**COVER PAGE AND TITLE PAGE**

**DECLARATION**

**CERTIFICATE**

**ABSTRACT-POORA NAHI HUA**

**ACKNOWLEDGEMENT**

**INDEX**

**LIST OF FIGURES**

**LIST OF TABLES**

**LIST OF SYMBOLS ABBREVATIONS AND NOMENCLATURE**

**CHAPTERS**

**CH1-Introduction-mention what we did in what chapter and Objectives**

**CH2-Related Studies**

**Chapter 3 - DESCRIPTION**

**3.1 Definitions**

**3.2- Design of Model**

**CH4-Experiments and evaluation metrics**

**Datasets**

**Environment**

**Algorithm**

**CH5-Performance results**

**APPENDICES**

**REFERENCES**

**Smart Agriculture Solutions using Machine Learning on Sensor Network Data**

1 Introduction 1

1.1 Overview . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.2 Motivation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.3 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

2 Related Work

3 Literature Review

WQI standard of WHO and Formula Derivation

Q value Normalization

Regression Algos

Mention Evaluation Metrics

Classification Algos

Evaluation Metrics

4 Proposed Work

4.1 Present Manual Workflow

4.1.1 FLOWCHART

4.1.2 Limitations

(GATHER SAMPLES FROM VARIOUS SOURCES->TEST IN LAB->RESULTS COSTLY and Time Consuming)

4.2 Proposed Workflow of the system

4.2.1 Datasets Used

4.2.2 Architecture

4.2.3 Reconstruction of Missing Data of Sensors

4.2.4 WQI prediction and Classification

4.2.5 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4.2.5 Algorithm and Architecture

5 Experiments and Evaluation Metrics 26

5.1 Experimental Environment . . . . . . . . . . . . . . . . . . . . . . . . . . 26

6 Conclusion and Future Work

**LIST OF SYMBOLS ABBREVATIONS AND NOMENCLATURE**

IOT-Internet of Things

ML-Machine Learning

WSN- Wireless Sensor Networks

SVR-Support Vector Regressor

MLP-Multilayer Perceptron

WQI-Water quality Index

ANN - Artificial Neural Network

SVM - Support Vector Machine

**LIST OF TABLES**

[Table 1] Weights of various parameters according to WQI-NSF

[Table 2] Classification of water sample based on WQI

[Table 3] Parameters measured by environment sensor

[Table 4] Water quality Datasets’ Parameters

**LIST OF FIGURES**

[Figure 1] Q values for various parameters

[Figure 2] Working of random forest regressor

[Figure 3] One hidden layer of MLP

[Figure 4] Example of an ANN

[Figure 5] Margins for an SVM trained with samples from two classes

[Figure 6] Visualization of a Random Forest

[Figure 7] Example of a decision tree

**CHAPTER 1 Introduction**

**1.1 Overview**

According to the economic survey of 2018[1] in India, around 50% of the workforce is involved in the agriculture sector. The contribution of the sector to the GDP is only 16% which has reduced significantly from 50% in 1950. The decline is not limited to India and is observed in the rest of the world. This low productivity depends on many factors, one of the major one is wastage of farming resources, money and time. Also, the majority of people involved in farming are from rural areas who are poor and have insufficient knowledge regarding farming practices. Each step in farming, be it preparation of soil, adding Fertilizers, irrigation or harvesting, requires proper analysis of soil nutrients, water quality, weather, sunlight etc for improving the productivity.

With improvement in technology, various farming equipment have been developed aiming to increase productivity and proper utilization of resources. The use of the Internet of Things, which consists of various devices connected through a network interacting with computers to transfer data, is becoming feasible with the internet and computers getting faster and cheaper. Various sensors can be placed in a farming environment to obtain large, real-time datasets such as irrigation water(pH, TDS, various chemical concentration), soil(moisture, pH, nutrients) , weather(temperature, humidity etc). These datasets can then be analyzed and used to train machine learning models that will help in proper utilization of resources, determining which crop could be suitable for a specific environment and increasing the overall efficiency of crop production.

Due to Soil and underwater sensors like moisture, pH, temperature, light, being cheaper, this work focuses on using soil, water and climate(obtained from public sources) to train various machine learning algorithms for helping in increased agricultural productivity. Chapter 2 focuses on related works that are being carried out independent to this project but some references are taken from those works with proper citations.

