Automatic Ontology matching using Heterogeneous evidences

[1]Purvi H. Bhensdadia, [2]Priyanka Raval

[1]ME Scholar of Computer Engineering Department, B.H. Gardi College of Engineering and Technology, Gujarat, India

[2]Associate Professor of Computer Engineering Department, B.H. Gardi College of Engineering and Technology, Gujarat, India

Abstract : Machine learning is a technology design to build intelligent systems. These systems also have the ability to learn from past experience and analyses historical data. It provides results according to its experience. Semantic Web will improve more relationship, trust and question. Semantics opens up the web of data to Artificial Intelligence to do thinking for us. Semantics uses ontology that enlists all the properties and inter-relationship of the entities in a particular domain. These ontologies are required to be matched with each other to provide semantic interoperability and then how semantic reasoners can be used for knowledge representation and reasoning. Matching ontologies based on string similarity solves the interoperability issues and thus helps ontology to communicate with each other. The main objective of this research work is to predict correct ontology matching using classification algorithm such as Naïve Bayes. To improve performance by giving multiple evidences by classifier.

Index Terms **– Naïve bayes classifier, Ontology Matching, Ontology, String similarity**

I. INTRODUCTION

Semantics is the study of meaning (within some context). It is a "system that enables machine to understand and respond to complex human requests based on their meaning". Semantic technology makes everything connected to anything and helps building smart data through intricate models for representing information.

Ontology is a formal shared conceptualization. [13]

Ontology consists of four aspects: [13]

1. Conceptualization: refers to the models which are conceptualized from the abstract phenomena in the objective world, and they are free from their specific environmental Conditions
2. Explicit: refers to the fact that both the type of the concept and constraints of the use of them have been explicit
3. Formal: refers that it can be managed by computer, and not entirely natural language expression
4. Share: refers that Ontology resembles the common recognition of knowledge, reflecting the commonly recognized set of concepts in the related areas, not unique to an individual, but accepted by all.

"The matching process can be seen as a function of which, from a pair of ontologies to match OA and OB, an input alignment A1, a set of parameters p and a set of oracles and resources r, returns an alignment A2 between these ontologies."

1.2 Ontology Matching

Measuring correspondence between words, sentence, paragraph and documents is vital segment in various errands such as information retrieval, document clustering, automatic essay scoring, machine translation etc. [3]. Overlap between domain ontologies can be avoided based on string matching similar labels and ontological structure & relations to find overlapping concepts. This overlap needs to be eliminated. Semantic heterogeneity handles variation in interpretation of meaning of entities and ambiguity among them. Ontology is a type of vocabulary that describes the particular domain and specifies meaning of terms used in that vocabulary. Different applications are not interoperable as they demand ontology belonging to different domains. Sematic heterogeneity is overcome in two steps: by matching two entities that determine a set of correspondence and interpret alignment according to the need of applications. Ontology matching finds correspondence between semantically related ontologies and hence is a solution to the problem. Set of correspondence are used in various applications like merging two ontologies, query answering and data translation. Knowledge and data can interoperate by matching two ontologies [2]

Among the several number of systems that have been developed in these recent years, selection is based on those which have repeatedly participated to the Ontology Alignment Evaluation Initiative (OAEI) tracks which has basis for comparisons and have corresponding factual publications. Among different matching systems that are currently used for ontology matching, AgreementMakerLight (AML) ranked first in terms of performance. [9] It takes as input ontologies and output and alignment which is set of correspondence between semantically related concepts. concepts used in matching ontology using AML starts with finding cementing relationship between all the classes entities of source and target ontology. This relationship can be categorized as equivalence, more or less general than. AML restrict these relationships to equivalence relationship which are called as mapping alignment between two input ontologies. Different matching algorithms are used which are called as matchers which assign a value in terms of percentage that reflects the semantic similarity between two entities of input ontologies. [15].

1.2.1 String Similarity Algorithms:

Word translation is fundamental part of text similarity. It is primarily used for sentence, paragraph and document similarity. Similarity between words is mainly of two types: lexical and semantic. If words have similar character sequence, they are lexically similar and if they have same meaning or used in same context then they are semantically similar. Different string-based algorithms are used for lexical similarity and corpus-based and knowledge based algorithms are used for semantic similarity. "A string metric is a metric that measures similarity or dissimilarity (distance) between two text strings for approximate string matching or comparison". "Corpus based similarity is a semantic similarity that determines the similarity between words according to information gained from large corpora" [20].

