

**Ministry of Higher Education and Scientific Research
20 Aout 1955 University – Skikda**



**Faculty of Science
Department of Computer Science
Specialty: MASTER 2 IA**

**Design and Development of a Configurable Framework for Soft
Sensor Generation in Wastewater Treatment: A Case Study of
BOD5 Prediction at the Sidi Merouane WWTP**

A project to obtain a Master's Degree

Presented by:

- **BEN LOUCIF Nour El Houda**
- **ZEGHDOUDI Achouak**

Supervised by:

- **Mr. TOUIL Ghassen**

Academic Year: 2025-2026

DEDICATION

All praise is due to Allah, it gives me great pleasure to dedicate this work, the fruit of my studies, to all my loved ones.

Especially:

I dedicate this modest work to the two people dearest to my heart, to whom I owe all my respect and gratitude for their upbringing, their unwavering support, and their love:

« To my father, who always told me that challenges aren't meant to be met but to be overcome, and that you must always be involved and dedicated to your work. And for the pride you show me every day. »

« To my mother, who taught me that in life nothing is impossible, there are only possibilities we don't know about, and who has always helped me find those possibilities. For never doubting myself and my abilities. »

To my brothers Anis, Abderraouf and Amer for their support.

To everyone who supported me until the completion of this work, I offer my thanks.

It is with great respect that we express our gratitude to our supervisor Mr. Touil Ghassen thank him for his assistance, his availability and his advice.

Nour El Houda

DEDICATION

All praise is due to Allah, whose blessings and guidance have enabled me to reach this milestone and accomplish this work.

I dedicate this achievement first and foremost to my beloved parents, the greatest source of support, strength, and motivation in my life.

To my father, who has always been there for me, ready to help and support me at every moment.

To my mother, whose love, care, and prayers have been my greatest blessing, and who often understands my needs before I even express them.

To my dear sister, my companion, my constant supporter, and my lifelong encourager.

To my brother, a source of happiness and joy in my life.

To my entire family, with all my love and appreciation. May Allah have mercy on those who left us before witnessing this moment.

To my friends, my companions, and my cousins, thank you for your support, kindness, and the beautiful memories we share.

I also dedicate this work to everyone who contributed to my journey, whether directly or indirectly. Every experience, challenge, and lesson helped shape the person I am today.

Finally, I dedicate this work to myself, for the patience, perseverance, and determination that made this achievement possible.

Achouak

AKNOWLEDGEMENT

First and foremost, we thank Allah, the Almighty, who gave us the patience and courage to complete this work, and who provided us with the ability to write and to think, and the strength to believe in it.

At the end of this work, we wish to express our gratitude and thanks to all those who contributed to the completion of this final year project.

We would like to extend our deepest appreciation to our supervisor, Mr. Touil Ghassen, for his valuable guidance, continuous support, and insightful advice throughout the realization of this project.

We would also like to express our heartfelt thanks to Mr. Haddam Salim, from hydraulic department whose encouragement, support, and constant motivation played a significant role in the success of this project.

Special thanks are also addressed to the staff of Larbi Ben M'hidi Station, particularly the station manager and the deputy manager, for their valuable help, availability, and collaboration.

We would also like to express our deepest gratitude to our parents and families for their unconditional love, endless support, encouragement, and sacrifices throughout our academic journey. Their faith in us has been a constant source of strength and motivation.

Our thanks also extend to all our professors who taught us and shared their experiences with us.

Finally, we would like to express our gratitude to everyone who contributed to this project, directly or indirectly, and to all those who generously shared their time, knowledge, ideas, and support with us throughout this journey.

RÉSUMÉ

La surveillance de la qualité des eaux usées est essentielle au bon fonctionnement des stations d'épuration et à la protection des milieux récepteurs. Parmi les principaux indicateurs utilisés pour évaluer la pollution organique, la demande biochimique en oxygène sur cinq jours (BOD5) occupe une place importante, car elle traduit la charge organique biodégradable présente dans les eaux usées. Toutefois, sa détermination conventionnelle repose sur une incubation de cinq jours en laboratoire, ce qui limite son utilisation pour une prise de décision rapide et pour le suivi en temps opportun des procédés de traitement.

