JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 10, 2020

PREDICTION OF STUDENT DROPOUT USING FEATURE SELECTION ALGORITHM

UBHAIDA ASLAM, DR RAVINDRA KUMAR GUPTA

Ubhaida Aslam,P.hd scholar Department of computer Applications and Sciences, Sarvepalli Radhakrishnan University, Bhopal, MP, INDIA.
(Email id:uaslambhat@gmail.com Whatsapp no: 6006264467)

Dr Ravindra Kumar Gupta, Sarvepalli RadhakrishnanUniversity, Bhopal, MP, INDIA.

ABSTRACT
Applicable feature recognizable proof has become a basic assignment to apply data mining calculations viably in true situations. Along these lines, many feature selection methods have been proposed to get the important feature subsets in the writing to accomplish their targets of order and grouping. This paper presents the ideas of feature pertinence, general strategies, assessment standards, and the attributes of feature selection lastly feature selection calculation (utilizing the chi square test) will be utilized on prediction of school dropouts. The objective of this paper is to discover comparable examples of utilization in the data assembled from datasets and in the end have the option to make predictions for every student dependent on different segment, scholastic and point of view characteristics. In conclusion data from the investigation could reveal insight into how to all the more likely help in danger students. We will finish up this work with genuine application (like early prediction of student dropouts), difficulties, and future research headings of feature selection utilizing filter method.
KEYWORDS*: Feature Selection, Filter method, Student dropout, Data mining, Machine learning.*

ABBREVIATIONS: *FS, Feature Selection ; FM, filter method; SD, Student Dropout*

1. INTRODUCTION

The wealth of data in contemporary datasets requests advancement of sharp calculations for finding significant data. Data models are developed relying upon the data mining errands, however normally in the territories of order, relapse and grouping. Regularly, pre-preparing of the datasets happens for two primary reasons:
1) Reduction of the size of the dataset so as to accomplish progressively effective investigation, and
2)Adaptation of the dataset to best suit the chose investigation method [1]
The subsequent explanation is increasingly significant these days and let me proceed with my exploration on this, There are two significant ways to deal with feature selection.
The first is Individual Evaluation, and the second is Subset Evaluation. Positioning of the features is known as Individual Evaluation . In Individual Evaluation, the heaviness of an individual feature is doled out as indicated by its level of significance.[3]
In Subset Evaluation, up-and-comer feature subsets are developed utilizing search methodology.
The general methodology for feature selection has four key strides as appeared in Figure 1.

- Feature Subset Generation
- Evaluation of Feature Subset
- Stopping Criteria
- Result Validation

Subset generation is a heuristic pursuit wherein each state determines a competitor subset for assessment in the inquiry space. Two essential issues decide the idea of the subset generation process.
In the first place, replacement generation chooses the pursuit beginning stage, which impacts the inquiry course. To choose the inquiry beginning stages at each state, forward, in reverse, compound, weighting, and arbitrary methods might be considered .
Second, scan association is answerable for the feature selection process with a particular procedure, for example, successive inquiry, exponential hunt [6] or irregular pursuit .
A recently produced subset must be assessed by a specific assessment standards. Accordingly, numerous assessment measures have been proposed in the writing to decide the integrity of the up-and-comer subset of the features. Base on their reliance on mining algorithms, assessment rules can be classified into gatherings: free and ward measures . Free measures abuse the fundamental attributes of the preparation data without including any mining algorithms to assess the integrity of a feature set or feature and subordinate measures include foreordained mining algorithms for feature selection to choose features dependent on the presentation of the mining algorithm applied to the chose subset of features[7].
At last, to stop the selection procedure, stop standards must be resolved. Feature selection process stops at approval method. It isn't the piece of feature selection process, yet feature selection method must be approve via completing various tests and correlations with recently settled outcomes or , genuine world datasets, or both.

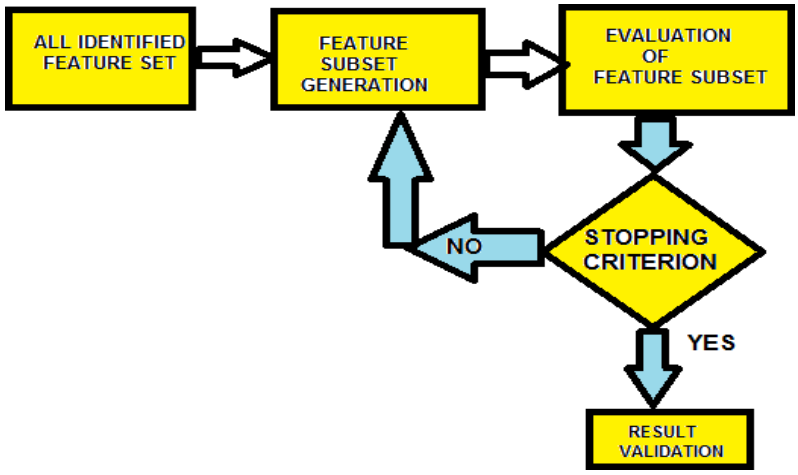


Figure 1:- Process of Feature Selection[2].