**OBJECTIVES**

**To be CTD.**

**Chapter 2 RELATED WORKS**

Many equipment have been developed for helping in agriculture that includes modern farming tools such as mower, sprayer, seed drill etc. Also, with recent developments in IOT and cloud computing, many new and cheap sensors have come out that can help in agriculture. Also, many methods have been developed which make use of these equipment for efficient farming. Precision agriculture[2] is the process developed which involves gathering various time series, geolocation and individual data, processing and analyzing them, and making use of them for improving productivity in agriculture. Any step of agriculture where data can be gathered can come under precision agriculture. For example, in irrigation systems data can be gathered and analyzed to find the quality of water, determining and classifying soil and fertilizing it accordingly which will help in better fertilizing.

Irrigation involves watering the soil after sowing the seed for proper growth of crops. Vij. et al[3] proposes the use of wireless sensor networks for measuring various parameters such as temperature, moisture, water level, weather etc. , passes it through Support vector regression(SVR), Random Forest Regression algorithms to classify soil type and predict the amount of water required for irrigation. Janani and Jebakumar[4] measure soil, plant and water data and pass it through a MLP to estimate the irrigation amount.

Before Irrigation, it is essential to measure the quality of water so as to prevent crops from getting damaged by polluted water. Orozco et al.[5] studied various samples of water for irrigation and developed a water quality index(WQI) measure. Lerios et al.[6] developed a method for prediction and classification of water according to parameters such as pH, FC etc. using various classifying techniques like Naive Nayes, Decision Tree, Random Forest, Gradient Boost and MLP.

Analysis of soil gives proper insights for fertilizing and irrigation of soil. Cai Y[7] uses a combination of meteorological and soil moisture data and uses a deep learning regressor network to determine weights for soil moisture prediction. Gholap[8] measures soil parameters like pH, Electrical Conductivity to classify soil using various algorithms and implementing automated soil sample classification. Suchitra et al.[11] classifies soil nutrients on the basis of mineral contents using classification techniques known as Extreme learning machines(feedforward neural networks) using various activation functions.

Sensors and wireless networks are susceptible to failure in a few cases, due to which some percentage of data may be missing. Gad I.[9] proposes the use of a deep learning imputation model using various optimizers (SGD, Adam, Rmsprop etc]) to compute missing parameters. Balducchi[10] uses a set of decision Tree and K nearest neighbours to predict the missing values.

The data gathered from farming sensors are also used in prediction of crop yield. Meeradevi[12] designed a system of WSN to gather various soil and irrigation equipment data along with time series data of crop yield, using that data to train a regression model to find relationships between various parameters and crop yield which would help in future prediction of crop yield.

**CHAPTER 3 LITERATURE REVIEW**

**Water Quality Index**

The Water Quality Index(WQI), formulated by Horton[13] is a linear function which gives the quality of water. The index is calculated using a weighted sum of various parameters such as pH, temperature etc. These parameters are selected such that they are available in all of the water sources that one wants to measure from. The WQI is calculated using four steps:

**Parameter Selection** **and Weights**

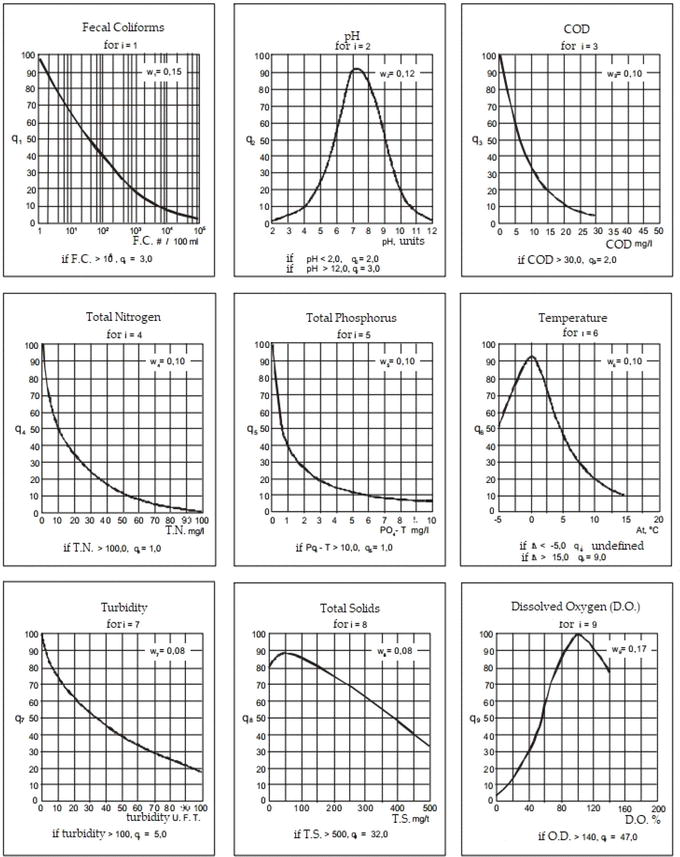
According to one of the most widely used index calculation, nine parameters are used to calculate WQI: Temperature, pH, turbidity, phosphate, Nitrate, total Solids, dissolved Oxygen(DO), biochemical Oxygen Demand(BOD) and fecal coliform. The weights for each of the parameters were obtained by the use of DELPHI technique[14] under WQI-NSF[15] as shown in table 1.

