Grenze International Journal of Engineering and Technology, Jan Issue

Estimation of Water Requirement for a Location using Machine Learning

Lavanya P[1], Nitin Krishna[2], Abhishek B J[3] and Bhaskar G[4]
[1-3]Student, Dept. Of CSE, Siddaganga Institute of Technology, Tumakuru, Karnataka, India.
{lavanya.cs051, kittynitin2000, abhisheksets}@gmail.com
[4]Assistant Professor, Dept. Of CSE, Siddaganga Institute of Technology, Tumakuru, Karnataka, India.
gbhaskar@gmail.com

***Abstract*—To most basic requirement of lifeline, water is getting scarce by day. Many reasons are contributing to the scarcity of water.Firstly, uneven rainfall over the past period of years. Secondly, mass movement of people from rural to semi-urban (or to urban areas) which increases the demand of water naturally.Thirdly,Industrial Evolution which demands pure water putting the agriculture domain's requirements at stake and lastly, ever reducing catchment area or storage area like ponds,lakes etc., due to urbanization.**
All these above factors also contribute to the depletion of ground water table.To address these issues, the government is planning to join two rivers or to channel the rivers to divert the water to the areas where water is scarce which demands high cost for implementation.Considering all these factors,this paper aims at studying the water requirements for a considered region by applying the Machine Learning Algorithms to the input factors like average rainfall in the region,storage capacity(lakes,ponds etc) and the per capital requirement of water.
The final outcome of this paper will be an analysis directing the authority like water board to take needy actions by forecasting the requirements.

***Index Terms*— Evaporation, Catchment area, Losses, Demand, Rainfall, Seepage, Percolation.**

I. INTRODUCTION

Water means life. Water is a prime natural resource. It is a basic need for humans and a precious asset that living beings have.Water is equally vital for the survival of plants and animals.Soil needs water for sustaining plants.The water cycle is essential for ecological balance too. Though a big portion of the Earth is covered with water, only a small portion water is capable for usage purpose. Every cell, organs and tissues in our body use water to help with temperature regulation, keeping hydrated and maintaining bodily functions. In addition, water acts as a lubricant and cushions your joints. Many industries require large quantities of water for processing, cooling, and diluting products. Examples of industries that consume large quantities of water include the paper industry, the food industry, and the chemical industry. Water is also used as an industrial solvent for the production of several commercially important products.

Almost all power plants that generate electricity employ water to spin turbines.Water is used to grow fresh produce and sustain livestock. The use of water in agriculture makes it possible to grow fruits and vegetables and raise livestock, which is a main part of our diet. All living organisms, plants, animals and human beings contain water. Almost 70% of our body is made up of water. Our body gets water from the liquids we drink and the food

Grenze ID: 01.GIJET.8.1.20

we eat. Nobody can survive without water for more than a week. All plants will die if they do not get water. This would lead to the death of all the animals that depends on plants for their food. So, the existence of life would come to an end. Although our planet earth is covered with 71% of water and 29% of the land.Among these rivers, ponds and underground water are the fresh water resources they occupy only 4% of total water and can be used directly. The fast-growing industries contaminating the water bodies which is affecting both humans, plants as well as marine life. The unequal distribution of water on the earth and its increasing demand due to the increasing population is becoming a concern for all. So, we need to be judicious and rational regarding the usage of water. Usually, the rain cycle happens and refills these ponds, lakes. Rain cycle includes evaporation of water, formation of clouds and finally the precipitation.Many years ago, rain water used to fill up all the storage areas like ponds and lakes that was sufficient for all domestic,industrial and agricultural activities. The entire population of that particular area generally depends on this kind of stored water for all their daily needs.Year by year population is increasing but the catchment area remains same or even decreasing in few areas. So, estimation and conservation of water is very important to meet the human's demand. In today's world, the rate of depletion of natural resources is increasing day by day. Water resource is one among the worst hit natural resources due to depletion of ground water, pollution and decreased rainfall in some areas. This paper concern about the knowledge need to conserve the rainwater collected in the reservoir, the necessary action that are needed to meet the demand for the local population. Our paper talks about these things in details.
The rest of the paper is organized as, Section II describes our Problem statement.Section III gives a glimpse on Literature survey. Section IV contains the Approach and our Proposed System which we taken to solve this problem using modern tools and technologies.Section V we give a Conclusion for this paper.