Les capteurs logiciels constituent une alternative prometteuse, car ils permettent d'estimer des variables difficiles, lentes ou coûteuses à mesurer à partir d'autres variables plus facilement accessibles. Dans le domaine du traitement des eaux usées, les modèles d'apprentissage automatique peuvent apprendre les relations entre les paramètres physicochimiques mesurés en routine et la BOD5, afin de fournir des estimations rapides susceptibles d'améliorer le suivi et l'aide à la décision. Néanmoins, de nombreux travaux existants se limitent à un modèle unique conçu pour une station et un jeu de données spécifiques. Ces solutions sont souvent difficiles à reproduire, à adapter, à comparer ou à étendre à d'autres paramètres cibles et à d'autres contextes de fonctionnement.

Ce travail propose la conception et l'implémentation d'un framework configurable destiné au développement de capteurs logiciels fondés sur les données dans le domaine du traitement des eaux usées. Le framework couvre les principales étapes du cycle de développement d'un capteur logiciel, notamment la préparation des données, le prétraitement, la configuration des expériences, l'entraînement des modèles de régression, l'évaluation des performances, la comparaison des modèles, l'interprétation des résultats et la prédiction à partir du modèle sélectionné. Un prototype fonctionnel a été développé afin de démontrer ces fonctionnalités.

La validation du framework a été réalisée à travers une étude de cas portant sur la prédiction de la BOD5 à la station d'épuration de Sidi Merouane, en Algérie. Des données physicochimiques historiques ont été utilisées pour entraîner et comparer cinq modèles de régression : la régression linéaire, la régression par forêts aléatoires, Extra Trees, XGBoost et CatBoost. Trois scénarios expérimentaux ont été considérés : la génération d'un capteur logiciel local pour la BOD5, la comparaison des modèles candidats et l'évaluation de différentes combinaisons de variables d'entrée. Le modèle CatBoost a obtenu les meilleures

performances sur l'ensemble de test, avec un coefficient de détermination R^2 de 0,9042, une erreur quadratique moyenne RMSE de 1,5205 mg/L et une erreur absolue moyenne MAE de 1,1270 mg/L. L'analyse fondée sur la méthode SHAP a montré que la DCO, la conductivité spécifique et le pH figurent parmi les variables les plus influentes.

Les résultats obtenus démontrent la faisabilité du framework proposé pour le développement de capteurs logiciels de BOD5 à partir de données issues de stations d'épuration. Le prototype doit toutefois être considéré comme un outil d'expérimentation et d'aide à la décision ; il ne remplace pas les analyses de laboratoire et ne constitue pas une intégration opérationnelle en temps réel. Les travaux futurs devront porter sur la validation à partir de plusieurs stations, l'utilisation de séparations chronologiques des données, le réentraînement périodique des modèles et l'intégration avec des systèmes réels de surveillance.

Mots-clés : capteur logiciel, prédiction de la BOD5, traitement des eaux usées, apprentissage automatique, framework configurable, CatBoost, interprétabilité des modèles, SHAP.

ABSTRACT

The monitoring of wastewater quality is essential for the effective operation of wastewater treatment plants and for the protection of receiving environments. Among the main indicators used to assess organic pollution, the five-day biochemical oxygen demand (BOD5) remains particularly important because it reflects the biodegradable organic load of wastewater. However, the conventional determination of BOD5 requires a five-day incubation period under controlled laboratory conditions. This delay limits its usefulness for timely operational decision-making and prevents its direct use for rapid process monitoring.

