hautamaki (2006), *Fast agglomerative clustering using a k-nearest neighbor graph*, IEEE transactions on pattern analysis and machine intelligence, vol. 28, no. 11, pp. 1875-1881.

[63] M. Brito, E. Chavez, A. Quiroz et al. (1997), *Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection,* Statistics & Probability Letters, Linear regression mathematical model ,vol. 35, no. 1, pp. 33- 42.

[64] Z. Hu, and R. Bhatnagar (2012), *Clustering algorithm based on mutual K-nearest neighbor relationships,* Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 5, no. 2, pp. 100-113.

[65] D. L. Davies, and D. W. Bouldin (1998), *A cluster separation measure*, IEEE transactions on pattern analysis and machine intelligence, no. 2, pp. 224-227.

[66] B. Rezaee (2010), *A cluster validity index for fuzzy clustering*, Fuzzy Sets and Systems, vol. 161, no. 23, pp. 3014-3025.

[67] P. J. Rousseeuw (1999), *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,* Journal of computational and applied mathematics, vol. 20, pp. 53-65.

[68] S. E. Schaeffer (2007), *Graph clustering*, Computer science review, vol. 1, no. 1, pp. 27-64.

[69] R. A. Jarvis, and E. A. Patrick (2013), *Clustering using a similarity measure based on shared near neighbors,* IEEE Transactions on computers, vol. 100, no. 11, pp. 1025-1034.

[70] Zou, Jie, Ling Xu, Mengning Yang, Xiaohong Zhang, and Dan Yang (2016), *Towards comprehending the non-functional requirements through Developers' eyes: An exploration of Stack Overflow using topic analysis*, Information and Software Technology,.

[71] Soonthornphisaj, Nuanwan & Tuomchomtam *Sarach* (2019), *Internet User Perception on Data Privacy Protection: Big Data Analytics,* on Twitter, 10.3233/FAIA190178.

[72] Shubham Bindal (2022), *Clustering in ML – Part 4: Connectivity Based Clustering*. Information systems.

[73] Graham (2016), *Better Bayesian Filtering,* Information systems.

[74] Yamada, Hiroyasu & Matsumoto, Yuji (2003), *Statistical dependency analysis with support vector machines*, Proc. Iwpt.

[75] James Clark (2003), *A neural network based approach to automated e-mail classification,* ieeexplore.ieee.org Web Intelligence.

[76] Madjid Tavana, Amir-Reza Abtahi, Debora Di Caprio, Maryam Poortarigh (2018), *An Artificial Neural Network and Bayesian Network model*, volume 275.

Appendix A - Attribute selection

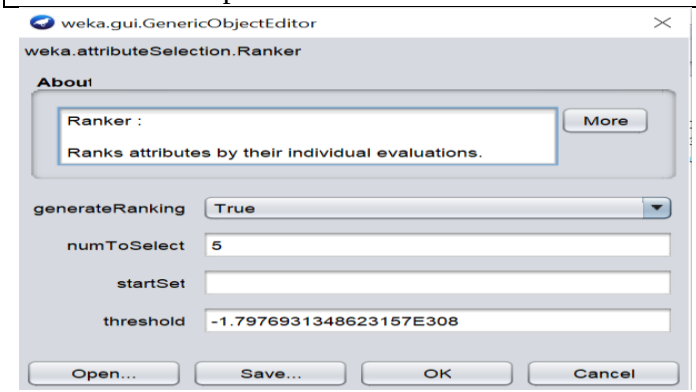
```
Attribute selection out put
=== Run information ===
Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     Fulldataset-weka.filters.unsupervised.attribute.Reorder-
R2,3,4,5,6,7,8,9,10,11,1-weka.filters.unsupervised.attribute.NumericToNominal-
Rlast-weka.filters.unsupervised.attribute.NumericToNominal-R3,4,5,6,7,9,10-
weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,7,8,9,11,10-
weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0-
weka.filters.unsupervised.attribute.Remove-R2
Instances:    24403
Attributes:   10
              Tags
              ViewCount
              AnswerCount
              CommentCount
              FavoriteCount
              A/N
              Id
              Reputation
              Response_01
              TagCount
Evaluation mode:    10-fold cross-validation
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===
average merit      average rank  attribute
 0.814 +- 0        1   +- 0       9 Response_01
 0.694 +- 0.001    2   +- 0       1 Tags
 0.358 +- 0.001    3   +- 0       2 ViewCount
 0.219 +- 0.001    4   +- 0       8 Reputation
 0.027 +- 0.001    5.1 +- 0.3     3 AnswerCount
 0.021 +- 0.003    5.9 +- 0.3     7 Id
 0.005 +- 0        7   +- 0       5 FavoriteCount
 0.003 +- 0        8   +- 0       4 CommentCount
 0     +- 0        9   +- 0      10 TagCount
```

