Drought Prediction System Using Data Mining

Dr. D. R. Ingle[1], Akanksha Achukola[2] , Shivani Kakade[3] , Khushali Chauhan [4]
Department of Computer Science Engineering
[1,2,3,4] Bharati Vidyapeeth College of Engineering, Navi Mumbai

Abstract- *Drought studies are important because of their influence on the society and the economy of any nation. An endeavor toward this idea would surely assist one with understanding the common procedure of drought for further work. The definitions of drought, classification of drought based on the influential factors, parameters affecting occurrence of drought, different drought assessment procedures available, studies based on statistical data have been covered. In order to improve the skill of spatial-temporal forecasts for drought and flood and identify the most important hazard factors, a data mining of precipitation in spatial and temporal has been executed in this study.*

We seek to improve the performance of the naive Bayes classifier by adding a discriminative model component to the naive Bayes. It can be referred to as modified version of the Naive Bayesian. The motive of this study is to plan and analyze the average precipitation, crop production and ground water levels for past many years and correlate them to predict the drought in forthcoming years using data mining algorithms.

Keywords- Data mining, drought disasters, precipitation, crop production, ground water level, Standardized Precipitation Index.

I. INTRODUCTION

Drought is a mind boggling climate peculiarity that happens much of the time and broadly in a given reality. In case for a ton of time this precipitation need is unnoticed, it may finish up being a calamitous occasion. In another words, dry spells are dubious and hard to anticipate.

As shown by the Indian Meteorological Department (IMD), dry season is a condition when the immovable precipitation is under 25 percent of its ordinary. Subsequently, it is basic to fathom distinctive thoughts of dry seasons before the social event people from this assignment present the profile of drought slanted regions. There are many different types of drought including: (a) Meteorological drought-Precipitation as the parameter, (b) Agricultural Drought - Soil moisture as the parameter, and (c) Hydrological-Surface/subsurface water level as the parameter. Some of the Key Indicators for Monitoring Drought are as follows: (a) Climate data (precipitation), (b) Stream flow, (c) Reservoir and lake levels, and (d) Soil moisture. Some of the drought assessment tools to be used in the system are as follows: (1) Ground Water Level Index: The level of ground water in meters, (2) Crop Production: The amount of crop produced in the region in a particular year, and (3) Standardized Precipitation Index (SPI): The Standardized Precipitation Index (SPI) is a tool which was developed primarily for defining and monitoring drought. It allows a person to determine the rarity of a drought of interest in any rainfall station with historical data at a given time scale (temporal resolution). It can also be used to determine anomalously wet event periods.

Drought results into water deficiency for some action, gathering or natural area. Dry spell is likewise connected with precipitation timing. Other climatic components are all the time connected with dry season, for example, high temperature, high wind and low relative dampness.

Droughts have pushed old human advancements to the edge of total collapse with starvations, nourishment shortage, and the loss of lives and property. Proof recommends that dry spell influenced zones are expanding all around. Moreover, numerous atmosphere expectations foresee an expansion in drought influenced areas. Drought monitoring data mining would in this manner be a champion among the most essential issues to examine. The accessibility of enormous datasets from satellite, atmosphere and natural perceptions is an open door for utilizing this reasonable critical thinking research. Before, constrained instruments were accessible for extricating and changing over these colossal datasets into significant data.

Studies reveal that a gap is obvious in having a reasonable idea for information recovery and coordination for improved drought modeling and prediction. This gap has often resulted in delayed information delivery to decision makers. Therefore, the target of this project is to create another insightful framework idea for drought data extraction and expectations. Moreover, for analyzing huge datasets available on various parameters affecting drought, the group members of this project need a good and sound technology or algorithm which will give importance to all the parameters and be able to

predict drought accurately. The way toward finding significant and important patterns/profiles and trends by shifting through data using pattern recognition technologies such as neural networks ,machine learning and genetic algorithms is known as Data mining . Nowadays, the technology can provide a great deal of information about agricultural practices, which could then be analyzed in order to find important information. Thus, this paper will be useful to find the pattern and relationship between several parameters affecting drought.