|  |  |
| --- | --- |
| Parameter | Weights |
| DO | 0.17 |
| Fecal coliforms | 0.15 |
| pH | 0.12 |
| BOD | 0.10 |
| Nitrate | 0.10 |
| Total phosphate | 0.10 |
| Temperature | 0.10 |
| Turbidity | 0.08 |
| Total solids | 0.08 |

**[Table 1] Weights of various parameters according to WQI-NSF**

**Q-value Normalization of Various Parameters**

The values of various parameters are normalized to be in the range of 0-100 for easy index calculation. The Q-values are calculated according to figure [1] .



**[Figure 1] Q values for various parameters**

**Formula for Water Quality Index**

After obtaining the weights and Q-values, WQI is calculated using the formula :

*Water Quality Index* **=**

Where, is the relative weight of feature *‘i’* obtained by dividing Each weight by total weight,

is the Q-value of i-th parameter

**Classification based in WQI**

Based on calculated WQI, the water sample can be classified for various uses. Meirels et al[23] proposed a water quality index as shown in [Table 2].

|  |  |  |  |
| --- | --- | --- | --- |
| WQI | Restrictions | Soil | Plant |
| 85–100 | No restrictions (NR) | It can be used for most soils with low probability of solidification and salinization | Most plants won’t be affected |
| 70–85 | Low restriction (LR) | Use for soil with fine texture or moderate permeability | Avoid use in plants with salt sensitivity |
| 55–70 | Moderate restriction (MR) | Can be used in soils with high or moderate permeability | Plants with moderate salt tolerance will be unaffected |
| 40–55 | High restriction (HR) | Can be used on soils with high permeability without layers of compaction. | It should be used to irrigate plants with moderate to high salt tolerance with special salinity control practices |
| 0–40 | Severe restriction (SR) | Use for irrigation under normal conditions should be avoided. | Avoided for all plants |

**[Table 2] Classification of water sample based on WQI**

**REGRESSION**

**Linear Regression**

Linear regression uses a linear approach to depict relationships between a scalar response and one or more variables[16]. This idea can be extended to predict multiple correlated dependent variables. The basic model of linear regression can be represented as:



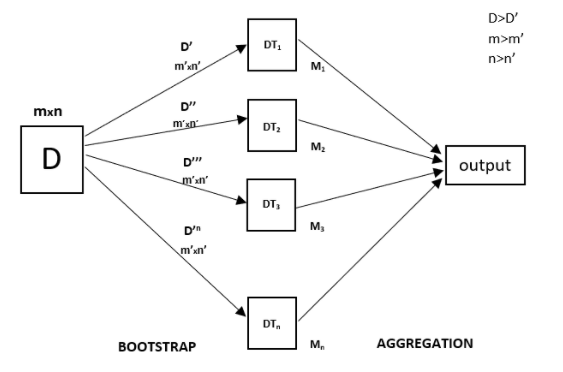
Mathematically it solves



This method is highly sensitive to the presence of outliers.

**Random Forest Regression**

Random forest regression uses the idea of random forests which is an ensemble learning method[17]. This method uses multiple decision trees to predict the output which in case of classification will be the output class and in case regression will be the output value. For regression the output is the mean value of the outputs. It relies on Bootstrap and Aggregation to compute results.