II. LITERATURE REVIEW

Ontology Alignment using Combined Similarity Method and Matching Method [8]

Ontology Alignment showed that between Ontology BibTex and IFLA has a relationship. This is evidenced by several entities of both Ontology are matching. And for writing bibliographies using Bibtex format or IFLA format, no need full change the contents of bibliography, because of the experimental results in this study show that there are similarities content between Bibtex format and IFLA format.

Steps of algorithm is given as follow:

1. Insert the two Ontology as input
2. After that do the similarity of the entities contained in the Ontology 1 and Ontology 2.
3. Having obtained the value of similarity between the entities, then selected which have a value greater than the threshold (In this research used a threshold = 0.5)
4. The results of similarity that has been sorted, and then do the matching process using string matching (one of them is Brute Force String Matching)
5. If the value of the result of matching is equal to 1, then the entity Ontology 1 and Ontology 2 have a relationship. Meanwhile, if the result of matching indicates a value equal to 0 then the entity of Ontology 1 and Ontology 2 do not have a relationship.

A Comparative Evaluation of String Similarity Metrics for Ontology Alignment [4]

In this research they perform mainly 8 primary category of entity names' variations classified:

Like, Syntactic similar but different naming conventions, Synonym, Word omissions, Abbreviations, Misspelling, Irregular.

In this paper giving the different string matching techniques like Character based string matching like Isub, Jaro-winkler, Levenstein distance and Token based string matching like jaccard similarity and also give the each string algorithm computational time. And also include that hybrid methods take more time compare to base method and not as much as improved.

The Comparative Analysis of Smith-Waterman Algorithm with Jaro-Winkler Algorithm for the Detection of Duplicate Health Related Records [7]

In this paper we learn brief Methodology about the Jaro-Winkler algorithm and Smith-Waterman algorithm for string matching. Main purpose of this paper is how accurate these algorithm are detecting duplicate word in large data set.

The Jaro-Winkler distance computes distance between two strings (string1 and string2), thereafter the characters in each string is matched and then transposed to find the Jaro distance. Thus, the Jaro distance allows only transposition of characters. While the Winkler modification is a formulation to compute distance and give a higher score if the prefix of starting words are similar in both strings.

An application of Levenshtein algorithm in vocabulary learning [10]

The paper presented new application of levenshtein algorithm. Levenshtein distance is a measure for the similarity of two strings. But new application of Levenshtein distance algorithm, for the selection of choices in a vocabulary quiz.

Levenshtein algorithm: $Sim(x, y) = 1 - \frac{LevenshteinDist(x, y)}{Max(|x|, |y|)}$

Chronic Kidney Disease Analysis Using Data Mining Classification Techniques [23]

This paper is to predict Chronic Kidney Disease (CKD) using classification techniques like Naive Bayes and Artificial Neural Network (ANN). The experimental results implemented in Rapid miner tool show that Naive Bayes produce more accurate results than Artificial Neural Network.

Table 1 Comparative Study of Matcher Algorithm [4][7][10]

Algorithm	Formula
I Sub	$sim(s1,s2) = comm(s1,s2) - diff(s1,s2) + wrinkler(s1,s2)$ $Comm(s1,s2) = 2*P\ i\ |maxComSubstring||s1|+|s2||$ $Diff(s1,s2) = uLens1*uLens2/p + (1-p) *$ $(uLens1 + uLens2 - uLens1 * uLens2)$
Jaro - Wrinkler	$dj = \frac{1}{3} * \left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-(t/2)}{m} \right)$ $djw = dj + \frac{Max(p)}{10} * (1- dj)$
Levenstein	$Sim(x, y) = 1 - \frac{LevenshteinDist(x, y)}{Max(|x|, |y|)}$
Q - Gram	$Sim(set1, set2) = \frac{2 * |set1 \cap set2|}{|set1| + |set2|}$