Patterns that might be utilized to anticipate drought, substantial authentic informational collections are fundamental to recognize connections between various climatic parameters and to recognize. It is thus critical to have an efficient way to extract information from large databases and to deliver relevant and effective information for drought risk management. Data mining is one of the newly developed techniques for these purposes. Along these lines, drought monitoring information mining would be a champion among the most imperative issues to evaluate. In the present investigation information mining is utilized to distinguish complex connections including air conditions that conceivably cause dry seasons over chosen districts of Maharashtra, India.

II. PROPOSED ALGORITHM

Traditional Naïve Bayesian algorithm

It is a characterization method dependent on Bayes' Theorem with a suspicion of autonomy among indicators. A Naive Bayes classifier accept that the nearness of a specific component in a class is irrelevant to the nearness of some other element. Innocent Bayes demonstrate is anything but difficult to fabricate and especially valuable for huge informational indexes. Alongside straightforwardness, Naive Bayes is known to beat even exceptionally modern grouping techniques.

Bayes hypothesis gives a method for computing back likelihood P(c|x) from P(c), P(x) and P(x|c). Take a gander at the condition beneath: $P(c \mid x) = P(x \mid c) * P(c)/P(x)$

Where, $P(c \mid x) = P(x1 \mid c) * P(x2 \mid c) * ... * P(xn \mid c) * P(c)$

Above,

P(c|x) is the back likelihood of class (c, target) given indicator (x, traits).

P(c) is the earlier likelihood of class.

P(x|c) is the probability which is the likelihood of indicator given class.

P(x) is the earlier likelihood of indicator.

How about we comprehend it utilizing a model. Beneath there is a preparation informational index of climate and comparing target variable 'Play' (proposing potential outcomes of playing). Presently, one needs to group whether players will play or not founded on climate condition. We should pursue the underneath ventures to perform it.

Stage 1: Convert the informational index into a recurrence table.

Stage 2: Create Likelihood table by finding the probabilities like Overcast likelihood = 0.29 and likelihood of playing is 0.64.

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Table No 1:Given Table

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Table No 2:Frequency Table

Likelihood Table				
Weather	No	Yes		
Overcast		4	= 4/14	0.29
Rainy	3	2	= 5/14	0.36
Sunny	2	3	= 5/14	0.36
All	5	9	= 4/14	0.29
	= 5/14	= 9/11		
	0.36	0.64		

Table No 2:Likelihood Table

Step 3: Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

Problem: Players will play if weather is sunny. Is this statement is correct?
One can solve it using above discussed method of posterior probability.

P(Yes | Sunny) = P(Sunny | Yes) * P(Yes) / P (Sunny)
Here, P (Sunny |Yes) = 3/9 = 0.33, P(Sunny) = 5/14 = 0.36, P(Yes)= 9/14 = 0.64

Now, P (Yes | Sunny) = 0.33 * 0.64 / 0.36 = 0.60, which has higher probability.

Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes.

Drawback of Naïve-Bayesian Algorithm:

One drawback of Naïve-Bayesian algorithm is that it doesn't take into consideration the dependency among the parameters. Moreover, the original traditional form of Naïve-Bayesian algorithm does not allow us to provide weighting factors to the parameters. Many a times the parameters have to be given a weighting factor to give more accurate predictions. This is not possible in the traditional model of Naïve-Bayesian algorithm.

So to overcome this drawback, modified version of Naïve-Bayesian algorithms are used.

Modified naïve Bayesian algorithm

The Naïve-Bayesian classifier considers all the parameters involved to produce the output.

In the traditional Naïve-Bayesian algorithm, all the parameters are given equal importance. Priorities cannot be assigned to the attributes. In our case, different parameters have different priorities i.e. the weighting factor. Rainfall affects drought more as compared to crop production or the ground water level in the region. So, what we are doing is assigning the weighting factors to each parameter. Using the traditional Naïve-Bayesian algorithm, accuracy of prediction was getting compromised.

