International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Chi-Square Method of Feature Selection: Impact of Pre-Processing of Data

Alisha Sikri[1], N. P. Singh[2], Surjeet Dalal[3]

Submitted: 01/11/2022 **Accepted**: 02/02/2023

Abstract

Feature selection is a technique of lowering computation and data collecting costs and rejecting the less significant or redundant factors/variable which in turn may also increase the efficiency of machine learning algorithms. One of the often-used method of feature selection for categorical data is Chi-Square method which is based on certain assumptions. Defilement of assumptions has an impact on the computed p-values which are surrogate to importance of the features. The main purpose of this study is to identify the impact of pre-processing of data keeping in view the assumptions of Chi-Square on raking of features. A secondary objective is to see how in-built sub-routines of computer languages such as Python or R are incorporating the assumptions of Chi-Square. Based on empirical evidence it was found that it is essential to pre-process the data to fulfill the assumptions of chi-square before subjecting it to analysis using either R or Python, or any other application available on web or otherwise.

Keywords : *Python, assumptions, Chi-Square, R programming.*

1. Introduction

In recent years, with machine-to-machine communication, amount of data collected, and number of features/ variables/ dimensions included in solving business objectives are very large [1]. The problem of high dimensionality or large variables of data can be resolved with feature selection techniques [2]. By doing so, one can identify important features that significantly influencing the model's performance. The main goals of "feature selection" are to remove redundant and unnecessary features from data that make the "knowledge discovery process" during the training phase more efficient in terms of requirement of resources. It also helps in reducing noise and increases the reliability of the training data in training models [3]. The researchers employ "Chi-square (CS) statistics for feature selection to select the variables which in turn enhance the performance of classification models [4]. CS feature selection method is based on the concept of dependence or independence of features with target class [5][6]. It is useful in analyzing categorical variables [7].

Another problem which always results in case of large features is sparsity. Sparsity is chi-square statistics-based methods results in to contingency tables which violate the assumption of "minimum number of expected frequencies". Many software's does not have features of merging ordered categories before using CS featureselection method. It has impact of the ranking of features. This study aims to observe the impact of merging various categories of features on values of Chi-square statistics to get statistically validated measures of dependency. The one computation of Chi-Square is done when the expectedfrequencies of all cells of contingency table are equal to and greater than 5 and other with natural distribution of cell frequencies wherein it could be less than 5 for few cells. With an objective to identify the impact of merging categories to satisfy the assumption of CS feature selection method, three data sets are analyzed. The main objective of this research paper is therefore to study the impact of data preprocessing on the rank(importance) of features calculated using Chi-Square statistics for a given dataset. The secondary objective is to see the features of in-built functions in the code of languages such as Python or R in the context of the pre-processing of data for using Chi-Square test statistics in relation of its assumptions.

[1] Research Scholar, Department of Computer Science and Engineering, SRM University, Delhi-NCR, India, Email: alisha.sikri92@gmail.com

[2] Professor, School of Business Management and Commerce, MVN University, (NCR), Palwal, Haryana, India. Email: dr.npsingh@mvn.edu.in

[3] Professor, Department of Computer Science and Engineering, Amity University, Gurugram, Haryana, India. Email: sdalal@ggn.amity.edu

The article is organized into five sections starting with the first section of introduction on the hidden problems of application of the Chi-Square for feature selection using available application sub-routines of software such as "R" and Python. Section 2 presents review of the literature about study. Section 3 is divided into two sub-sections. The description of dataset used in the study is detailed in first subsection and research procedure used in second sub-section. The results of the computation in terms of Chi-Square values, dimensions of the matrix, and ranks of the features are given in section 4. Last section 5 summarizes the findings of the study and suggestions for future research