II. Problem Statement

As we already discuss about the water scarcity.The groundwater resources come under increasing pressure due to over-reliance and unsustainable consumption, water in wells, ponds and tanks dry up. We majorly depending on the rainwater that get collected in the containment area. Due to urbanisation the catchment area deceasing day by day. Population explosion is also a main reason for the water scarcity. The calculation for analysing the demand for the population by taking the required details as it is a time consuming and one who calculate has to be gather all the necessary details correctly to get the accurate results. So to make this job easy we can use the modern tool like Machine Learning that gives accurate result in few seconds.

III. Literature Survey

David Seckler said that many countries are entering an era of severe water shortage, their analysis shows that around 50 percent of the increase in demand for water by the year 2025 can be met by increasing the effectiveness of irrigation. While some of the remaining water development needs can be met by small dams and conjunctive use of aquifers, medium and large dams will almost certainly also be needed.One of the problems with the supply side approach is that the criterion for water scarcity is based on a country's AWR without reference to present and future demand or needs for water. In their work they give us a brief description of how the rain water gets collected, the factors come into play for successive storage of water.Their analysis of AWR for all 118 countries and the methodology to meet the future water demand helped us to take this project [1].
Mekonnen and Hoekstra said scarcity of freshwater increasing globally the reason for this water crises as reported by the World Economic Forum that the population explosion which leads to rising global demand for water and more, these things help us to get to know about the variants in water scarcity in different levels like low, moderate, significant, and severe water scarcity that the number of people facing all over the world. The statistical measure given that about 71% of the global population (4.3 billion people) lives under conditions of moderate to severe water scarcity. About 66% (4.0 billion people) lives under severe water scarcity. Its concerns about protecting ecosystem by meeting the demand for freshwater by maintaining blue water footprints within maximum sustainable levels per catchment and asked for governments and companies to formulate water footprint benchmarks based on best available technology to improve the storage to meet the demand [2].

IV. Approach and Proposed System

Machine learning (ML) algorithm is an important component in structure–activity modelling and analysis of compounds.Machine Learning is a subfield of artificial intelligence. It focuses mainly on the designing of systems, thereby allowing them to learn and make predictions based on some experience which is data in case of machines. Machine learning enables computers to act and make data-driven decision rather than being explicitly

programmed to carry out a certain task, these programs are designed to learn and improve over time when exposed to new data. This finds results from the extraction of dataset this means that the machine cannot only find the rules for optimal behaviour but also can adapt to the changes in the world many of the algorithms involved have been known for decades and centuries even thanks to the advances in computer science and parallel computing. They can scale up to massive volumes of data. The working of machine learning begins by using a labelled or unlabelled training data set to produce a model new input data is introduced to the machine learning algorithm and it make prediction based on the model. The prediction is evaluated for accuracy and if the accuracy is acceptable, the machine learning algorithm is deployed. If the accuracy in not acceptable then the model in trained again and again to improve it.Machine Learning have three kinds supervised, unsupervised and reinforcement learning.*Supervised learning* in thiswe use labelled dataset here for each input instances we are having an expected output the associated value can be discrete representing a category or can be a real or continuous value in either case.This kind of techniques are useful to give output for unseen input. *Unsupervised learning* is used to train the models using unlabelled dataset and here we know only the input but not the exact output. Machine learning includes algorithms like: -

Logistic Regression: This algorithm used to solve classification problems. The algorithm is best suits when you are working on discrete/categorical values like either *Yes* or *No*,0 or 1, *True* or *False*. It's similar to linear regression but instead of a straight line we fit an "S" shaped logistic curve function. It transforms its output using logistic sigmoid function to return a probability value. This predictive analysis algorithm is based on the concept of probability. The hypothesis of logistic regression tends to limit the cost function between 0 and 1 [7].

Support vector Machine: It is a *Supervised learning* algorithm. This divides the n-dimensional space into classes so that we can easily put new data value in the correct category in future. In other words, it is a representation of examples as points in space that are mapped so that the points of different categories are separated by a gap as wide as possible. The boundary is called Hyperplane. The space between to hyperplane is called *Margin*. The data points which are around the hyperplane are called *Support vector*. In some cases, the support vector machines are not so efficient in that situation we can use kernel to transform the data into high dimensional space. This helps us to easily segregate the data points. We have three kinds of kernels in this support vector machine algorithms such as *Linear kernels*, *Polynomial kernels*, *Radial Basis Function kernel (RBF)* [5].The Linear kernel this can be used as normal dot product between any two observations so the product between two vectors is the sum of the multiplication of each pair of input values.RBF kernel it is commonly used in SVM classification can map the space in infinite dimensions.