Appendix B - Attribute selection

```
=== Run information ===
Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     Fulldataset-weka.filters.unsupervised.attribute.Reorder-
R1,2,3,4,5,6,7,9,10,11,8-weka.filters.unsupervised.attribute.Remove-R2-3-
weka.filters.unsupervised.attribute.NumericToNominal-R1,2,3,4,5,6,7,8-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R4-
weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
Instances:    24403
Attributes:   6
              AnswerCount
              CommentCount
              FavoriteCount
              Reputation
              TagCount
              A/N
Evaluation mode:    10-fold cross-validation
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit      average rank  attribute
 0.219 +- 0.001    1   +- 0      4 Reputation
 0.027 +- 0.001    2   +- 0      1 AnswerCount
 0.005 +- 0        3   +- 0      3 FavoriteCount
 0.003 +- 0        4   +- 0      2 CommentCount
 0     +- 0        5   +- 0      5 TagCount
```

Appendix C - CFS Subset Eval (Remove unwanted attributes)

```
=== Run information ===
Evaluator:    weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search:       weka.attributeSelection.BestFirst -D 1 -N 5
Relation:     Fulldataset-weka.filters.unsupervised.attribute.Reorder-
R1,2,3,4,5,6,7,9,10,11,8-weka.filters.unsupervised.attribute.Remove-R2-3-
weka.filters.unsupervised.attribute.NumericToNominal-R1,2,3,4,5,6,7,8-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R4-
weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
Instances:    24403
Attributes:   6
              AnswerCount
              CommentCount
              FavoriteCount
              Reputation
              TagCount
              A/N
Evaluation mode:    evaluate on all training data
=== Attribute Selection on all input data ===
Search Method:
              Best first.
              Start set: no attributes
              Search direction: forward
              Stale search after 5 node expansions
              Total number of subsets evaluated: 16
              Merit of best subset found:    0.039

Attribute Subset Evaluator (supervised, Class (nominal): 6 A/N):
              CFS Subset Evaluator
              Including locally predictive attributes
Selected attributes: 4 : 1
                     Reputation
```

