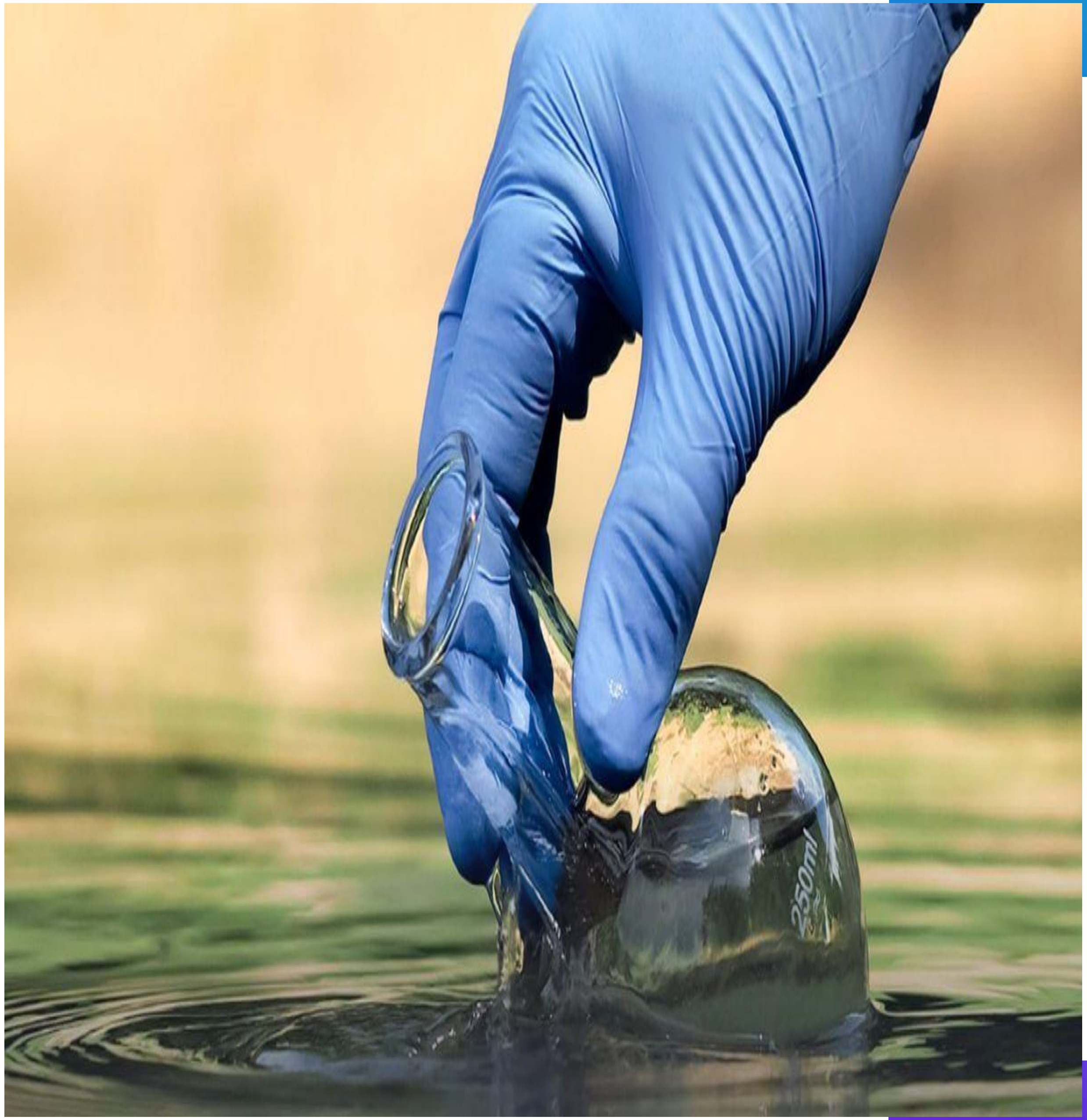


WATER QUALITY ANALYSIS

BATCH MEMBER

611721205701:DEEPAK KUMAR P

PHASE 4 SUBMISSION DOCUMENT

PROJECT TITLE:*WATER QUALITY ANALYSIS*

PHASE 4:*DEVELOPMENT PART 2*

TOPIC:*creating visualizations and building a predictive model.*

WATER QUALITY ANALYSIS

INTRODUCTION

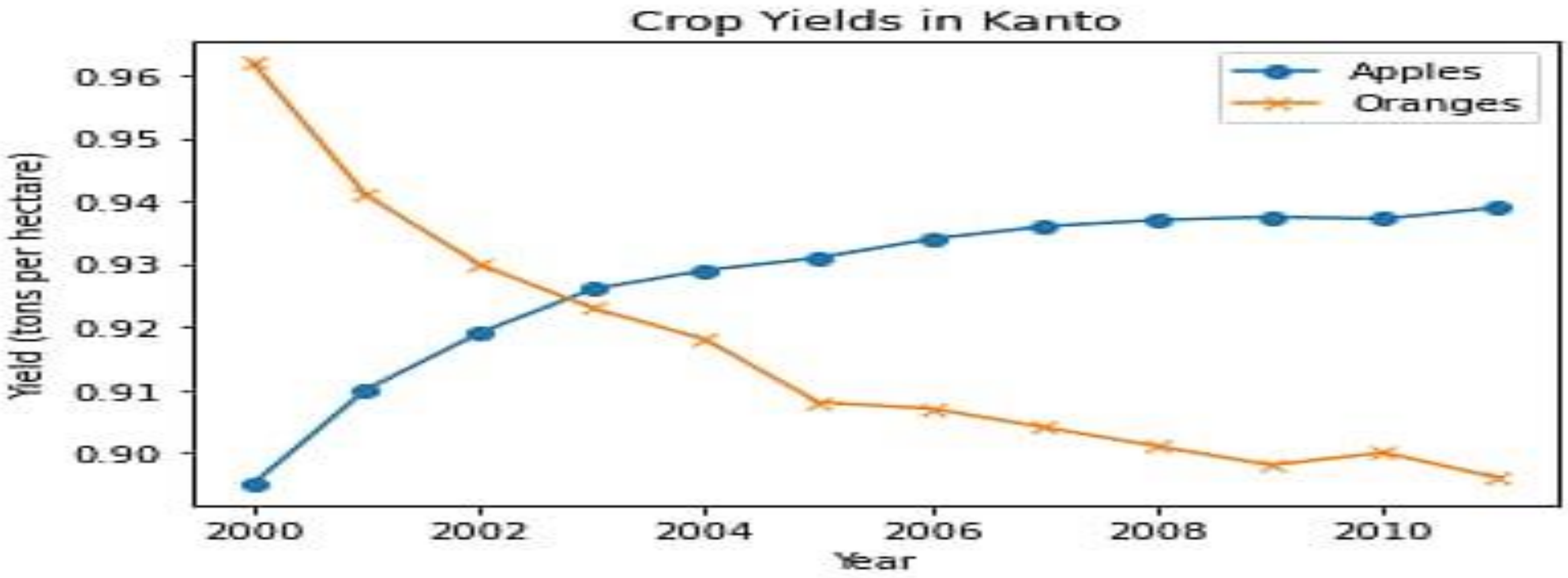
- The process of building a predictive model involves defining the problem, collecting and preprocessing data, performing exploratory analysis, selecting and training the model, evaluating and refining its performance, and finally deploying it into production.