Naive Bayes: Naive Bayes classifiers are a collection of classification algorithms based on *Bayes Theorem*. It is not a single algorithm but a family of algorithms where all of them share a common principle i,e...every pair of features being classified is independent of each other. The main logic behind this algorithm is that the presence of a particular feature of a class is unrelated to the presence of any other feature, given the class variable. The advantage of using this algorithm is that it helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object [6].

K-Nearest-Neighbours: K-Nearest Neighbour is a *Supervised* machine learning algorithm.That stores all the available cases and classify the new data or case based on a similarity measure, it used in a search application when we want to search in similar items, where '*K*' means to the number of nearest neighbours which are voting class of new data or testing data.It is a simple algorithm which uses entire dataset in its training phase whenever a prediction is required for the unseen data it searches through the entire trained dataset for K most instances and the data with most similar instance is finally returned as the prediction [8].

Decision Tree: This algorithm comes under *Supervised machine learning* algorithm. It is the graphical representation of all the possible solutions to a decision. A tree like structure gets created in which each node in the tree acts as a test case for some attribute, each edge descending from the node corresponding to the possible answers to the test case. This process is recursive in nature. There are two nodes, root node used to make any decision and have multiple branches whereas leaf nodes are the output or target variable [9].
Random Forest: This algorithm more similar to decision tree algorithm. Here it builds multiple decision trees and merges them together. Here the output is more accurate and stable. Most of the time the random forest algorithm is built on the bagging method that is based on the idea that the combination of learning model increases the overall results. If you are combining the learning from different models and then clubbing at

together what it will do is, it will increase the overall results. At the end the final result is taken based on majority voting of all the tress that generated [4].
The above algorithms are supervised learning classification algorithms where we used when we have a dataset with labelled data.

V. METHODOLOGY

In Figure 1. we represented flow diagram which represent our dataset which having attributes that includes name of the containment area,the year, average rainfall, catchment area,losses, area population and the storage capacity. From these data we further calculate the actual yield which the exact amount of water that will be get collected in the storage. Then we calculate the yield after losses by considering yield and losses, which is the practically collected water after deduction from losses. We calculate the demand from population where we calculate the amount of water required per head per annum. Then by taking yield after losses, demand and capacity we give these parameters to our machine learning model so that our model is ready whenever a new data input came then the model compare each with a decision tree by taking majority voting concept for our target variables *yes* or *no*. If the result is *YES,* we face the scarcity of water so that we need to take additional measurement like increasing the catchment area or by increasing the storage capacity.

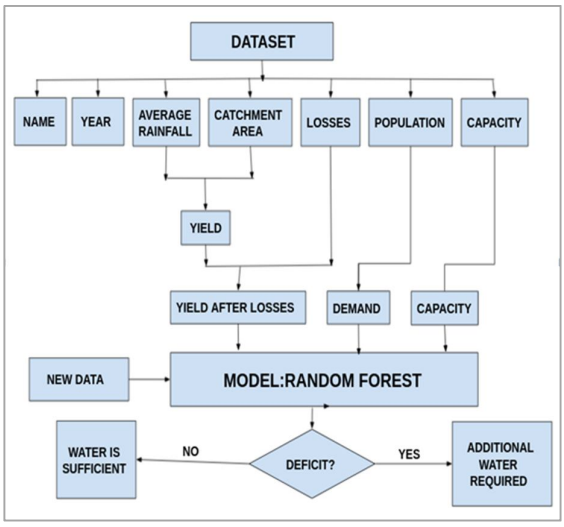

Fig 1. Flow Diagram

If result is *NO,* we are safe that the water is sufficient for that area.

