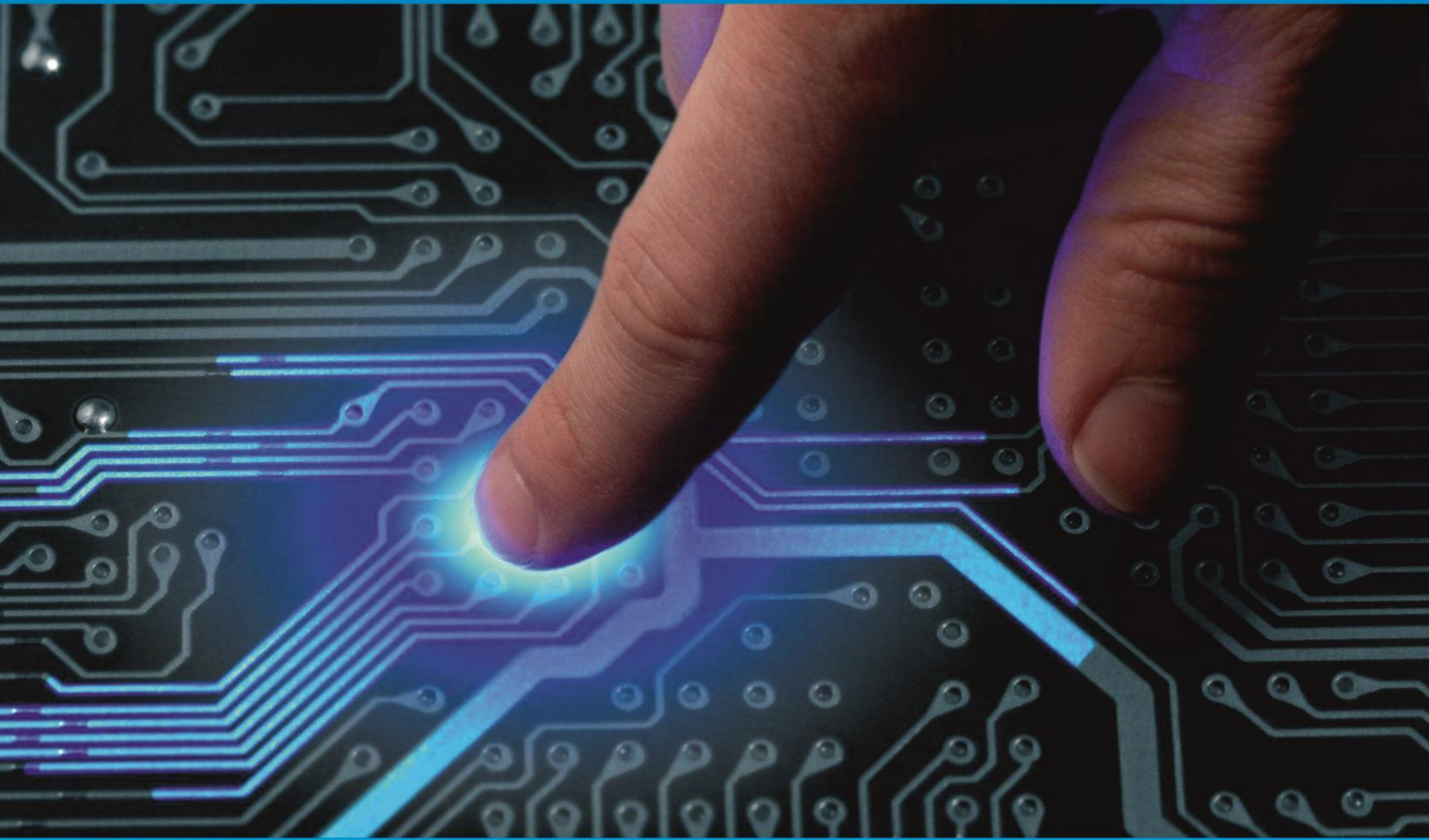
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 1, January 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

| DOI: 10.15680/IJIRCCE.2021. 0901007 |

Analysis of Water Pollutants in River Ganga using Data Mining Approach

Prakhar Gautam

Assistant Professor, Dept. of Computer Science and Engineering, Saraswati Institute of Engineering and Technology, Jabalpur, India

ABSTRACT: For numerous reasons, the rivers are drying and getting polluted which is very harmful for our Environment as well as forliving beings. Species are dying, loss of habitats due to human activities is becoming dreadful for the environment and for the future human race. The River Ganga is being polluted in large scale and has become a serious concern for everyone. River Ganga is sacred in Hindu Religion. The scale at which pollution level is increasing is alarming, it will be difficult for us to save our sacred Ganga. In this paper, it has been tried to uncover those causes which are polluting the river Ganga. The research data has been taken from various sources such as Government Open Dataset, Online forums. Newspapers, and carried out an analysis to understand the real causes behind Ganga Pollution and what can be the measures to prevent it. The analysis is performed using Data mining approach which is the extraction of useful and hidden information that cannot be discovered easily. Mining is performed on the dataset and the causes are uncovered. To perform Data Mining, RapidMiner, a data mining software platform is used which is an open-source software and a widely used one. The pollution causes are sub-classified by using Naive Bayes Classifier. To check the accuracy of the system, the data is trained and tested respectively as the predictions are made for the pollution causes based on given pollution details. The paper is concluded with the causes of pollution and the feasible solutions to it.