**[Figure 2] Working of random forest regressor**

**Gradient Boosting Regressor**

Gradient Boosting Algorithm uses weak learners and makes changes in it to construct a strong learner[18]. Gradient boosting uses decision trees as their weak learners. Regression trees are used for the weak learners, and these regression trees output real values. It uses an additive model that allows for optimization of differentiable loss function. It is quite powerful but is prone to overfitting.

**Polynomial Regression**

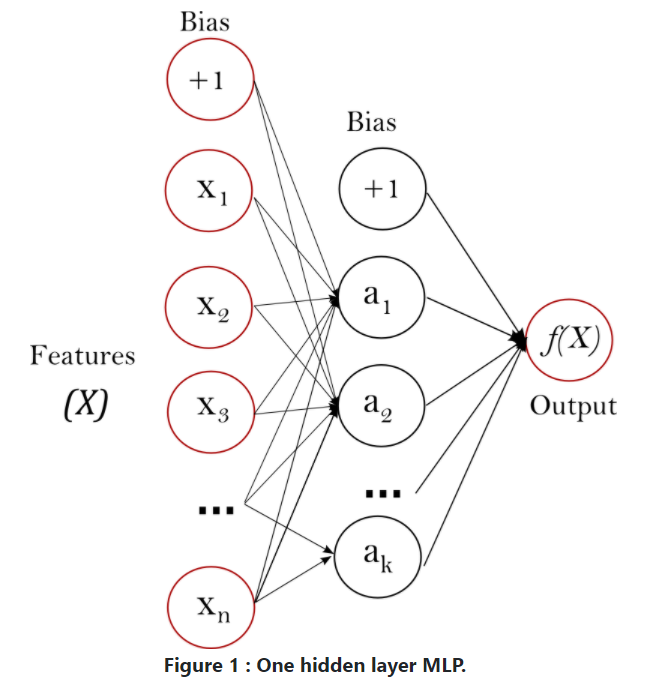
This technique is used when the relationship between input variables and output is non linear[19]. The computation statistically is similar to the linear regression method, due to which it is also known as the special case of multiple linear regression.

It uses the given below equation:

****

**MLP Regression**

MLP(Multilayer perceptron) represents a feedforward artificial neural network where each perceptron has a linear function and an activation function[20]. These perceptrons are the building blocks of the neural network.Multi-layer perceptron (MLP) is a conventional model of neural net, which is mostly used for classification, but it can be used for regression as well by not using an activation function in the perceptron.

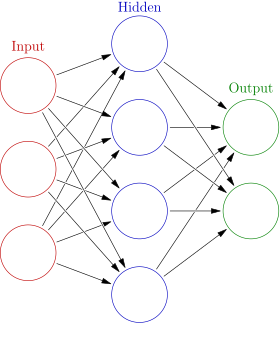
****

**[Figure 3] One hidden layer of MLP**

**CLASSIFICATION**

**ANN**

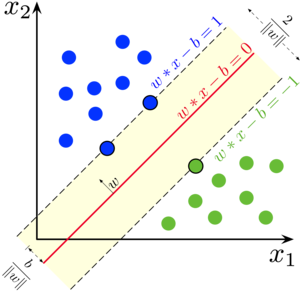
Artificial Neural Networks are inspired by the structure of our brain where dendrites receive the message which is passed through axon[21]. In ANN, neurons are responsible for receiving the input and producing the output after applying the activation function. Each neuron has a weight which increases or decreases as the learning proceeds. Typically, neurons are aggregated into layers. Different layers may perform different transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times.



**[Figure 4] Example of an ANN**

**SVM**

SVM presents one of the most robust prediction methods. Its objective is to find a hyperplane with maximum margin separation which distinctly classifies data points[22].Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. It can be used for classification, regression and also for outlier detection.