VI. DATASET

The dataset is done by collecting data from water resource department of Karnataka and also from several weather forecasting websites. The dataset contains 700 rows and 12 columns. We find out the main parameters that affects the storage of rain water and taken out attributes as shown in Figure 2.

	name	year	avg_rain_fall(m)	catchment_area(m2)	population	losses	yield(tmc)	yield_after_losses(tmc)	demand(tmc)	capacity_of_lake(tmc)	deficit
0	bugudnalli	1996	0.80	36654548	64000	0.10	0.442728	0.398456	0.111435	0.2345	no
1	bugudnalli	2002	0.52	35484829	98000	0.16	0.187841	0.157787	0.170634	0.2338	yes
2	bugudnalli	1998	0.86	36248646	76000	0.12	0.570534	0.502070	0.132329	0.2343	no
3	bugudnalli	2016	0.57	32680000	180000	0.23	0.223195	0.171861	0.313410	0.2230	yes
4	bugudnalli	2000	0.92	35898984	86000	0.14	0.631206	0.542837	0.149740	0.2341	no

Name of containment area, year, average rainfall, catchment area, losses, population of that particular area, capacity of the reservoir, yield after losses, demand and deficit. All these attributes help us in predicting sufficiency/deficit of water in particular area.

A brief introduction of attributes:

1. *Name*:- It is the name of the water storage area where rain water get collected and stored.
2. *Year*:-It is the year that we are estimating the sufficiency of water.
3. *Average rainfall*:-Amount of precipitation of rainwater that we expect per annum.
4. *Catchment area*:-It is the area from which rainfall flows into a river ,lake or reservoir.
5. *Population*: -The entire population situated around the reservoir.
6. *Losses*:-The percentage of water lost due to evaporation and seepage/percolation of water into ground.
7. *Yield*:- It is the actual amount of water that should be get collected before deduction of losses occurs.
8. *Yield after Losses***:-**It is the amount of water that will be stored after deduction of losses due to evaporation and percolation.
9. *Demand***:-**It is the estimated water demand is required by the entire population of that particular area per annum.
10. *Deficit***:-**It indicates whether we will face shortage of water or not.If the value is Yes, it means we may face the shortage of water,if it is NO the water is sufficient for the usage.

In every Machine Learning project, we usually follow some steps for better understanding and to get exact results, the very first step is *Data Collection* we collect the data from Karnataka water resource department and other weather forecasting websites and stored all data in an excel sheet. Our next step is *Data Wrangling* in this phase we filter out our dataset from NaN*(Not a Number)* values or any blank space if present this will avoid from unnecessary output, we got balanced dataset also as shown in Figure 3. In our next phase Training and Testing phase we divide the whole data into two parts in some ratio(we done in 80:20 ratio) for training and testing purpose.Finally, we *check the accuracy* of the model, from this we will get to know how accurately that the model can predict.

Before training our model, we need to analyse how data attributes correlate to each other for this we plot Correlation Matrix using Heatmap. The correlation matrix is used to find the pairwise correlation of all columns in the dataframe. If any NaN or categorical values are present in the dataset it automatically excludes them. It gives us information about the interdependence of various attributes. The interdependence is ranged from 0 to 1 where 1 indicates that those two attributes are strongly dependent on each other and 0 shows that those two parameters are completely independent of each other. Hence all the diagonal elements of the matrix will obviously be 1 as it indicates complete self-dependency.

From Figure 3 we showed that our data set is balanced which means that our data set approximately contain equal proportion of target value YES and NO.

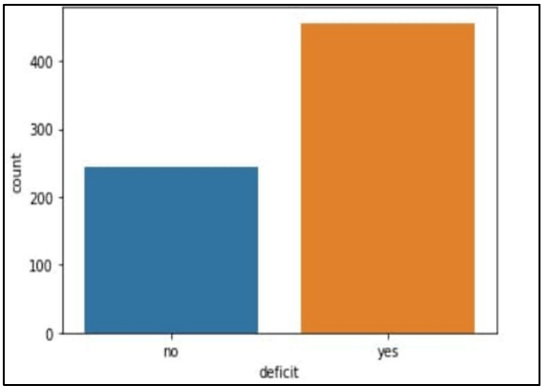

Fig 3: Balanced Dataset

Balanced dataset helps to improve the working efficiency of the model and it avoids the overfitting problem. So, the model gives accurate results for the new entries.