III. RESEARCH METHODOLOGY

3.2 Proposed approach

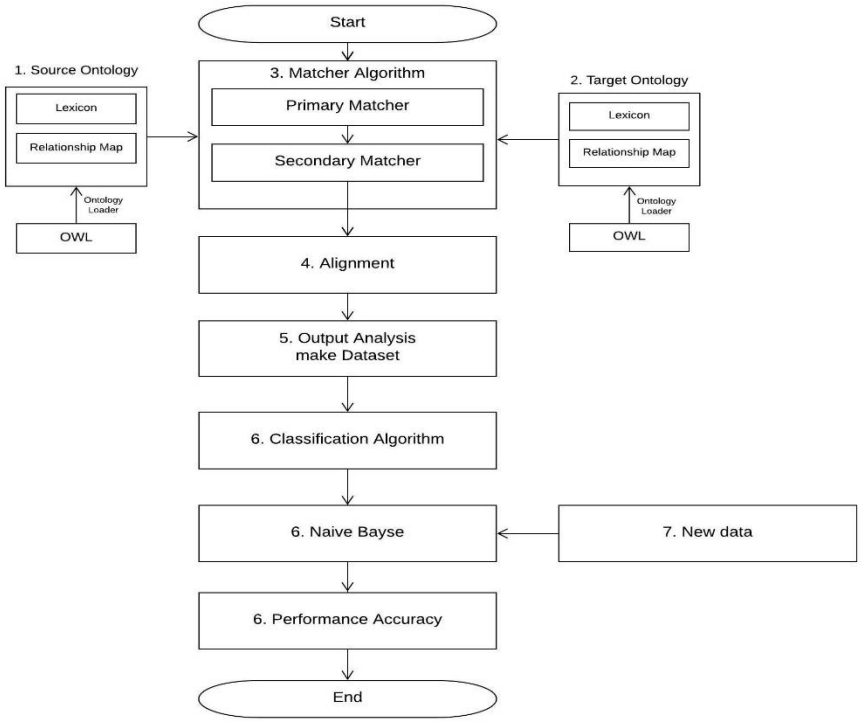

Figure 3.2.1 Proposed Work

3.2.1 Input Ontologies:
First the most important module is Ontology Loading module. Here the two input ontologies (Source Ontology and Target Ontology) are giving input in AgreementMakerLight tool. After that the Ontology Loader differentiate the Lexicon and Relationship Map in the input Ontologies.

3.2.2 Matcher Algorithm:
Matchers are the algorithms which compare two ontologies and return an Alignment between them. AML divides Matcher into two categories:
1. Primary Matcher
2. Secondary Matcher
In the Primary Matchers there will be 3 Matcher algorithms that is the Lexical Matcher, the Mediating Matcher and the Word Matcher and in the Secondary Matcher there will be Parametric String Matcher.

3.2.3 Alignment:
Alignment is a data structured which is used by the ontology matching module to store mapping between the input ontologies. Alignment was used only to store the final output of a matching algorithm or combination of algorithms.

3.2.4 Classification
It maps data into predefined groups or classes. In classification the classes are indomitable before examining the data thus it is often mentioned as supervised learning. Classification is the process which classifies the collection of objects, data or ideas into groups, the members of which have one or more characteristic in common. In this research work Naïve Bayes to classify different stages of Ontology Matching from the dataset. The main objective of this research work is to predict correct ontology matching using classification algorithm such as Naïve Bayes. To improve performance by giving multiple evidences by classifier.

3.2.5 Naïve Bayes
A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". This restricted individuality assumption infrequently clutches true in real world applications, hence the characterization as Naive yet the algorithm inclines to perform well and learn rapidly in various supervised classification problems [6]. An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

3.2.5.1 Naïve Bayes Algorithm
Step 1: Load the Conference domain dataset.
Step 2: Store the future matrix (X) and response vector (Y)
X = conference.data Y = conference.target
Step 3: Splitting X and Y into training and testing sets
Step 4: Training the model on training set
Step 5: Making prediction on the testing set
Step 6: Comparing actual Output values with Predicted values

3.2 Implementation

The main objective of this research is to build Intelligent Ontology matching Prediction System that gives accurate prediction using historical database. To develop this system, Attributes such as Label 1, Label 2, I sub, Levenstein, Jaro_Winkler, Q-gram, and Actual Output are used. Classification techniques Naive Bayes are used. For mapping two ontologies AgreementMakerLight is used and for Software Development Orange 3.20 is used. We will work on these tools for future work and try to make these tools work in better way and make increase accuracy.