KEYWORDS: Ganga, Pollution Data, River pollution, Data Mining, Rapid Miner.

I. INTRODUCTION

The Ganges or Ganga is the largest river in India. Not only the river is symbol of spiritual faith to many, but it also provides water to more than 11 states accounting to more than 40% of the Indian population [1]. But, today the situation is getting worse as the sacred river is being polluted at a remarkably high level[2]. From Human Sewage to Industrial waste, everything is being dumped into River Ganga which is making the riverhighly polluted. In this Paper, Data Mining has been applied to understand and uncover the real causes behind the rapid increaseof pollution level in the river Ganga [3][4]. What are the causes which has made River Ganga so polluted and what measures can be taken to prevent it? This analysis has been performed in our research [5][6]. The data is taken from Open Gov dataset website, Forums, Newspapers and several other sources. The data mining can be explained in the following diagram:

A. DATA MINING

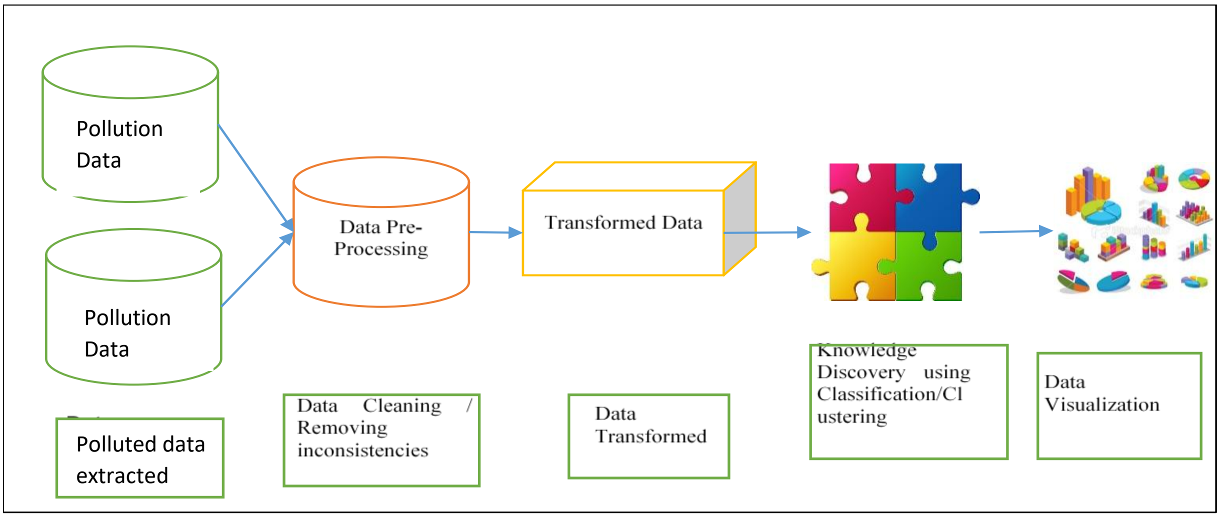


Figure 1: Data Mining Diagram

The data mining is a part of a process known as Knowledge Discovery in Databases (KDD) though the terms are used interchangeably [7]. Firstly, the data is taken (extracted) from various sources and is pre-processed where the inconsistencies are removed followed by transformation of data into one common format where finally the Data mining is applied to uncover the hidden and useful information [8]. This data can be visualized through Pie charts and Bar charts for further analysis [9]. To implement the data mining task, RapidMiner software is used which is an open-source software widely used for data mining tasks [10]. The pollution 'causes' are sub-classified using Naive Bayes Classifier.

II. LITERATURE REVIEW

A.K Shukla et. al proposed a research paper on, "SURFACE WATER QUALITY ASSESSMENT OF GANGA RIVER BASIN, INDIA USING INDEX MAPPING" [11] study of Ganga Pollution in which the Authors have integrated Water Quality Index with Geographic Information System to understand the health status and pollution causing agents in River Ganga. The quality is measured using Overall Index of Pollution (OIP) with reference to the different geographic River basins. Different monitoring stations are selected such as Rishikesh, Varanasi, Kanpur and Allahabad. Dataset is taken from Government of India open dataset. OIP value is calculated for all the monitoring stations using pollution index and number of parameters. Varanasi station is found to the have the highest OIP value which makes it highly polluted. The limitation of this paper is that only one parameter is used to define pollution in river Ganga. No focus has been made on different causes of Ganga pollution. No real time data has been taken for analysis.

A.K Bisht et. al proposed a research paper on, "DEVELOPMENT OF AN AUTOMATED WATER QUALITY CLASSIFICATION MODEL FOR THE RIVER GANGA" [12]. In this paper, the researchers have developed anautomated classification model where the water quality is classified using different parameters through Data mining approach and the model is developed so that the water quality can be automatically classified in future. Weka is an open-source data mining tool which is used for implementation. Classification is done using Decision Tree classifier. The data for mining is taken from Government of India dataset. The limitation of this research work is that no comparison has been made with other classifiers such as K-NN, Naive Bayes, etc. No focus has been made on the causes of pollution in river Ganga.

Munafo et.al proposed a research work on, "RIVER POLLUTION FROM NON-POINT SOURCES: A NEW SIMPLIFIED METHOD OF ASSESSMENT" [13].In this paper, PNPI (Potential Non-Point Index) has been defined to measure the pollution level in Rivers. The research work is performed in a River in Italy. The PNPI tool is defined as a new index for assessment in pollution level around a land area taking three parameters as Land Cover Indicator, Runoff Indicator, Distance Indicator to assess the pollution level around land area surrounding the river basins. The PNPI values are calculated using these parameters. Itdetermines what could be the pollution level based onthe use of land. The limitation of this work is that real time data is not used. No classification or mining is performed. No focus has been made on the causes of river pollution and impact of pollutants. No accuracy has been defined for the new index.