****

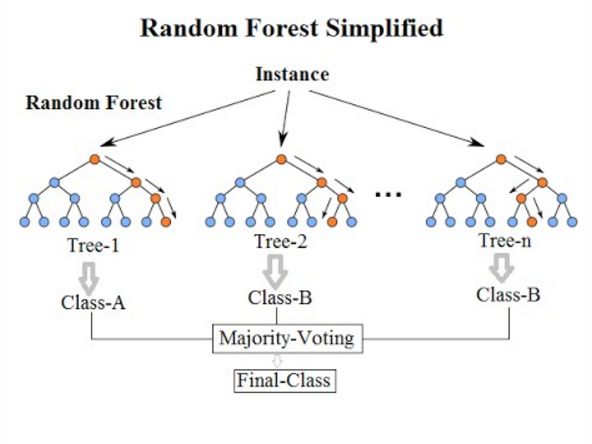
**[Figure 5] Margins for an SVM trained with samples from two classes**

**GRADIENT BOOST CLASSIFIER**

Gradient Boosting Algorithm uses weak learners and makes changes in it to construct a strong learner[18]. Gradient boosting uses decision trees as their weak learners. It uses an additive model that allows for optimization of differentiable loss function. It is quite powerful but is prone to overfitting.

**RANDOM FOREST CLASSIFIER**

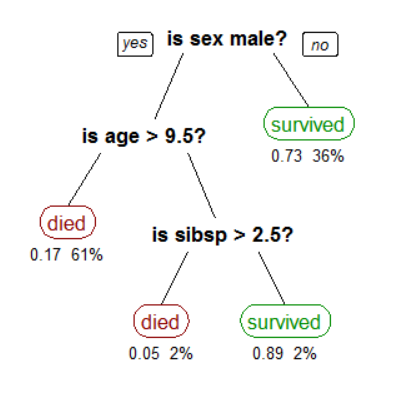
The random forest classifier uses multiple decision trees as an ensemble[17]. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction. The classifier is based on the idea that a group of classifiers outperforms a single classifier.The reason for this is that trees protect each other from their individual errors.



**[Figure 6] Visualization of a Random Forest Model**

**DECISION TREE**

A decision tree is a flowchart like structure where nodes represent an if else condition, each branch represents the outcome and leaf nodes represent the actual class label[24]. Here, the path from root to leaf gives us the classification rule for that class. Commonly used algorithms for splitting are: Gini impurity, Chi-Square and Information Gain.



**[Figure 7] Example of a decision tree**

**DATASETS(Change Headings)**

**Environment Data**

**Source**:[Environmental Sensor Telemetry Data](https://www.kaggle.com/garystafford/environmental-sensor-data-132k)

The data is generated by three identical sensor networks. Each of the sensor sets are placed under different environmental conditions:

1. Network 1- cooler and humid
2. Network 2- highly variable temperature and humidity
3. Network 3- warmer and dryer

The following are the parameters measured by the sensor networks:

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Description** | **Units** |
| deviceID | Unique device ID | string |
| co | Carbon Monoxide | ppm(%) |
| humidity | humidity | percentage |
| temp | Temperature | Celsius |
| smoke | smoke | ppm(%) |
| lpg | Liquid Petroleum Gas | ppm(%) |

**[Table 3] Parameters measured by environment sensor**

**Indian Water Quality**

**Source:** [**National Water Quality Monitoring Programme (NWMP)**](http://www.cpcbenvis.nic.in/water_quality_data.html)

Data gathered by various underwater sensors and lab tests by Indian Government under NWQMP from various water bodies. This dataset contains the following parameters:

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Description** | **Units** |
| Temp | Temperature | Degree Celsius |
| D.O. | Total Dissolved Oxygen | mg/l |
| PH | Potential Hydrogen | None |
| Conductivity | Electrical Conductivity | µmho/ Cm |
| B.O.D | Biological Oxygen Demand | mg/l |
| Nitrate | Total Nitrate(NO3-) | mg/l |
| FC | Fecal Coliform | MPN/100ml |
| TC | Total Coliform | MPN/100ml |

**[Table 4] Water quality Datasets’ Parameters**

**Prediction Of Missing Data**

Data Kya hai

Miss kaise Ho sakta hai

Humne kaise use kiya

**REFERENCES**

[1] India economic survey 2018

<https://www.financialexpress.com/budget/india-economic-survey-2018-for-farmers-agriculture-gdp-msp/1034266/>

[2] Precision Agriculture, International Society for Precision Agriculture

<https://www.ispag.org/about/definition>

[3] Anneketh Vij, Singh Vijendra, Abhishek Jain, Shivam Bajaj, Aashima Bassi, Arushi Sharma, IoT and Machine Learning Approaches for Automation of Farm Irrigation System, Procedia Computer Science, Volume 167, 2020, Pages 1250-1257, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.440.