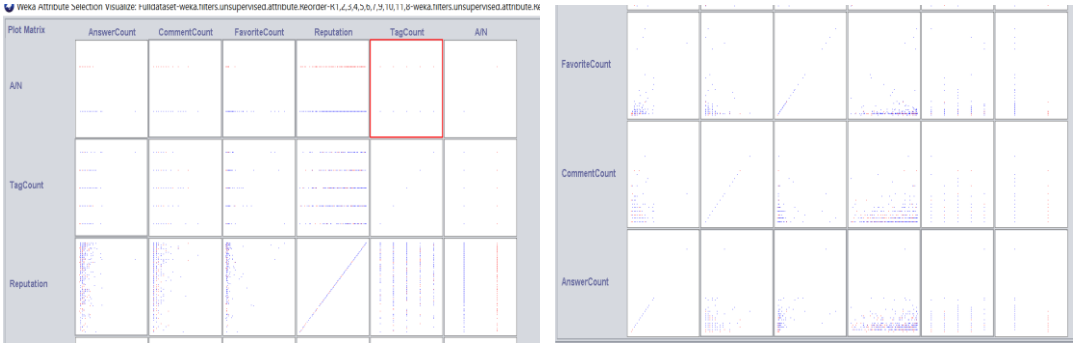


Appendix D- Selected the highest important data set

```
=== Run information ===
Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N 5
Relation:     Fulldataset-weka.filters.unsupervised.attribute.Reorder-
R1,2,3,4,5,6,7,9,10,11,8-weka.filters.unsupervised.attribute.Remove-R2-3-
weka.filters.unsupervised.attribute.NumericToNominal-R1,2,3,4,5,6,7,8-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R4-
weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
Instances:    24403
Attributes:   6
              AnswerCount
              CommentCount
              FavoriteCount
              Reputation
              TagCount
              A/N
Evaluation mode:    evaluate on all training data
=== Attribute Selection on all input data ===
Search Method:
        Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 6 A/N):
        Information Gain Ranking Filter
Ranked attributes:
 0.213138  4 Reputation
 0.026565  1 AnswerCount
 0.004456  3 FavoriteCount
 0.002734  2 CommentCount
 0.000464  5 TagCount
Selected attributes: 4,1,3,2,5 : 5
```

Selected the highest important data set

Visualize reduced data



Classify use to create model in **naïve bayes.**
Use test option **cross validation or Training set or percentage split (training 80 testing 20%). Use supply test set in** evaluation Cross validation
Label attribute A/N
----------Classifier output----------

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     Fulldataset-weka.filters.unsupervised.attribute.Reorder-
R1,2,3,4,5,6,7,9,10,11,8-weka.filters.unsupervised.attribute.Remove-R2-3-
weka.filters.unsupervised.attribute.NumericToNominal-R1,2,3,4,5,6,7,8-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R4-
weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
Instances:    24403
Attributes:   6
              AnswerCount
              CommentCount
              FavoriteCount
              Reputation
              TagCount
              A/N
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier

                Class
Attribute            A       N
              (0.75)  (0.25)
================================
AnswerCount
  0                3.0   255.0
  1             5146.0  2067.0
  2             4680.0  1513.0
  3             3015.0   990.0
  4             1922.0   471.0
  5             1140.0   281.0
  6              730.0   180.0
  7              457.0   143.0
  8              312.0    83.0
  9              198.0    48.0
  10             159.0    31.0
  3             3015.0   990.0
  4             1922.0   471.0
  5             1140.0   281.0
  6              730.0   180.0
  7              457.0   143.0
```

34	2.0	1.0
35	2.0	1.0
36	1.0	2.0
38	3.0	1.0
39	5.0	1.0
43	2.0	1.0
44	2.0	1.0
55	1.0	3.0
67	2.0	1.0
73	2.0	1.0
[total]	18298.0	6191.0
CommentCount		
0	11255.0	3419.0
1	2628.0	996.0
2	1820.0	720.0
3	1017.0	371.0
4	573.0	270.0
5	342.0	144.0
6	250.0	91.0
7	131.0	58.0
8	58.0	21.0
9	57.0	23.0
10	38.0	15.0
[total]	18281.0	6176.0
FavoriteCount		
0	580.0	231.0
1	3651.0	1217.0
2	1651.0	514.0
3	1017.0	289.0
.		
.		
1492	2.0	1.0
1669	1.0	2.0
1728	1.0	3.0
[total]	10525.0	3104.0
……		
Reputation		
1	14.0	2.0
41	1.0	5.0
43	2.0	3.0
54	1.0	3.0
491077	24.0	18.0
806159	15.0	2.0
[total]	20916.0	8811.0
TagCount		
1	1780.0	613.0
2	5146.0	1586.0
3	5966.0	2028.0

```
 4              3662.0  1304.0
 5              1705.0   621.0
 [total]       18259.0  6152.0

Time taken to build model: 0.03 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances       18767               76.9045 %
Incorrectly Classified Instances      5636               23.0955 %
Kappa statistic                          0.2454
Mean absolute error                      0.3166
Root mean squared error                  0.4002
Relative absolute error                 83.9718 %
Root relative squared error             92.1869 %
Total Number of Instances            24403

=== Detailed Accuracy By Class ===
               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.941    0.742    0.790      0.941    0.859      0.278    0.746     0.891     A
               0.258    0.059    0.596      0.258    0.360      0.278    0.746     0.509     N
Weighted Avg.  0.769    0.570    0.741      0.769    0.733      0.278    0.746     0.795

