PROJECT REPORT

\

**WATER QUALITY ANALYSISANDPREDICTION**

JCT COLLEGE OFENGINEERING

TECHNOLOGY

# INTRODUCTION

# PROJECT OVERVIEW

With the rapid increase in the volume of data on the [aquatic environment](https://www.sciencedirect.com/topics/earth-and-planetary-sciences/aquatic-environment), machinelearninghasbecomeanimportanttoolfordataanalysis,classification,andprediction. Unlike traditional models used in water-related research, data-drivenmodels based on machine learning can efficiently solve more complex nonlinearproblems. In water environment research, models and conclusions derived frommachine learning have been applied to the construction, monitoring, simulation,evaluation, and optimization of various [water treatment](https://www.sciencedirect.com/topics/earth-and-planetary-sciences/water-purification) and management systems.Additionally,machinelearningcanprovidesolutionsforwater [pollutioncontrol](https://www.sciencedirect.com/topics/earth-and-planetary-sciences/pollution-control), [waterqualityimprovement](https://www.sciencedirect.com/topics/earth-and-planetary-sciences/water-quality-improvement),and [watershed](https://www.sciencedirect.com/topics/earth-and-planetary-sciences/watershed) ecosystemsecuritymanagement. In this review, we describe the cases in which machine learningalgorithms have been applied to evaluate the water quality in different waterenvironments, such as surface water, groundwater, drinking water, sewage, andseawater.Furthermore,weproposepossiblefutureapplicationsofmachinelearningapproachestowater environments.

**1.2.PURPOSE**

Hence, rapid industrial development has prompted the decay of water quality at adisturbing rate. Furthermore, infrastructures, with the absence of public awareness,andlesshygienicqualities,significantlyaffectthequalityofdrinkingwater.Infact,the consequences of polluted drinking water are so dangerous and can badly affecthealth,theenvironment,andinfrastructures.AspertheUnitedNations(UN)report,about 1.5 million people die each year because of contaminated water-drivendiseases. In developing countries, it is announced that 80% of health problems arecaused by contaminated water. Five million deaths and 2.5 billion illnesses arereportedannually.Suchamortality rateishigherthandeathsresultingfromaccidents,crimes,andterroristattacks.

Massive increases in population, the industrial revolution, and the use of fertilizersand pesticides have led to serious effects on the.

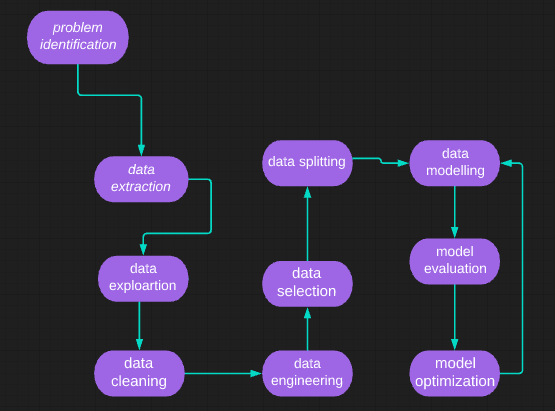
# EXISITNG PROBLEM

the main problem lies here. For testing the water quality we have to conduct labtestsonthewaterwhichiscostlyandtime-consumingaswell.So,inthispaper,weproposeanalternativeapproachusingartificialintelligencetopredictwaterquality.ThismethodusesasignificantandeasilyavailablewaterqualityindexwhichissetbytheWHO(WorldHealthOrganisation).Thedatatakeninthispaperistakenfromthe PCPB India which includes 3277 examples of the distinct wellspring. In thispaper, WQI(Water Quality Index) is calculated using AI techniques. So in futurework,wecanintegratethiswithIoTbasedframeworktostudylargedatasetsandtoexpandourstudytoalargerscale.Byusingthatitcanpredictthewaterqualityfastand more accurately than any other IoT framework. That IoT framework systemuses some limits for the sensor to check the parameters like ph, Temperature,Turbidity,andsoon.AndfurtherafterreadingthisparameterpassthesereadingstotheArduino microcontrollerandZigBeehandset forfurtherprediction

# PROJECTDESIGN

* 1. **DATAFLOWDIAGRAM**

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFDcandepicttherightamountofthesystemrequirementgraphically.Itshowshowdataentersandleavesthesystem,whatchangestheinformation,andwhere data isstored.



# SOLUTIONANDTECHNINCALARCHITECTURE

Therearebasically10steps formaking our modelpredictthewaterqualityofthewatersamples.Thosestepsare:-

1. *ProblemIdentification*

In this step, we identify the problem which is solved by our model. So the problem to be solved by our model is water qualityprediction usingadataset.

1. *DataExtraction:-*

In this, we extract the data from the internet to train our data and predict the water quality. So for that, we take the CPCB(CentralPollution Control Board India) dataset which contains 3277 instances of 13 different wellsprings which are collected between 2014to2020.

1. *DataExploration:-*

In this step, we analyze the data visually by comparing some parameters of water with the WHO standards of water. It gives a slightoverviewof the data.

1. *DataCleaning*

In this step, we clean that data like if there are some missing values in it so we replace them with mean and remove noise from thedata..

1. *DataSelection*

In this step, we select the data types and source of the data. The essential goal of data selection is deciding fitting data type, source,and instrumentthatpermitagentsto respondtoexplorequestionssufficiently

1. *Data Splitting*

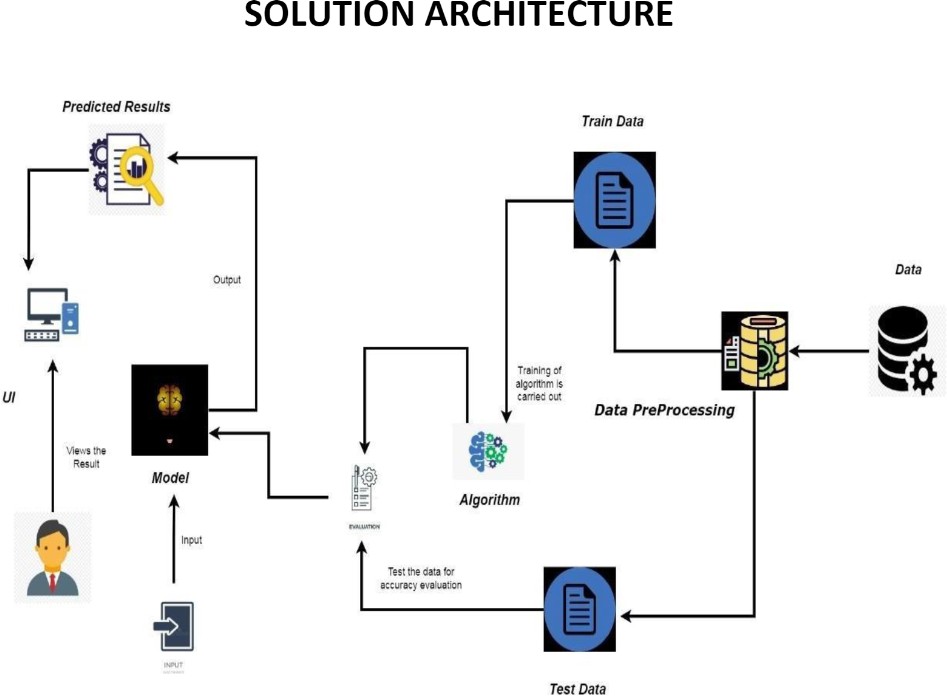
Inthisstep,wedividethedatasetintosmallersubsetsforeasingthecomplexity.Normally,withatwo-sectionsplit,onesectionisutilizedtoassessortestthe informationandtheother topreparethemodel.

1. *DataModeling*

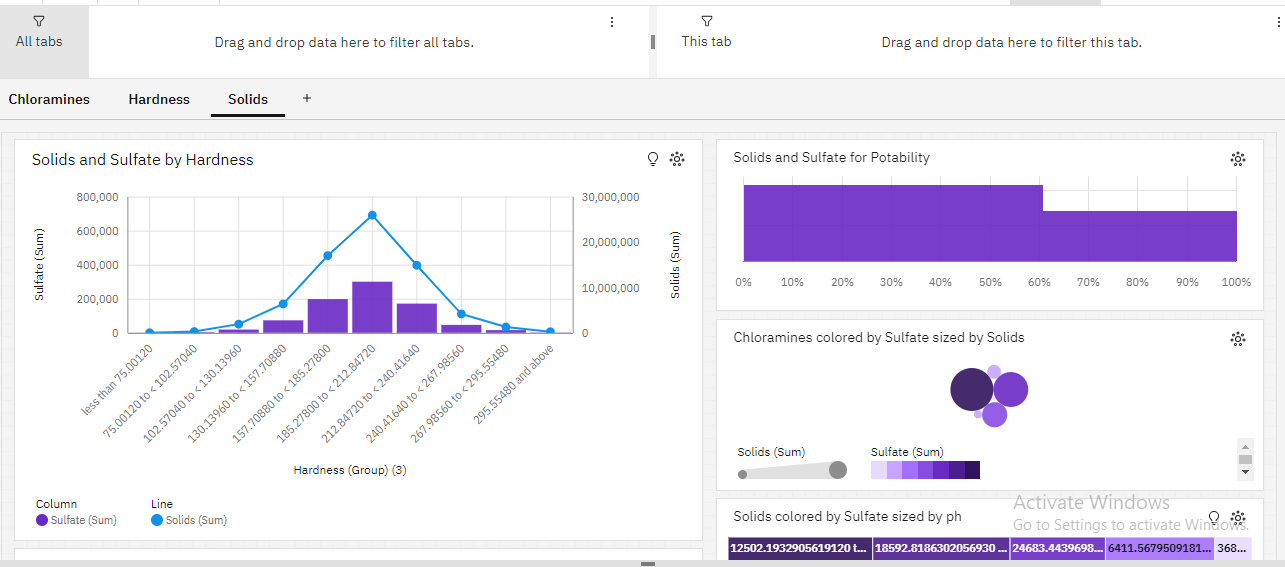
Inthisstep,wecreateagraphofthedatasetforvisualrepresentationofdataforbetterunderstanding.ADataModelisthistheoreticalmodel thatpermitsthefurther structureofconceptual modelsandto setconnections between data.

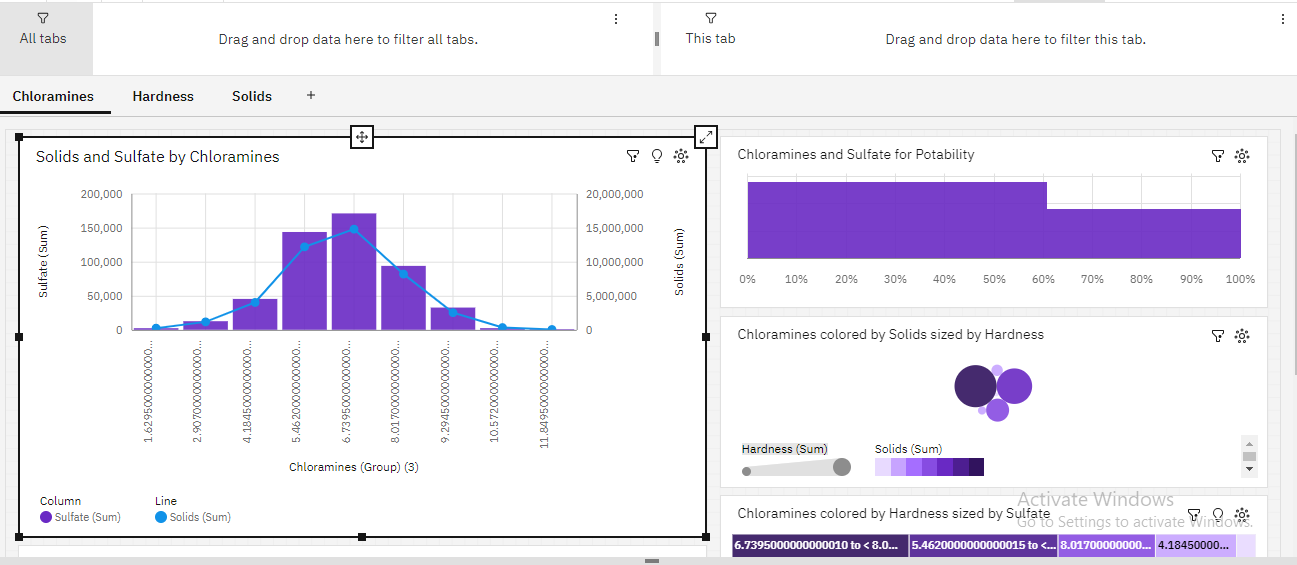
1. *ModelEvaluation*

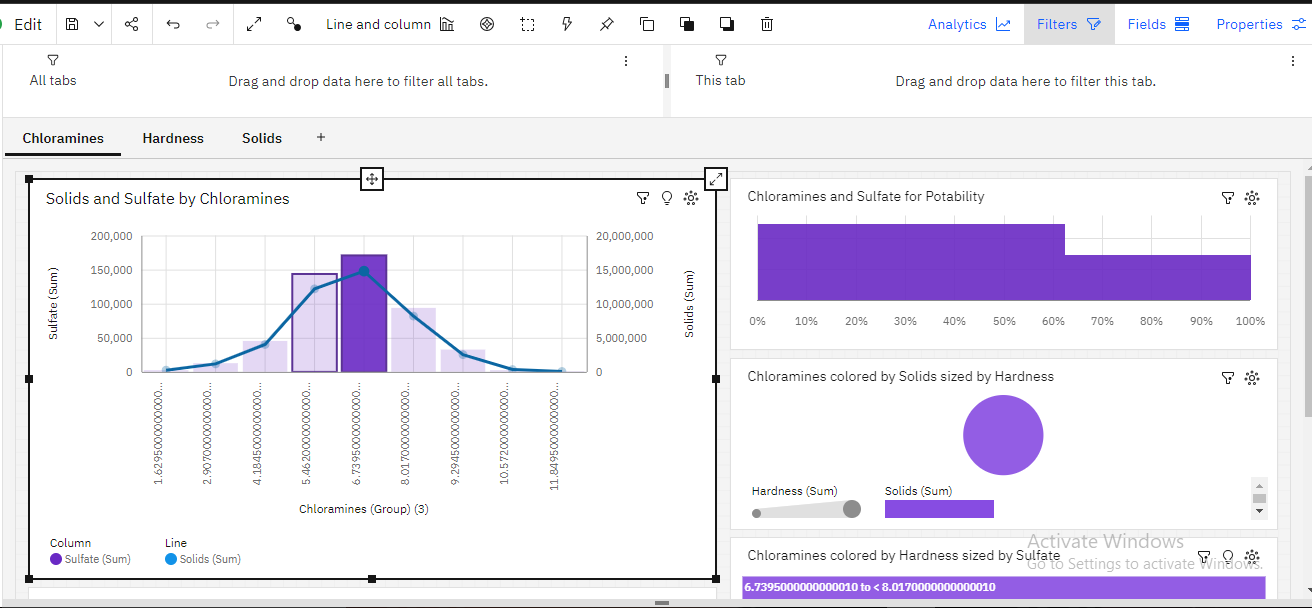
ModelEvaluationisafundamentalpieceofthemodelimprovementprocess.Inthisstep,weevaluateourmodelandcheckhowwellourmodeldointhe future.



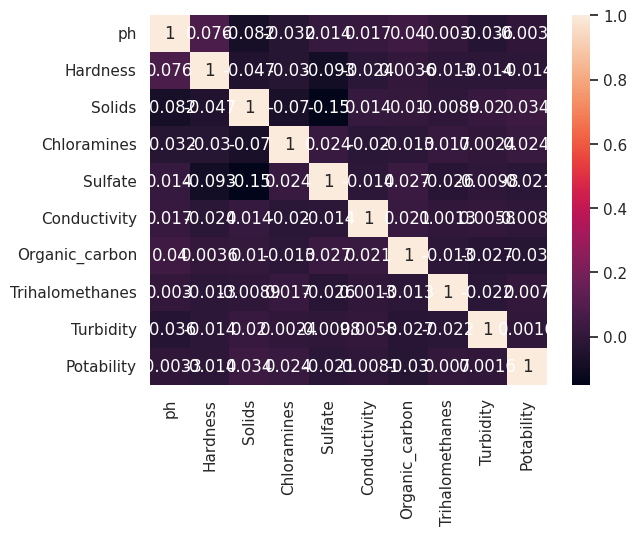
# REPORTSFROMIBM COGNOS

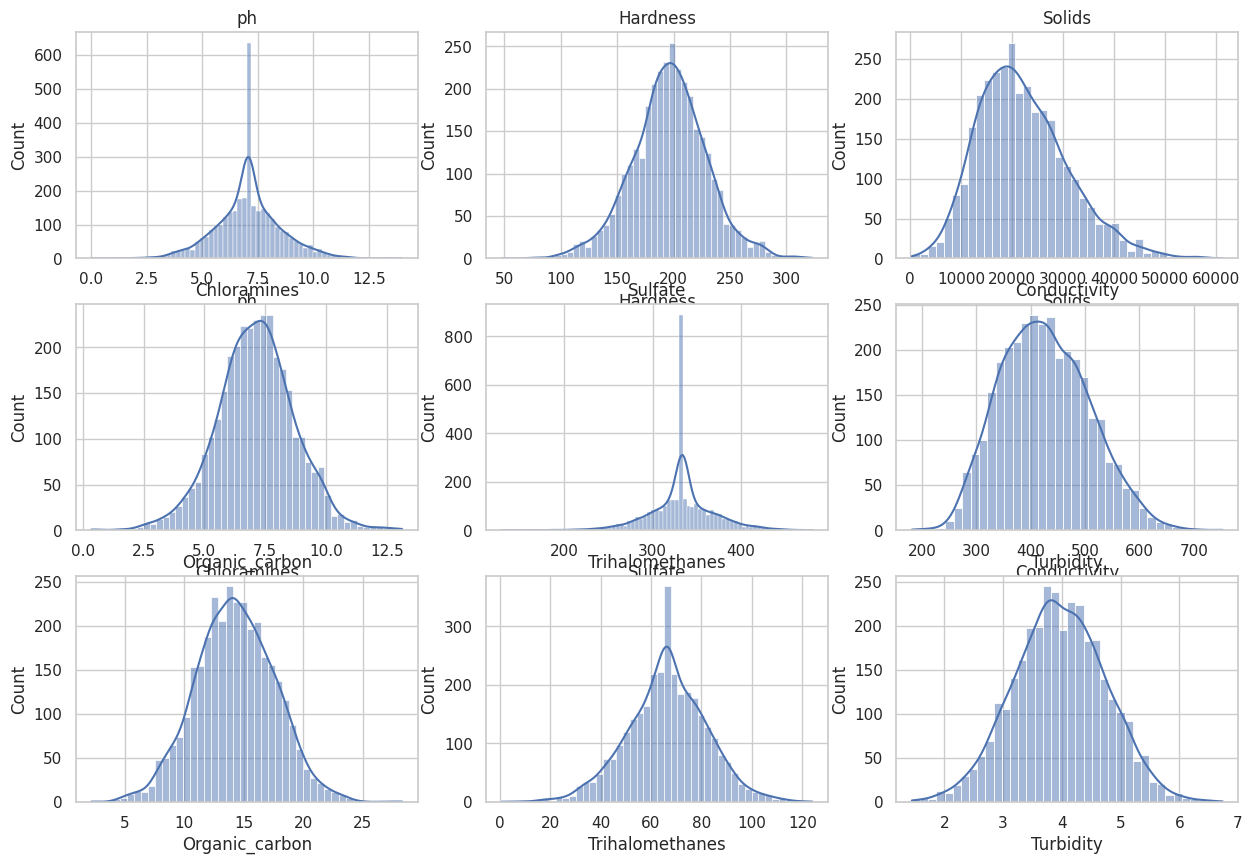
****



****

1. **ANALYTICS FROM G COLAB**

****

****

1. **CODINGANDSOLUTIONS**

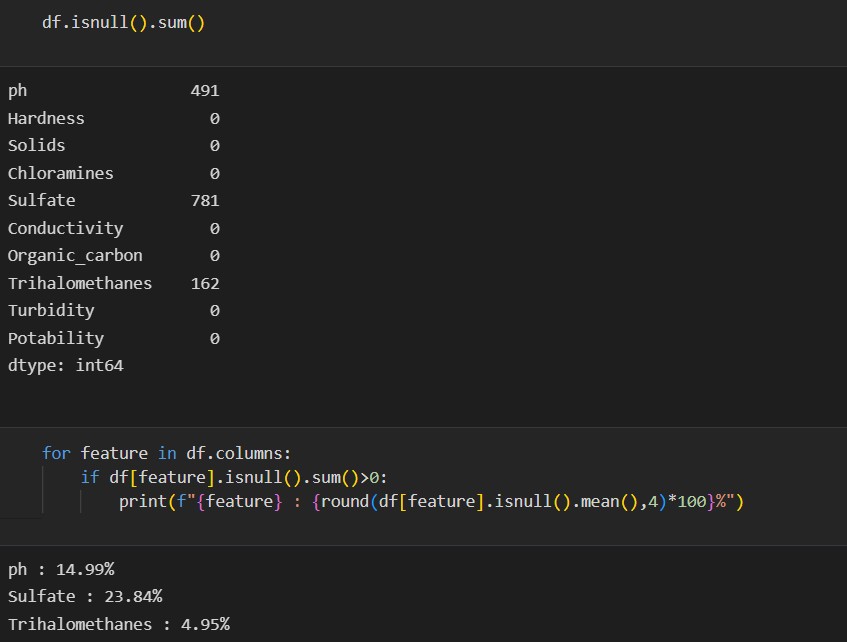
# FEATURE1

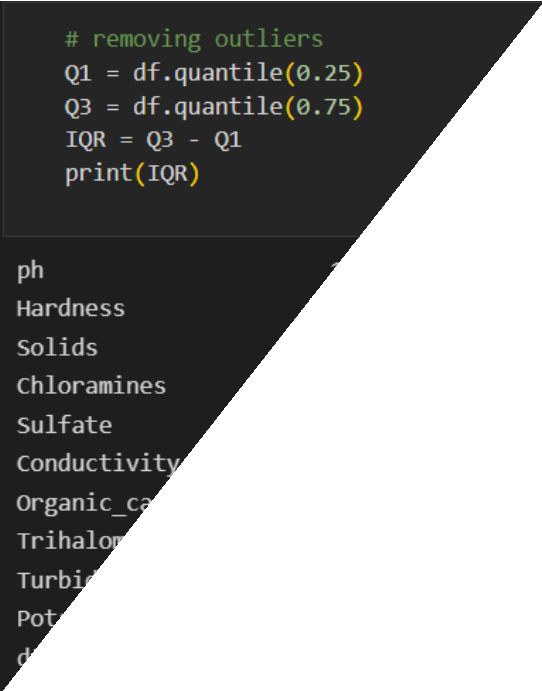
**Datacollectionandcreation**

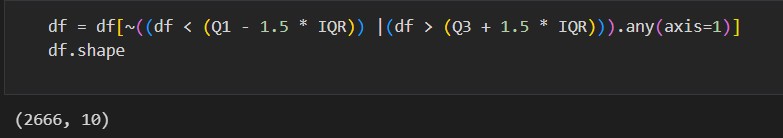
Data mining techniques require domain knowledge in order to generate predictions. For water quality applications, it is vital tounderstand how various water quality parameters influence water quality. This information can come from a domain expert orhistoricaldatacollections.Fortheforecastingtask,twotypesofdatasets wereused:acarefullycreatedhugesyntheticdatasetandan available realdataset

# DataPreprocessing

The processing phase is very important in data analysis to improve the data quality. In this phase, the WQI has been calculated fromthemostsignificantparametersofthedataset.Then,watersampleshavebeenclassifiedonthebasisoftheWQIvalues.Forobtainingsuperioraccuracy,the -score methodhasbeenusedas adatanormalization technique.

**ce****dw**





# FeatureEngineering:

RemovingOutliersusingoutlierTechnique:WaterQualityIndexCalculation

Tomeasurewaterquality,WQIisusedtobecalculatedusingvariousparametersthatsignificantlyaffectWQ[40–42].Inthisstudy,a published dataset is considered to test the proposed model, and seven significant water quality parameters are included. The WQIhasbeencalculatedusingthefollowingformula:

where: is the total number of parameters included in the WQI calculations is the quality rating scale for each parametercalculatedbyequation([2](https://www.hindawi.com/journals/abb/2020/6659314/#EEq1))below,andistheunitweight for eachparameter calculatedbyequation([3](https://www.hindawi.com/journals/abb/2020/6659314/#EEq2)).



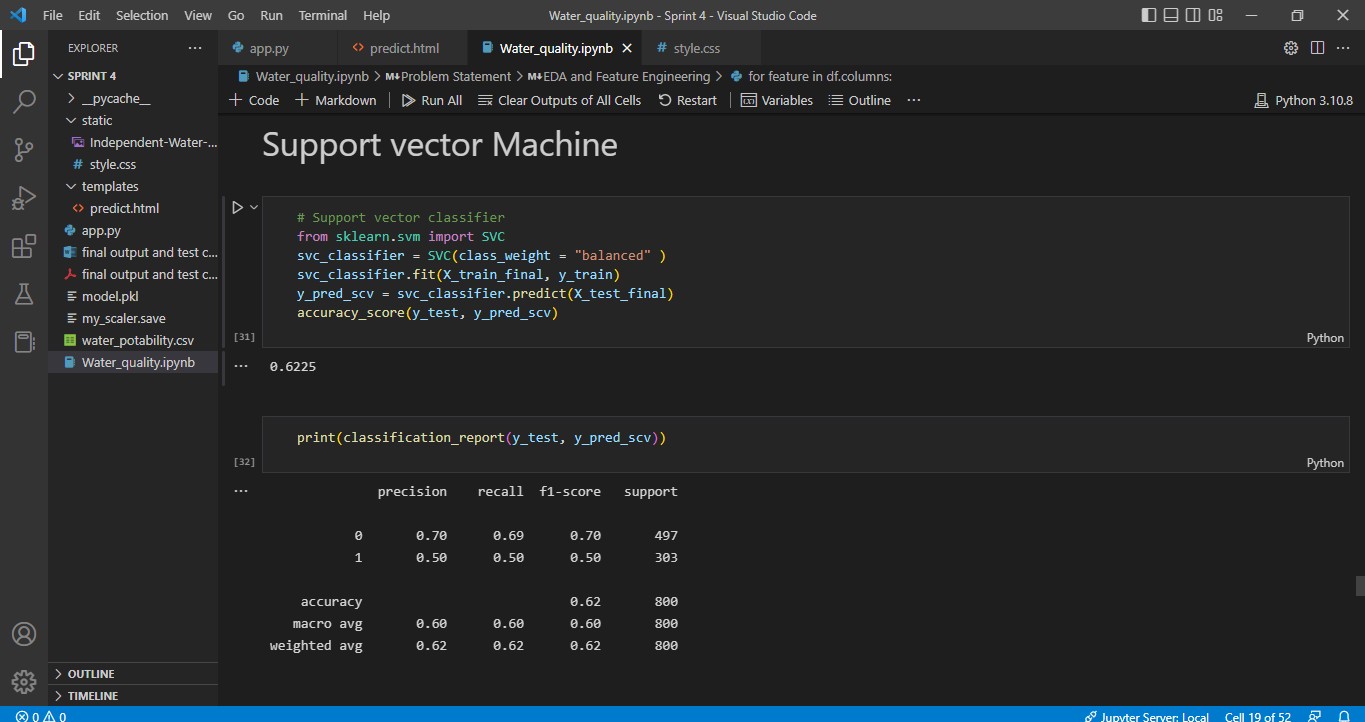
where: is the measured value of parameterin the tested water samples is the ideal value of parameterin pure water (0 for allparameters exceptand),andis the recommended standardvalueof parameter(asshownin Table[1](https://www.hindawi.com/journals/abb/2020/6659314/tab1/))



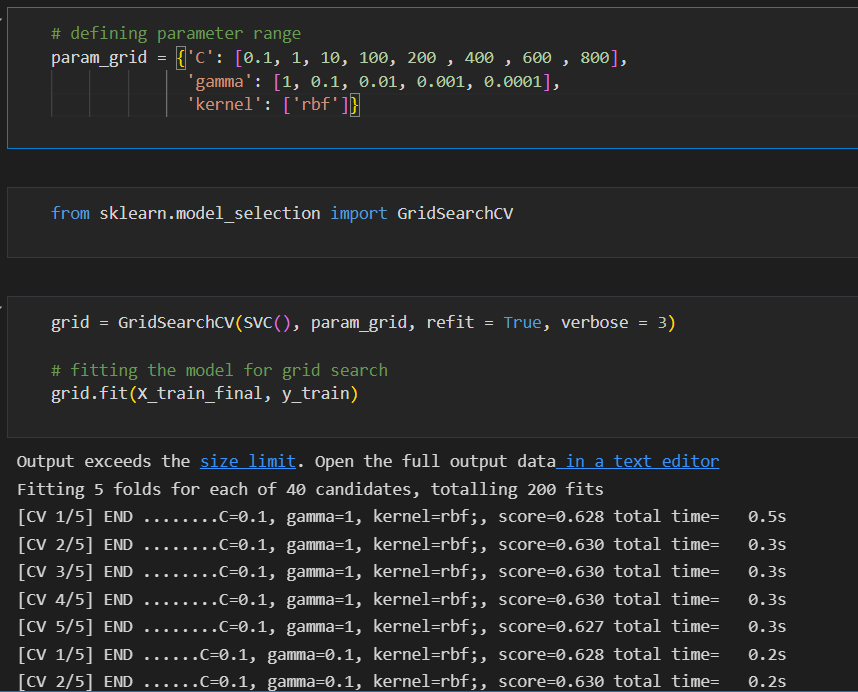
# FEATURE2

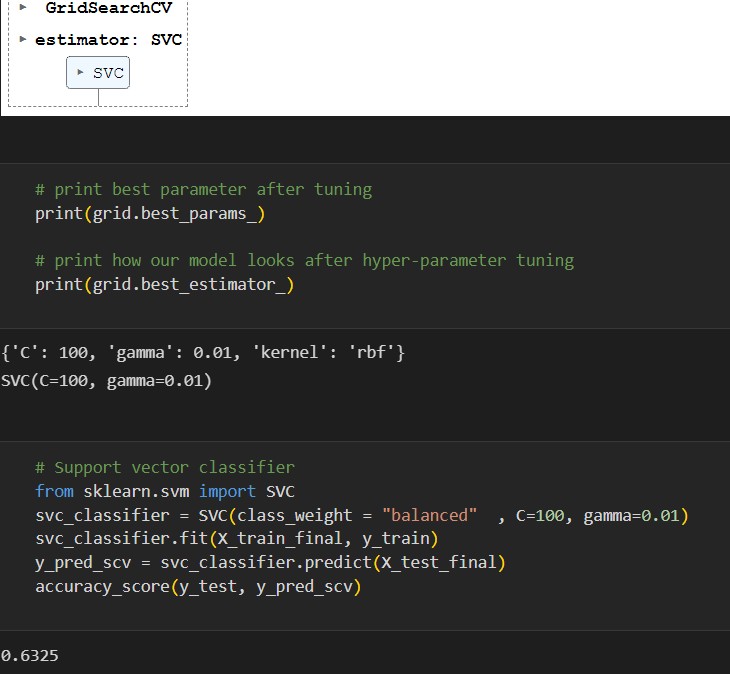
Performance Measures Results True Positives (TP) are when the model predicts the positive class properly. True Negatives (TN) isoneofthecomponentsofaconfusionmatrixdesignedtodemonstratehowclassificationalgorithmswork.Positiveoutcomesthatthemodel predicted incorrectly are known as False Positives (FP). False Negatives (FN) are negative outcomes that the model predictsnegative class. Accuracy is the most basic and intuitive performance metric, consisting of the ratio of successfully predictedobservationstototalobservations.

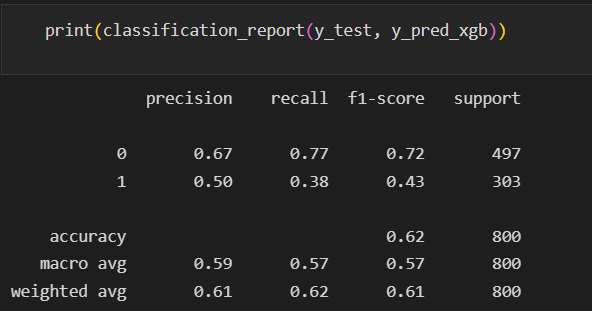
Accuracy=TP+TN/(TP+FP+FN+TN)



TheSVMmodel wasdevelopedin1995byCorinnaCortesandVapnik.It hasseveral uniquebenefitsinsolving smallsamples,andnonlinear and high-dimensional pattern recognition. It can be extended to function in the simulation of other machine learningproblems. It uses the hyperplane to separate the points of the input vectors and finds the needed coefficients. The best hyperplane isthe line with the largest margin, which is meant the distance between the hyperplane and the nearest input objects. The input pointsdefined in the hyperplane are called *support vectors*. In this work, the linear SVM model along with the Gaussian radial basisfunction (equation ([17](https://www.hindawi.com/journals/abb/2020/6659314/#EEq4))) is usedto classifythetestedwatersamplesbasedon their quality.

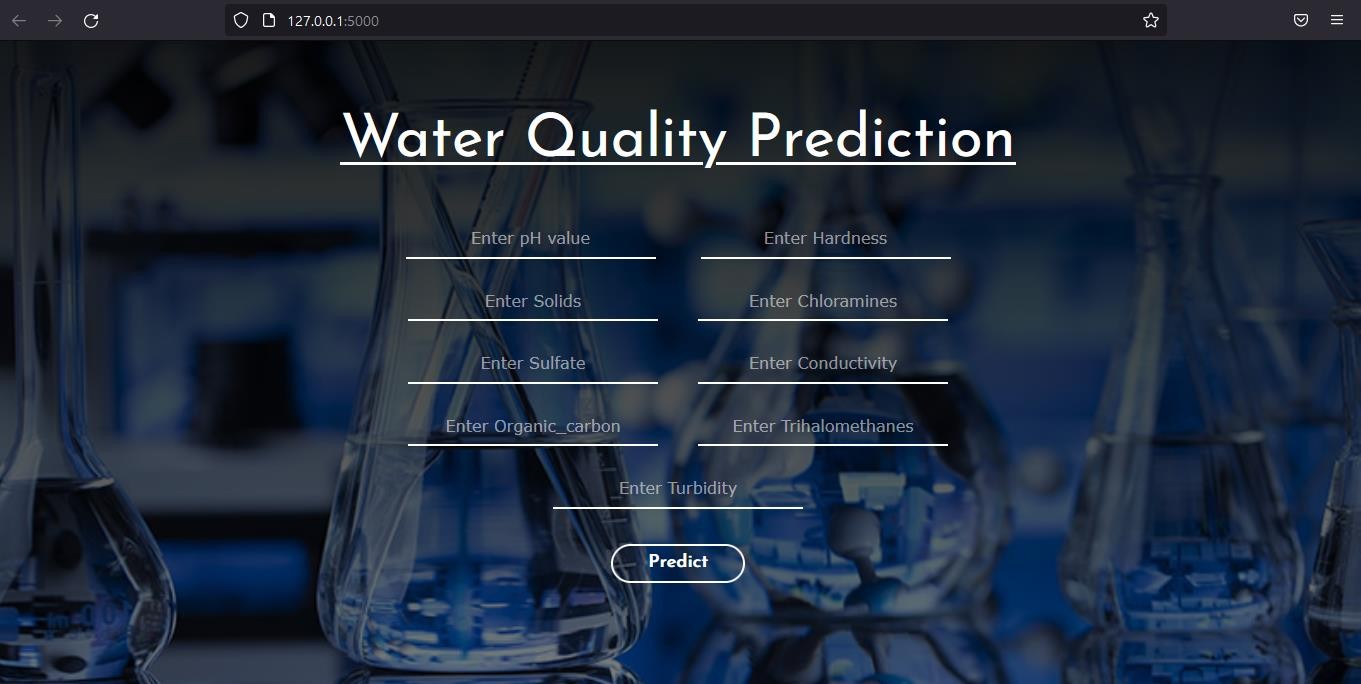
**HyperParametertuningwithSupportVectorM****achines(SVM):**



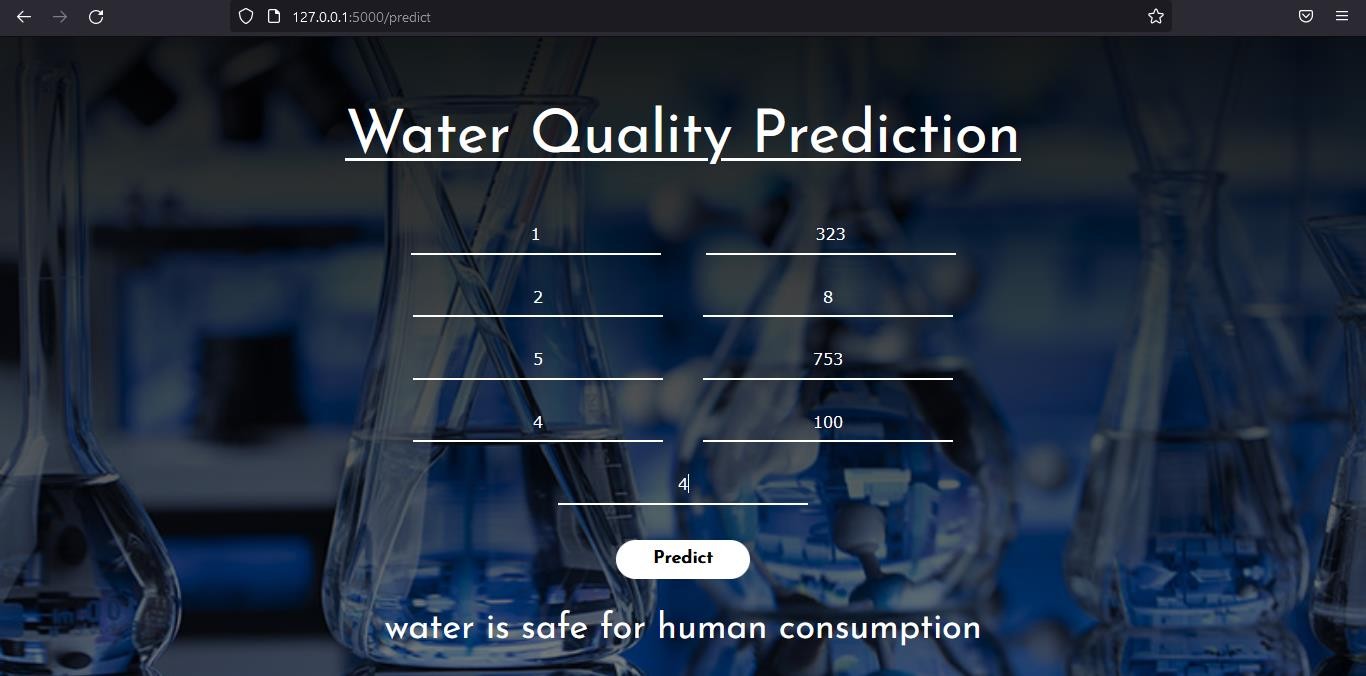


# TESTING

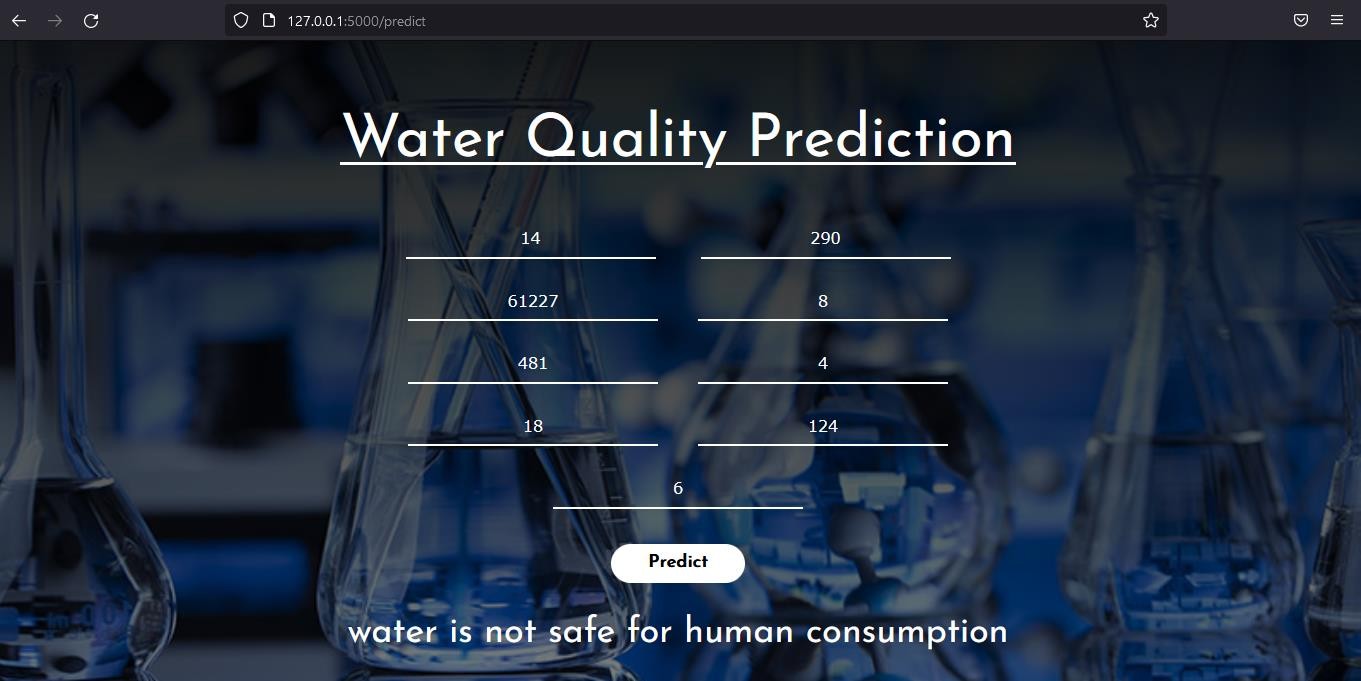
* 1. **HomePage**



# TESTCASE-1



* + 1. **TESTCASE-2**



# USERACCEPTANCE TESTING

* + 1. **PurposeofDocument**

Thepurposeofthisdocumentistobrieflyexplainthetestcoverageandopenissuesofthe[ProductName]projectatthetimeof the release toUserAcceptance Testing(UAT).

# DefectAnalysis

Thisreportshowsthenumberofresolvedorclosedbugs ateach severitylevel,andhowtheywereresolved

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Resolution** | **Severity1** | **Severity2** | **Severity3** | **Severity4** | **Subtotal** |
| ByDesign | 10 | 4 | 2 | 3 | 20 |
| Duplicate | 1 | 0 | 3 | 0 | 4 |
| External | 2 | 3 | 0 | 1 | 6 |
| Fixed | 11 | 2 | 4 | 20 | 37 |
| NotReproduced | 0 | 0 | 1 | 0 | 1 |
| Skipped | 0 | 0 | 1 | 1 | 2 |
| Won'tFix | 0 | 5 | 2 | 1 | 8 |
| Totals | 24 | 14 | 13 | 26 | 77 |

# TestCaseAnalysis

Thisreportshowsthenumberoftestcasesthathavepassed,failed,anduntested

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Section** | **TotalCases** | **NotTested** | **Fail** | **Pass** |
| PrintEngine | 7 | 0 | 0 | 7 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ClientApplication | 51 | 0 | 0 | 51 |
| Security | 2 | 0 | 0 | 2 |
| OutsourceShipping | 3 | 0 | 0 | 3 |
| ExceptionReporting | 9 | 0 | 0 | 9 |
| FinalReportOutput | 4 | 0 | 0 | 4 |
| VersionControl | 2 | 0 | 0 | 2 |

# RESULT

* 1. **PERFORMANCEMETRICS**

Forvalidatingthedevelopedmodel,thedatasethasbeendividedinto70%trainingand30%testingsubsets.TheSVM,Xgboost,andRandomForestwereutilizedforthewaterqualityclassificationprediction

**SO,WEARE GOINGTOUSESVC**

Performance Measures Results True Positives (TP) are when the model predicts the positive class properly. True Negatives (TN) isoneofthecomponentsofaconfusionmatrixdesignedtodemonstratehowclassificationalgorithmswork.Positiveoutcomesthatthemodel predicted incorrectly are known as False Positives (FP). False Negatives (FN) are negative outcomes that the model predictsnegative class. Accuracy is the most basic and intuitive performance metric, consisting of the ratio of successfully predictedobservationstototalobservations.

**Accuracy=TP+TN/(TP+FP+FN+TN)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SN. | Algorithm | Type | ACCURACY | Precision | Recallf1-Score |
| 1 | RANDOMFOREST | 58.5 | 0.42 | 0.38 | 0.40 |
| 2 | XGBOOST | 61.7 | 0.43 | 0.12 | 0.18 |

**Table1.ComparisonofalgorithmsSN.**

# ADVANTAGES

Whether it be for groundwater, surface water or open water, there are a number of reasons why it is important for you to undertakeregularwaterqualitytesting.Ifyou’rewantingtocreateasolidfoundationonwhichtobuildabroaderwatermanagementplan,theninvesting in water quality testing should be your first point of action. This testing will also allow you to adhere to strict permitregulations andbe incompliance with Australianlaws.

Identifyingthehealthofyourwaterwillhelpyoutodiscoverwhereitmayneedsomehelp.Ultimately,findingasourceofpollution,or remaining proactive with your monitoring will enable you to save money in the long term. The more information that you canobtainwillassistyouwithyourdecisiononwhatproductyoumayneedtoimprovetheconditionofyourwater.Simplyguessingandbuying products based on a hunch or a general trend is ill-advised, as each body of water has unique properties that can only bediscovered throughtesting.

Measuringtheamountofdissolvedoxygeninyourwaterisanotherimportantadvantageofwaterqualitytesting,astypicallythelessoxygen,thehigherthewatertemperature,resultinginamoreharmfulenvironmentforaquaticlife.Theselevelsdofluctuateslightlyacross the seasons, but regular monitoring of your water quality will allow you to discover trends over time, and whether there areotherfactorsthatmay be contributingtotheresultsyoudiscover.

# DISADVANTAGES

* TrainingnecessarySomewhat difficulttomanageovertimeandwith largedatasets
* Requiresmanualoperationtosubmitdata,someconfigurationrequired
* Costly,usuallyonlyfeasibleunderExchangeNetworkgrantsTechnicalexpertiseandnetworkserverrequired
* RequiresmanualoperationtosubmitdataCannotrespondtodataqueriesfromothernodes,andthereforecannotinteractwiththeExchange Network Technicalexpertiseandnetwork server required

# CONCLUSION

Potability determines the quality of water, which is one of the most important resources for existence. Traditionally, testing waterquality required an expensive and time-consuming lab analysis. This study looked into an alternative machine learning method forpredicting water quality using only a few simple water quality criteria. To estimate, a set of representative supervised machinelearning algorithms was used. It would detect water of bad quality before it was released for consumption and notify the appropriateauthoritiesItwillhopefullyreducethenumberofindividualswhodrinklow-qualitywater,loweringtheriskofdiseasesliketyphoidanddiarrhea.Inthiscase,usingaprescriptiveanalysisbasedonprojectedvalueswouldresultinfuturecapabilitiestoassistdecisionandpolicymakers.