In real world we won't get data to compare them directly some parameters require modification. In Machine Learning also it is easy to train a model with numerical data. In some cases, to predict the results forwe need to apply some equations and we need to do conversion some times.In our project also we make use of some calculation before prediction. The last four columns are our main attributes we derived those attributes using formulas for calculating yield which means the calculated amount of water should be get stored in the containment area which is calculate using average rainfall and catchment area from using (1),

$$Yield = (0.87 \times Average\ Rainfall - 0.3025) \times Catchment\ Area \times 35.314(10^{-9}) \quad (1)$$

Yield after losses as the water get evaporated due to external factors like evaporation, seepage/percolation of water into underground,we taken water which left after the such losses from using (2),

$$Yield\ After\ Losses = Yield\ \times (1 - Losses) \qquad (2)$$

Demand for the population which we get to know how much is the demand rate/water requirement is there in that particular area,from using (3).

$$Demand = (Population \times 135 \times 365 \times (10^{-9})) \div 28.3 \qquad (3)$$

Then we applied again correlation matrix to computed data,which is as shown in Figure 4.

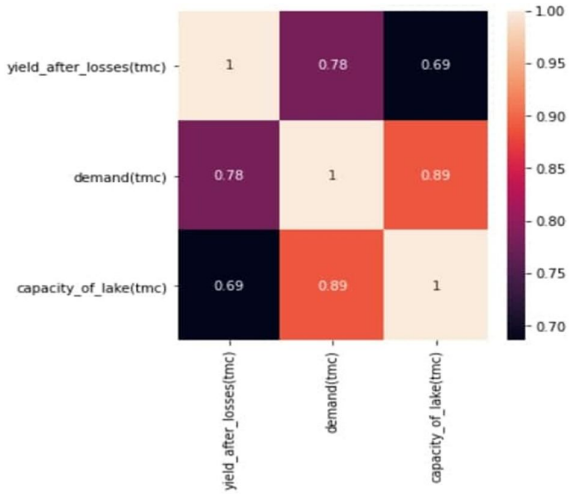

Figure 4: Correlation Matrix

Now our dataset is ready to train with the model,we have applied different classification algorithms on our dataset which we can see the accuracy of those algorithms in Figure 5.that clearly shown that the results that we obtained for each algorithm and the respective accuracy we get is shown, for SVC(*support vector classifier*) kernel sigmoid gives an accuracy of 0.59,SVC kernel RBF[5] gives an accuracy of 0.67,Naive Bayes[6] gives an accuracy of 0.67,Logistic Regression[7] gives an accuracy of 0.68, KNN(*k-Nearest-Neighbours*) [8] gives an accuracy of 0.96,Random Forest[4] gives an accuracy of 0.98 and Decision Tree[9] gives a accuracy of 1.

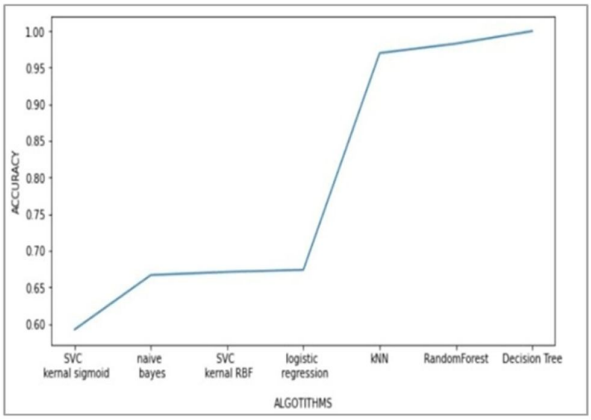

Figure 5: Accuracy V/s Different Algorithms

Even though among all algorithm decision tree gives maximum accuracy,this model get overfitted.So, to avoid this problem we taken random forest algorithm.Random forest avoids the overfitting problem as the accuracy

also high. It is flexible to both classification and regression problems and works well for our categorical dataset.Finally, we have chosen *Random Forest algorithm* to train our model.

VII. RESULTS

As you can see from the Figure 6.1 the website takes inputs like name of the reservoir, year, amount of annual rainfall, catchment area, population of that area, amount of losses occurred, capacity of reservoir. For the inputs(Figure 6.1)the results are shown in Figure 6.2 where it tells that the water gets collected are sufficient to meet the demand of the nearby population.