3.2.1 DATA SOURCE

The data set consists of 3 types of attributes: Input, Key & Predictable attribute which are listed below. The dataset consists of total 560 records in Ontology matching database. The total records are divided into two data sets one is used for training and second testing.

Figure.3.2.1. Data set of conference Domain

Output.csv - Excel (Product Activation Failed)

	Sr. No.	Label 1	Label 2	matching method output Isub	evensteiro-	Winkl	Q - gram	actual output Output
	1	Conference	Conference_volume	0	0	0	0	1
	2	Preference	Review_preference	0	0	0	0	1
	3	Author	Regular_author	0	0	0	0	1
	4	Person	Person	0.99	0.99	0.99	0.99	1
	5	email	has_an_email	0	0	0	0	1
	6	Co-author	Contribution_co-author	0.7556	0.7556	0.7556	0.7556	1
	7	PaperAbstract	Abstract	0	0	0	0	1
	8	Document	Conference_document	0.6887	0.6887	0	0.6887	1
	9	Review	Review	0.9888	0.9888	0.9888	0.9888	1
	10	Conference	Conference	0.9801	0.9801	0.9801	0.9801	1
	11	ProgramCommittee	Program_committee	0.9801	0.9801	0.9801	0.9801	1
	12	Chairman	Chair	0.6755	0	0.7351	0	1
	13	SubjectArea	Topic	0	0	0	0	1
	14	assignedByReviewer	invited_by	0	0	0	0	1
	15	assignExternalReviewer	invites_co-reviewers	0	0	0	0	1
	16	user	track workshop chair	0.99	0.99	0.99	0.99	0
	17	paper	paper	0.9824	0.9824	0.9824	0.9824	1
	18	reviewer	reviewer	0.9801	0.9801	0.9801	0.9801	1
	19	program committee member	committee member	0.6801	0.6801	0.6801	0.6801	0
	20	external reviewer	conference participant	0	0.99	0	0.99	0
	21	chairman	conference applicant	0	0.99	0	0.99	0
	22	conference member	conference contributor	0	0.99	0	0.99	0
	23	author	active conference participant	0	0	0.7638	0	0
	24	conference chair	conference part	0	0	0.7505	0	0
	25	conference member	conference fees	0	0	0.7455	0	0