AABHA SARGAONKAR et. al proposed a research work as, "DEVELOPMENT OF AN OVERALL INDEX OF POLLUTION FOR SURFACE WATER BASED ON A GENERAL CLASSIFICATION SCHEME IN INDIAN CONTEXT "[14].In this paper, the researchers have determined and calculated the OIP (Overall Index of Pollution) by taking certain parameters such as Dissolved Oxygen, biochemical Oxygen Demand, Hardness, TDS, etc. The water quality is classified into Excellent, Acceptable, Slightly polluted, Heavily Polluted, etc. The reference standards are taken from European Standards, WHO for acceptable water quality. OIP values are calculated for River Ganga, Yamuna. The limitation of this paper is that no methodology is used for implementation. Only OIP values are calculated. Again,no focus has been made on the different causes of water pollution.

Harlieen Bindra et. al proposed a research paper titled as, "APPLICATION OF CLASSIFICATION TECHNIQUES FOR PREDICTION OF WATER QUALITY OF 17 SELECTED INDIAN RIVERS [15] ".In this paper, the predictions for quality of water are made using different classification techniques for 17 Indian rivers where the comparison is made for the water quality in 2008 and 2011. It is observed that the water quality has highly deteriorated in 3 years. Weka, which is a data mining tool is used for implementation. The dataset is taken from National Data Sharing and Accessibility Policy. Naive Bayes classifier, ID3 and J48 classifiers are used for classification. Different classes for water quality are made such as Excellent, Good, Average and Fair. The training is performed and then predictions are made for the classes defined by the classifier. The results are obtained as more classes for Average and Fair water quality has been predicted. The limitation on this paper is that different classifiers such as Decision Tree, K-NN are not implemented. No real time dataset has been used for analysing the causes of pollution in Indian Rivers. No solutions have been suggested for the river pollution.

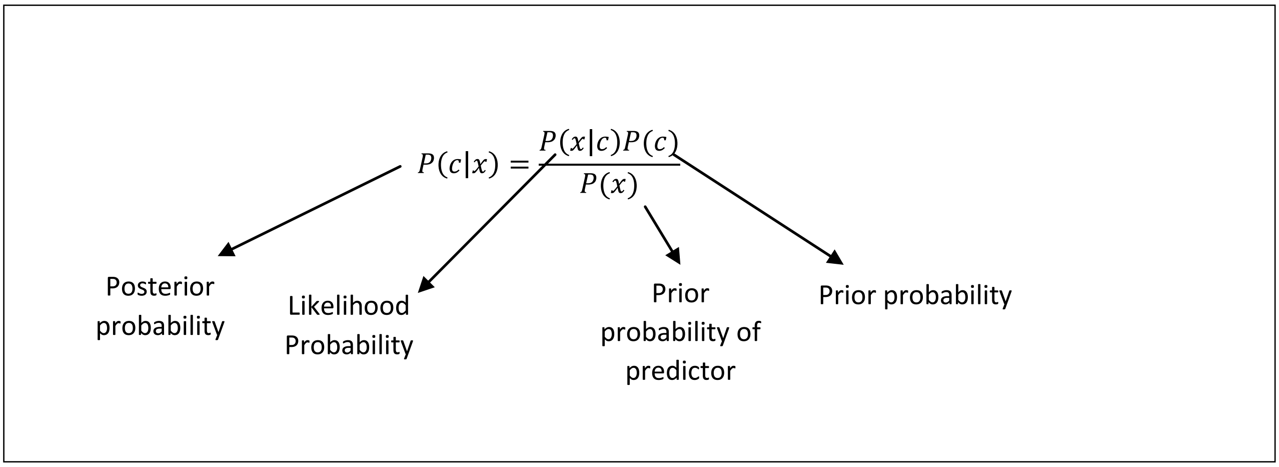
III. METHODOLOGY

A.NAIVE BAYES CLASSIFIER:

Naive Bayes classifier is a Data Mining Classification Algorithm. The basis for this algorithm is Probability [16]. The causes for Ganga pollution are the different classes defined by the classifier. Every cause of pollution has defined a probability [17]. The pollution details text is defined, and the prediction is made by the classifier [18]. The class having highest probability is predicted as the target class. It is defined as 'Naïve' because one class is totally unaware of the features of another class. One class feature is totally independent of another class.

B.EQUATION:

Naive Bayes equation is defined as [19]: -



$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

There are several parameters defined in above equation:

P(C|X) → It is the Posterior Probability defined as the output probability.

P(x) → It is apriori probability for the feature (x).

P(c) → Apriori Probability for class (c)

P(x|c) →Apriori Probability for feature upon class.

Naive Bayes Classifier is used to classify different causes of pollution in river Ganga [20].