=== Confusion Matrix ===

    a     b   <-- classified as
 17180  1074 |     a = A
  4562  1587 |     b = N
```

Appendix F- Naive Bayes Classifier

Short version first and last part

```
=== Run information ===
Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     Fulldataset-weka.filters.unsupervised.attribute.Reorder-
R1,2,3,4,5,6,7,9,10,11,8-weka.filters.unsupervised.attribute.Remove-R2-3-
weka.filters.unsupervised.attribute.NumericToNominal-R1,2,3,4,5,6,7,8-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R4-
weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
Instances:    24403
Attributes:   6
              AnswerCount
              CommentCount
              FavoriteCount
              Reputation
              TagCount
              A/N
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier
                Class
Attribute            A       N
              (0.75)  (0.25)
================================
AnswerCount
  0               3.0           255.0
  1            5146.0    2067.0
```

Appendix G-Naïve Bayes for training dataset

```
=== Run information ===
Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     Fulldataset-weka.filters.unsupervised.attribute.Reorder-
R1,2,3,4,5,6,7,9,10,11,8-weka.filters.unsupervised.attribute.Remove-R2-3-
weka.filters.unsupervised.attribute.NumericToNominal-R1,2,3,4,5,6,7,8-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R4-
weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
Instances:    24403
Attributes:   6
              AnswerCount
              CommentCount
              FavoriteCount
              Reputation
              TagCount
              A/N
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier
                Class
Attribute            A       N
              (0.75)  (0.25)
================================
Answer Count
 0                3.0   255.0
 1             5146.0  2067.0
```

Appendix H -2. test Naïve bayes with percentage split 66%

Time taken to test model on test split: 0.1 seconds

=== Summary ===

Correctly Classified Instances	6413	77.293 %
Incorrectly Classified Instances	1884	22.707 %
Kappa statistic	0.2405	
Mean absolute error	0.3167	
Root mean squared error	0.3971	
Relative absolute error	84.3292 %	
Root relative squared error	92.1883 %	
Total Number of Instances	8297	

=== Detailed Accuracy By Class ===

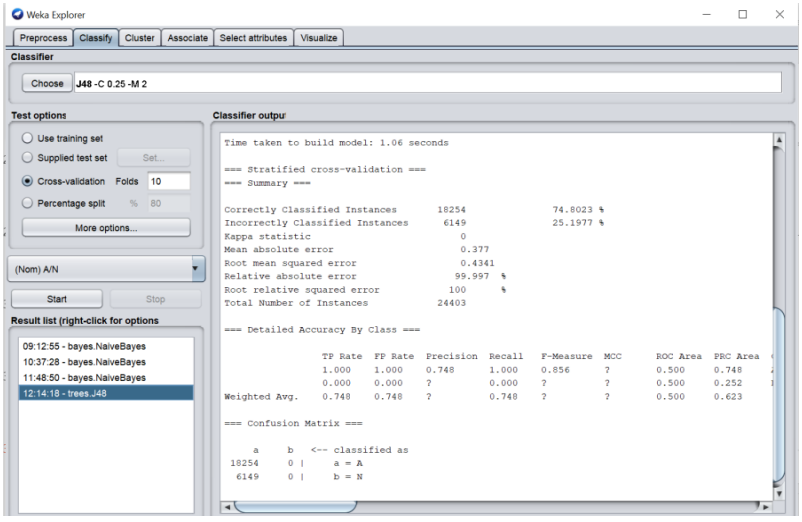
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.943	0.750	0.794	0.943	0.862	0.273	0.746	0.893	A
	0.250	0.057	0.591	0.250	0.352	0.273	0.746	0.504	N
Weighted Avg.	0.773	0.579	0.744	0.773	0.737	0.273	0.746	0.797	

=== Confusion Matrix ===

a	b	<-- classified as
5902	354	a = A
1530	511	b = N

Decision tree model

\ with cross validation10



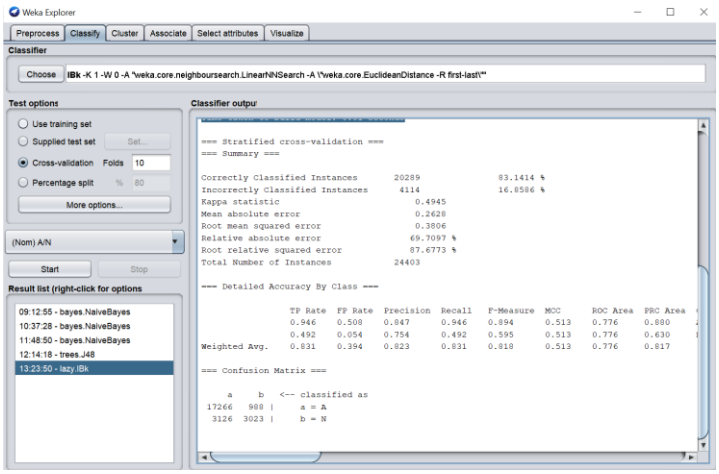
Appendix I- Decision tree model for with cross validation10