2. Literature Review

This section reviews the research studies of feature selection methods wherein the "chi-square" test statistics is used to identify the features in order of their relevance for further analysis such as classification, clustering etc. The objective of the review of literature is to identify the importance and weaknesses of Chi-Square based feature selection methods. Putra & Wardhani (2019) [8] elaborated the importance of "Chi-square feature selections" in identifying features that are not very important. The study aims to adjust theconsequence of "Chi-square based feature selection" on the efficiency of the "Naïve Bayes algorithm" in analyzing the sentiment of documents. They concluded that the analysis without feature selection attained 73.33% accuracy, 100% precision, and 65.21% recall. With the "Chi-square feature selection" it attained 93.33% accuracy, 93.33% precision, and 93.33% recall. (Nihan,2020) [9] explained the Chi- square test statistics and used "Chi-square tests statistics" for "homogeneity", "goodness of fit", and "independence" He claimed that the Chi-square statistic is a powerful tool for testing hypotheses pertaining to variables with nominal measurements.

(Mirkin,2001) [10] applied the "Chi-Square test" of homogeneity and independence. (Goodman & Kruskal,1979) [11] concluded that there is no convincing defense of Chi-Square statistics as a measure of association. (Bolboaca et al.,2011*)* [12] highlighted the common problems in applying chi-square test for "goodness-of-fit", "homogeneity" and "independence test". They mentioned that there are problems with the Chi-Square test when applied to compare observed and theoretical distributions. They pointed out that defining frequency classes is key problem.

(Cai *et al.,* 2021) [14] studied the "Chi-Square (CHI) feature selection" algorithm in terms of its flaws, new direction of CHI based algorithms, development of a new algorithm based on the coefficient of variances, and verification of usefulness of the new algorithm with two different datasets. (Zhai *et al.*, 2018) [15] proposes a method of extracting feature words based on chi-square statistics for using classification algorithms such as naïve Bayes & support vector machines. The authors also mentioned the various feature selection techniques and elaborated on the need to use a relevant feature selection technique to compress its dimensions.

(Pintas*et al.,* 2021) [17] elaborated the various issues involved in the feature selection for text classification and these can be mentioned as measure feature selection and these are the variable entropy, computing the correlation with the target, etc. The next is a subset search which aims to discover the best subset of features to be used in classification. The third issue is "Globalization". One way to overcome the issue of "globalization" is to use a specific set of features for each class/label. The last issue is an ensemble, as explained by the authors, each "FS method" has itspros and cons. Therefore, combining two or more strategies can produce better outcomes than utilizing them individually (Gulia*et al*., 2014) [19] presented the"computational intelligence techniques for Liver Patient Classification". The authors in this paper applied 5 classification algorithms namely "Multi-Layer Perceptron", "J-48", "Random Forest", "Support Vector Machine", and "Bayesian Network" algorithms have been measured for comparing their performance based on "ILPD (Indian Liver Patient Dataset)". The authors explained the dataset in detail. The authors also carried out the analysis afterapplying feature selection using a greedy stepwise approach. After applying the feature selection, the dataset comprised only significant attributes. (Matchima et *al*., 2018) [23] Yate's correction and William's correlation were two other ways that were compared for effectiveness using simulation data after an effort was made to first design a correction method for maintaining a continuity value of the chi-square. The comparison was made using various conditions. The authors used three imputation techniques to address the chi- square test issue of tiny, predicted cell frequency.

(Dahiya et al.,2016) [26] uses five feature selection methods, i.e., chi-square, gain ratio, information gain, relief F and asymmetrical uncertainty for further application of four classification models for credit risk evaluation.

(Dahiya et al.,2017) [25] uses two feature selection methods, i.e., chi-square and principal component analysis (PCA) for ranking and selecting the important features from the data set for further analysis using classification algorithms for credit risk evaluation. They did not mention about weaknesses or strength of any of these methods.

(Bachri et al.,2017) [27] used Chi Square test to select the feature in student academic record data to be the input of Artificial Neural Network (ANN) models for prediction whether a student will graduate or not with better accuracy. They did not mention about data preparation of software used.

Mahmood (2017) [28] applied two feature selection methods, i.e., Chi-square and Relief-F for facial expression recognition for further analysis with Support Vector Machine, K-Nearest, Decision Tree, and Radial Base Function classifiers.