RAPIDMINER is a data mining software platform which is used to implement data mining research work [21]. It is an open-source software which makes it freely available for use. The data mining work is been implemented in RapidMiner using different set of operators [22]. The operators are connected in following ways:

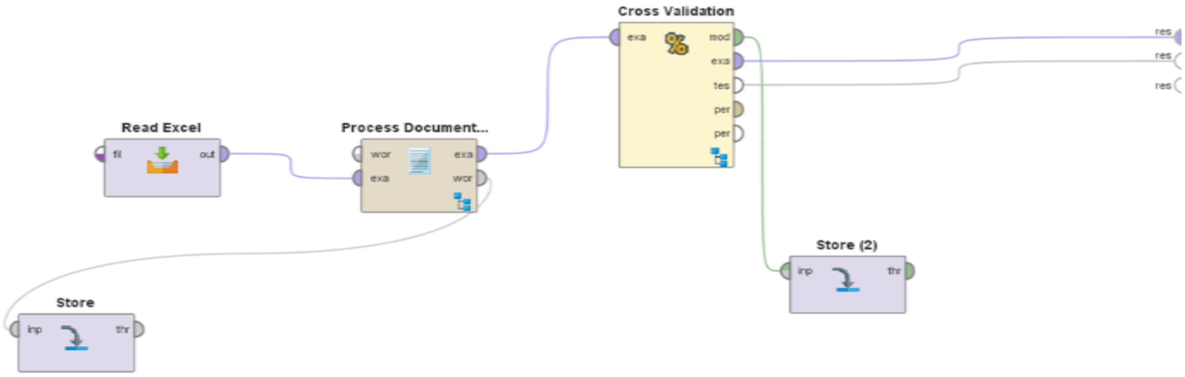
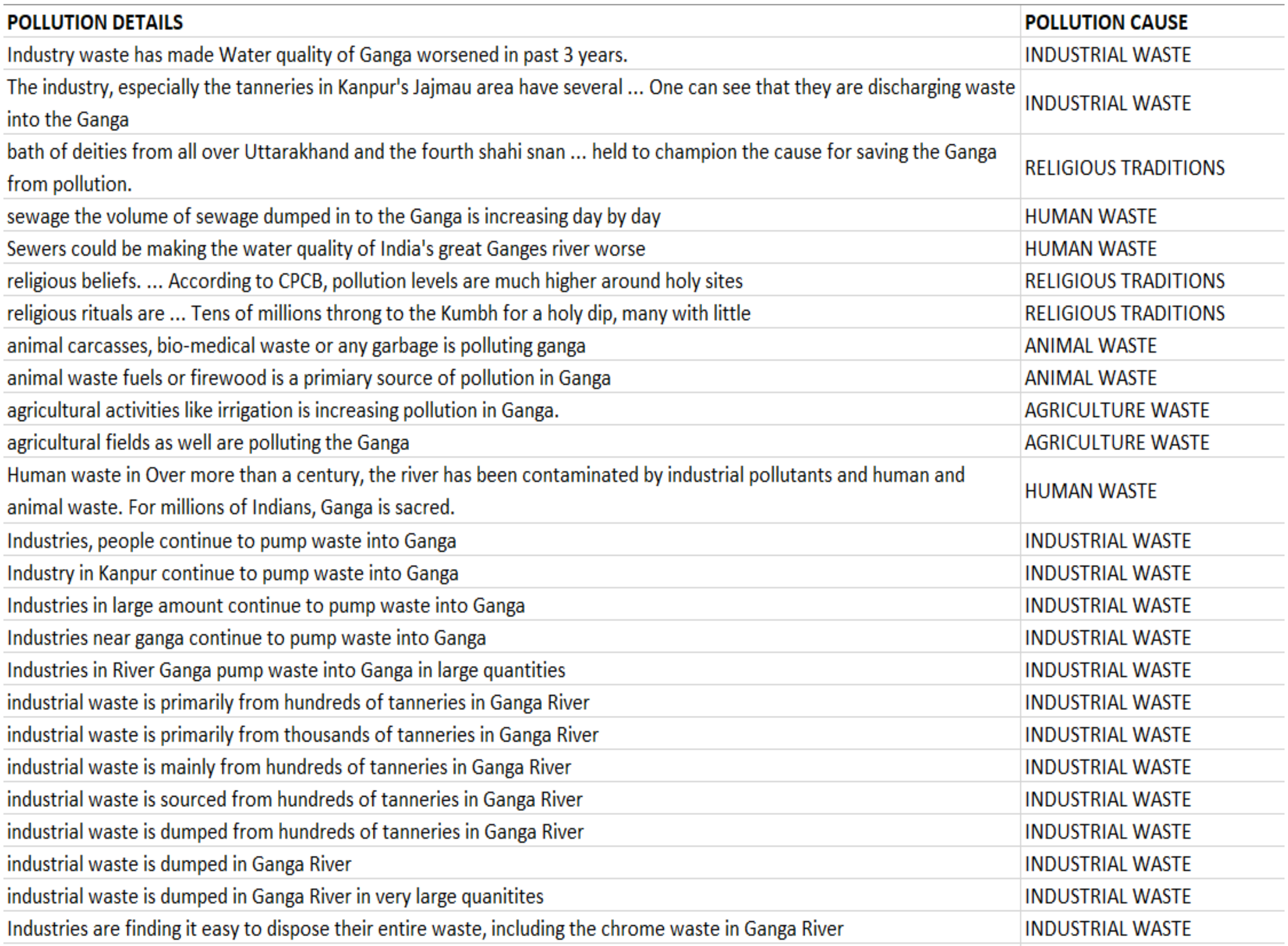


Figure 2 : RapidMiner main process

Read Excel operator is used to read the Excel file which is the Training dataset file. Process documents from data pre-processes the data by removing the inconsistencies. Cross Validation checks the accuracy of the system. Store operators stores the wordlist which is later retrieved using the retrieve operator.