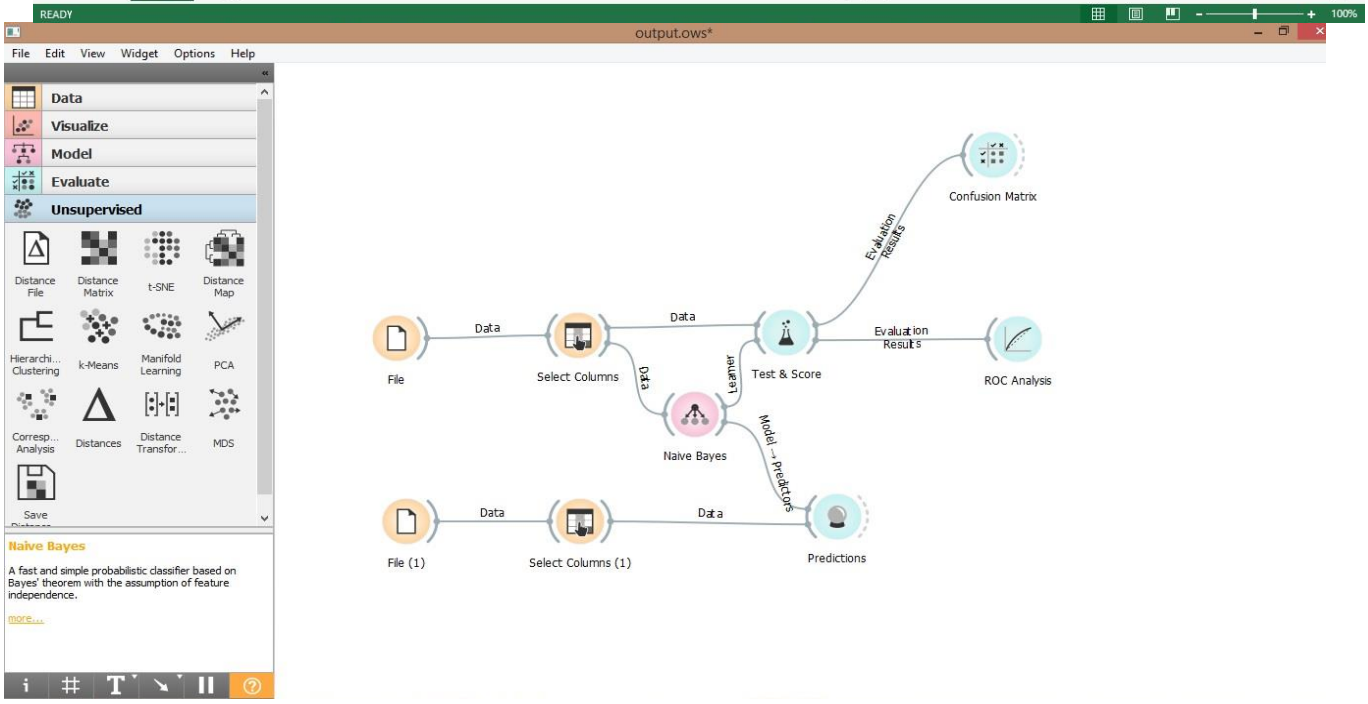

Figure.3.2.2. Orange tool (Workflow)

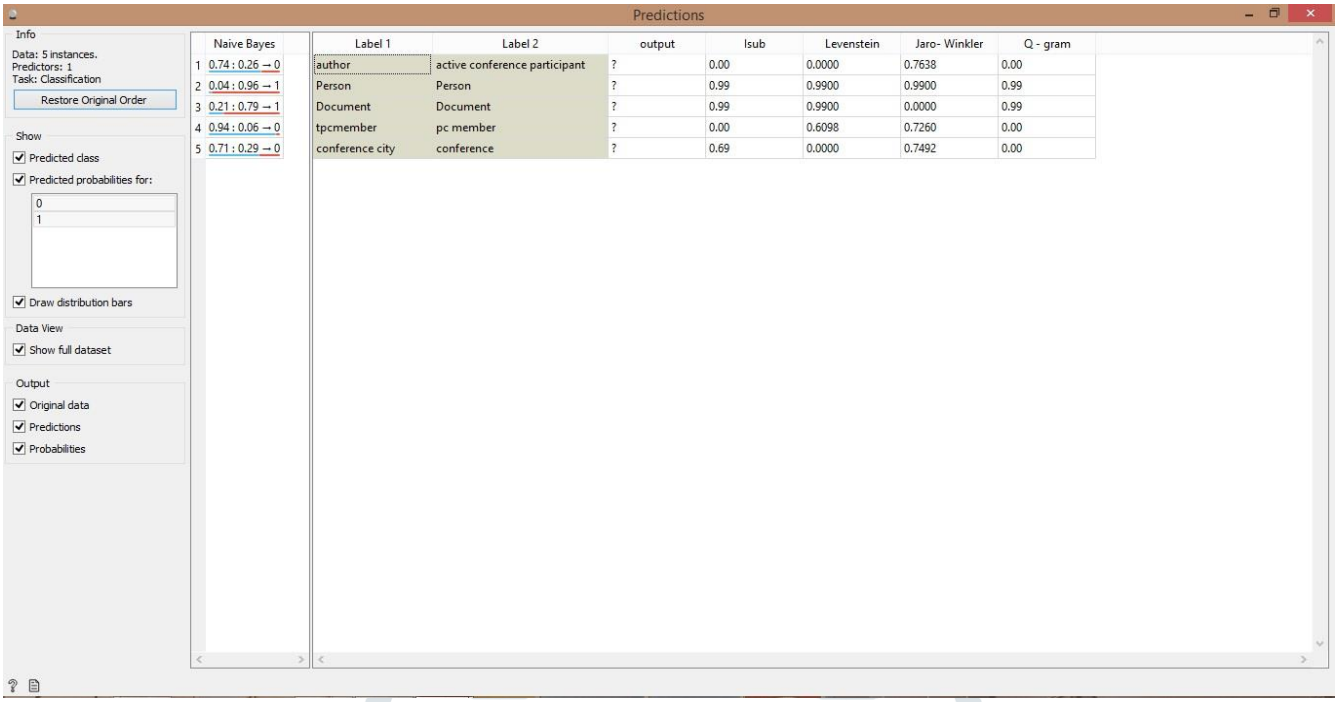

Figure.5.2.3. New data Prediction

IV. FUTURE WORK

I will use different classifier in future like decision tree and neural network and I will generated graphs by comparing accuracy provided by classifier and I will find which classifier is the best for ontology matching.