Some of the common feature selection techniques are "Mutual Information (MI)", "Information Gain (IG)", "Gini Coefficient (GI)", "Document Frequency (DF)", "Word Frequency (TF)", "Chi-Square Statistics (CHI)", etc. And as per the analysis done by the authors as well as the literature studied, the "Chi-Square statistics" is one of the most operative feature selection techniques.

Table no.1 summarizes the Literature of Review in terms of algorithms are used by different authors.

3. Research Methodology

The Context: Chi-Square test/statistics are used to determine the importance of features or feature selection in theapplication of machine learning algorithms. For this purpose of feature ranking researchers are using built-in functions of R and Python which may be termed a black-box approach.

This section presents the research methodology followed in the paper to study/analyze the impact of merging categories of features. Furthermore, it mentions the attributes and their details. The next sub-section, section 3.2 explains in detail the research procedure carried out in this research.

Table 1: Summary of literature review

Author	Ch-Square (CS)	Mutual Information (MI)	Information Gain (IG)	Symmetrical Uncertainty (SU)	Gain Ratio (GR)	Relief	Greedy	Gini Index	PCA
(Sulistiani & Tjahyanto,2017) [1]	✓		✓		✓				
(Harbil,2019) [2]	✓		✓					✓	
(Kumar & Sree,2014) [3].	✓		✓	✓	✓	✓			
(Bahassine et al., 2015) [4].	✓	✓	✓						
(Rachburee & Punlumjeak, 2015) [5]	✓		✓				✓		
(Deeplakshmi &Velmurugan, 2016) [6].	✓	✓	✓	✓					
Hazra & Gogtay,2016) [7].	✓								
(Putra & Wardhani, 2019) [8]	✓								
Nihan,2020) [9]	✓								
(Cai et al., 2021) [14]	✓	✓	✓					✓	
(Dahiya et al.,2017) [25]	✓		✓	✓	✓	✓			
(Dahiya et al.,2017) [26]	✓								✓
(Otong Saeful Bachri et al.,2017) [27]	✓								
Mahmood (2017) [28]	✓					✓			
(Mahmood., & Abdulrazzaq, 2022) [2	✓					✓			
(Zhai et al.,2018) [30]	✓								
(Alshaer et al., 2020) [31]	✓								
(Thabtah et al.,2009) [32]	✓								
(Bachi et al., 2017) [33]	✓								

Dataset and attribute information

To attain the objectives this study presents the analysis of three datasets. The detailed instances of thefeature of these datasets are:

a) Dataset 1: Dress Dataset

The Data set contains 13 attributes of women's dresses and their recommendations to prospective buyers according to their sales. The dataset has 500 records in all. The sales are monitored on an alternate day basis.

The dataset consists of 12 independent variables/features/predictors and one target variable. The details are given below in Table no. 2.

Table 2. Attributes/Feature description of dress dataset

Features	Type	Description
Style	Categorical variable	The style of dress is categorized to be of 8 types, and they are mentioned as follows: Bohemia, brief, cute, party, sexy, vintage, and work.
Price	Ordered variable	The prices of the dress are categorized into 5 categories and these categories are Low, average, medium, high, and very high.
Rating	Ordered variable	The dresses are given a rating and this rating varies from 1 to 5.
Size	Ordered variable	The sizes of the dresses are 5 which are: S, M, L, XL, and free size.
Season	Categorical variable	The attribute season is categorized into 4 types, and these are Autumn, winter, spring, and summer.

Neckline	Categorical variable	The Neckline is also one of the attributes of dresses and it is divided into 15 categories, and these areO-neck, boat-neck, bowneck, slash-neck, sweetheart, turndowncollor, v-neck., etc.
Sleeve-Length	Categorical variable	Another attribute of dresses is the length of the sleeve, and these are categorized asFull, half sleeves, short, sleeveless, and three quater.
Waiseline	Categorical variable	Waiseline is another attribute that defines the waiseline of the dresses and these can be of 5 different types, and these are: dropped, empire, natural, princess, and null.
Material	Categorical variable	This attribute defines the dress material used and these areBroadcloth, chiffon, microfibre, micro silk, mix, nylon, partyyster, rayon, silk.
Fabric Type	Categorical variable	This attribute defines the type of fabric, and this can be of 8 types Broadcloth, Chiffon, jersey, Sattin, Worsted.
Decoration	Categorical variable	This attribute defines the type of decoration used on the dresses and these areApplique, beading, bow, broadcloth, button, hpartyLowout, pockets, ruflfe, sashes, sequined, and Patte
Pattern type	Categorical variable	This attribute defines the type of pattern, and this can be of 4 types Animal, Dot, Patchwork, Print, Spartyid, Striped