Consider the case where rainfall was high and the crop production and ground water level were low for a given year for a district. Using the traditional Naïve-Bayes method, we were getting wrong prediction. Since the rainfall is very high for that year, drought should not occur. But as the traditional method is not using weighting factors and giving equal importance to all the parameters we were getting the result as drought for the following year. So, to overcome this drawback of traditional Naïve-Bayes method, we modified it slightly to get more accuracy. We have assigned the weighting factors for all the attributes affecting the occurrence of the drought in the dataset. Rainfall, as it affects the occurrence of drought more as compared to crop production and ground water is given more weighting factor.

On Modified-naïve Bayesian method, we are giving more importance to the annual rainfall as drought is affected by rainfall in a major way. Crop production and ground water level doesn't affect the occurrence of drought as much as the annual rainfall. The weighting factors are given to the parameters in the following way:

Alpha= Sum of the parameters.
Weighting factor for Total Rainfall =2/Alpha.
Weighting factor for Crop Production =1/Alpha.
Weighting factor for Groundwater level=1/Alpha.

III. METHODOLOGY

So as to build up a clever framework idea for dry spell data extraction and expectations from parameters like yearly precipitation, crop generation and ground water level, a canny model is built up, a framework with GUI was planned, and forecast for dry spell was finished. The following four processes are performed to get the result:

1. Data-Set Selection.
2. Cleaning and Pre-processing
3. Classification.
4. Prediction.

Detailed explanation of the above processes is given below:

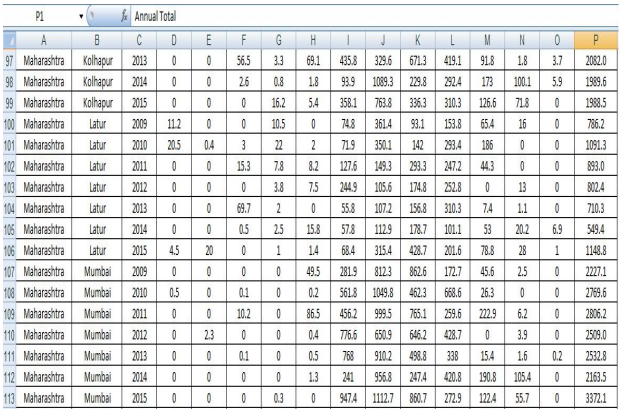
Data-Set Selection

Input Selection is the first process. In this, first we have to browse and select the input for the process. Input of the process is dataset. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows. In our process, Dataset is collected from the website or collected manually.
We are taking the following parameters for predicting drought:

- Annual Rainfall
- Crop Production

- Ground Water Level

The above parameters data is obtained for all the districts of the state of Maharashtra. The rainfall is obtained in the form of month-wise data. This month-wise data is converted into annual rainfall by performing the addition operation so that preprocessing can be done on the data.