```
=== Run information ===
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Fulldataset-weka.filters.unsupervised.attribute.Reorder-
R1,2,3,4,5,6,7,9,10,11,8-weka.filters.unsupervised.attribute.Remove-R2-3-
weka.filters.unsupervised.attribute.NumericToNominal-R1,2,3,4,5,6,7,8-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R4-
weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
Instances:    24403
Attributes:   6
              AnswerCount
              CommentCount
              FavoriteCount
              Reputation
              TagCount
              A/N
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
: A (24403.0/6149.0)
Number of Leaves  :         1
Size of the tree :  1
Time taken to build model: 1.06 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances       18254               74.8023 %
Incorrectly Classified Instances      6149               25.1977 %
Kappa statistic                          0
Mean absolute error                      0.377
Root mean squared error                  0.4341
Relative absolute error                 99.997  %
Root relative squared error            100      %
Total Number of Instances            24403

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    1.000    0.748      1.000    0.856      ?        0.500     0.748     A
               0.000    0.000    ?          0.000    ?          ?        0.500     0.252     N
Weighted Avg.  0.748    0.748    ?          0.748    ?          ?        0.500     0.623

=== Confusion Matrix ===
    a     b   <-- classified as
 18254     0 |     a = A
  6149     0 |     b = N
```

Appendix J- KNN IBk model

```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A
"weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R
first-last\""
Relation:     Fulldataset-weka.filters.unsupervised.attribute.Reorder-
R1,2,3,4,5,6,7,9,10,11,8-weka.filters.unsupervised.attribute.Remove-R2-3-
weka.filters.unsupervised.attribute.NumericToNominal-R1,2,3,4,5,6,7,8-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R4-
weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
Instances:    24403
Attributes:   6
              AnswerCount
              CommentCount
              FavoriteCount
              Reputation
              TagCount
              A/N
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
Time taken to build model: 0.01 seconds
=== Stratified cross-validation  Summary ===

Correctly Classified Instances       20289               83.1414 %
Incorrectly Classified Instances      4114               16.8586 %
Kappa statistic                          0.4945
Mean absolute error                      0.2628
Root mean squared error                  0.3806
Relative absolute error                 69.7097 %
Root relative squared error             87.6773 %
Total Number of Instances            24403

=== Detailed Accuracy By Class ===
               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.946    0.508    0.847      0.946    0.894      0.513    0.776     0.880     A
               0.492    0.054    0.754      0.492    0.595      0.513    0.776     0.630     N
Weighted Avg.  0.831    0.394    0.823      0.831    0.818      0.513    0.776     0.817

=== Confusion Matrix ===
    a     b   <-- classified as
 17266   988 |     a = A
  3126  3023 |     b = N
```

Appendix K- KNN model for training data



=====KNN model for training data 80%=====

=== Run information ===

Scheme: weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Relation: Fulldataset-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,7,9,10,11,8-weka.filters.unsupervised.attribute.Remove-R2-3-weka.filters.unsupervised.attribute.NumericToNominal-R1,2,3,4,5,6,7,8-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R4-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
Instances: 24403
Attributes: 6
AnswerCount
CommentCount
FavoriteCount
Reputation
TagCount
A/N
Test mode: split 80.0% train, remainder test
=== Classifier model (full training set) ===
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
Time taken to build model: 0.01 seconds

=== Evaluation on test split ===
Time taken to test model on test split: 7.22 seconds

=== Summary ===

Correctly Classified Instances 4075 83.487 %
Incorrectly Classified Instances 806 16.513 %

```
Kappa statistic                  0.4944
Mean absolute error              0.2606
Root mean squared error          0.3764
Relative absolute error         69.6387 %
Root relative squared error     87.5678 %
Total Number of Instances        4881

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.946    0.509    0.852      0.946    0.896      0.511    0.779     0.887     A
                 0.491    0.054    0.746      0.491    0.593      0.511    0.779     0.632     N
Weighted Avg.    0.835    0.398    0.826      0.835    0.822      0.511    0.779     0.825

=== Confusion Matrix ===

   a    b   <-- classified as
 3489  199 |    a = A
  607  586 |    b = N
```

Appendix L - Classes to clusters evaluation on training data

```
=== Run information ===

Scheme:       weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter
1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Relation:     Edited Dataset_summary v1-
weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0-
weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
Instances:    6147
Attributes:   10
              Score (partition)
              ViewCount / ( Partition)
              AnswerCount
              CommentCount
              AskerId
              Reputation
              TagCount
              TitleLength
Ignored:
              Technology
              AorN
Test mode:    Classes to clusters evaluation on training data

=== Clustering model (full training set) ===
EM
==
Number of clusters selected by cross validation: 6
Number of iterations performed: 30
                        Cluster
Attribute                    0      1      2      3      4      5
                        (0.14)  (0.01)  (0.02)  (0.27)  (0.26)   (0.3)
============================================================
==============
Score (partition)
  mean                  0.0087   0.005  0.0595  0.0038  0.0048  0.0033
  std. dev.             0.0049  0.0021  0.0966  0.0009  0.0016  0.0005

ViewCount / ( Partition)
  mean                  0.0106  0.0027  0.0946  0.0013   0.003  0.0005
  std. dev.             0.0105  0.0053  0.1365  0.0011  0.0024  0.0004

AnswerCount
  mean                  0.0849  0.0323  0.1945  0.0399  0.0502  0.0255
  std. dev.             0.0553  0.0202    0.16  0.0247  0.0307  0.0145

CommentCount
  mean                   0.073  0.1241  0.1598       0  0.0954  0.0371
  std. dev.             0.0822  0.1384  0.1958  0.0002  0.1021  0.0501
```

```
AskerId
  mean                0.0014  0.2071  0.0014  0.0011  0.0019  0.0018
  std. dev.           0.0013  0.3944  0.0013  0.0008  0.0019  0.0018

Reputation
  mean                0.1018  0.4075  0.0849  0.0641  0.0334  0.0258
  std. dev.           0.1007  0.2248  0.1377  0.0721  0.0333  0.0266

TagCount
  mean                0.4893  0.4639  0.5153  0.4798  0.5386  0.4539
  std. dev.           0.2857  0.2349  0.3198   0.281  0.2775  0.2754

TitleLength
  mean                0.3117  0.3315  0.3246  0.3224  0.3095  0.3253
  std. dev.           0.1372  0.1881  0.1424  0.1578  0.1396  0.1492

Time taken to build model (full training data) : 48.89 seconds
=== Model and evaluation on training set ===
Clustered Instances

0       860 ( 14%)
1        41 (  1%)
2       143 (  2%)
3      2597 ( 42%)
4      1428 ( 23%)
5      1078 ( 18%)

Log likelihood: 21.77031

Class attribute: AorN
Classes to Clusters:

   0    1    2    3    4    5  <-- assigned to cluster
 860   41  143 2597 1428 1078 | N

Cluster 0 <-- No class
Cluster 1 <-- No class
Cluster 2 <-- No class
Cluster 3 <-- N
Cluster 4 <-- No class
Cluster 5 <-- No class

Incorrectly clustered instances :         3550.0  57.7517 %
```