There are three general approaches for feature selection. To begin with, the Filter Approach exploits the general attributes of preparing data with free of the mining algorithm.
Second, the Wrapper Approach investigates the connection among importance and ideal feature subset selection. It scans for an ideal feature subset adjusted to the particular mining algorithm [9].
What's more, third, the Embedded Approach is finished with a particular learning algorithm that performs feature selection during the time spent preparing.
In this paper we will utilize the main approach and actualize its algorithm on prediction of student dropout utilizing chi square test.

2. RELATED WORK

Prediction of Student dropout utilizing feature selection Algorithm is generally new zone for the examination, Although some work has been done on featue selection algorithm yet to actualize it in foreseeing the early student drop out is new commitment to the data mining region.
A portion of the works that have examined on feature selection algorithm include:
El-Halees [4] has structured a clear contextual investigation that utilizes scholastic data handling to research the changed conduct of researchers learning. the objective of his investigation is to call attention to how accommodating data handling will be utilized in instructive movement to improve student's exhibition. They applied strategies info} mining to reveal large databases like affiliation rules and grouping rules utilizing a call tree, pack and anomaly investigation.
Bharadwaj and Pal [5] arranged the novel approach abuse the decision tree methodology for grouping to pass judgment on the student's presentation. This contextual investigation is to work out the data that portrays students' exhibition inside the completion semester assessment. This investigation was very useful to detect the dropout's student in A prior stage and students United Nations office need unique consideration and license the coach to require prior regard for the researchers.
Chang, Verhaegen, and Duflou, 2014, Guyon and Elisseeff, 2003; Kohavi and John, 1997, H. M. Harb and M. A. Moustafa Scheirer, H.Liu and H.Motoda, and others have contributed in the significant field.

3. METHODOLOGY

This paper has been preoccupied from the theory work of the creator which depends on both essential and auxiliary data and data. With the end goal of research one areas of jammu and Kashmir in India has been chosen, to be specific Anantnag which is instructively typical and is second evolved locale in Kashmir division of the state having education pace of 64.32%. The data gathered from the region show net enrolment proportion (GER) of student's up to the class twelfth and furthermore gathered sexual orientation insightful, data of drop out students.
For the assortment of essential data we have directed a field overview of the family unit of drop-out students subsequent to gathering auxiliary data and data from schools data.
For the assortment of auxiliary data and data we have visited ZONAL EDUCATION OFFICEs, different higher optional schools in Anantnag area of jammu and Kashmir and counseled diverse library Published sources on the issue of dropout, Director school training Kashmir and JKBOSE (jammu and Kashmir State leading body of School Education Kashmir) site data have additionally been used.
We got different reasons why students can't proceed with higher education(why they fizzled) lastly dissuades more reactions were broke down utilizing chi square test of feature selection algorithm on the data gathered to foresee whether these features were critical or not.

ISSN- 2394-5125 VOL 7, ISSUE 10, 2020

4. FILTER METHODS

There are different methods in machine learning to determining whether our info features are pertinent to the result to be anticipated or not. Positioning strategies are utilized as guideline rules in Filter method.

The factors are appointed a score utilizing an appropriate positioning basis and the factors having score underneath some edge esteem are expelled. These methods are computationally less expensive, maintains a strategic distance from over fitting yet Filter methods disregard conditions between the features. Consequently, the chose subset probably won't be ideal and an excess subset may be gotten. One of the essential filter feature selection algorithms is examined underneath alongside usage

4.1 CHI-SQUARE TEST

Filter feature selection methods are those that utilization some factual strategies (like Chi-Squared test) considering the data sort of the info and target variable to assess the connection between them. The Pearson's Chi-Squared test, or just Chi-Squared test, is named after the mathematician Karl Pearson. It is likewise called an "integrity of fit" measurement. A chi-square ($\chi 2$) measurement is a test that estimates how desires contrast with real watched data (or model outcomes). The data utilized in ascertaining a chi-square measurement must be irregular, crude, fundamentally unrelated, drawn from autonomous factors, and drawn from a huge enough example. Chi-square tests are regularly utilized in theory testing.
The Formula for Chi-Square Is

$$\chi_c^2 = \sum Ei(Oi - Ei)^2/E_i \qquad (i)$$

Where: c=Degrees of freedom

O=Observed value(s)

E=Expected value(s)

5. IMPLEMENTATION OF CHI SQUARE TEST ON SCHOOL DATA TO PREDICT STUDENT DROPOUT.

For the assortment of essential data we have directed a field review of the family unit of drop-out students subsequent to gathering optional data and data from schools data. For the assortment of optional data and data we have visited ZONAL EDUCATION OFFICEs, different higher auxiliary schools in Anantnag locale of jammu and Kashmir and counseled distinctive library Published sources on the issue of dropout, Director school instruction Kashmir and jkbose site data have likewise been used.
At that point we got a data of 10364 students of locale/District Anantnag who showed up in class twelfth assessment held in December 2019 conducted by jammu and Kashmir state leading group of school training. Among them 8342 students have a place with Govt. Higher secondary's(There are 42 Govt. Higher secondary's in the area).
Out of 8342 Govt. school students 4589 passed (55.01%) and 3753(44.99%) students neglected to clear the assessment.
For our exploration reason we chose data of two higher secondary's one from provincial region and one from urban zone.
All out students of these higher optional's who showed up (appeared) in class twelfth assessment 2019 = 282

Number of students passed=152
Number of students fizzled/FAILED =130
What's more, subsequently they won't have the option to go for school (advanced education). When directed the overview of these disappointment students ,we got different responses(features) from students, guardians, instructors and academicians.
The responses generated were noted as below (Reason of failure).

- Poverty
- Negative behaviour of teacher
- Early marriage
- Hartal/strikes
- Carelessness of parents
- Illness of parents
- Orphanage

be that as it may, lion's share of these students, guardians, teachers and academicians worried upon poverty, negative behavior of teacher and early marriage(for example the rate for reactions of poverty, negative behavior of teacher and early marriage was more when contrasted with rest of the features).

ISSN- 2394-5125 VOL 7, ISSUE 10, 2020

Let us consider an exceptionally straightforward dataset with just two segments. We will see whether Gender is connected/subordinate/related to the Reason of Failure.

Gender	Reason of Failure
Male	Poverty
Female	Early Marriage
Male	poverty
Male	NEGATIVE BEHAVIOUR OF TEACHER
…….	……

Table 1:- simple dataset with only two columns

Let us do Hypothesis Test (measurable method) which assesses two proclamations (speculation) and figures out which articulation is valid.
Let Null Hypothesis (beginning proclamation) be indicated as H0
Interchange Hypothesis(usually correlative to the first) be signified as H1.
For our model, the theory are:
H0: The features Gender and Reason Of Failure are free (which implies they are not related).
H1: Gender and Reason Of Failure are Reason Of Failure (which implies they are related).
Let $\alpha = 0.05$
Lower α values are commonly favored which might be in the scope of 0.01 to 0.10.
where α (alpha) is a proportion of centrality level for example how certain we need to be about our outcomes.
The noteworthiness level is utilized to decide if the invalid theory ought to be dismissed or not.
For the invalid theory to be dismissed the p-worth ought to be not exactly the noteworthiness level.

NOW WE CREATE A CONTINGENCY TABLE

A table demonstrating the recurrence circulation of one variable in lines and another in segments, used to examine the relationship between's the two factors is known as a possibility table (otherwise called a cross arrangement or crosstab).

	Poverty	Early Marriage	NEGATIVE BEHAVIOUR OF TEACHER	Row total
Male	20	20	25	65
Female	25	30	10	65
Column Total	45	50	35	130

Table 2:- Contingency table

We can see that out of 130 candidates, in the dataset, there are 20, 20 and 25 guys having reason of disappointment Poverty, Early Marriage and NEGATIVE Behavior OF TEACHER individually. Thus 25, 30 and 10 Females are keen on Poverty, Early Marriage and NEGATIVE Behavior OF TEACHER respectively.The values in the table are called as watched values.

CALCULATE EXPECTED FREQUENCY

We figure the normal recurrence mean every cell. The equation to tally expected recurrence is:
E = (push complete * segment all out)/grand aggregate
The Expected recurrence for first cell (for example Male-poverty) will be:
$E_1 = (65 * 40)/130 = 20$
We figure Expected Frequency for rest of the cells and get the table underneath where esteems in sections '[]' signify anticipated frequencies:

	Poverty	Early Marriage	NEGATIVE BEHAVIOUR OF TEACHER	Row total
Male	20[22.5]	20[25]	25[17.5]	65
Female	25[22.5]	30[25]	10[17.5]	65
Column Total	45	50	35	130

Table 3:- Expected Frequencies

ISSN- 2394-5125 VOL 7, ISSUE 10, 2020

CALCULATE THE CHI-SQUARE VALUE (CHI-SQUARE STATISTIC)

The formula to calculate Chi-square value or χ^2 is:

$$\chi_c^2 = \sum Ei(Oi-Ei)^2/E_i$$

where χ isn't the English alphabet we know yet the Greek letters in order Chi.

As appeared, χ^2 is the summation of the squared distinction among Observed and Expected frequencies partitioned by the Expected recurrence for all the cells. The estimations are demonstrated as follows:
χ^2 = ((20-22.5)^2/22.5) + ((20-25)^2/25) + ((25-17.5)^2/17.5) + ((25-22.5)^2/22.5) + ((30-25)^2/25) + ((10-17.5)^2/17.5)

$\chi^2 = 0.277 + 0 + 3.214 + 0.277 + 0 + 3.214$

$\chi^2 = 6.982$

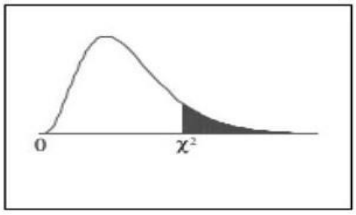
CALCULATE DEGREES OF FREEDOM

The degrees of opportunity can be determined as:
df = (total_rows - 1) * (total_cols - 1)
There are 2 lines and 3 segments in the possibility table, thus our degrees of opportunity is 2.

FIND P-VALUE

We can see the Chi Square distribution tables

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766

ISSN- 2394-5125 VOL 7, ISSUE 10, 2020

Table 4 :- chi square Distribution Table (which was used to calculate p value)

To discover the p-value utilising the Chi Square and degrees of opportunity esteems.
The degrees of opportunity esteem (2) on the left, track with its column to the nearest number to the Chi-Square worth (6.982), and then check the relating number in the principal line to get the p-esteem which is 0.025 There are numerous sites that figure the p-esteem yet we utilised the above table to discover the p value.

6. RESULTS AND DISCUSSIONS

In straightforward words if our p value is not exactly the essentialness value we dismiss the Null Hypothesis and if our p value is more prominent than the centrality value we don't dismiss it.
Since 0.025 is not exactly our essentialness value of 0.05 we dismiss the Null Hypothesis which implies that there is as relationship among Gender and the Reason of disappointment.
Subsequently these the two features are adequate to discover the student Dropout.

7. CONFLICT OF INTEREST:

The creators articulate that we have no beyond reconciliation conditions with rest of the Research.

8. REFERENCES

[1] Parneet Kaur,Manpreet Singh,Gurpreet Singh Josan, “Order and prediction based data mining algorithms to anticipate moderate students in training segment”, 3rd Int. Conf. on Recent Trends in Computing, Vol 57, 2015, pp. 500-508.

[2] J. Tang, S. Alelyani, and H. Liu, "Feature Selection for Classification: A Review," in: C. Aggarwal (ed.), Data Classification: Algorithms and Applications. CRC Press, 2014

[3]. A. Mueen, B. Zafar, and U. Manzoor, Demonstrating and Predicting Students' Academic Performance Using Data Mining Techniques,
International Journal of Modern Education and Computer Science, 8:36, 2016.

[4]. Alaa el-Halees (2009) Mining Students Data to Analyze e-Learning Behavior: A Case Study.

[5]. Bharadwaj B.K. and Pal S. "Mining Educational information/data to Analyze Students? Performance", International Journal of Advance Computer Science and Applications (IJACSA), Vol. 2, No. 6, pp. 63-69.

[6] E. Scheirer and M. Slaney, “Construction and evaluation of a powerful /robust multifeature music/speech discriminator,” in Proc. ICASSP’ 97, Apr. 1997, vol. II, pp. 1331–1334.

7] Daily News and Analysis, India, online: http://www.dnaindia.com/india/report_rte-report-carddropout-rate-in-schools-falls_1669959 ,, April 1 (2012**)** accessed on October 30, 2012

[8] Study on Feature Selection Methods/Techniques in Educational Data Mining, M. Ramaswami and R. Bhaskaran,Vol1,Dec,2009.

[9] “feature selection for high-dimensional data: a fast correlation-based filter solution”, lei yu and huan liu.

[10] “A review of feature selectionmethods/techniques in bioinformatics”, yvan saey, in aki inza and pedro larran aga. vol. 23 no. 19 2007, pages
2507–2517