Literacy Rate Analysis Dashboard

Kavita Sheoron

ABSTRACT: Education is the foremost important tool for change of the society and betterment of nation. Proficiency and level of training are fundamental pointers of the level of improvement accomplished by a general public. Spread of literacy is by and large connected with vital attributes of present day development for example, modernization, urbanization, trade and industrialization. Literacy shapes a vital contribution to generally improvement of society empowering them to understand their social, political and social condition better and react to it appropriately. Better education and literacy prompt a more noteworthy mindfulness and furthermore contributes in enhancement of economical and social conditions. Ministry of Human Resource Development (DISE) releases a data on literacy rate each year which can be exceptionally valuable in examining different elements influencing education rate of a state or an area. An all around structured dashboard that exhibits the best possible examination of the information will give a reasonable picture of proficiency in different locales of India. Data to be analyzed is handled and cleaned to draw out the most imperative and significant features. The data at that point analyzed gives the last outcome which is presented on dashboard making it easy to understand and comprehend.

KEYWORDS: Indian literacy, literacy rate, data pre-processing, data analysis, machine learning

1 INTRODUCTION

Literacy is characterized as the capability to read and compose a basic message in any language. A more expansive translation is literacy as apprehension and competence in a specific area. The key to literacy is a fundamental comprehension of composed content, capacity to comprehend someone else talking and comprehension and ability to write. Reading and writing are foundation skills. Not solely are they needed for additional study, they're conjointly crucial in helping us to know and interact with the world around us. Literacy in india is marked with an excellent amount of regional variation from one half to another. The regional differences in literacy levels within the nation has resulted from the regional diversity in various cultural,economical and social factors beside a marked distinction within the historical expertise of various regions. India's illiteracy is a prime concern that has numerous factors connected to it. Illiteracy in India is majorly involved with completely different sorts of disparities that exist within the country. There are gender disparity, income variance, state variation, caste disproportion, technological hurdles which forms the literacy rates that exist within the country. So, study and analysis of literacy data of India is needed to supply a timely and sophisticated basis for serving to planning and management of education services and to ascertain or contribute to an education system for assortment, organization and utilization of education data.

2. BACKGROUND

This report uses the Ministry of Human Resource Development (DISE) data for Literacy rate in India. Few notable sources which helped in developing the approach followed includes Literacy Rate Analysis (IJSER), Census of India: Literacy and Education Level. Analysis of raw data is not available and proper visualization at one single place is not there. Hence this dashboard extends the dataset and provide extensive functionality.

3. SOLUTION APPROACHES

Following are the means in moving toward the analysis of the data.

3.1 PREPARATION OF DATASET

Data preprocessing is a data mining procedure that includes transforming crude data into a comprehensible format. Real-world data is frequently inadequate, conflicting, and lacking in certain behaviors, and is probably going to contain numerous blunders. Data preprocessing is evidenced technique for resolving such conflicts. Data preprocessing constructs raw data for additional processing. Data experiences a progression of ventures amid preprocessing:

- Data Cleaning: Data is cleaned through processes like filling in missing values, smoothing the noisy information, or resolving the inconsistencies within the data.
- Data Integration: Data with various portrayals are assembled and clashes inside the data are settled.
- Data Reduction: This progression intends to exhibit a diminished portrayal of the data in data distribution center.

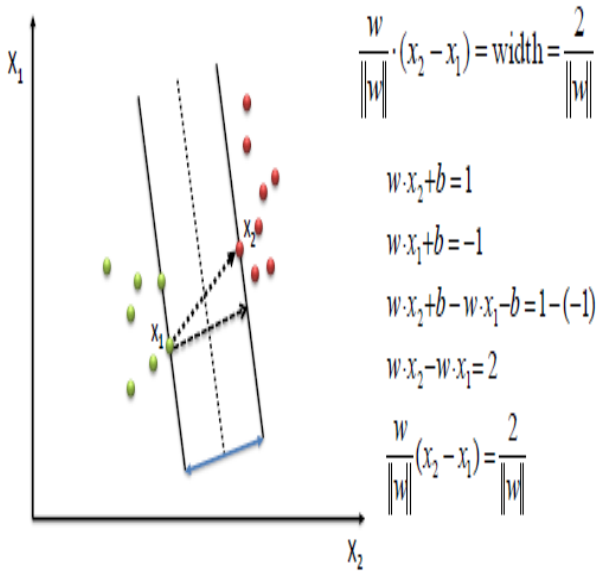
3.2 PRINCIPAL COMPONENT ANALYSIS

The fundamental thought of PCA is to decrease the dimensionality of data index comprising of numerous factors corresponded with one another, either intensely or gently, while holding the variety present in the dataset, up to the most extreme degree.The same is done by transforming the variables to a new set of variables, which are known as the principal components and are orthogonal, requested with the end goal that the maintenance of variety present in the first factors diminishes as we move down in the request.