b) Dataset 2: Indian Patient Liver Dataset

The second dataset is the “Indian Liver Patient dataset” available from the “UCI Machine Learning” Repository. This dataset has 583 records and 10 attributes/features. There are 167 non-liver patient records and 416 liver patient record in this data set.

The data set was gathered in India's north-eastern state of Andhra Pradesh.A class label called "Selector" is used to create groups“liver patient or not”. There are 441 male patient records and 142 female patient records in this data set. The description of the attribute of the dataset is given in Table no. 3.

Table3. Attribute description (Indian liver patient dataset)

S.No	Parameter/Attribute Name	Data type	Attribute Description
1.	age	Integer	“Age of the Patient”
2.	gender	String	“Gender of the Patient”
3.	tot_bilirubin	Real	“Total Bilirubin”
4.	direct_bilirubin	Real	“Direct Bilirubin”
5.	tot_proteins	Integer	“Total Proteins present in patient”
6.	albumin	Integer	“Albumin amount of the patient”
7.	ag_ratio	Integer	“Albumin and Globulin Ratio”
8.	sgpt	Integer	“Alamine Aminotransferase”
9.	Sgot	Real	“Aspartate Aminotransferase”
10.	Alkphos	Real	“Alkaline Phosphatase”
Source:		https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)	

c) Dataset 3: German Credit Dataset

The third dataset used is the German Credit-dataset. The description of this dataset is as follows:

The data is publicly available “German dataset” which is used from the “UCI Machine Learning Repository” [13]. It is the qualitative “real-world credit dataset” with 20 predictor variables that keep track of each loan applicant's identifying details and financial background.One class variable has two possible values: “good” or “bad”.The dataset's qualitative elements predominate, however some are also numerical.1,000 instances make up the dataset, 700 of which are application cases that have been approved and 300 of which have been rejected.

The statistics of the data are given below in Table no. 4.

Table 4. Statistics of German Dataset

Features	Min	Max	Mean	Standard Deviation
Account Balance	1	4	2.577	1.257
Duration of Credit(month)	4	72	20.903	12.479
Payment Status of Pervious Credit	1	5	2.545	1.083
Purpose	0	10	2.828	2.744
Credit Amount	250	18424	3271.248	2822.75
Value Savings/Stocks	1	5	2.105	1.58
Length of current employment	1	5	3.384	1.20
Instalment per cent	1	4	2.973	1.11
Personal status	1	4	2.682	0.708
Guarantors	1	3	1.145	0.477

Duration in a current address	1	4	2.845	1.103
Most valuable available asset	1	4	2.358	1.050
Age (years)	19	75	35.542	11.35
Concurrent Credits	1	3	2.675	0.705
Type of apartment	1	3	1.928	0.530
No. of Credits at this Bank	1	4	1.407	0.577
Occupation	1	4	2.904	0.653
No. of dependents	1	2	1.155	0.362
Source:		https://www.kaggle.com/datasets/uciml/german-credit		

a. Research Methodology of Research Data Analysis

This section details the procedure of calculating chi-square statistics in python and excel including the preparation of data with three namely: (i) original categories of features; (ii) reducing the categories of features meeting the assumptions of estimated frequencies and (iii) further reduction of categories to three. The code for this algorithm isas under:

```
Import pandas as pd From
scipy import stats
crosstab = pd.crosstab(df['Age'], df['SELECTOR'])
stats.chi2_contingency(crosstab)
```

Computation of chi-square statistics in a spreadsheet with three sets of data, The formula of Chi-Square statistics used is as under[6][7][9][16][21]:

$$Chi - Square = \sum_{i=1}^{n} (O_i\text{-}E_i)^2/E_i \qquad (1)$$

Where “O_i”: The observed count ineach cell, and “E_i”: The expected count in that cell.