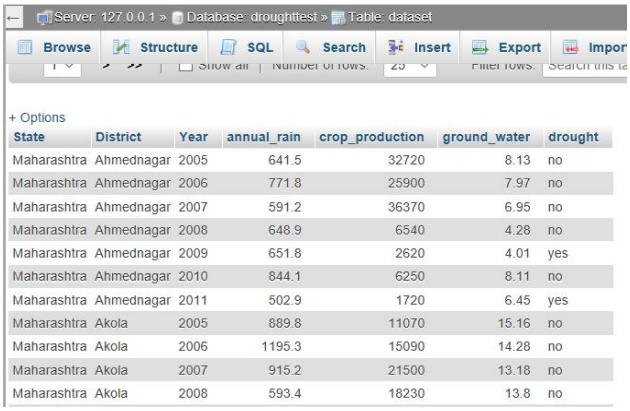
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
97	Maharashtra	Kolhapur	2013	0	0	56.5	3.3	69.1	435.8	329.6	671.3	419.1	91.8	1.8	3.7	2082.0
98	Maharashtra	Kolhapur	2014	0	0	2.6	0.8	1.8	93.9	1089.3	229.8	292.4	173	100.1	5.9	1989.6
99	Maharashtra	Kolhapur	2015	0	0	0	16.2	5.4	358.1	763.8	336.3	310.3	126.6	71.8	0	1988.5
100	Maharashtra	Latur	2009	11.2	0	0	10.5	0	74.8	361.4	93.1	153.8	65.4	16	0	786.2
101	Maharashtra	Latur	2010	20.5	0.4	3	22	2	71.9	350.1	142	293.4	186	0	0	1091.3
102	Maharashtra	Latur	2011	0	0	15.3	7.8	8.2	127.6	149.3	293.3	247.2	44.3	0	0	893.0
103	Maharashtra	Latur	2012	0	0	0	3.8	7.5	244.9	105.6	174.8	252.8	0	13	0	802.4
104	Maharashtra	Latur	2013	0	0	69.7	2	0	55.8	107.2	156.8	310.3	7.4	1.1	0	710.3
105	Maharashtra	Latur	2014	0	0	0.5	2.5	15.8	57.8	112.9	178.7	101.1	53	20.2	6.9	549.4
106	Maharashtra	Latur	2015	4.5	20	0	1	1.4	68.4	315.4	428.7	201.6	78.8	28	1	1148.8
107	Maharashtra	Mumbai	2009	0	0	0	0	49.5	281.9	812.3	862.6	172.7	45.6	2.5	0	2227.1
108	Maharashtra	Mumbai	2010	0.5	0	0.1	0	0.2	561.8	1049.8	462.3	668.6	26.3	0	0	2769.6
109	Maharashtra	Mumbai	2011	0	0	10.2	0	86.5	456.2	999.5	765.1	259.6	222.9	6.2	0	2806.2
110	Maharashtra	Mumbai	2012	0	2.3	0	0	0.4	776.6	650.9	646.2	428.7	0	3.9	0	2509.0
111	Maharashtra	Mumbai	2013	0	0	0.1	0	0.5	768	910.2	498.8	338	15.4	1.6	0.2	2532.8
112	Maharashtra	Mumbai	2014	0	0	0	0	1.3	241	956.8	247.4	420.8	190.8	105.4	0	2163.5
113	Maharashtra	Mumbai	2015	0	0	0	0.3	0	947.4	1112.7	860.7	272.9	122.4	55.7	0	3372.1

Fig 1:Dataset collected

Cleaning and Preprocessing:

Data cleaning is a procedure used to decide mistaken, fragmented or irrational data and after that improve the quality through correcting of detected errors and omissions. Ordinarily Dataset preprocessing is the strategy for cleaning the dataset. Usually data is deficient of: attribute with values, attribute with values of interest, or containing just data with mistakes or exceptions and containing inconsistencies in codes or names. In this we will remove this sort of happening in the dataset. Dispensing with the undesirable esteem or images or characters in the dataset. Repetitive information from the tables is expelled and we get a precise data. We have the drought dataset of all the districts in Maharashtra consisting of the total rainfall, ground water and crop production as the parameters and the class drought which is classified as yes or no. The CSV dataset consists of continuous values which has to be converted into discrete values. The CSV file is loaded into the database first.

This CSV file consists of the rainfall of all the districts month-wise. We are calculating the total rainfall for each district. The total rainfall is in continuous manner. This form of dataset cannot be used in Modified Naïve-Bayesian. So, it is converted into labels namely Low, Medium and High using analysis techniques.



Server: 127.0.0.1 » Database: droughttest » Table: dataset

Browse | Structure | SQL | Search | Insert | Export | Impor

+ Options

State	District	Year	annual_rain	crop_production	ground_water	drought
Maharashtra	Ahmednagar	2005	641.5	32720	8.13	no
Maharashtra	Ahmednagar	2006	771.8	25900	7.97	no
Maharashtra	Ahmednagar	2007	591.2	36370	6.95	no
Maharashtra	Ahmednagar	2008	648.9	6540	4.28	no
Maharashtra	Ahmednagar	2009	651.8	2620	4.01	yes
Maharashtra	Ahmednagar	2010	844.1	6250	8.11	no
Maharashtra	Ahmednagar	2011	502.9	1720	6.45	yes
Maharashtra	Akola	2005	889.8	11070	15.16	no
Maharashtra	Akola	2006	1195.3	15090	14.28	no
Maharashtra	Akola	2007	915.2	21500	13.18	no
Maharashtra	Akola	2008	593.4	18230	13.8	no

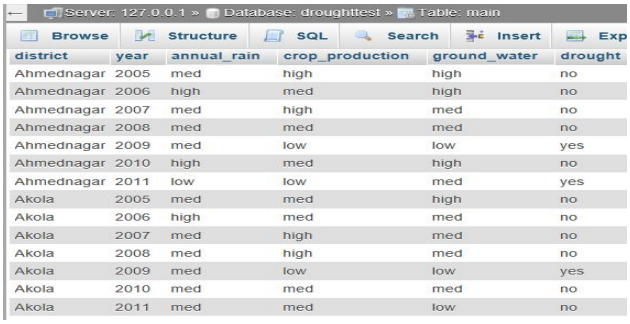
Fig. 2: Prepocessing

Classification

Classification is a way of categorizing the data (records) for an attribute. The choice of classification system is critical to information displayed by a map. We are going to classify the dataset based on the different parameters such as precipitation index, crop production and ground water level in the districts. We will get various classified tables based on the assessment tool. This will be passed as input to calculate the probability of drought in the prediction. The CSV file is converted into a table in a MySQL database. The continuous values have to be classified into classes depending on the needs.

We are considering three classes namely Low, Medium and High. So the total rainfall, crop production and ground water level is converted into classes as Low, Medium and High.

This classified data is passed and Naïve-Bayes is applied on this dataset.



Server: 127.0.0.1 » Database: droughttest » Table: main

Browse | Structure | SQL | Search | Insert | Exp

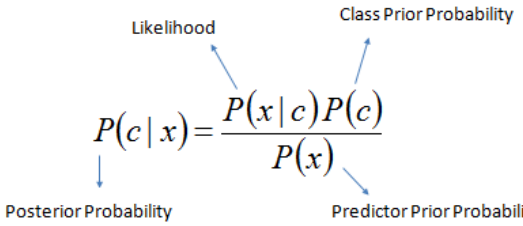
district	year	annual_rain	crop_production	ground_water	drought
Ahmednagar	2005	med	high	high	no
Ahmednagar	2006	high	med	high	no
Ahmednagar	2007	med	high	med	no
Ahmednagar	2008	med	med	med	no
Ahmednagar	2009	med	low	low	yes
Ahmednagar	2010	high	med	high	no
Ahmednagar	2011	low	low	med	yes
Akola	2005	med	med	high	no
Akola	2006	high	med	med	no
Akola	2007	med	high	med	no
Akola	2008	med	high	med	no
Akola	2009	med	low	low	yes
Akola	2010	med	med	med	no
Akola	2011	med	med	low	no

Fig 3: Classification

Prediction

A Bayes classifier is a basic probabilistic classifier dependent on applying Bayes' hypothesis (from Bayesian insights) with solid (innocent) autonomy suppositions. A

progressively expressive term for the fundamental likelihood model would be "autonomous element demonstrate". Contingent upon the exact idea of the likelihood show, innocent Bayes classifiers can be prepared in all respects effectively in a managed getting the hang of setting. In our execution procedure, we actualize the dry season look into by utilizing the directed arrangement calculation called adjusted Naïve Bayes Classification calculation. In numerous reasonable applications, parameter estimation for gullible Bayes models utilizes the strategy for greatest probability. Hence it is reasonable to add some type of "weighting" factor to take this into account.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Where
P(C|X) = probability of the class given the attribute
P(X|C) = probability of the attribute given the class.
P(C)=Probability of the class.
P(X)=Probability of the attribute irrespective of the class.
The classified dataset is passed to the Modified Naïve-Bayesian classifier.

IV. EXPECTED RESULT

After passing the classified dataset to the modified Naïve-Bayes class, the prediction of drought takes place in the following manner:

- P (drought |yes)>P (drought |no)
  Then Drought will occur in the next year.

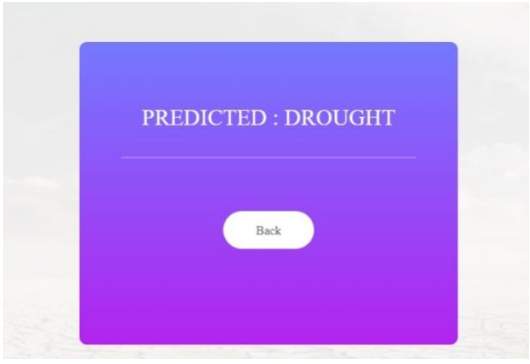


Fig 1: Drought predicted

- P (drought |yes) < P (drought |no)

Then Drought will not occur in the next year.

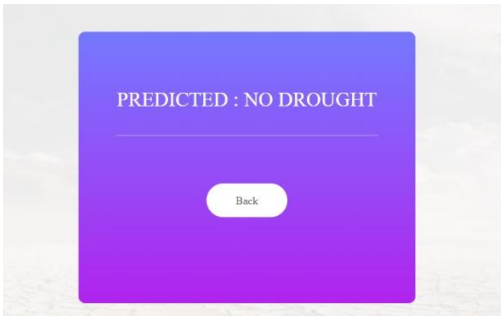


Fig 2: No drought predicted

With this small tweaking and modification in the traditional method, we could achieve greater accuracy.

V. CONCLUSION

In order to develop an efficient and smart system for drought information extraction and predictions from parameters like annual rainfall, crop production and ground water level, an intelligent model is developed, a system with GUI was designed, and prediction for drought was done. The developed concepts may serve as a starting point for future full system developments. Because drought is related to different factors, its modeling and prediction were found to be challenging in the past.

In line with this, our research has resulted in the following contributions:

1. The key attributes that characterize drought are identified and used to model drought.
2. The concept of the Drought Object extraction from all the classified attributes.
3. The basic concept of prediction of drought is clearly illustrated and evaluated. The output of this new concept is based on the freely available attributes.

The Modified Naive Bayes algorithm takes into consideration the importance of the attributes which help in predicting the drought correctly as compared to the traditional Bayesian algorithm and in improving the exactness of the system. More weightage (weighing factor) being given to the rainfall attribute, since the values of this parameter is closely related to drought's outcome compared to crop production and ground water levels. In this way, the members of this project group will help to increase the accuracy of the system and predict drought correctly.

ACKNOWLEDGMENT

We would like to express our gratitude to our HOD as well as our Guide Prof. Dr. D.R.INGLE for his invaluable support, encouragement, supervision and useful suggestions throughout this project work. His ethical help and nonstop direction empowered us to finish our work effectively and standard proposals made our work simple and capable.

We are grateful for the cooperation and constant encouragement from our coordinator Prof. Rahul Patil for advising and helping throughout the project.

Having endured the experience, there were many who helped us in our project and we very much like to thank everyone.

REFERENCES

[1] Xiaotian Gu and Ning Li, "Study of droughts and floods predicting system based on Spatial-temporal Data Mining"-2012 IEEE,6th International Conference on New Trends in Information Science, Service Science and Data Mining.
[2] K. Sriram and K. Suresh, "Machine Learning Perspective for Predicting Agricultural Droughts Using Naive Bayes Algorithm" Department of Computer Science and Engineering, KCG College of Technology, Chennai, India.
[3] Can be accessed at: https://www.idosi.org/mejsr/mejsr24(IIECS)16/28.pdf
[4] Yirong Shen and Jing Jiang, "Improving the Performance of Naive Bayesian Algorithm", CS224N Spring 2003
[5] Getachew Berhan, Shawndra Hill, Tsegaye Tadesse, and Solomon Atnaf, "Drought Prediction System for Improved Climate Change Mitigation".
[6] T. Tadesse, M. Haile, G. Senay, D. Brian, W. Cody, and L. Knutson, "The need for integration of drought monitoring tools for proactive food security management in sub-Saharan Africa," Nat. Res. Forum, vol. 32, no. 4, pp. 265–279, 2008.
[7] R.KALPANA Dr.S.ARUMUGAM, "A Significant Review of Different Drought Indices for Predicting Agricultural Droughts", Nandha Engineering College Nandha Educational Institutions.
[8] Scikit learn for Datasets Available At: http://scikit-learn.org
[9] Google Translate Available At: https://translate.google.com/#auto/en/A