A.TRAINING DATASET

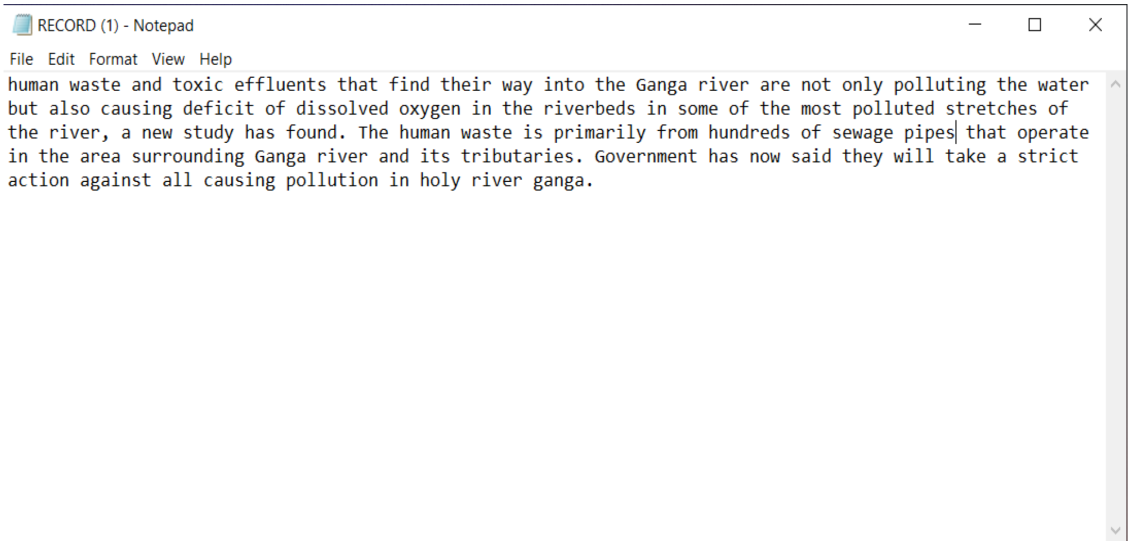
The training is performed to train the system for making predictions. The system is trained against known outputs in an excel file. 5000 + records are taken for training the system. Two columns are defined: First is the “Pollution text” which defines the pollution details, and the second column is “label”, the cause of the pollution which is defined as the label class. This can be understood from the following diagram:



POLLUTION DETAILS	POLLUTION CAUSE
Industry waste has made Water quality of Ganga worsened in past 3 years.	INDUSTRIAL WASTE
The industry, especially the tanneries in Kanpur's Jajmau area have several ... One can see that they are discharging waste into the Ganga	INDUSTRIAL WASTE
bath of deities from all over Uttarakhand and the fourth shahi snan ... held to champion the cause for saving the Ganga from pollution.	RELIGIOUS TRADITIONS
sewage the volume of sewage dumped in to the Ganga is increasing day by day	HUMAN WASTE
Sewers could be making the water quality of India's great Ganges river worse	HUMAN WASTE
religious beliefs. ... According to CPCB, pollution levels are much higher around holy sites	RELIGIOUS TRADITIONS
religious rituals are ... Tens of millions throng to the Kumbh for a holy dip, many with little	RELIGIOUS TRADITIONS
animal carcasses, bio-medical waste or any garbage is polluting ganga	ANIMAL WASTE
animal waste fuels or firewood is a primiary source of pollution in Ganga	ANIMAL WASTE
agricultural activities like irrigation is increasing pollution in Ganga.	AGRICULTURE WASTE
agricultural fields as well are polluting the Ganga	AGRICULTURE WASTE
Human waste in Over more than a century, the river has been contaminated by industrial pollutants and human and animal waste. For millions of Indians, Ganga is sacred.	HUMAN WASTE
Industries, people continue to pump waste into Ganga	INDUSTRIAL WASTE
Industry in Kanpur continue to pump waste into Ganga	INDUSTRIAL WASTE
Industries in large amount continue to pump waste into Ganga	INDUSTRIAL WASTE
Industries near ganga continue to pump waste into Ganga	INDUSTRIAL WASTE
Industries in River Ganga pump waste into Ganga in large quantities	INDUSTRIAL WASTE
industrial waste is primarily from hundreds of tanneries in Ganga River	INDUSTRIAL WASTE
industrial waste is primarily from thousands of tanneries in Ganga River	INDUSTRIAL WASTE
industrial waste is mainly from hundreds of tanneries in Ganga River	INDUSTRIAL WASTE
industrial waste is sourced from hundreds of tanneries in Ganga River	INDUSTRIAL WASTE
industrial waste is dumped from hundreds of tanneries in Ganga River	INDUSTRIAL WASTE
industrial waste is dumped in Ganga River	INDUSTRIAL WASTE
industrial waste is dumped in Ganga River in very large quanitites	INDUSTRIAL WASTE
Industries are finding it easy to dispose their entire waste, including the chrome waste in Ganga River	INDUSTRIAL WASTE