Soft sensors provide a promising alternative by estimating difficult-to-measure variables from other physicochemical parameters that are more readily available. In wastewater treatment, machine learning models can learn the relationships between routinely measured variables and BOD5, thereby providing rapid estimates that may support monitoring and decision-making. Nevertheless, many existing studies focus on a single model developed for a particular dataset and treatment plant. Such solutions are often difficult to reproduce, adapt, compare, or extend to other target variables and operating conditions.

This work proposes the design and implementation of a configurable framework for the development of data-driven soft sensors in wastewater treatment. The framework supports the main stages of the soft-sensor development process, including dataset preparation, preprocessing, experimental configuration, regression-model training, performance evaluation, model comparison, interpretation of results, and prediction using a selected model. A functional prototype was implemented to demonstrate these capabilities.

The framework was validated through a case study involving BOD5 prediction at the Sidi Merouane wastewater treatment plant in Algeria. Historical physicochemical data were used to train and compare five regression models: Linear Regression, Random Forest Regression, Extra Trees Regression, XGBoost, and CatBoost. Three experimental scenarios were considered: the generation of a local BOD5 soft sensor, the comparison of candidate models, and the evaluation of different input-variable combinations. CatBoost achieved the best performance on the test set, with an R^2 of 0.9042, an RMSE of 1.5205 mg/L, and an MAE of 1.1270 mg/L. SHAP-based analysis indicated that COD, specific conductivity, and pH were among the most influential variables.

The results demonstrate the feasibility of the proposed framework for developing BOD5 soft sensors from treatment-plant data. The prototype should be regarded as a decision-support and experimentation tool; it does not replace laboratory analysis or constitute a live operational deployment. Future work should focus on validation using multiple treatment plants, chronological data splits, periodic retraining, and integration with real monitoring systems.

Keywords: soft sensor, BOD5 prediction, wastewater treatment, machine learning, configurable framework, CatBoost, model explainability, SHAP.

ملخص

تُعد مراقبة جودة المياه العادمة أمراً أساسياً لضمان حسن تشغيل محطات معالجة المياه المستعملة وحماية الأوساط البيئية المستقبلية. ومن بين أهم المؤشرات المعتمدة لتقييم التلوث العضوي، تبرز الحاجة البيوكيميائية إلى الأكسجين خلال خمسة أيام، والمعروفة اختصاراً بـBOD5، باعتبارها مؤشراً يعكس الحمل العضوي القابل للتحلل البيولوجي في المياه العادمة. غير أن القياس التقليدي لهذا المؤشر يعتمد على حضانة مخبرية تستغرق خمسة أيام، وهو ما يحد من فائدته في اتخاذ القرارات السريعة وفي المتابعة الآنية لعمليات المعالجة.

تمثل المستشعرات البرمجية حلاً واعداً لهذه الإشكالية، إذ تسمح بتقدير المتغيرات التي يصعب قياسها أو التي تتطلب وقتاً وتكلفة مرتفعين، انطلاقاً من متغيرات أخرى تكون أكثر سهولة في القياس والتوفر. وفي مجال معالجة المياه العادمة، يمكن لنماذج تعلم الآلة أن تتعلم العلاقات بين المتغيرات الفيزيوكيميائية المقاسة بشكل روتيني بقيمة BOD5، من أجل تقديم تقديرات سريعة يمكن أن تدعم المراقبة واتخاذ القرار. ومع ذلك، فإن العديد من الدراسات الحالية تقتصر على نموذج واحد مخصص لمحة معينة ومجموعة بيانات محددة، مما يجعل هذه الحلول صعبة التكرار أو التكيف أو المقارنة أو التوسيع لتشمل متغيرات مستهدفة وسياقات تشغيلية أخرى.