$$C_x = covariance\ matrix\ of\ original\ data\ set\ X$$
$$C_y = covariance\ matrix\ of\ transformed\ data\ set\ Y$$
$$such\ that,$$
$$Y = PX$$
$$For\ simplicity,\ we\ discard\ the\ mean\ term\ and\ assume\ the\ data\ to\ be\ centered.\ i.e.\ X = (X - \bar{X})$$
$$So,\ C_x = \frac{1}{n}XX^T$$
$$C_y = \frac{1}{n}YY^T$$
$$= \frac{1}{n}(PX)(PX)^T$$
$$= \frac{1}{n}PXX^TP^T$$
$$= P(\frac{1}{n}XX^T)P^T$$
$$= PC_xP^T$$

3.3SUPPORT VECTOR REGRESSION

- Support Vector Machine can likewise be utilized as a regression technique, keeping up all the fundamental features that describe the algorithm (maximal edge). The Support Vector Regression (SVR) utilizes indistinguishable standards from the SVM for characterization, with just a couple of minor contrasts.. As a matter of first importance, since output is a real number it turns out to be hard to foresee the current data, which has interminable conceivable outcomes. On account of relapse, an edge of tolerance (epsilon) is set in estimate to the SVM which would have officially asked for from the issue. Be that as it may, other than this reality, there is likewise an increasingly entangled reason, the calculation is progressively muddled along these lines to be taken in thought.



3.4 K-NEAREST NEIGHBOUR (K-NN)

K nearest neighbors is a simple algorithm that is used for both classification and regression predictive problems. KNN stores all available cases and predict the numerical target based on a similarity measure (e.g., distance functions). KNN has been utilized in factual estimation and pattern recognition.A straightforward execution of KNN regression is to figure the mean of the numerical focus of the K closest neighbors.

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

3.5 LINEAR REGRESSION

Regression is a technique for modelling an objective esteem dependent on independent indicators. This strategy is for the most part utilized for estimating and discovering circumstances and logical results connection between variables. Regression techniques for the most part vary dependent on the quantity of independent variable and the sort of connection between the independent and dependent variables.Basic linear regression is a kind of regression technique where the quantity of independent factor is one and there is a straight connection between the independent(x) and dependent(y) variable.

$$h_\theta(x) = \theta^T x = \sum_{i=0}^{n} \theta_i x_i --(1)$$

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} --(2)$$

θ : *parameters need to find*
x : *input (features)*
y : *output (target)*
m : *number of samples*
α : *learning rate*

4. PROPOSED APPROACH

The following steps were decided to be used:-

- Removing the irrelevant features and combining features that can be clubbed together.
- Plotting heat map and removing highly correlated features.
- Applying model and tune the parameters to predicting result as per the requirement.

Linear regression was chosen as it provided best accuracy and fits best to the given dataset. Hyperparameter tuning fine tunes the model to give highly accurate result. Tools used for this analysis are Scikit-Learn and Numpy libraries using Python programming language.

5. ANALYSIS

This project deals with the analysis of Literacy Rate in different states of India based on 680 factors. This dataset contains information about the year 2015-16 and was published by HRD Ministry of India. We are focusing at finding top five factors and the least five factors that influence the literacy rate of given state. Also we predicted the literacy rate based on features and compared them with the available literacy rates and it was found to be accurate up to 93%. In our effort we have tried to predict the Literacy Rates of each state using reduced set of features. Heat maps were plotted to remove highly correlated features. PCA enabled the processes of feature elimination and feature extraction. Feature elimination is simply limiting final feature space to the most important ones, whereas feature extraction creates new independent variables by using a logical combination of old somewhat correlated variables. By this methodology we are able to reduce the features and dimension of dataset. The table below lists a subset of the final selected features.

	statcd	area_sqkm	tot_population	urban_population	grwoth_rate	sexratio
0	1	222236.0	12548.93	20.05	23.71	883.0
1	2	55673.0	6856.51	8.69	12.81	974.0
2	3	50362.0	27704.24	29.82	13.73	893.0
3	4	114.0	1054.69	76.66	17.10	818.0
4	5	53483.0	10116.75	21.54	19.17	963.0

	sc_population	st_population	male_literacy_rate	female_literacy_rate	\
0	7.4	11.9	78.26	58.01	
1	25.2	5.7	90.83	76.60	
2	31.9	0.0	81.48	71.34	
3	18.9	0.0	90.54	81.38	
4	18.8	2.9	88.33	70.70	