Figure 6.1:Input given to test the model

Figure 6.2:Result of given input(Figure 6.1)

For the inputs given in Figure 7.1 one can see the results Figure 7.2 that it shows the water is not sufficient to meet the demand. By using these results one can take further action and act accordingly.

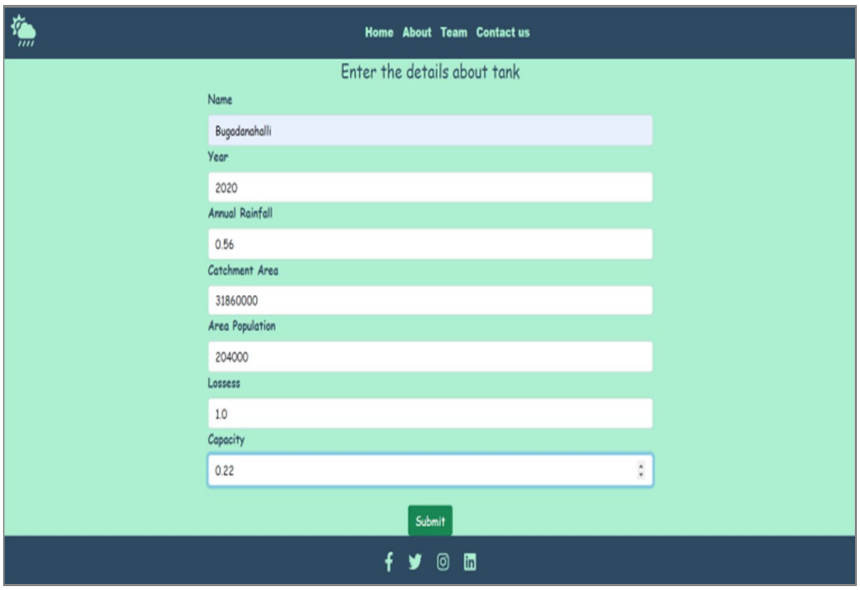

Figure 7.1:Input given to test the model

Figure 7.2:Result of given input(Figure 7.1)

VIII. CONCLUSION

Water is one of the basic needs for every human being. We use water for our daily activities. In practical three-fifth of world population depends on the water get stored from rainwater.Our work helps to take an analysis on the water collected in containment area through rainwater is sufficient for the usage of that particular area population or not, by taking few details about external and internal factors. In this project we used Machine Learning which is the most powerful and popular trending technology that helped us to solve problem easily and also benefit for making predictions that helps to take better decisions in future. Our work purely dedicated to the water resource department to make use of this for the estimation of water requirement and to take additional actions in case the water collected in reservoir fails to meet the demand.

REFERENCES

[1] Seckler, D. W. (1998). *World water demand and supply, 1990 to 2025: Scenarios and issues* (Vol. 19). Iwmi.

[2] Mekonnen, M. M., & Hoekstra, A. Y. (2016). Four billion people facing severe water scarcity. *Science advances*, *2*(2), e1500323.

[3] Sammut, C., & Webb, G. I. (Eds.). (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

[4] Biau, G., &Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197-227.

[5] Sánchez A, V. D. (2003). Advanced support vector machines and kernel methods. *Neurocomputing*, *55*(1-2), 5-20.

[6] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46).

[7] Menard, S. (2002). *Applied logistic regression analysis* (Vol. 106). Sage.

[8] Zhang, H., Berg, A. C., Maire, M., & Malik, J. (2006, June). SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (Vol. 2, pp. 2126-2136). IEEE.

[9] Freund, Y., & Mason, L. (1999, June). The alternating decision tree learning algorithm. In *icml* (Vol. 99, pp. 124-133).

[10] Holovaty, A., & Kaplan-Moss, J. (2009). *The definitive guide to Django: Web development done right*. Apress.

[11] Forcier, J., Bissex, P., & Chun, W. J. (2008). *Python web development with Django*. Addison-Wesley Professional.

[12] Chodorow, K. (2013). *MongoDB: the definitive guide: powerful and scalable data storage*. " O'Reilly Media, Inc.".

[13] Dede, E., Govindaraju, M., Gunter, D., Canon, R. S., & Ramakrishnan, L. (2013, June). Performance evaluation of a mongodb and hadoop platform for scientific data analysis. In *Proceedings of the 4th ACM workshop on Scientific cloud computing* (pp. 13-20).