VISUALITATION LIBRARIES

- Python offers several plotting libraries, namely Matplotlib, Seaborn and many other such data visualization packages with different features for creating informative, customized, and appealing plots to present data

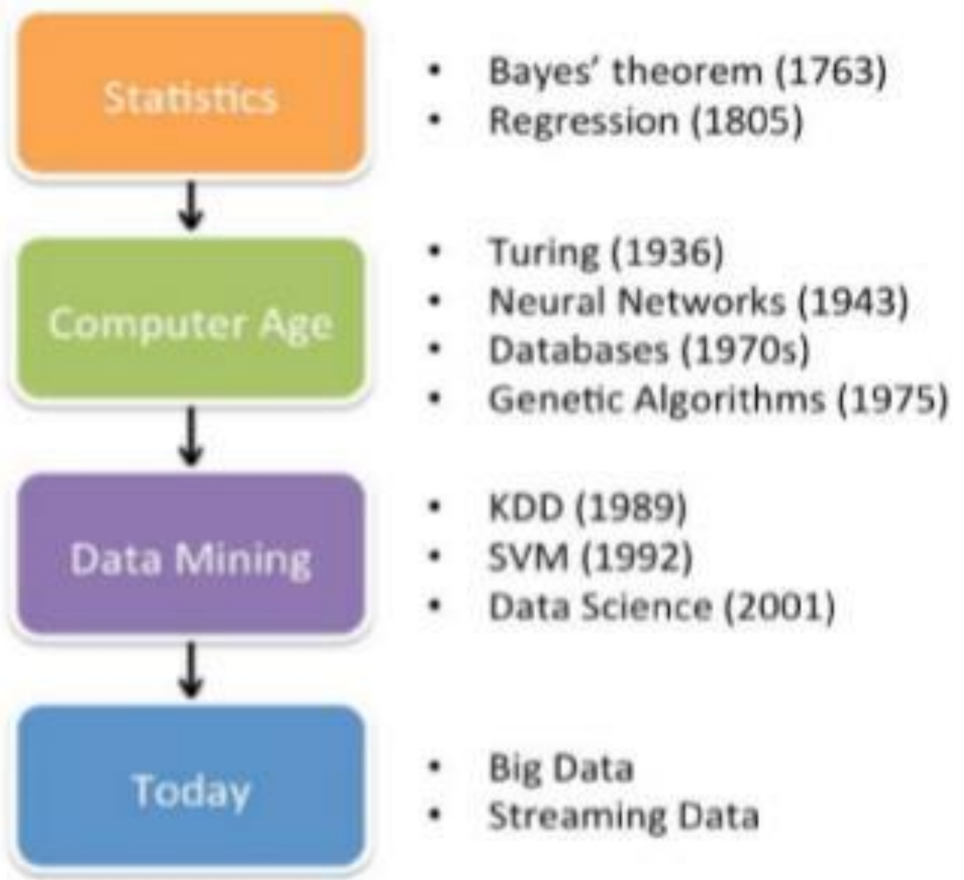
```python
plt.plot(years, apples, marker='o')
plt.plot(years, oranges, marker='x')

plt.xlabel('Year')
plt.ylabel('Yield (tons per hectare)')

plt.title("Crop Yields in Kanto")
plt.legend(['Apples', 'Oranges'])
```

```
<matplotlib.legend.Legend at 0x2054a8940d0>
```

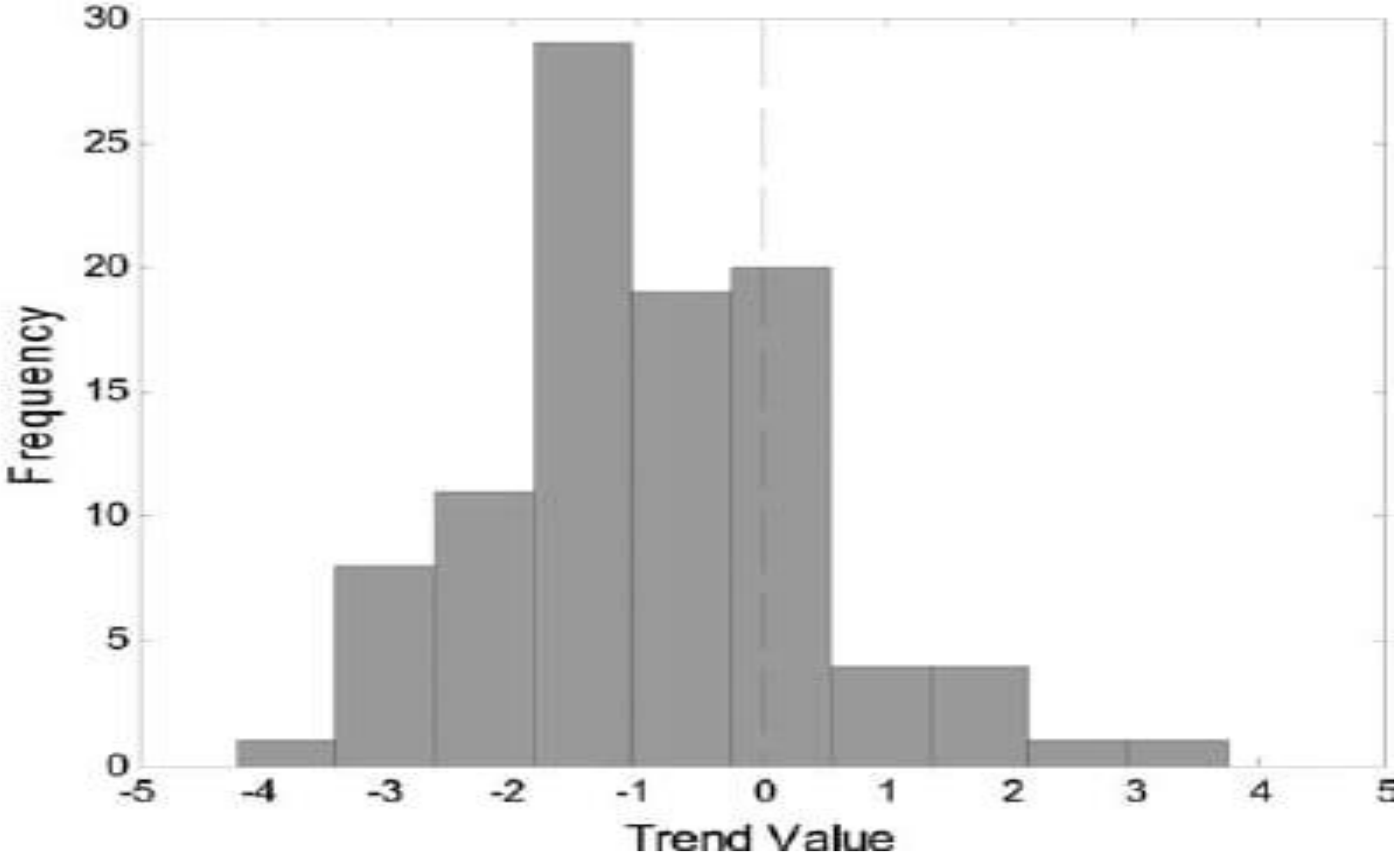


DATA MINING

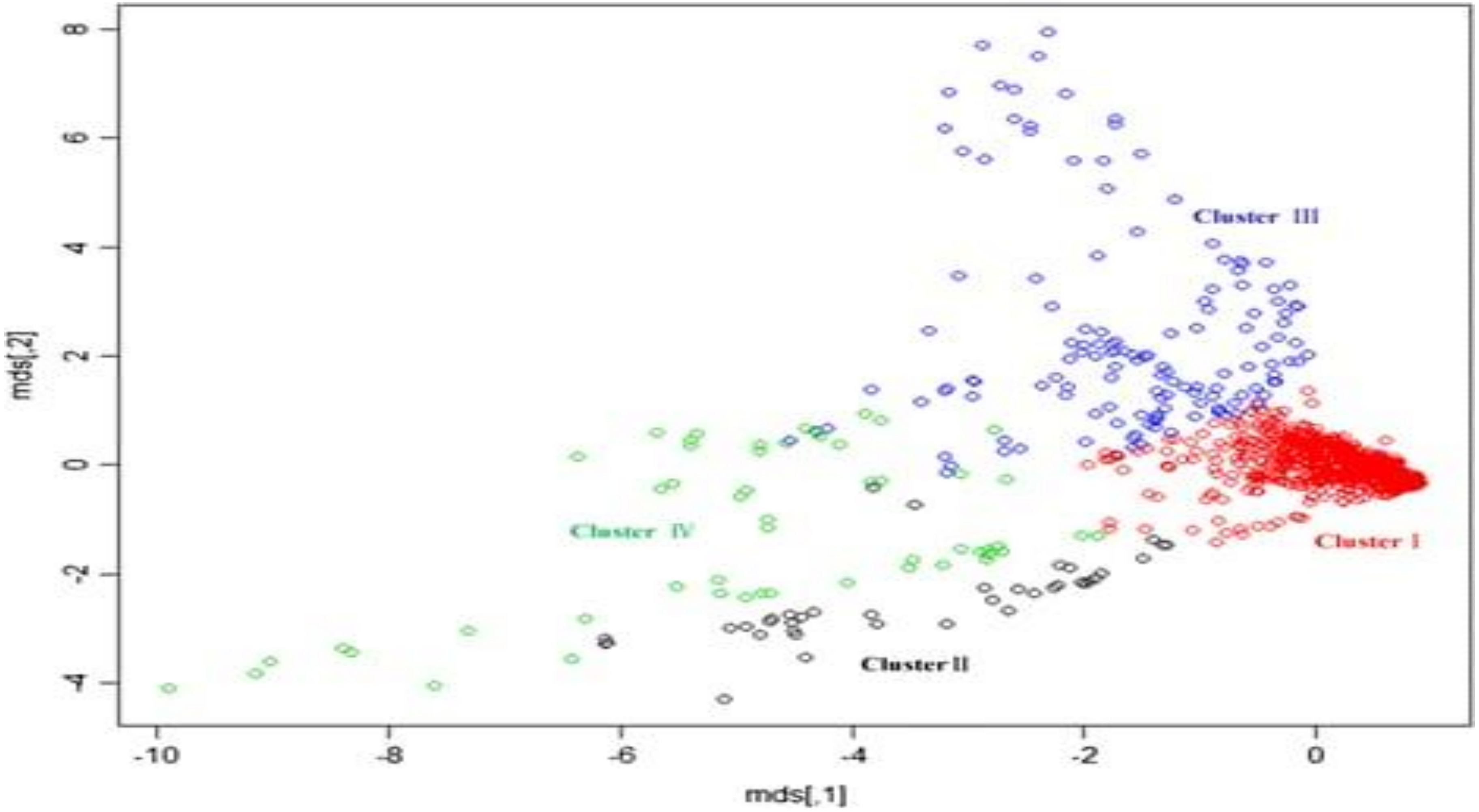
- Formally, the beginning of data analysis field had begun at the moment humanity started mak-
- ing simplest analysis of the surrounding environment by watching and manually interacting
- with nature. For data mining itself, there are some more or less consistent and defined events
- in history that are associated with the birth of the discipline, such as publishing of Bayes' the-
- orem (which describes the probability of an event, based on conditions that might be related
- to the event) by Thomas Bayes’ in 1763 and first regression analysis by Adrien-Marie Legen-
- dre and Carl Friedrich Gauss in 1805 (Figure 1.1). (Li 2015


Statistics
Bayes' theorem (1763)
Regression (1805)
Computer Age
Turing (1936)
Neural Networks (1943)
Databases (1970s)
Genetic Algorithms (1975)
Data Mining
KDD (1989)
SVM (1992)
Data Science (2001)
Today
Big Data
Streaming Data

HISTOGRAM GRAP FOR WATER QUALITY ANALYSIS



SCATTER PLOTS

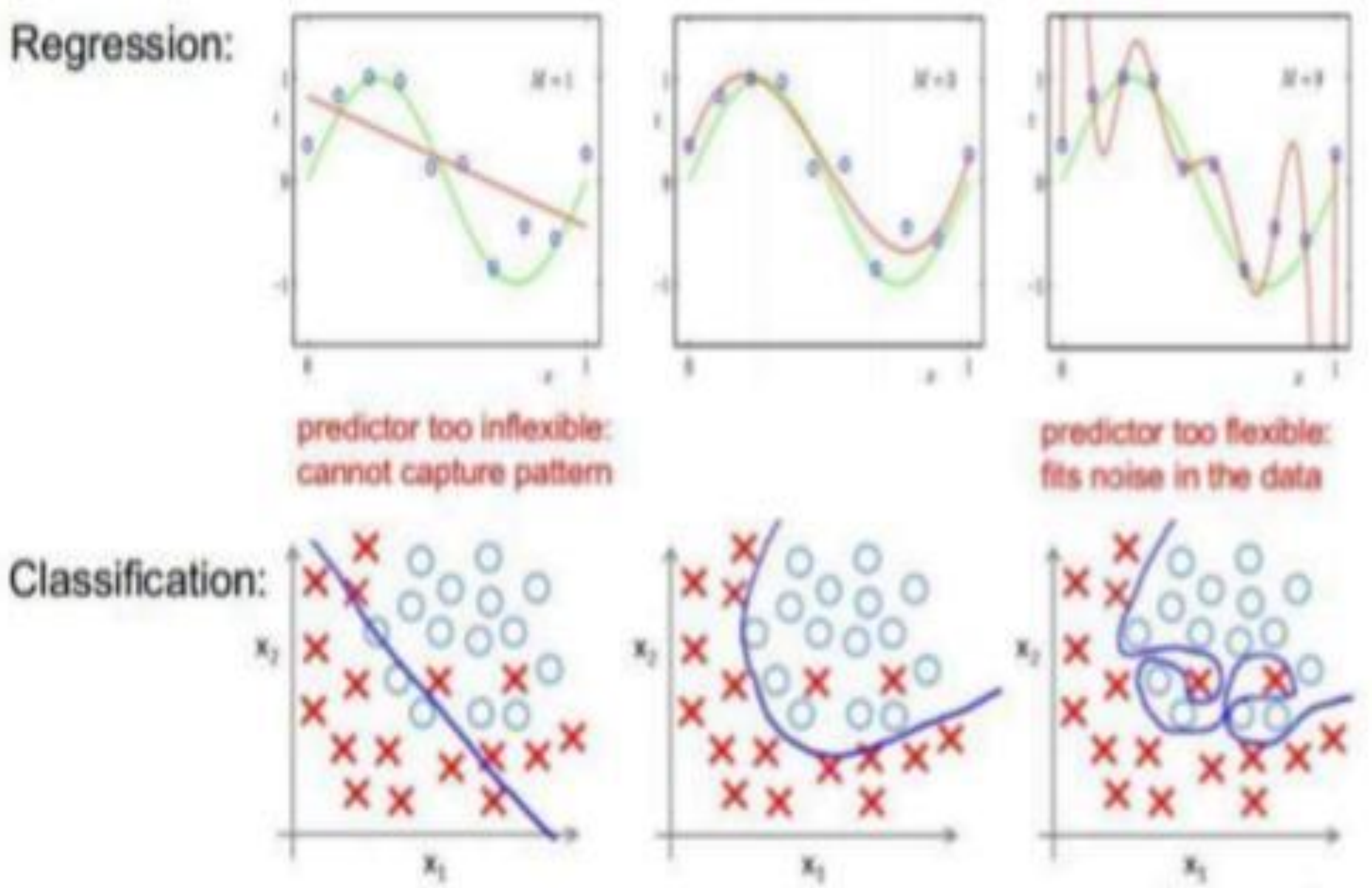


OVERFITTING

- Overfitting is a common problem among a vast majority of the machine learning algorithms
- and it comes from the origins of these tools. While fitting a model, algorithm tries to
- minimize the error function of the model, which is the difference between the estimator and
- what is estimated (Lebanon 2010).

THREE TYPICAL TYPES OF MODEL FIT

- **1. Predictor too inflexible** – these models are underfitted, which means that the
- formula describes data poorly and mean square error (MSE) is significantly high.
- These models can't provide sufficial degree of accuracy
- **2. Middle pictures** – this model has nearly a perfect fit. The MSE is not absolutely
- zero; however, this model will provide an appropriate function to the given data
- and will be able to make suitable predictions based on new data (or data not
- included in training set)
- **3. Predictor is too flexible** – in these cases, we can see a typical example of
- overfitting. The model is too complicated and it is able to minimize MSE nearly to
- zero, so it seems to fit the training set alone perfectly. However, it will describe
- random errors or noise and thus show pretty bad results on the n


Regression:
predictor too inflexible:
cannot capture pattern
predictor too flexible:
fits noise in the data
Classification:

CALCULATING ACCURACY

- Goals of the evaluation stage can’t be effectively met without proper ways to interpret the
- achieved results. For different models and different types of tasks, various techniques may be
- used for this purpose. During this research, the following ones were used: confusion matrix,
- root mean square error and node purity.
- Confusion matrixes are one of the most commonly used tools for evaluating a performance of
- classification models. It contains information about the actual and predicted classes.

MATRIX AND METHODS

+ DATA

The data used for this research was generated during European STREAMES (STream REAch Management, an Expert System) project, which is an international enterprise for the development of a knowledge-based environmental decision support system to assist water managers with their decision-making tasks. The core of the project itself involved the evaluation of the effect of substantial nutrient loads on the overall water quality and ecological status of stream ecosystems. Empirical data for the knowledge base come from several streams located throughout Europe and Israel, with emphasis on streams from the Mediterranean region. These data comprise several types of variables, including physical, chemical and biological parameters. (Vellido, et al. 2007

Table 2.1 Catchment characteristics of the chosen streams (Vellido, et al. 2007)

Stream	Dominant Geology	Climate	Catch ment area (km^2)	Stream length (km)	Altitudinal range (m a.s.l.)	Land-use (%)		
						Arable and grass-land	Forest and open land	Urban
Tordera (Spain)	Siliceous	Mediterra-nean	80.2	21.7	1100-190	10.8	87.4	1.8
Grandola (Portugal)	Siliceous	Mediterra-nean	54.9	40.1	258-11	15.8	83.0	1.2
Apose-lemis (Greece)	Calcareous	Mediterra-nean	19.6	4.6	902-240	42.3	57.1	0.4
Montagut (France)	Calcareous	Atlantic	12.9	8.0	620-320	49.4	50.6	0
Bagnatore (Italy)	Calcareous	Mediterra-nean	11.0	5.0	828-470	59.7	36.0	4.4
Erpe (Germany)	Siliceous	Sub-continental	207	20.0	65-38	60.0	21.0	19.0

Gurri (Spain)	Calcareous	Mediterra-nean	37.7	14.3	1140-503	60.7	35.2	4.0
Lezat (France)	Calcareous	Atlantic	226.1	44.0	620-207	79.0	20.9	0.1
Demnitzer (Germany)	Siliceous	Mediterra-nean	15.0	6.2	67-60	100	0	0

For this particular study, out of all 52 variables, the most significant 29 variables were chosen during the data preparation process. These variables are presented in Table 2.2.

Table 2.2 List of the 29 variables selected for the study, grouped by their topology (Vellido, et al. 2007)

Type	Variable	Description
Ion Concentrations (chemical)	Cations	$Na^{+} + K^{+} + Mg^{2+} + Ca^{2+} + NH^{+}{}_{4}$ (Concentration in meq/l)
	Anions	$Cl^{-} + SO^{-}{}_{4} + NO^{-}{}_{3}$ (Concentration in meq/l)
	Alkalinity	(Concentration in meq/l)
Nutrient Concentration (chemical)	NH_4^{+}-N	Ammonium (concentration in mgN/l)
	NO_3^{-}-N	Nitrate (concentration in mgN/l)
	PO_4^{3}-N	Phosphate (concentration in mgP/l)
	D.O.C.	Dissolved Organic Carbon (Concentration in mg/l)
	Conductivity	In µS/cm
	D.I.N.	Dissolved Inorganic Nitrogen (in mgN/l)
Hydrological, Hydrau-lic & Morphologic (physical)	Depth	Wet channel average depth (m)
	Wet Perimeter	Cross-sectional area divided by depth
	Substrate Ratio	Percentage of (Cobbles þ Pebbles) substrata, divided by percentage of (Gravel þ Sand þ Silt) substrata
	Wet Perimeter: Depth Ration	Ratio between Wet Perimeter and average Depth (unitless)
	K1	Water transient storage exchange coefficient: from water column to transient storage zone (in s^{-1})
	K2	Water transient storage exchange coefficient: from transient storage zone to water column (in s^{-1})
	Transient Storage Ratio	K1/K2

	Froude number	$v/(g*D)^{1/2}$, where v is Average Water Velocity as defined below, g is the gravitational acceleration and D is the hydraulic depth
	Reynolds number	$(v*D)/KV$, where v and D as above and KV is the kinematic viscosity
	Discharge	In m^3/s
	Average Water Velocity	In m/s
	Manning's Coefficient	$(h^{2/3}*s^{1/2})/v$, where v as above, h is the wet channel depth and s is the reach slope
Stream Metabolism & Biofilm (biological)	Respiration	Daily rate of ecosystem respiration (in g O_2/m^2)
	G.P.P.	Daily rate of gross primary production (in g O_2/m^2)
	G.P.P.:R	G.P.P. to Respiration ratio (unitless) per day
	Daily Light (P.A.R.)	In mol/m^2
	Temperature	Average temperature at midday (in ^{O}C)
	D.O. Range	Daily variation in dissolved oxygen concentration (in mg O_2/l)
	Chlorophyll	In mg/m^2
	Biomass	In $mgAFDM/m^2$ (AFDM: Ash-Free Dry Mass)

The chosen dataset contains an average level of 5.3% missing values. Some of the variables were dropped due to high amount of the missing data, which makes imputation process useless and some of the variables were dropped due to their obvious meaningless for the analysis. In addition to the explained variables, some basic information about the measurements was included in order to show the examples of fitting the classification algorithm: *season* of the measurement and *land use* (forested or agricultural).

MODEL AND SOFTWARE

- There is a huge variety of machine learning algorithms and tools existing nowadays. In the
- following subchapters, the following algorithms used in this research are covered: support
- vector machines, random forests, artificial neural networks (used for classification, regression,
- variable importance tasks), k-nearest neighbours (used for data imputation) and k-means clus-
- tering (used for unsupervised classification).

SUPPORT VECTOR MACHINE

- Support vector machine is one of the basic algorithms and in this research is used mostly as a
- baseline in order to be able to compare the performances of the models. The core of this algo-
- rithm refers to the family of linear models. The model is trained by transfering of the original
- vector in the space of higher dimension and search for dividing hyperplane with the maximum
- gap in this space. Two parallel hyperplanes are constructed on both sides of the hyperplane
- separating classes. Separating hyperplane is a hyperplane that maximizes the distance to two
- parallel hyperplanes. The algorithm works on the assumption that the greater the difference
- and the distance between these parallel hyperplanes, the smaller the average error of the classifier

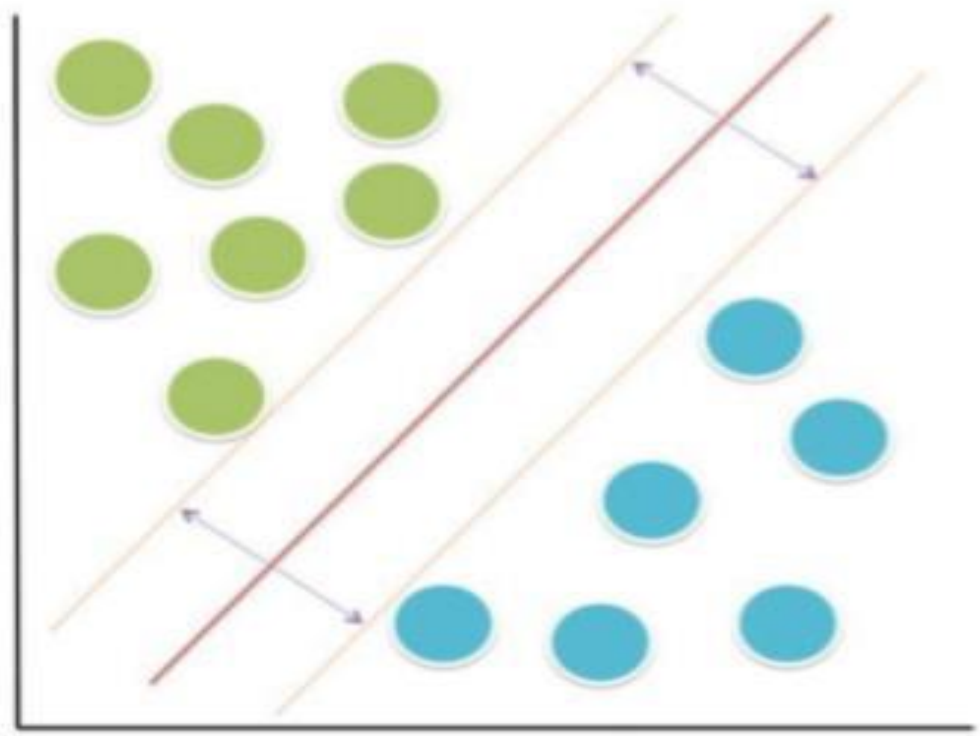


Figure 2.1 General graphical scheme of SVM algorithm. The model based on two hyperplanes with the maximum distance separates *green* and *blue* classes

RONDAM FOREST

+ Random forest (RF) is a relatively new model, developed by Leo Breiman and Adele Cutler
+ in the end of 90s. The general graphical scheme of the RF algorithm is sketched in Figure 2.2.
+ At each split of the observed sample data, a random subset of variables is selected and the
+ process is repeated until the specified number of decision trees is generated. Each tree is built
+ from a bootstrap sample drawn with replacement from the observed data, and the predictions
+ of all trees are finally aggregated through majority voting. A feature of RFs is the definition of
+ an out-of-bag (OOB) error, which is calculated from observations that were not used to build
+ a particular tree; it can thus be considered as an internal cross-validation error measure. This
+ is an important feature for the type of experiments carried out in this study, because it simpli-
+ fies the otherwise cumbersome cross-validation procedures that would be required if alterna-
+ tive classification methods such as, for instance, support vector machines or artificial neural
+ networks were used. (Breiman 2001, Shkurin and Vellido 2016)

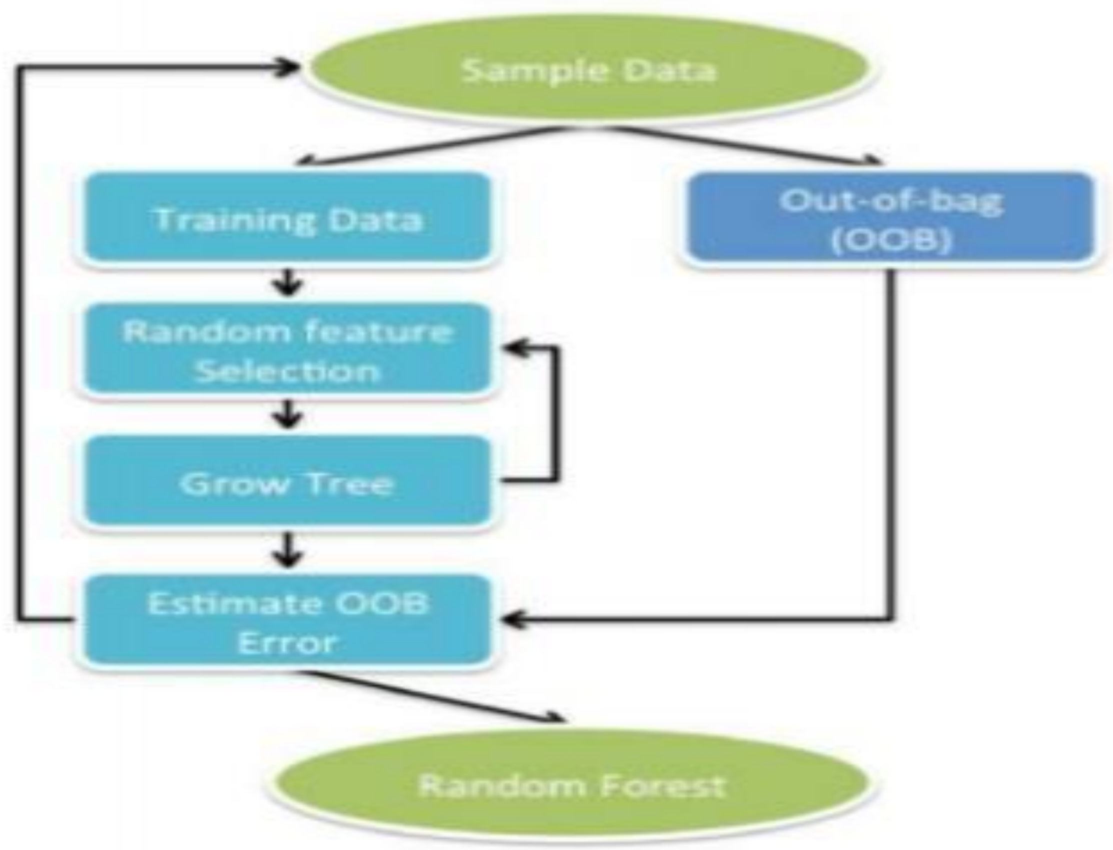


Figure 2.2 General graphical scheme of RF algorithm

ARTIFICIAL NEURAL NETWORK

- The artificial neural networks (ANN) are one of the most popular machine learning models
- nowadays, with a huge variety of possible applications, including regression, classification, image recognition etc, introduced in 1943 by neurophysiologist Warren McCulloch and math-
- ematician Walter Pitts (Warren and Pitts 1943). The basic of this model is in building several
- layers that are made up of a number of interconnected nodes, containing the activation func-
- tion. The training set is presented to a model through input layer; one or more hidden layers
- perform processing by the system of weighted connections, taking each of the inputs for cal-
- culation, and finally output layer gives the fitted function

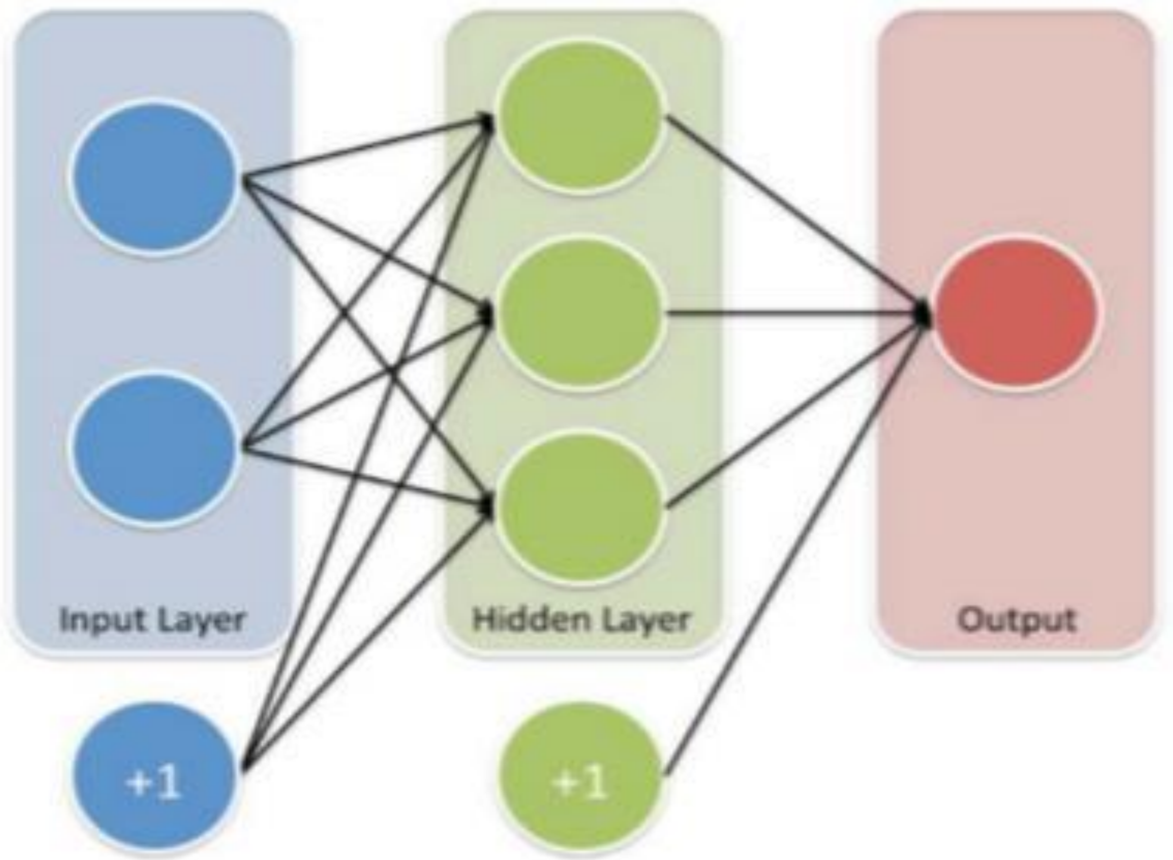


Figure 2.3 General graphical scheme of ANN model; *+1 nodes* represent bias units that help shifting the activation function depending on the task

K-MEANS CLUSTERING

- One of the useful analytical tools is unsupervised clusterization, where the data is classified
- by the algorithm into specified amount of classes based on internal patterns. It can be used to
- search for the subtypes and subclasses for researched process, value or compound. (Likas,
- Vlassis and Verbeek 2003)
- The data is classified firstly by setting k centroids, which will be the core to the searched clas-
- ses. Then, the grouping is done by minimizing the sum of squares of distances (analogue to
- MSE) between data and the corresponding cluster centroid, as shown on the Figure 2.5. At
- each iteration cluster centre is recalculated until the best position is reached. (Teknomo 2007)

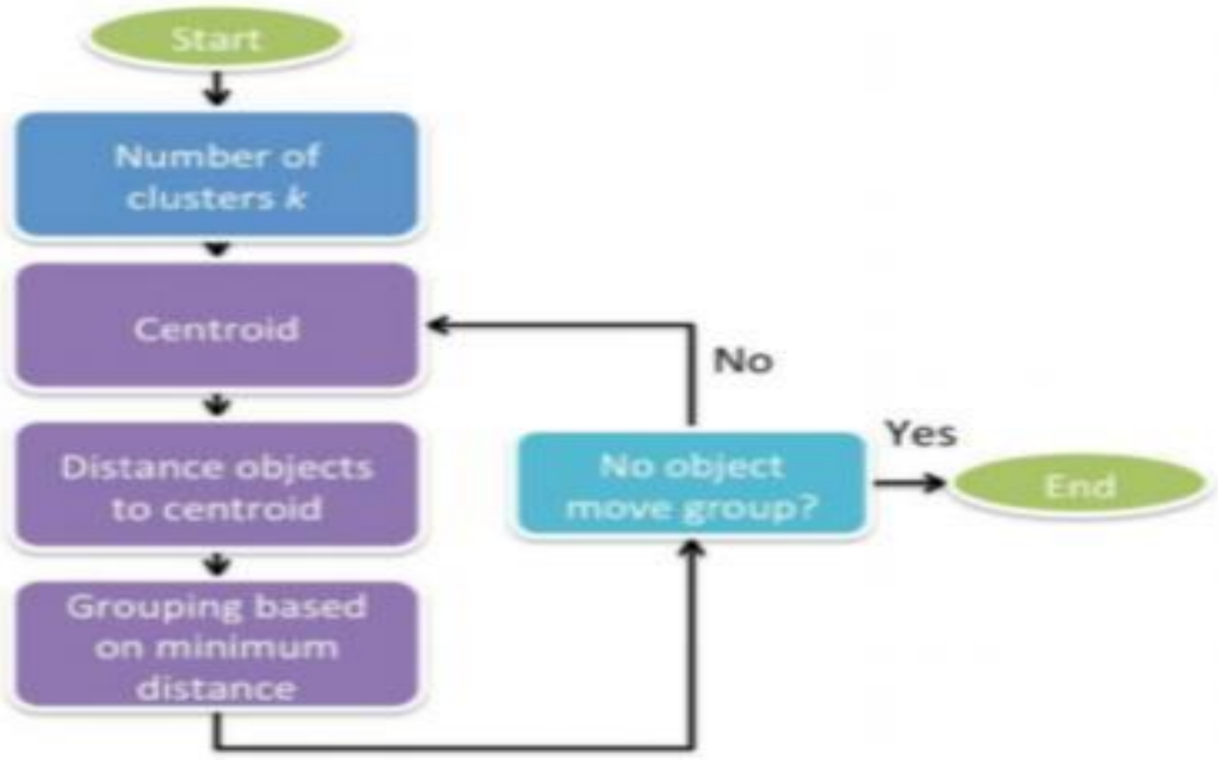


Figure 2.5 k-means clustering algorithm

DATA IMPUTATION

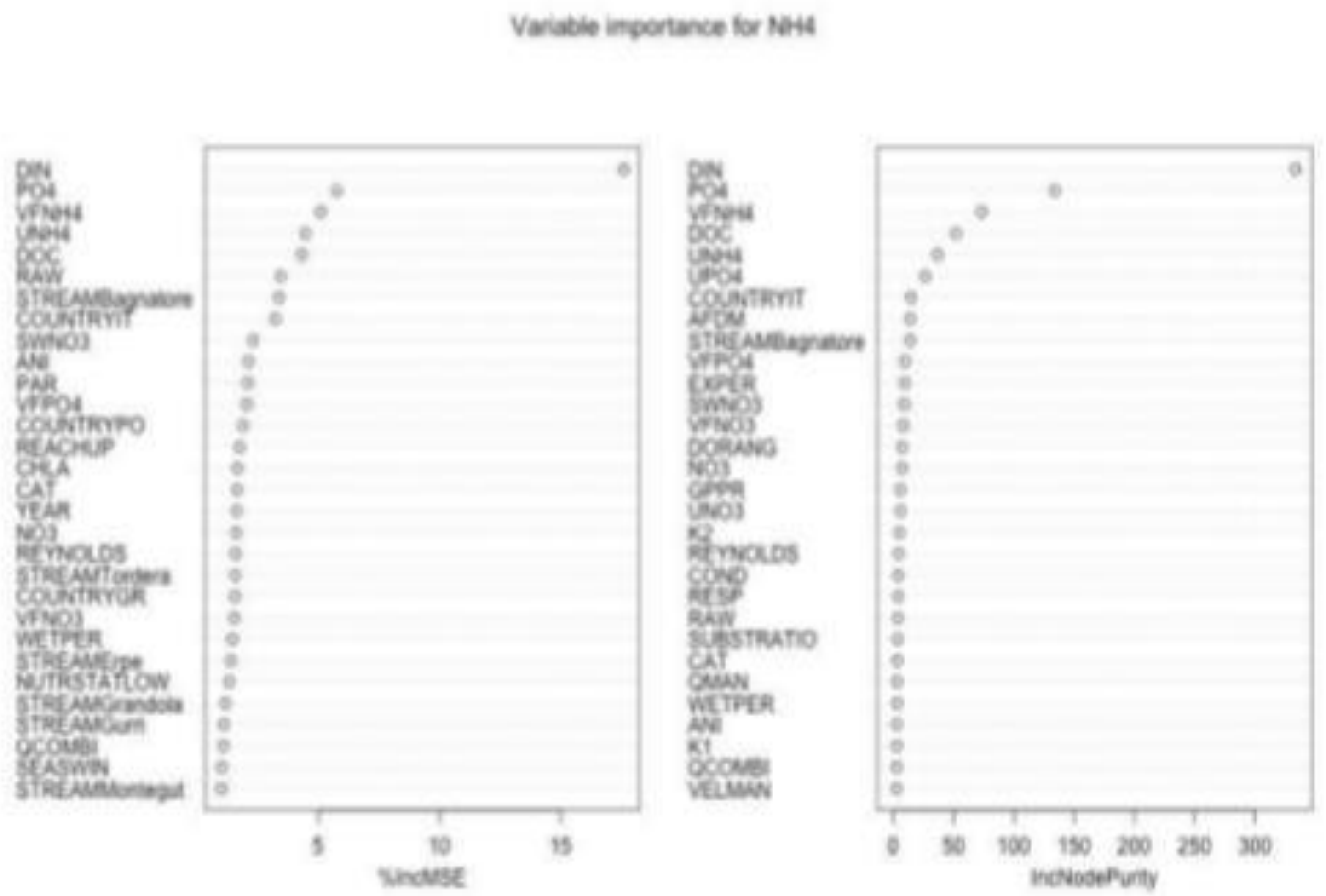
+ First of all, the data was imputed. In the original data, missing values are presented as “0” and
+ “-1” values. After these were assigned to NA, KNN with 2, 3 and 4 neighbours were used for
+ imputation of the results. The accuracies of the models based on these numbers of neighbours
+ for NH_4, NO_3 and PO_4 values are presented

Table 3.1 Results of data imputation with different values for *k*

Measurement	*k* value of neighbours	Accuracy of variables explained (in %)
NH_4	2	48.73
	3	48.6
	4	48.76
NO_3	2	76.18
	3	79
	4	78.71
PO_4	2	64.17
	3	64.23
	4	61.94

NH_4

- For NH4, the initial model using random forest had an RMSE of 1.8449 and accuracy of
- 48.6%. The variable importance of all the measurements were subtracted

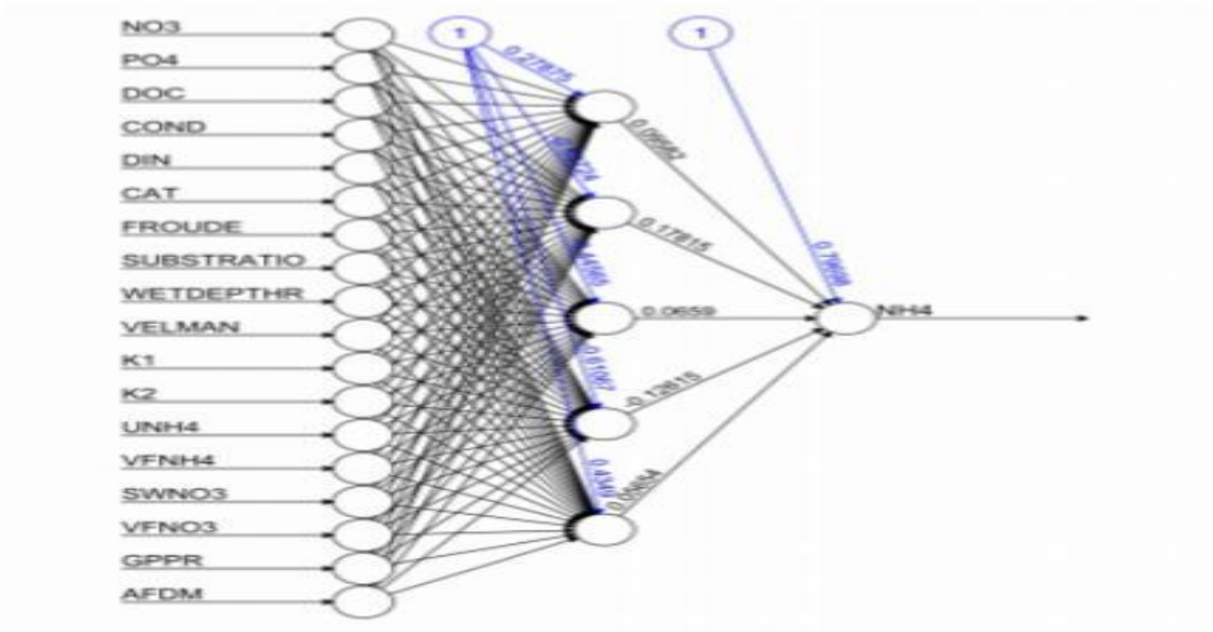


Variable importance for NH4
DIN
PO4
VFNH4
UNH4
DOC
RAW
STREAMBagnatore
COUNTRYIT
SWNO3
ANI
PAR
VFPO4
COUNTRYPO
REACHUP
CHLA
CAT
YEAR
NO3
REYNOLDS
STREAMTordera
COUNTRYGR
VFNO3
WETPER
STREAMErpe
NUTRSTATLOW
STREAMGrandola
STREAMGurri
QCOMBI
SEASWIN
STREAMMontegut
5
10
15
%IncMSE
DIN
PO4
VFNH4
DOC
UNH4
UPO4
COUNTRYIT
AFDM
STREAMBagnatore
VFPO4
EXPER
SWNO3
VFNO3
DORANG
NO3
GPPR
UNO3
K2
REYNOLDS
COND
RESP
RAW
SUBSTRATIO
CAT
QMAN
WETPER
ANI
K1
QCOMBI
VELMAN
0
50
100
150
200
250
300
IncNodePurity

Table 3.2 Results of the regression training before and after variable importance analysis for all the algorithms in RMSE for NH_4

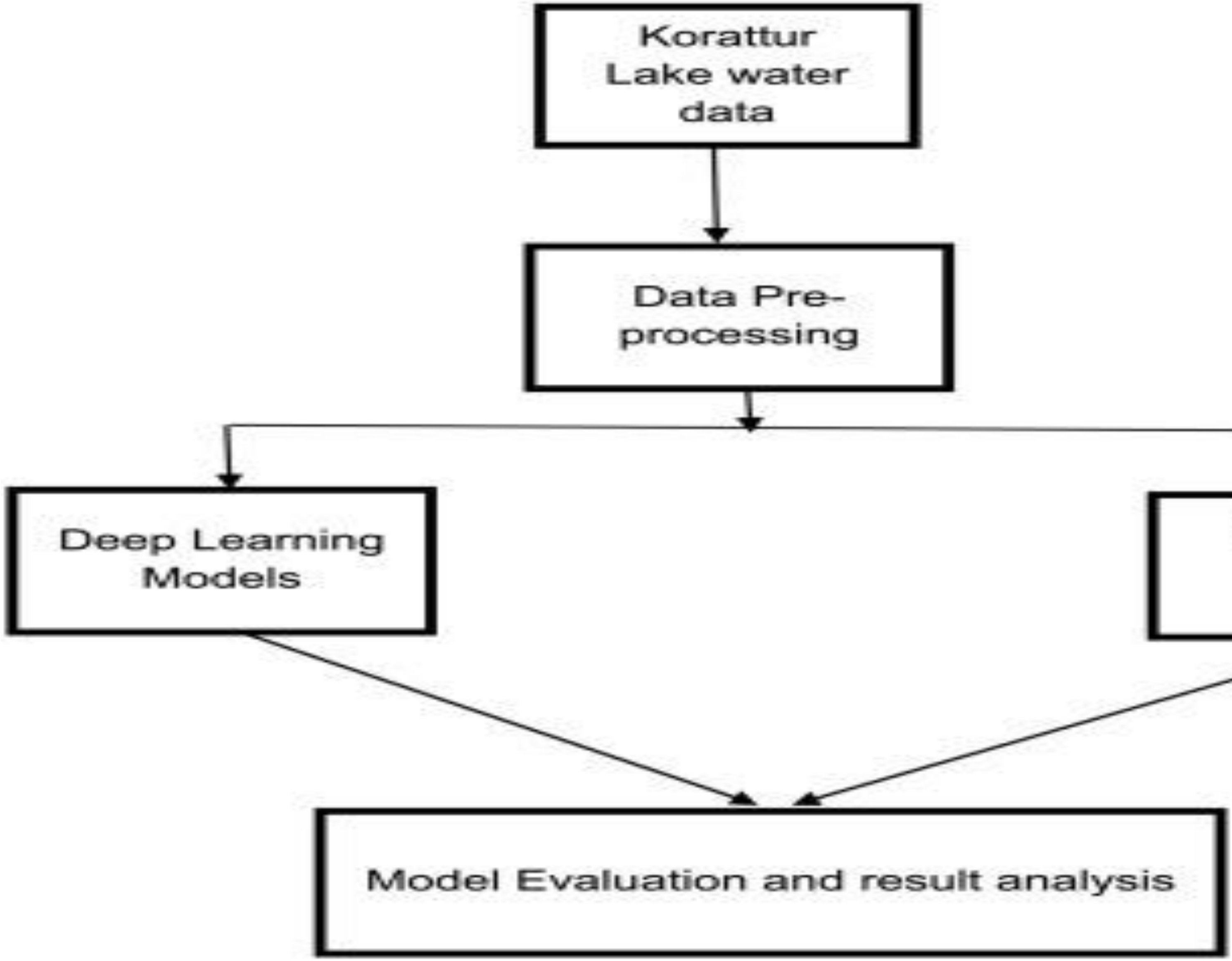
Model	RMSE of the model before Variable Importance analysis (relative)	RMSE of the model after Variable Importance analysis (relative)
RF	1.8449	1.7888
ANN	2.9297	2.8804
SVM	5.1370	4.7132

In addition to that, *devtools* package in R provides the possibility to visualize neural network models, showing its scheme and weights on each step. On Figure 3.2 one can see an example of such model for the dataset created after variable importance analysis.



PREDICTIVE MODEL

- Water quality has a direct impact on public health and the environment. Water is used for various practices, such as drinking, agriculture, and industry.


Korattur Lake water data
Data Pre-processing
Deep Learning Models
Auto Deep Learning Models
Model Evaluation and result analysis

CONCLUSION

- Overall, the goals defined for this research were reached and the examples of the application
- of machine learning models are presented, covering most of the aspects of the average re-
- search working in the field of artificial intelligence for environmental sciences tasks. This
- work also reveals the importance of consulting data scientists before starting of the monitor-
- ing, since data sets unsuitable for requested tasks is a common problem.
- Generally, regression models were able to show the consistent trend and overall correlation
- between each other, even though for some of the measurements they give models of poor
- quality. Random forests (RF) show the best performance and are advised for scientists and
- engineers working with environmental data. Artificial neural networks (ANN) are another
- alternative, though their performance is inferior and they are prone to overfitting. Support
- vector machines (SVM) are the good example for the cases where a baseline model is needed,
- being one of the basic algorithms.

- Water quality varies considerably at different geographical locations.
- Fish can use some water supplies considered impaired for human use, even some saline waters have aquaculture potential.
- Water quality affects growth and well being of fish. Therefore, water quality should be of great importance to the aquaculture.
- Water quality varies with time to time and therefore requires regular monitoring.
- It is equally important to know how to interpret the water quality parameters that are measured to maintain the health and well being of their fish stock.