1. **APPENDIX**

**REQUIREMENT.TXT**

Flask == 2.2.2

joblib == 1.2.0

numpy ==1.23.4

pandas ==1.5.1

scikit-learn ==1.1.3

xgboost ==1.7.1

gunicorn ==20.1.0

matplotlib ==3.6.2

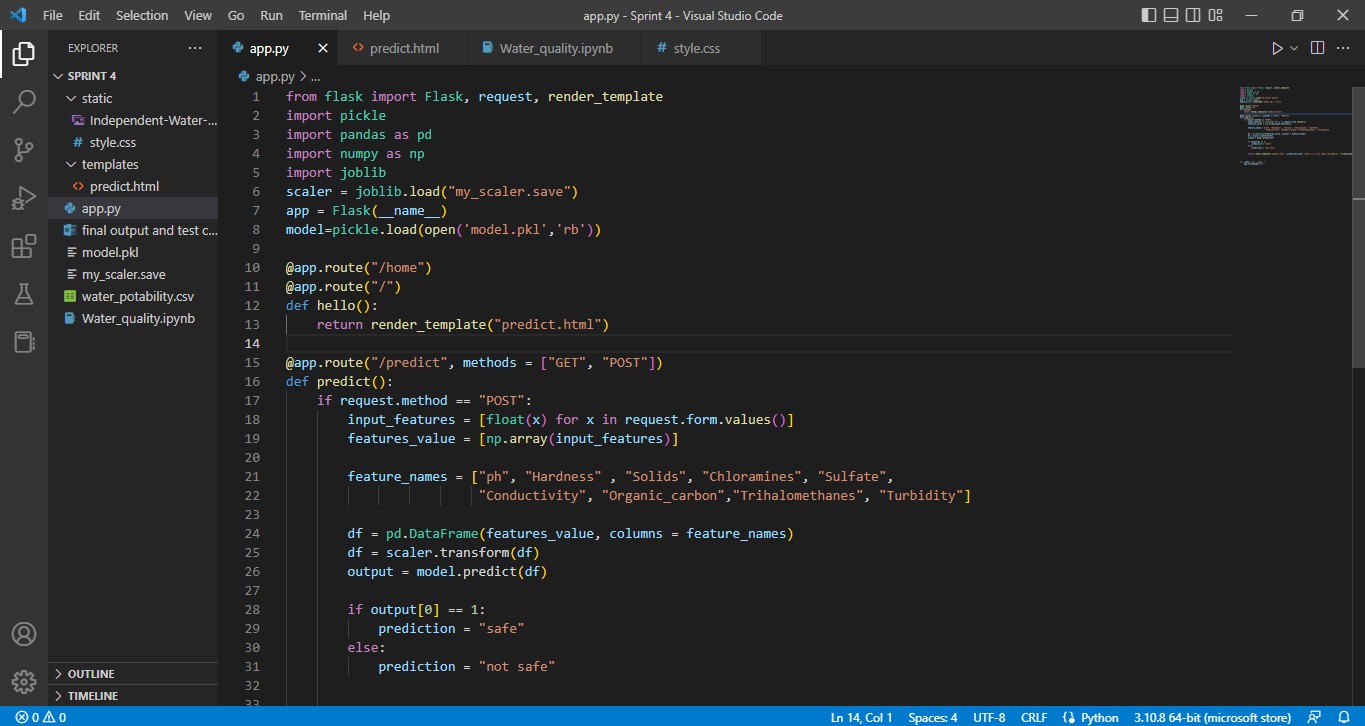
seaborn == 0.12.1gevent

requests

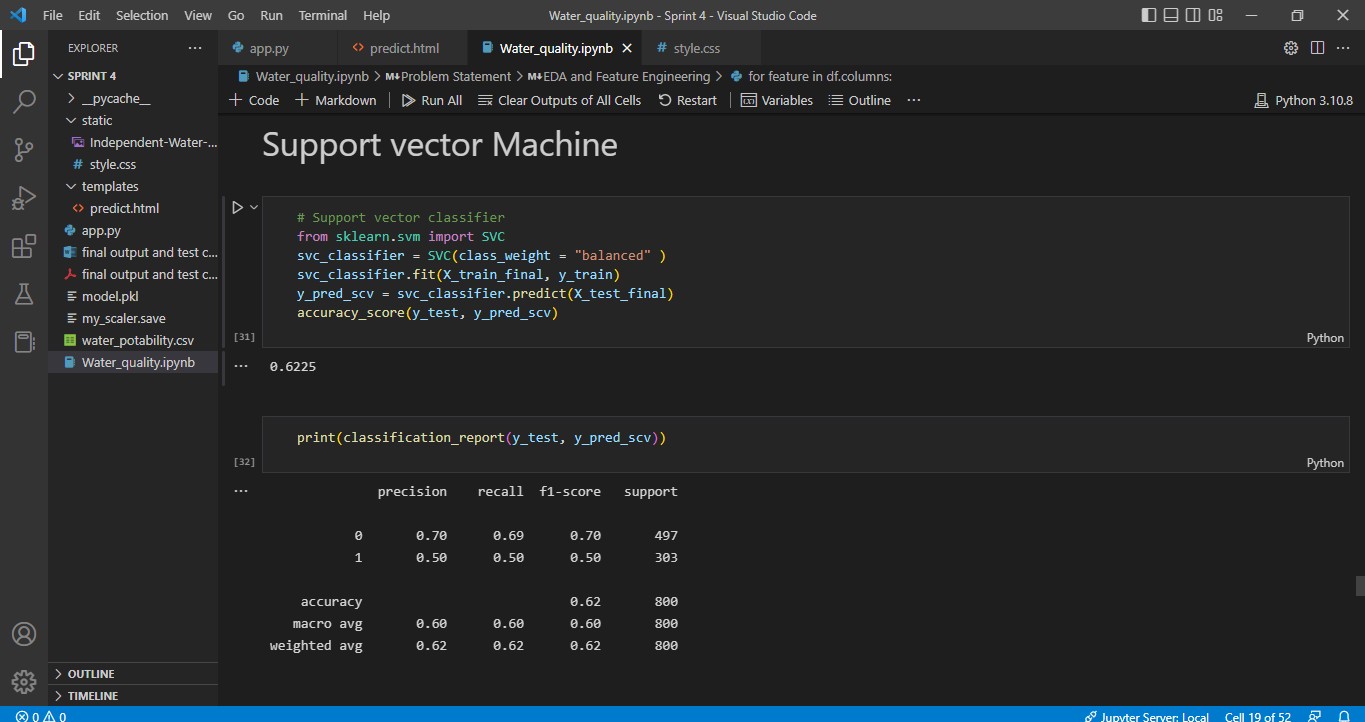
IBM cognos

flask-cors==3.0.10

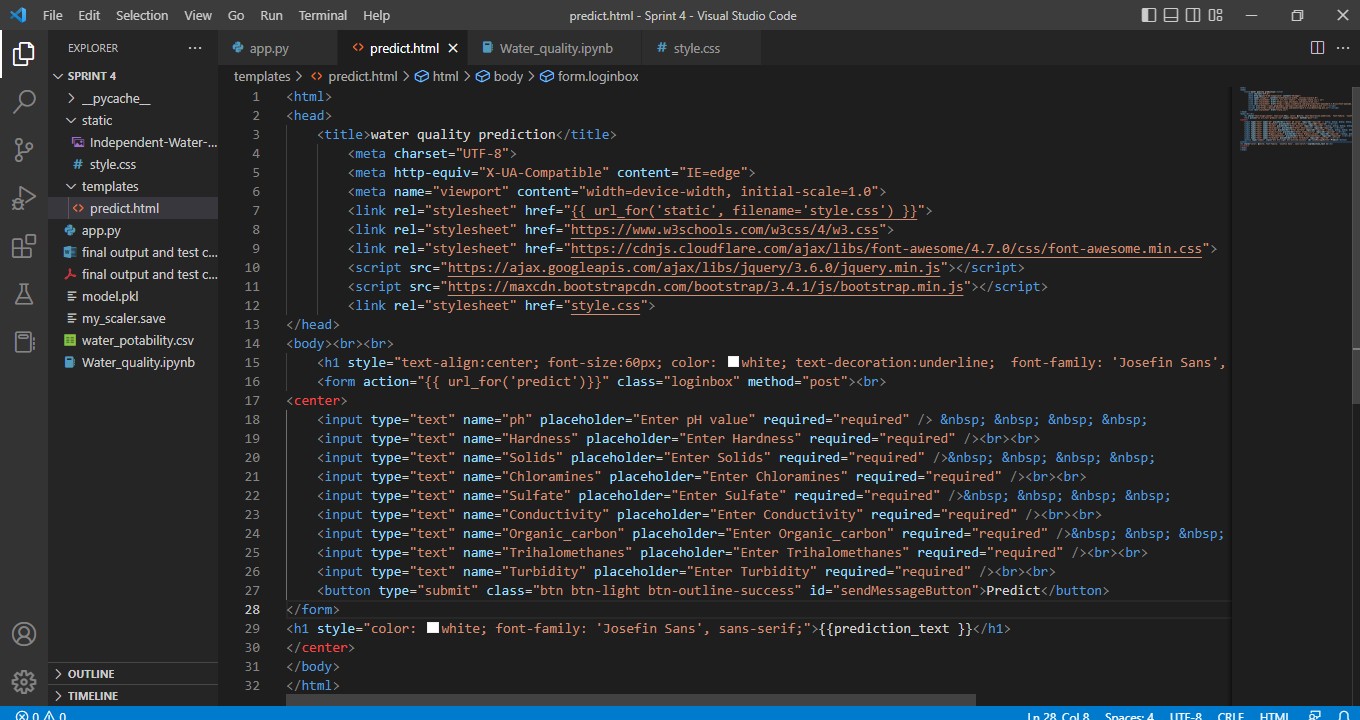
**APP.PY**



**WATERQUALITY.IPYNB**



**HOME.HTML**



GOOGLE DRIVE LINK: <https://colab.research.google.com/drive/1jL-UrE7Jdvc4gF25-dC_OI9TD6eS2YIO?usp=sharing>

IBM Link:

https://eu2.ca.analytics.ibm.com/bi/?perspective=dashboard&pathRef=.public\_folders%2FNM%2FNew%2Bdashboard&action=view&mode=dashboard&subView=model0000018b8931f1d4\_00000000