Assumptions of Chi-Square:

If the following three suppositions are true, the chi-square for 2x2 contingency tables is the fair test ofthe implication of the difference between expected and observed raw frequencies. This test is used to assess raw frequencies for two binary variables [22]:

i. The frequencies for the category variables make up the nominal scales.

ii. Every observation stands alone from every other one.

iii. The expected frequencies are, as a rule, equal to or greater than 5.

The concept of the Yates correction is elaborated in the article. This correction reduces the magnitudeof the calculated chi-square, rendering it less likely to be found significant [24].

4. Results

This section gives the computational results of the chi-square technique when calculated in 3 scenarios. It can be seen in the data those categories have a very low frequency of occurring which might affect the ranks calculated using the feature selection techniques and this can hamper the performance results. So, to avoid this to happen, the categories with low frequency of occurrence can be merged with the ones which have similar characteristics with them. In this research work firstly, the results are exhibited without any merging, then the categories with a frequency of occurrence less than or equal to 5 are merged with other categories having similar characteristics with them, and lastly, all the categories are merged to give a maximum of only 3categories.

Table 5 below displays the CS results of excel and Python of the dress dataset.

Table 5: Chi-Square values and ranks (Dress dataset)

Features	Chi-Square test statistic (Python)	Chi-square test statistic(Excel)	Rank	Chi-square test statistics for data satisfying Chi-Square assumptions (Excel)	Rank	Chi-square test statistics converted crosstab (3*2) (Excel)	Rank
Style	24.05(8x2)	24.05(8x2)	2	18.44(6x2)	4	8.09(3x2)	4
Price	19.97(5x2)	19.97(5x2)	5	14.74(4x2)	6	6.83(3x2)	6
Rating	13.47(13x2)	13.47(13x2)	8	9.11(10x2)	8	1.65(3x2)	11
Size	2.14(5x2)	2.14(5x2)	12	1.43(4x2)	12	0.67(3x2)	12
Season	26.00(4x2)	26.00(4x2)	3	26.00(4x2)	3	25.87(3x2)	1
Neckline	22.82(7x2)	22.82(7x2)	4	5.12(5x2)	10	3.17(3x2)	7
Waiseline	3.26(3x2)	3.26(3x2)	11	2.99(2x2)	11	2.99(2x2)	8
Sleevelength	12.51(5x2)	12.51(5x2)	9	12.67(5x2)	7	2.15(3x2)	10
Material	16.12(9x2)	16.12(9x2)	7	41.43(6x2)	1	7.32(3x2)	5
Fabric Type	18.71(5x2)	18.71(5x2)	6	17.67(4x2)	5	16.88(3x2)	3
Decoration	29.10(10x2)	29.10(10x2)	1	26.51(8x2)	2	2.98(3x2)	9
Pattern.Type	5.95(6x2)	5.95(6x2)	10	5.95(6x2)	9	16.88(3x2)	2
Note: Value in parenthesis is the dimensions of the matrix							

Table no. 6 below displays the CS results of excel and Python of the Indian liver patient dataset.

Table 6. Chi-Square values and ranks (Indian liver patient dataset)