Figure 3: Training Dataset

B.TESTING DATASET



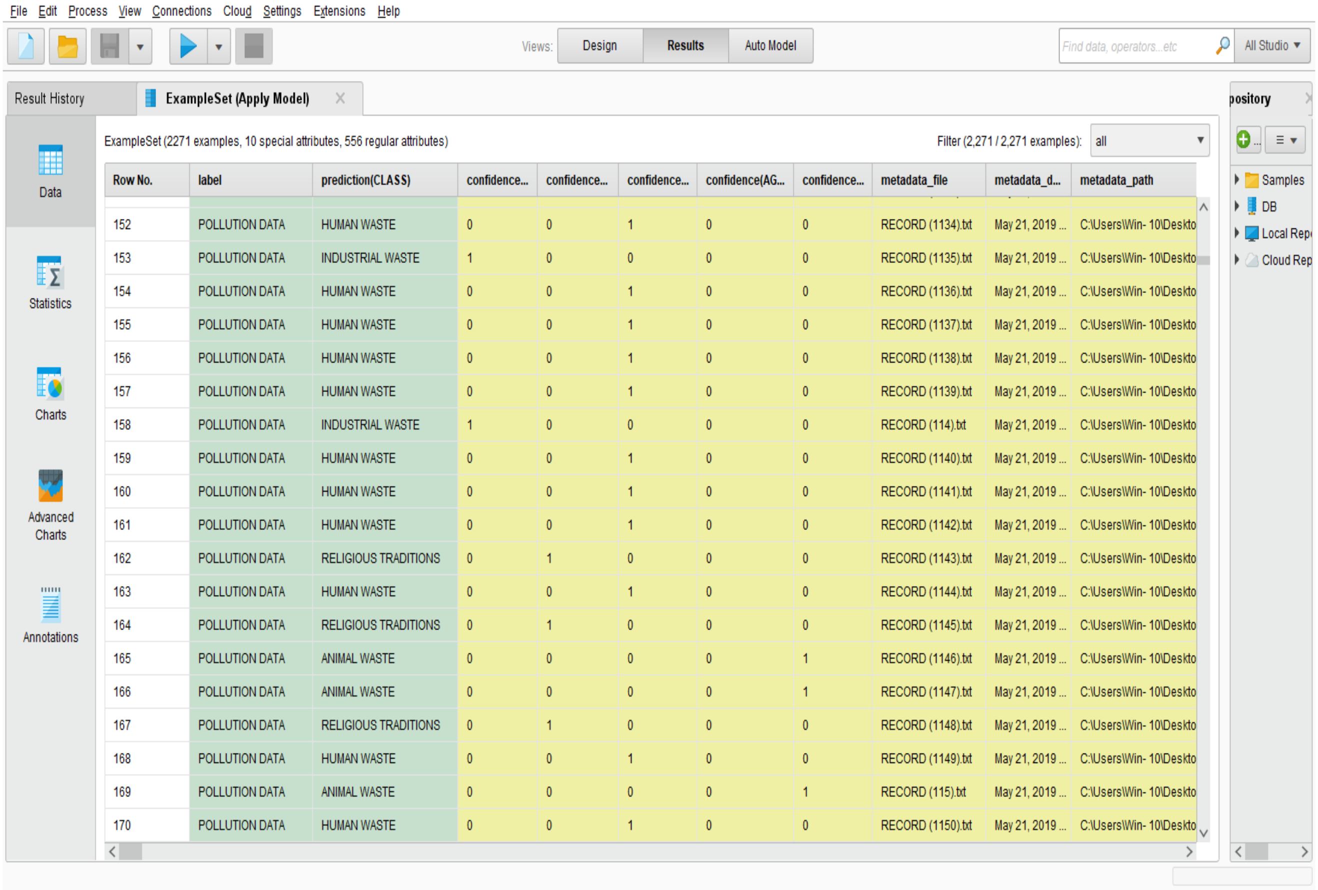
```
human waste and toxic effluents that find their way into the Ganga river are not only polluting the water
but also causing deficit of dissolved oxygen in the riverbeds in some of the most polluted stretches of
the river, a new study has found. The human waste is primarily from hundreds of sewage pipes that operate
in the area surrounding Ganga river and its tributaries. Government has now said they will take a strict
action against all causing pollution in holy river ganga.
```

Figure 4 Testing dataset

Testing dataset is applied after the training phase is completed. 2000 + records are being taken for performing testing. The accuracy is checked on the basis of how well and correct the system has performed in making predictions. This helps in uncovering the most common causes of pollution in River Ganga. 2000+Notepad Text files store the testing dataset, one text file for each testing record.

IV. RESULT & ANALYSIS

The finalresults are obtained by running the processes in RapidMiner. Two processes are executed for Training and Testing in RapidMiner. Various sources are included for taking the dataset such as Open Government Dataset Portal, E-news, E-Newspapers, Forums, Offline direct people to people connect, etc. The analysis is performed to check the common causes of pollutants found in Holy River Ganga. The desired result obtained are as follows:



ExampleSet (2271 examples, 10 special attributes, 556 regular attributes)