REFERENCES

[1] Sheena Angra, Sachin Ahuja : "Machine Learning and its Applications" IEEE-2017
[2] Shvaiko, P., & Euzenat, J. "Ontology matching: state of the art and future challenges" IEEE -2013
[3] Gomaa, W. H., & Fahmy, A. "A survey of text similarity approaches" IJCA-2013
[4] Yufei Sun, Liangli Ma, Shuang Wang : "A Comparative Evaluation of String Similarity Metrics for Ontology Alignment" jics-2015
[5] Kazar Okba, Saouli Hamza, Benfanatki Hind, Alaoui Amira, Bourakach Samir : "Semantic natural language translation based on ontologies combination" IEEE-2017
[6] Samira Babalou, Mohammad Javad Kargar, Seyyed Hashem Davarpanah : "Large-Scale Ontology Matching: a Review of the Literature" IEEE-2016
[7] Israel Edem Agbehadji, Hongji Yang, Simon Fong, Richard Millham : "The Comparative Analysis of Smith-Waterman Algorithm with Jaro-Winkler Algorithm for the Detection of Duplicate Health Related Records" IEEE - 2018
[8] Didih Rizki Chandranegara, Riyanarto Sarno: "Ontology Alignment using Combined Similarity Method and Matching Method" IEEE-2016
[9] Cheatham, M., Dragisic, Z., Euzenat, J., Faria, D., Ferrara, A., Flouris, G. & Lambrix, P. "Results of the ontology alignment evaluation initiative 2015". In 10th ISWC workshop on ontology matching (OM) (pp. 60-115). No commercial editor.-2015
[10] Alexandru Ene, Andrei Ene : "An application of Levenshtein algorithm in vocabulary learning", IEEE – 2017
[11]Teodor Petrican, Ioan Salomie, Ionut Anghel, Tudor Cioara, Marsal Antal, Ciprian Stan : "Ontology-Based Skill Matching Algorithms" IEEE-2017
[12]Xingsi Xue, Jeng-Shyang Pan : "An Overview on Evolutionary Algorithm based Ontology Matching" ISSN-2018
[13] Xiyin Liu, Lijun Cao, Wei Dai : "OVERVIEW ON ONTOLOGY MAPPING AND APPROACH" IEEE-2011
[14] Inne Gartina Husein, Saiful Akbar, Benhard Sitohang, Fazat Nur Azizah : "Review of Ontology Matching with Background Knowledge" IEEE-2016
[15] Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I. F., & Couto, F. M. "The agreementmakerlight ontology matching system" In OTM Confederated International Conferences Springer Berlin Heidelberg.-2013
[16] Charis Fauzan, Riyanarto Sarno, Nurul Fazrin Ariyani : "Structure based ontology matching with fraud detection techniques" ICTS-2017
[17] Siham AMROUCH, Siham MOSTEFAI : "Survey on the literature of ontology Mapping, Alignment and Merging" IEEE-2012
[18] Kaladevi Ramar, Geetha Gurunathan : "Technical Review on Ontology Techniques" Medwell Journals-2016
[19]http://www.srmuniv.ac.in/sites/default/files/files/Semantic%20web%20Technology,%20Layered%20Architecture,%20RDF%20and%20OWL%20representation.pdf
[20] Gomaa, W. H., & Fahmy, A. A. "A survey of text similarity approaches" IJCA -2013
[21] http://oaei.ontologymatching.org/2018/
[22] Maqsud S. Kukasvadiya, Dr.Nidhi H. Divecha : "Analysis of Data Using Data Mining tool Orange" IJEDR-2017
[23] Veenita Kunwar, Khushboo Chandel, A. Sai Sabitha, Abhay Bansal : "Chronic Kidney Disease Analysis Using Data Mining Classification Techniques" IEEE – 2016