	...	boys_toilet_all	gsch_all	bsch_1	bsch_2	bsch_all	\
0	...	3940	180.0	6.0	3.0	97.0	
1	...	3728	54.0	2.0	21.0	26.0	
2	...	8766	378.0	4.0	95.0	133.0	
3	...	157	3.0	1.0	0.0	1.0	
4	...	3205	259.0	3.0	85.0	108.0	

	co_sch_all	cwsn_toilet_1	cwsn_toilet_2	cwsn_toilet_3	cwsn_toilet_4
0	3925	50	7	152	28
1	3702	115	572	103	275
2	8660	523	1000	495	1016
3	156	37	0	34	0
4	3145	104	114	21	24

[5 rows x 50 columns]

Later different models were trained and applied to find the best fitting model and then that model is used to find the features which affect the literacy rate the most. The following figure represents a glimpse of operational dashboard which shows the feature values of Kerala and list the five most significant features and also the least significant features influencing literacy in that state.

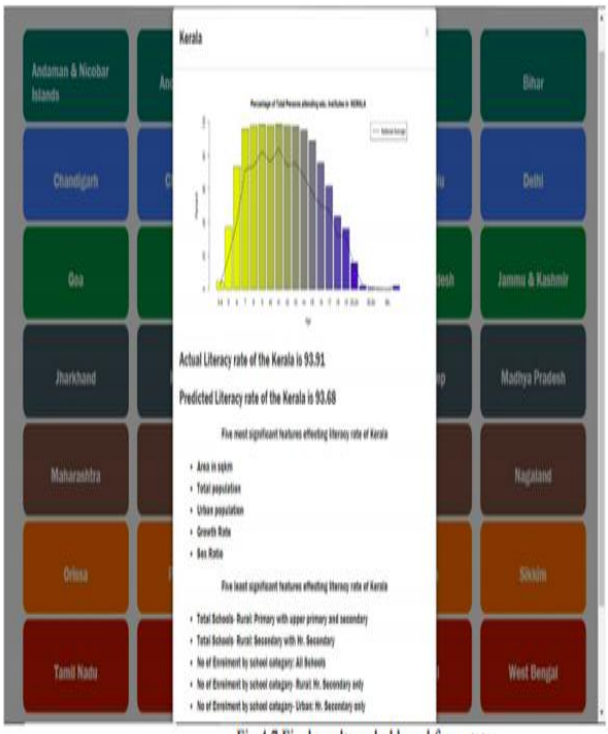


Fig 4.7 Final result on dashboard for a state

6. CONCLUSION

In India, Kerala has the highest literacy rate followed by by Lakshadweep and Mizoram whereas the states Bihar, Arunachal Pradesh and Rajasthan are bottom three. The major factors which affects the literacy rate in most of the states are :-

1. Area of state
2. Total population
3. Sex ratio
4. Female literacy rate

Least affecting features to the literacy rate are:-

1. Total number of teachers
2. Number of enrollments by category
3. School playground facility
4. Total boys and girls school

REFERENCES

[1]. Isabelle Guyon,"An Introduction to Variable and Feature Selection" in Journal of Machine Learning Research 3 (2003) 1157-1182,2013.

[2]. Tarun Verma, "Literacy Rate Analysis" in International Journal of Scientific & Engineering Research Volume 3, Issue 7, July-2012.

[3]. JinsongLeng, "A Wrapper-Based Feature Selection for Analysis of Large Data Sets" in 3rd International Conference on Computer and Electrical Engineering (ICCEE 2010), 2010.

[4]. Aparna Samudra, "Trends and Factors affecting Female Literacy-An inter-district study of Maharashtra" in International Journal of Gender and Women's Studies, June 2014.

[5]. The World Bank, "Education in India", September 20, 2011.

[6]. EemeliLepp¨aaho, "GFA: Exploratory Analysis of

Multiple Data Sources with Group Factor Analysis" in Journal of Machine Learning Research 18 (2017) 1-5, April 2017.

[7]. 2001 Census Data, "Literacy and Level of Education" by Govt. of India.

[8]. Brijesh Kumar Baradwaj, "Mining Educational Data to Analyze Students‟ Performance" in (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, June, 2011.

[9]. Deepak Talwar, Dr.Meenu, "An Analysis of Literacy Rate In Haryana" in Journal of Business Management & Social Sciences Research (JBM&SSR) Volume 3, No.7, July 2014.

[10]. Navjeet Kaur, "Literacy Rate and Gender Gap in Scheduled Castes in India" in ICFAI National College, Patiala.

[11]. PradipChouhan, "A study on literacy and educational attainment of scheduled castes population in Maldah District of West Bengal, India" in Journal of Geography and Regional Planning Vol. 6(1), pp. 19-30, February, 2013

[12]. Jason Brownlee, "How to predict classification or regression outcomes with scikit-learn models in Python" (article).