Filter (2,271 / 2,271 examples): all

Row No.	label	prediction(CLASS)	confidence...	confidence...	confidence...	confidence(AG...	confidence...	metadata_file	metadata_d...	metadata_path
152	POLLUTION DATA	HUMAN WASTE	0	0	1	0	0	RECORD (1134).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
153	POLLUTION DATA	INDUSTRIAL WASTE	1	0	0	0	0	RECORD (1135).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
154	POLLUTION DATA	HUMAN WASTE	0	0	1	0	0	RECORD (1136).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
155	POLLUTION DATA	HUMAN WASTE	0	0	1	0	0	RECORD (1137).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
156	POLLUTION DATA	HUMAN WASTE	0	0	1	0	0	RECORD (1138).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
157	POLLUTION DATA	HUMAN WASTE	0	0	1	0	0	RECORD (1139).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
158	POLLUTION DATA	INDUSTRIAL WASTE	1	0	0	0	0	RECORD (114).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
159	POLLUTION DATA	HUMAN WASTE	0	0	1	0	0	RECORD (1140).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
160	POLLUTION DATA	HUMAN WASTE	0	0	1	0	0	RECORD (1141).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
161	POLLUTION DATA	HUMAN WASTE	0	0	1	0	0	RECORD (1142).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
162	POLLUTION DATA	RELIGIOUS TRADITIONS	0	1	0	0	0	RECORD (1143).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
163	POLLUTION DATA	HUMAN WASTE	0	0	1	0	0	RECORD (1144).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
164	POLLUTION DATA	RELIGIOUS TRADITIONS	0	1	0	0	0	RECORD (1145).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
165	POLLUTION DATA	ANIMAL WASTE	0	0	0	0	1	RECORD (1146).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
166	POLLUTION DATA	ANIMAL WASTE	0	0	0	0	1	RECORD (1147).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
167	POLLUTION DATA	RELIGIOUS TRADITIONS	0	1	0	0	0	RECORD (1148).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
168	POLLUTION DATA	HUMAN WASTE	0	0	1	0	0	RECORD (1149).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
169	POLLUTION DATA	ANIMAL WASTE	0	0	0	0	1	RECORD (115).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto
170	POLLUTION DATA	HUMAN WASTE	0	0	1	0	0	RECORD (1150).txt	May 21, 2019 ...	C:\Users\Win- 10\Deskto

Figure 5: Results for Pollution data in RapidMiner.

As it can be seen from the above results, the system has made accurate and correct predictions for the class. The pollution data is classified into most common pollution causes in River Ganga such as Human Sewage Waste, Religious Traditions, Industrial waste, Animal waste, and Agricultural Waste. Correct predictions were made when the text was as “Sewage waste is dumped into River Ganga” is predicted as “HUMAN SEWAGE WASTE.”

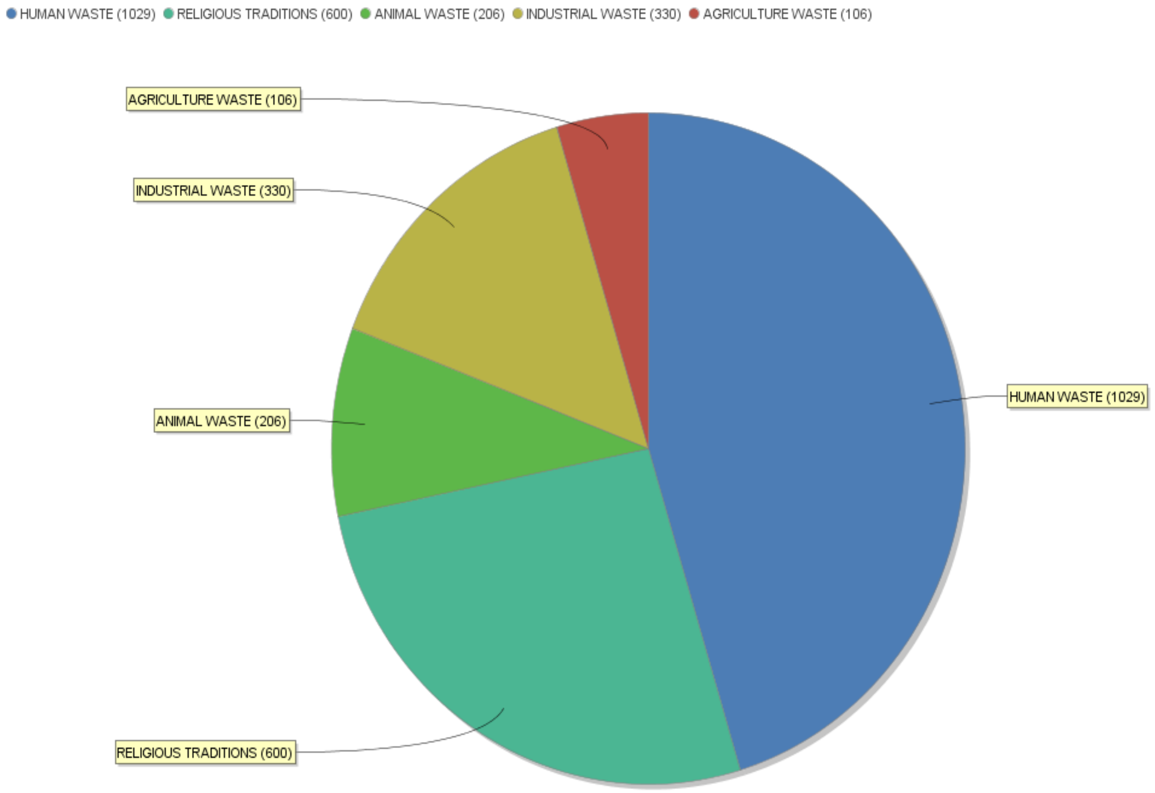


Figure 6: Result Pie Chart

The analysis is taken for 2000+ records, where the results obtained are as follows:

HUMAN SEWAGE WASTE	**1029 no.s. (51.45%)**
RELIGIOUS TRADITIONS	**600 no.s (30%)**
INDUSTRIAL WASTE	**330 no.s (16.5%)**
ANIMAL WASTE	**206 no.s (10.3%)**
AGRICULTURAL WASTE	**106 no.s (5.3 %)**

The above analysis shows that majority of the waste is dumped from Human Sewage and Religious Traditions such as Bathing, and waste dumped while performing Religious rituals and Traditions. Industrial waste, Animal waste and Agricultural waste also contribute considerable amount accounting to Ganga pollution.