Features	Chi-Square test statistic (Python)	Chi-square test statistic(Excel)	Rank	Chi-square test statistics for data satisfying Chi-Square assumptions (Excel)	Rank	Chi-square test statistics converted crosstab (3*2) (Excel)	Rank
age	32.00(10x2)	32.00(10x2)	3	27.70(7x2)	4	19.68(3x2)	5
gender	3.54(2x2)	3.95(2x2)	9	3.95(2x2)	9	3.95(2x2)	9
tot_bilirubin	29.00(9x2)	29.00(9x2)	4	29.00(6x2)	3	29.00(3x2)	3
direct_bilirubin	37.41(10x2)	37.41(10x2)	2	37.41(8x2)	2	37.41(3x2)	2
tot_proteins	1.35(4x2)	1.35(4x2)	10	1.71(3x2)	10	0.83(3x2)	10
Albumin	9.62(3x2)	9.62(3x2)	8	9.62(3x2)	8	9.62(3x2)	8
ag_ratio	17.50(5x2)	17.50(5x2)	6	12.59(2x2)	7	12.59(2x2)	7
sgpt	14.94(8x2)	14.94(8x2)	7	14.94(4x2)	6	14.94(3x2)	6
Sgot	20.39(9x2)	20.39(9x2)	5	20.05(2x2)	5	20.39(2x2)	4
Alkphos	45.82(10x2)	45.82(10x2)	1	42.94(3x2)	1	42.17(3x2)	1

Note: Value in parenthesis is the dimensions of the matrix

Table no.7 below displays the CS results of excel and Python of the Credit dataset.

Table 7. Chi-Square values and ranks (Credit Dataset)

Features	Chi-Square test statistic (Python)	Chi-square test statistic(Excel)	Rank	Chi-square test statistics for data satisfying Chi-Square assumptions (Excel)	Rank	Chi-square test statistics converted crosstab (3*2) (Excel)	Rank
Account Balance	123.72(4x2)	123.72(4x2)	1	123.72(4x2)	1	120.84(3x2)	1
Payment Status of previous credit	61.69(5x2)	61.69(5x2)	2	61.69(5x2)	2	40.92(3x2)	2
Value savings/stocks	36.09(5x2)	36.09(5x2)	5	35.95(5x2)	4	35.49(3x2)	3
Length of current employment	18.36(5x2)	18.36(5x2)	10	18.36(4x2)	10	1.29(3x2)	17
Instalment percent	5.47(4x2)	5.47(4x2)	15	5.47(4x2)	15	5.33(3x2)	12
Sex and Marital status	9.60(4x2)	9.60(4x2)	12	9.60(4x2)	12	7.06(3x2)	9
Guarantors	6.64(3x2)	6.64(3x2)	13	6.64(3x2)	13	6.64(3x2)	11
Duration in a current address	0.74(4x2)	0.74(4x2)	19	0.74(4x2)	19	0.0(3x2)	20
Most valuable available asset	23.71(4x2)	23.71(4x2)	8	23.71(4x2)	7	0.28(3x2)	18
Concurrent credits	12.83(3x2)	12.83(3x2)	11	12.83(3x2)	11	12.83(3x2)	5
Type of apartment	18.67(3x2)	18.67(3x2)	9	18.67(3x2)	9	18.67(3x2)	4
No. of credits at this bank	2.67(4x2)	2.67(4x2)	16	2.33(3x2)	16	2.33(3x2)	14
Occupation	1.88(4x2)	1.88(4x2)	17	1.88(4x2)	17	1.82(3x2)	15
No. of dependents	0.009(2x2)	0.009(2x2)	20	0.009(2x2)	20	0.009(2x2)	19
Telephone	1.32(2x2)	1.32(2x2)	18	1.32(2x2)	18	1.32(2x2)	16
Foreign Workers	5.82(2x2)	5.82(2x2)	14	5.82(2x2)	14	5.82(2x2)	10
Purpose	33.35(10x2)	33.35(10x2)	6	26.49(7x2)	6	10.65(3x2)	8
Age	24.05(13x2)	24.05(13x2)	7	23.04(10x2)	8	11.30(3x2)	7
Credit Amount	37.12(9x2)	37.12(9x2)	4	35.87(7x2)	5	11.84(3x2)	6
Duration of Credit (Month)	49.82(12x2)	49.82(12x2)	3	45.83(8x2)	3	5.02(3x2)	13

Note: Value in parenthesis is the dimensions of the matrix

5. Conclusions And Future Scope

3 datasets were analysed with a view to have empirical evidences how important is pre-processing of data in the context of using in-built functions for data analysis software's i.e., Python and R. It is observed that in-built functions of Python & R do not automatically take care of assumptions of Chi-Square test statistics. Therefore, the values calculated with in-built functions are not realistic if assumptions of Chi-Square are not met. It is shown empirically in Table no. 5,6, and 7 that the values of Chi-Square statistics calculated with excel spreadsheet without checking for assumption are equal to the values calculated with Python function mentioned below:

From scipy import stats crosstab = pd.crosstab(df['Predictor Variable], df['Target Variable']) stats.chi2_contingency(crosstab)

However, the values of Chi-Square statistics with pre-processed data or data satisfying assumptions of Chi-Square test are quite different from the values calculated with the Python function. It is suggested that in-built functions of Python should not be used on original data which may not satisfy the assumptions of Chi-Square specifically the assumption of threshold value of estimated frequencies of each cell.

References

[1] Sulistiani, H., & Tjahyanto, A. (2017). Comparative analysis of feature selection method to predict customer loyalty. *IPTEK the Journal of Engineering*, *3*(1), 1-5.

[2] Al-Harbi, O. (2019). A comparative study of feature selection methods for dialectal Arabic sentiment classification using support vector machine. *arXiv preprintarXiv:1902.06242*.

[3] Kumar, C. S., & Sree, R. J. (2014). Application Of Ranking Based Attribute Selection Filters to Perform Automated Evaluation of Descriptive Answers Through Sequential Minimal Optimization Models. ICTACT Journal on Soft Computing, 5(1).

[4] Bahassine, S., Madani, A., Al-Sarem, M., & Kissi, M. (2020). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University-Computer and Information Sciences*, *32*(2), 225-231.

[5] Rachburee, N., & Punlumjeak, W. (2015, October). A comparison of feature selection approach between greedy, IG-ratio, Chi-square, and mRMR in educational mining. In *2015 7th international conference on information technology and electrical engineering (ICITEE)* (pp. 420-424). IEEE.

[6] Rafei, N. S. I. M., Hassan, R., Saedudin, R. R., Raffei, A. F. M., Zakaria, Z., & Kasim, S. (2019). Comparisonof feature selection techniques in classifying stroke document. *Indonesian Journal of Electrical Engineering andComputer Science*, *14*(3), 1244-1250.

[7] Hazra, A., & Gogtay, N. (2016). Biostatistics series module 1: Basics of biostatistics. *Indian Journal of Dermatology*, *61*(1), 10.

[8] Putra, A. E., & Wardhani, L. K. (2019, November). Chi-Square Feature Selection Effect on Naive Bayes Classifier Algorithm Performance For Sentiment Analysis Document. In *2019 7th International Conference on Cyber and IT Service Management (CITSM)* (Vol. 7, pp. 1-7). IEEE.

[9] Nihan, S. T. (2020). Karl Pearsonâ€™ s chi-square tests. *Educational Research and Reviews*, *15*(9), 575-580.

[10] Mirkin, B. (2001). Eleven ways to look at the chi- squared coefficient for contingency tables. *The American Statistician*, *55*(2), 111-120.

[11] Goodman, L. A., & Kruskal, W. H. (1979). Measures of association for cross classifications. In *Measures of association for cross classifications* (pp. 2-34). Springer, New York, NY.

[12] Bolboacă, S. D., Jäntschi, L., Sestraş, A. F., Sestraş, R. E., & Pamfil, D. C. (2011). Pearson-Fisher chi-square statistic revisited. *Information*, *2*(3), 528-545.

[13] Asuncion, A., & Newman, D. (2007). UCI machinelearning repository.

[14] Cai, L. J., Lv, S., & Shi, K. B. (2021). Application of an improved CHI feature selection algorithm. *Discrete dynamics in nature and society*, *2021*.

[15] Zhai, Y., Song, W., Liu, X., Liu, L., & Zhao, X. (2018, November). A chi-square statistics-based feature selection method in text classification. In *2018 IEEE 9th International conference on software engineering and service science (ICSESS)* (pp. 160-163). IEEE.

[16] Bisht, N., Ahmad, A., & Bisht, S. (2016). Applicationof feature selection methods and ensembles on network security dataset. *International Journal of Computer Applications*, *135*(11), 1-5.

[17] Pintas, J. T., Fernandes, L. A., & Garcia, A. C. B. (2021). Feature selection methods for text classification: a systematic literature review. *Artificial Intelligence Review*, *54*(8), 6149-6200.

[18] Muthuselvan, S., Rajapraksh, S., Somasundaram, K., & Karthik, K. (2018). Classification of liver patient dataset using machine learning algorithms. *Int. J. Eng. Technol*, *7*(3.34), 323.

[19] Gulia, A., Vohra, R., & Rani, P. (2014). Liver patient classification using intelligent techniques. *International Journal of Computer Science and Information Technologies*, *5*(4), 5110- 5115.

[20] Ramana, B. V., Babu, M. S. P., & Venkateswarlu, N. B.(2011). A critical study of selected classification algorithms for liver disease diagnosis. *International Journal of Database Management Systems*, *3*(2), 101-114.

[21] Kim, H. Y. (2017). Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restorative dentistry & endodontics*, *42*(2), 152-155.

[22] Brown, J. D. (2013). Chi-square and related statistics for 2× 2 contingency tables. *Testing and Evaluation SIG*, 33.

[23] Matchima, K., Vongprasert, J., & Chutiman, N. (2018). The Development of a Correction Method for Ensuring a Continuity Value of The Chi-square Test with a Small Expected Cell Frequency. *Naresuan University Journal: Science and Technology (NUJST)*, *26*(1), 98-105.

[24] Peritz, E., & Haviland, M. G. (1992). Yates's correction for continuity and the analysis of 2× 2 contingency tables. *Statistics in medicine*, *11*(6), 845-847.

[25] Dahiya, S., Handa, S. S., & Singh, N. P. (2017). A feature selection enabled hybrid-bagging algorithm for credit risk evaluation. Expert Systems, 34(6),e12217. doi:10.1111/exsy.12217.

[26] Dahiya, S., Handa, S. S., & Singh, N. P. (2016). A Rank Aggregation Algorithm for Ensemble of Multiple Feature Selection Techniques in Credit Risk Evaluation. (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 5, No. 9, 2016.1-8

[27] Bachri, O. S., Kusnadi, M. H., & Nurhayati, O. D. (2017). Feature selection based on CHI square in artificial neural network to predict the accuracy of student study period. International Journal of Civil Engineering and Technology, 8(8).

[28] Mahmood, M.R. (2020). Two Feature Selection Methods Comparison Chi-square and Relief-F for Facial Expression Recognition, Journal of Physics: Conference Series 1804 (2021) 012056, doi:10.1088/1742-6596/1804/1/012056.

[29] Mahmood, M. R., & Abdulrazzaq, M. B. (2022). Performance evaluation of chi-square and relief-F feature selection for facial expression recognition. Indonesian Journal of Electrical Engineering and Computer Science, 27(3), 1470- 1478.

[30] Zhai, Y., Song, W., Liu, X., Liu, L., & Zhao, X. (2018, November). A chi-square statistics-based feature selection method in text classification. In 2018 IEEE 9th International conference on software engineering and service science (ICSESS) (pp. 160-163). IEEE.

[31] Alshaer, H. N., Otair, M. A., Abualigah, L., Alshinwan, M., & Khasawneh, A. M. (2021). Feature selection method using improved CHI Square on Arabic text classifiers: analysis and application. Multimedia Tools and Applications, 80(7), 10373-10390.

[32] Thabtah, F., Eljinini, M., Zamzeer, M., & Hadi, W. (2009, January). Naïve Bayesian based on Chi Square to categorize Arabic data. In Proceedings of the 11th international business information management association conference (IBIMA) conference on innovation and knowledge management in twin track economies, Cairo, Egypt (pp. 4-6).

[33] Bachri, O. S., Kusnadi, M. H., & Nurhayati, O. D. (2017). Feature selection based on CHI square in artificial neural network to predict the accuracy of student study period. International Journal of Civil Engineering and Technology, 8(8).

[34]