[4] Janani M and Jebakumar R. A Study on Smart Irrigation Using Machine Learning. Cell & Cellular Life Sciences Journal ISSN: 2578-4811, 2019, DOI:10.23880/cclsj-16000141

[5]De La Mora-Orozco C, Flores-Lopez H, Rubio-Arias H, Chavez-Duran A, Ochoa-Rivero J. Developing a Water Quality Index (WQI) for an Irrigation Dam. *Int J Environ Res Public Health*. 2017;14(5):439. Published 2017 Apr 29. doi:10.3390/ijerph14050439

[6] Jefferson L. Lerios, Mia V. Villarica, Pattern Extraction of Water Quality Prediction Using Machine Learning Algorithms of Water Reservoir,International Journal of Mechanical Engineering and Robotics Research Vol. 8, No. 6

[7] Research on soil moisture prediction model based on deep learning

Cai Y., Zheng W., Zhang X., Zhangzhong L., Xue X.

(2019) *PLoS ONE*, 14 (4) , art. no. e0214508

[8]Gholap, Jay & Ingole, Anurag & Gohil, Jayesh & Gargade, Shailesh & Attar, Vahida. (2012). Soil Data Analysis Using Classification Techniques and Soil Attribute Prediction. International Journal of Computer Science Issues. 9.

[9] Gad, I., Hosahalli, D., Manjunatha, B.R. et al. A robust deep learning model for missing value imputation in big NCDC dataset. Iran J Comput Sci (2020). <https://doi.org/10.1007/s42044-020-00065-z>

[10] Fabrizio Balducci, Donato Impedovo \* ID and Giuseppe Pirlo Dipartimento di Informatica, Università degli studi di Bari Aldo Moro, 70125 Bari, Italy; fabrizio.balducci@uniba.it (F.B.); giuseppe.pirlo@uniba.it (G.P.)

[11] M.S. Suchithra, Maya L. Pai, Improving the prediction accuracy of soil nutrient classification by optimizing extreme learning machine parameters, Information Processing in Agriculture, Volume 7, Issue 1, 2020, Pages 72-82, ISSN 2214-3173, https://doi.org/10.1016/j.inpa.2019.05.003.

[12] Meeradevi and Monica R Mundada, Automated Control System for Crop Yield Prediction using Machine Learning Approach, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 2 (2019) pp. 480-484

[13]

[13]Horton RK. An index number system for rating water quality. Journal of Water Pollution Control Federation. 1965;37(3):300-305

[14]Grisham, Thomas. (2009). The Delphi technique: A method for testing complex and multifaceted topics. International Journal of Managing Projects in Business. 2. 10.1108/17538370910930545.

[15]Brown RM, McClelland NI, Deininger RA, Tozer RG. A water quality index—Do we dare? Water & Sewage Works. 1970;117:339-343

[16] Francis Galton. "Regression Towards Mediocrity in Hereditary Stature," *Journal of the Anthropological Institute*, 15:246-263 (1886)

[17] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

[18] Friedman, J. H. (February 1999). ["Greedy Function Approximation: A Gradient Boosting Machine"](https://statweb.stanford.edu/~jhf/ftp/trebst.pdf)

[19] [Gergonne, J. D.](https://en.wikipedia.org/wiki/Joseph_Diaz_Gergonne) (November 1974) [1815]. "The application of the method of least squares to the interpolation of sequences". *Historia Mathematica* (Translated by Ralph St. John and [S. M. Stigler](https://en.wikipedia.org/wiki/Stephen_M._Stigler) from the 1815 French ed.). **1** (4): 439–447. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1016/0315-0860(74)90034-2](https://doi.org/10.1016%2F0315-0860%2874%2990034-2)

[20] Günther, F.; Fritsch, S. Neuralnet: Training of neural networks. R J. 2010, 2, 30–38.

[21] Daniel G.G. (2013) Artificial Neural Network. In: Runehov A.L.C., Oviedo L. (eds) Encyclopedia of Sciences and Religions. Springer, Dordrecht. <https://doi.org/10.1007/978-1-4020-8265-8_200980>

[22] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

[23] Meireles ACM, Andrade EM, Chaves LCG, Frischkorn H, Crisostomo LA. A new proposal of the classification of irrigation water. Revista Ciência Agronômica. 2010;41(3):349-357. DOI: 10.1590/S1806-66902010000300005

[24] Fürnkranz J. (2011) Decision Tree. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8\_204