V. CONCLUSION AND FUTURE WORK

This paper is concluded with the work performed in RapidMiner of pollution Data which contains the causes for pollution in River Ganga. The data was collected from various sources and an analysis has been done on the work performed by Government and NGO's for cleaning up and avoiding the pollution in Holy River. The analysis includes classification of most common causes causing pollution in Ganga and its tributaries. In future, this work can be further extended by increasing the dataset and classifying the pollution level from Gomukh to GangaSagar. The research can also be extended by applying different classifiers such as ID3, Support Vector Machine, etc. and by working on the pollution in Yamuna and other rivers. The work can also be implemented using different platforms such as R, KNIME, TIBCO, etc.

REFERENCES

1. Chaudhary, Meenakshi, and Tony R. Walker. "River Ganga pollution: Causes and failed management plans (correspondence on Dwivedi et al. 2018. Ganga water pollution: A potential health threat to inhabitants of Ganga basin. Environment International 117, 327–338)." *Environment international* 126 (2019): 202-206.
2. Bhargava, Devendra Swaroop. "Use of water quality index for river classification and zoning of Ganga River." *Environmental Pollution Series B, Chemical and Physical* 6.1 (1983): 51-67.
3. Joshi, Dhirendra Mohan, Alok Kumar, and Namita Agrawal. "Assessment of the irrigation water quality of river Ganga in Haridwar District." *Rasayan J Chem* 2.2 (2009): 285-292.
4. Singh, Pramod, Dhruv Sen Singh, and Uma Kant Shukla. "Ganga: The Arterial River of India." *The Indian Rivers*. Springer, Singapore, 2018. 75-92.
5. Kaushal, Nitin, et al. "Towards a healthy Ganga-improving river flows through understanding trade-offs." *Frontiers in Environmental Science* 7 (2019): 83.
6. Tan, Pang-Ning. *Introduction to data mining*. Pearson Education India, 2018.
7. Anoopkumar M and A. M. J. M. Z. Rahman, "A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration," *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, Ernakulam, 2016, pp. 122-133.
8. Aggarwal, Charu C., and ChengXiang Zhai, eds. *Mining text data*. Springer Science & Business Media, 2012.
9. Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "The KDD process for extracting useful knowledge from volumes of data." *Communications of the ACM* 39.11 (1996): 27-34.
10. Sharma, Prerna, and Anubha Kaushik. "Drivers of Ecosystem Change: A Case Study of River Ganga." (2018).
11. A. K. Shukla, C. S. P. Ojha and R. D. Garg, "Surface water quality assessment of Ganga River Basin, India using index mapping," *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Fort Worth, TX, 2017, pp. 5609-5612.
12. Bisht A.K., Singh R., Bhatt A., Bhutiani R. (2018) Development of an Automated Water Quality Classification Model for the River Ganga. In: Bhattacharyya P., Sastry H., Marriboyina V., Sharma R. (eds) Smart and Innovative Trends in Next Generation Computing Technologies. NGCT 2017. Communications in Computer and Information Science, vol 827. Springer, Singapore.
13. Michele Munafo et. al River pollution from non-point sources: a new simplified method of assessment, Journal of Environmental Management, Volume 77, Issue 2, 2005.
14. Sargaonkar, A. & Deshpande, V. Environ Monit Assess (2003) 89: 43.
15. Bindra H., Jain R., Singh G., Garg B. (2019) Application of Classification Techniques for Prediction of Water Quality of 17 Selected Indian Rivers. In: Balas V., Sharma N., Chakrabarti A. (eds) Data Management, Analytics and Innovation. Advances in Intelligent Systems and Computing, vol 808. Springer, Singapore.
16. Aggarwal, Charu C., ed. *Data classification: algorithms and applications*. CRC press, 2014.
17. Rish I. An empirical study of the naive Bayes classifier. InIJCAI 2001 workshop on empirical methods in artificial intelligence 2001 Aug 4 (Vol. 3, No. 22, pp. 41-46).
18. Friedman, Nir, Dan Geiger, and Moises Goldszmidt. "Bayesian network classifiers." *Machine learning* 29.2-3 (1997): 131-163.
19. Tan, Songbo. "An effective refinement strategy for KNN text classifier." Expert Systems with Applications 30.2 (2006): 290-298.

20. Islam, Mohammed J., et al. "Investigating the performance of naive-bayes classifiers and k-nearest neighbor classifiers." 2007 International Conference on Convergence Information Technology (ICCIT 2007). IEEE, 2007.
21. Jungermann, Felix. "Information extraction with rapidminer." Proceedings of the GSCL Symposium'Sprachtechnologie und eHumanities. 2009.
22. Kotu, Vijay, and Bala Deshpande. Predictive analytics and data mining: concepts and practice with rapidminer. Morgan Kaufmann, 2014.
23. Verma, Tanu, R. Renu, and D. Gaur. "Tokenization and filtering process in RapidMiner." International Journal of Applied *Information Systems* 7.2 (2014).

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING