الجمهورية الجزائرية الديمقراطية الشعبية

REPUPLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التعليم العالي والبحث العلمي

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHARCHE SCIENTIFIQUE

المدرسة الوطنية العليا للفلاحة الحراش – الجزائر–

ECOLE NATIONALE SUPERIEURE AGRONOMIQUE EL-HARRACH – ALGER

Département : Génie rural القسم : الهندسة الريفية

Spécialité : sciences de l'eau التخصص : علم المياه

Graduation Thesis

To obtain the Master's Degree in water engineering

Theme

Water quality classification using machine learning

Presented by: DIB Nadjet **Graduation date: 06/10/2024**

In front of a jury composed of:

President:

Mrs. GUEDIOURA Ilham MCA.ENSA.

Supervisor:

Mr. HARTANI Tarik Professor, ENSA.

Examiner:

Mr.DELLI Reda MCD, ENSA.

Examiner:

Mrs. ARAB Zahira MCB, ENSA.

Promotion: 2019 /2024

LIST OF CONTENT

GENERAL INTRODUCTION 18
CHAPTER ONE:OVERVIEW ABOUT WASTEWATER 23
I.1.Introduction: 24
I.2.Definition of wastewater: 24
I.3.Sources of wastewater: 24
I.3.1. Domestic/Municipal Wastewater: 24
I.3.2. Industrial Wastewater: 25
I.3.3. Stormwater Runoff: 25
I.3.4. Infiltration/Inflow: 25
I.3.5. Agricultural Wastewater: 25
I.3.6. Hospital and Pharmaceutical Wastewater: 26
I.4.Nature of pollution: 26
I.4.1.Organic Pollution 27
I.4.2 Microbiological Pollution 27
I.4.3 Mineral Pollution (Chemical pollution): 27
I.4.4. Physical Pollutants: 28
I.5.Importance of Water Treatment: 29
I.6. Physico-chemical Quality of Water: 29
I.6.1 Physico-chemical Parameters 29
I.6.1.1. Temperature: 29
I.6.1.2.Electrical Conductivity (EC): 29
I.6.1.3.Dissolved Oxygen (DO): 30
I.6.1.4.Suspended Solids (SS): 30
I.6.1.4.1.Volatile Suspended Solids (VSS): 30
I.6.1.4.2.Mineral Suspended Solids (MSS): 30
I.6.1.5.Turbidity: 31
I.6.1.6.pH (Hydrogen Potential): 31
I.6.1.7 Redox Potential: 31
I.6.1.8. Biochemical Oxygen Demand (BOD5): 31
I.6.1.9. Chemical Oxygen Demand (COD): 31
I.6.1.10. Biodegradability Coefficient (COD/BOD5): 32
I.6.1.11. Total Nitrogen (TN): 33
I.6.1.12. Total Phosphorus (TP): 34

I.6.1.13. Nitrates (NO3-): 34
I.6.1.7.14.Sulfates: 34
I.6.1.7.15.Potassium: 35
I.6.2.Microbial parameters: 35
I.6.2.1.Bacteria: 35
I.6.2.2.Protozoa: 36
I.6.2.3.Viruses: 36
I.6.2.4.Fungi: 36
I.7.Impact of water pollution on the public health (human health): 37
I.8.Impact of water pollution on the environment: 39
I.9.Wastewater Treatment process: 40
I.10.The use of wastewater in agriculture: 42
I.11. Sociocultural Aspects 45
I.12. Economic and financial considerations 46
I.13. Policy and Institutional Aspects 47
I.14.Project planning criteria: 47
CHAPTER TWO:THE USE OF AI AND TECHNOLOGY IN WATER MANAGEMENT 49
II.1.Introduction: 50
II.2. Artificial intelligence: 50
II.2.1.Definition: 50
II.2.2.Types of AI: 50
II.2.3.Branches of AI: 53
II.3.Machine Learning: 53
II.3.1.Definition: 53
II.3.2. Machine Learning Procedure: 54
II.3.3 Requirements for Effective Machine Learning Systems: 54
II.3.4.Machine learning approaches: 55
II.3.5.Deep learning comparison with Conventional Machine learning techniques: 56
II.4.Machine Learning and Deep Learning use in wastewater Management: 57
II.4.1.Some Literature review about water quality prediction: 57
II.4.2.Some Literature review about IoT sensor for water quality: 62
II.4.3.Some Literature review about PRIMA Project: 65
II.4.3.1.Document 1: 65
II.4.3.2.Document 2: 67

CHAPTER THREE: MATERIALS AND METHODS 73
III.1. Part water quality prediction (AI and Platform): 74
III.1.1.Data Collection: 74
III.1.1.1.Presentation of WWTP located in Boumerdes center (city FOES): 74
III.1.1.2.Technical Characteristics: 75
III.1.1.3.Design parameters related to pollution……..……..76
III.1.2. Loading the dataset: 76
III.1.3. Data Preprocessing (preparation) and normalization: 77
III.1.4.Data Visualization or Statistical study based on this parameters: 79
III.1.5.Training and testing the data by a machine learning models and algorithms: 79
III.1.6.Implementation of functions in the coding for the integration of irrigation, environment standards and recommendations: 80
III.2.Part of PRIMA PROJECT: 83
III.3.Water quality detector (IOT based water quality monitoring system): 92
III.3.1.Hardware step (Circuit and connections): 92
III.3.2.Software Step (programming) 96
CHAPTER FOUR :RESULTS AND DISSCUSION 100
IV.1.Water Quality Prediction: 101
IV.1.1.Correlation matrix of all parameters of 240 dataset: 101
IV.1.2.Histogrames and wrong values: 104
IV.1.3.Matrix of one thousound: 108
IV.1.4. Correlation matrix of 3600 dataset: 110
IV.1.5.Histograms of three features 111
IV.1.6-Boxplots of 3 features: 113
IV.1.7.Statistical characteristics and metrics values of water quality parameters 116
IV.1.8.Results of training and testing the model: 118
IV.2. Results of water quality detector: 119
IV.2.1.Circuit diagram design 119
IV.2.2.Coding related to 3 sensors Software design part (programing) 121
IV.2.3.Explanation of coding part related to the three sensors: 123
IV.3.RESULTS PRIMA 129
IV.4.Discussion of results: 130
V. Bibliographic References: 142
VI.LIST OF ANNEXES: 152
VII. Business Plan : 160
VIII. BMC: 207

ABSTRACT

A wastewater treatment plant (WWTP) is an essential part of the entire water cycle, which reduces concentrations of pollutants in the environment. To enhance the monitoring and control of WWTP efficiency , researchers developed different models and systems, This study presents the application of Machine learning-based (ML) Artificial intelligence techniques (AIT) such as Random Forest algorithm (RF), Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost) to design an automatic classifier for water quality and determine the appropriate destination for the treated wastewater ,providing justifications and direct recommendations based on international standards and thresholds, for this purpose, dataset consisting of 3600 values related to domestic WW was utilized, with the outputs categorized into two classes influent not pure water (untreated WW) and effluent pure water (treated WW) . Approximately 240 data points were sourced from Algerian records, spanning ten years of monthly data. The influent parameters including Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD), and Total Kjeldahl Nitrogen (TKN) were used as inputs for wastewater quality prediction, identified as the most predictive features through the correlation-based feature selection (CFS) method, sufficient data and correct values .The machine learning models were trained via 60% values of the dataset ,with their accuracy tested on the remaining 40%.From the results of the test .Random forest with the accuracy of 99,8% is found to be the most successful model although all models have excellent accuracy because in this case the effective features were just three and the data was simple, it seen that SVM model is the fastest technique although Random forest close results to SVM but it seems that the training speed of XGBoost is approximately 7 times longer than SVM. Moreover, different functions are then integrated to determine whether this predicted wastewater suitable for agriculture or environment or unsuitable for them both ,providing reasons and recommendations or advices which empower us to create a platform of digital water prediction by the implementation of machine learning coding ,The promising results obtained paved the way for forecasting the performance of WWTP operations by the prediction of water quality , optimizing the reuse of treated WW on agriculture and swiftly address process anomalies before they escalate into more severe issues thereby enabling informed decision-making by water system managers.

Keywords: Wastewater (WW), Wastewater treatment plant (WWTP), Machine learning (ML), Artificial intelligence techniques (AIT), algorithm, Support vector machine (SVM), Random forest (RF), Extreme Gradient Boosting (XGBoost), Water quality prediction, dataset

Résumé

Une station d'épuration des eaux usées (WWTP) est une partie essentielle du cycle de l'eau, qui réduit les concentrations de polluants dans l'environnement. Pour améliorer le suivi et le contrôle de l'efficacité des STEP, les chercheurs ont développé différents modèles et systèmes. Cette étude présente l'application de techniques d'intelligence artificielle (AIT) basées sur l'apprentissage automatique (ML) telles que l'algorithme foret aléatoire (RF) , la machine à vecteurs de support (SVM) et le Extreme Gradient Boosting (XGBoost) pour concevoir un classificateur automatique de la qualité de l'eau et déterminer la destination appropriée des eaux usées(WW) traitées, en fournissant des justifications et des recommandations directes basées sur les normes et seuils internationaux. À cette fin, un ensemble de données composé de 3600 valeurs liées aux eaux usées domestiques a été utilisé, avec les résultats catégorisés en deux classes : eaux usées influentes non pures (eaux usées non traitées) et eaux effluentes pures (eaux usées traitées). Environ 240 points de données ont été obtenus à partir de dossiers algériens, couvrant dix ans de données mensuelles. Les paramètres influents, y compris la demande biologique en oxygène (DBO5), la demande chimique en oxygène (DCO) et l'azote total Kjeldahl (NTK) ont été utilisés comme entrées pour la prédiction de la qualité des eaux usées, identifiés comme les caractéristiques les plus prédictives par la méthode de sélection des caractéristiques basée sur la corrélation (CFS), données suffisantes et valeurs correctes. Les modèles d'apprentissage automatique ont été entraînés via 60 % des valeurs de l'ensemble de données, leur précision testée sur les 40 % restants. À partir des résultats du test, RF avec une précision de 99,8 % s'est révélé être le modèle le plus performant bien que tous les modèles aient une excellente précision, car dans ce cas, les caractéristiques effectives n'étaient que trois et les données étaient simples. Il semble que le modèle SVM soit la technique la plus rapide bien que Random Forest obtienne des résultats proches de SVM, mais il semble que la vitesse d'entraînement de XGBoost soit environ 7 fois plus longue que celle de SVM. De plus, différentes fonctions sont ensuite intégrées pour déterminer si ces eaux usées prédites sont adaptées à l'agriculture ou à l'environnement ou inadaptées à ces deux usages, en fournissant des raisons et des recommandations ou des conseils qui nous permettent de créer une plateforme de prédiction numérique de l'eau par la mise en œuvre du codage de l'apprentissage automatique. Les résultats prometteurs obtenus ont ouvert la voie à la prévision des performances des opérations de STEP (WWTP) en prédisant la qualité de l'eau, en optimisant la réutilisation des eaux usées traitées en agriculture et en abordant rapidement les anomalies

de processus avant qu'elles ne se transforment en problèmes plus graves, permettant ainsi une prise de décision éclairée par les gestionnaires des systèmes d'eau.

Mots-clés : Eaux usées (WW), Station d'épuration(WWTP), Apprentissage automatique(ML), techniques de l'Intelligence artificielle (AIT), Algorithme, machine à vecteurs de support (SVM), foret aléatoire (RF), Extreme Gradient Boosting (XGBoost), Prédiction de la qualité de l'eau, Ensemble de données.

الملخص

تعتبر محطة معالجة مياه الصرف الصحي جزءًا أساسيًا من الدورة المائية، حيث تقلل من تركيزات الملوثات في البيئة. لتحسين مراقبة وكفاءة محطات معالجة مياه الصرف الصحي، قام الباحثون بتطوير نماذج وأنظمة مختلفة. تقدم هذه الدراسة تطبيق تقنيات الذكاء الاصطناعي القائمة على التعلم الآلي مثل خوارزمية الغابة العشوائية(**RF**) ، وآلة دعم المتجهات (**SVM**)، والتعزيز المتدرج الأقصى (**XGBoost**) لتصميم مصنف تلقائي لجودة المياه وتحديد الوجهة المناسبة للمياه المعالجة، مع تقديم تبريرات وتوصيات مباشرة استنادًا إلى المعايير الدولية والحدود القصوى. لتحقيق هذا الهدف، تم استخدام مجموعة بيانات مكونة من 3600 قيمة تتعلق بمياه الصرف الصحي المنزلية، مع تصنيف النتائج إلى فئتين: المياه غير النقية الداخلة (مياه الصرف غير المعالجة) والمياه النقية الخارجة (المياه المعالجة). تم الحصول على حوالي 240 نقطة بيانات من السجلات الجزائرية، والتي تغطي عشر سنوات من البيانات الشهرية. تم استخدام المعلمات الداخلة، بما في ذلك الطلب البيولوجي على الأكسجين(**BOD**) ، والطلب الكيميائي على الأكسجين(**COD**) ، والنيتروجين الكلي (**TKN**) كمدخلات لتوقع جودة مياه الصرف الصحي، والتي تم تحديدها كأكثر الخصائص التنبئية من خلال طريقة اختيار الخصائص المستندة إلى الارتباط(**CFS**) ، مع وجود بيانات كافية وقيم صحيحة. تم تدريب نماذج التعلم الآلي باستخدام 60% من قيم مجموعة البيانات، وتم اختبار دقتها على الـ 40% المتبقية. من نتائج الاختبار، تبين أن خوارزمية الغابة العشوائية كانت الأكثر نجاحًا بدقة 99.8% رغم أن جميع النماذج حققت دقة ممتازة، لأن الخصائص الفعالة في هذه الحالة كانت ثلاثة فقط والبيانات كانت بسيطة. يبدو أن نموذج **SVM** هو الأسرع رغم أن خوارزمية الغابة العشوائية أعطت نتائج قريبة من**SVM** ، ولكن يبدو أن سرعة تدريب **XGBoost** أطول بحوالي 7 مرات من **SVM.** علاوة على ذلك، تم دمج وظائف مختلفة لتحديد ما إذا كانت مياه الصرف الصحي المتوقعة مناسبة للزراعة أو البيئة أو غير مناسبة لهما معًا، مع تقديم أسباب وتوصيات أو نصائح تمكّننا من إنشاء منصة للتنبؤ الرقمي بالمياه من خلال تطبيق برمجة التعلم الآلي. مهدت النتائج الواعدة الطريق لتوقع أداء عمليات محطات معالجة مياه الصرف الصحي من خلال التنبؤ بجودة المياه، وتحسين إعادة استخدام مياه الصرف الصحي المعالجة في الزراعة، ومعالجة التشوهات العملية بسرعة قبل أن تتفاقم إلى مشاكل أكثر خطورة، مما يتيح اتخاذ قرارات مدروسة من قبل مديري نظم المياه.

الكلمات المفتاحية: مياه الصرف الصحي(**WW**)، محطة معالجة مياه الصرف الصحي(**WWTP**)، التعلم الآلي(**ML**)، تقنيات الذكاء الاصطناعي(**AIT**)، خوارزمية، آلة دعم المتجهات (**SVM**)، الغابة العشوائية(**RF**)، التعزيز المتدرج الأقصى(**XGBoost**)، توقع جودة المياه، مجموعة البيانات.