يقترح هذا العمل تصميم وتنفيذ إطار عمل قابل للتهيئة لتطوير المستشعرات البرمجية المعتمدة على البيانات في مجال معالجة المياه العادمة. يغطي هذا الإطار المراحل الرئيسية لدورة تطوير المستشعر البرمجي، بما في ذلك إعداد البيانات، والمعالجة المسبقة، وتهيئة التجارب، وتدريب نماذج الانحدار، وتقييم الأداء، ومقارنة النماذج، وتفسير النتائج، وإجراء التنبؤات باستخدام النموذج المختار. وقد تم تطوير نموذج أولي وظيفي لإثبات هذه الإمكانيات. تم التحقق من صحة الإطار المقترح من خلال دراسة حالة تخص التنبؤ بقيمة BOD5 في محطة معالجة المياه المستعملة بسيدي مروان في الجزائر. وقد استُخدمت بيانات فيزيوكيميائية تاريخية لتدريب ومقارنة خمسة نماذج انحدار، وهي: الانحدار الخطي، والانحدار بالغابات العشوائية، وExtra Trees، وXGBoost، وCatBoost. كما تم اعتماد ثلاثة سيناريوهات تجريبية تشمل: إنشاء مستشعر برمجي محلي للتنبؤ بـBOD5، ومقارنة النماذج المرشحة، وتقييم مجموعات مختلفة من متغيرات الإدخال. وقد حقق نموذج CatBoost أفضل أداء على مجموعة الاختبار، حيث بلغ معامل التحديد R^2 قيمة 0.9042، وبلغت قيمة الجذر التربيعي لمتوسط مربع الخطأ RMSE مقدار 1.5205 mg/L، بينما بلغت قيمة متوسط الخطأ المطلق MAE مقدار 1.1270 mg/L. كما أظهر التحليل المعتمد على تقنية SHAP أن الطلب الكيميائي على الأكسجين DCO، والموصلية النوعية، ودرجة الحموضة pH، تعد من أكثر المتغيرات تأثيراً في عملية التنبؤ.

تؤكد النتائج المتحصل عليها قابلية تطبيق الإطار المقترح لتطوير مستشعرات برمجية خاصة بالتنبؤ بـBOD5 انطلاقاً من بيانات محطات معالجة المياه المستعملة. ومع ذلك، ينبغي اعتبار النموذج الأولي أداة للتجريب ودعم

القرار، إذ لا يعوض التحاليل المخبرية ولا يمثل نظاماً تشغيلياً مدججاً في الزمن الحقيقي. ويمكن أن تركز الأعمال المستقبلية على التحقق من الإطار باستخدام بيانات من عدة محطات، واعتماد تقسيمات زمنية للبيانات، وإعادة تدريب النماذج بشكل دوري، وربط الإطار بأنظمة المراقبة الفعلية.

الكلمات المفتاحية: مستشعر برنجي، التنبؤ بـBOD5، معالجة المياه العادمة، تعلم الآلة، إطار عمل قابل للتهيئة، CatBoost، قابلية تفسير النماذج، SHAP.

TABLE OF CONTENTS

Dedication.....	i
Acknowledgment.....	iii
Résumé	iv
Abstract.....	vi
ملخص	viii
AGENERAL INTRODUCTION	1
Chapter 1: Wastewater Treatment Context and the BOD5 Monitoring Challenge.....	4
1.1 Introduction	5
1.2 Wastewater	5
1.2.1 Definition of wastewater	5
1.2.2 Wastewater Classification	5
1.2.3 Domestic effluent	6
1.2.4 Effluents from industrial, municipal or craft establishments	6
1.2.5 Agricultural effluents.....	6
1.2.6 Effluents of natural origin	6
1.3 Types of wastewater treatment	6
1.3.1 Centralized wastewater treatment.....	6
1.3.2 On-site (or individual or autonomous) wastewater treatment	7
1.4 Presentation of the Sidi Merouane wastewater treatment plant	7
1.4.1 Technical characteristics and basic sizing data for the WWTP	7
1.5 Treatment Processes at the Sidi Merouane Wastewater Treatment	9
1.5.1 Pumping Station	9
1.6 Parameters of treated water	18
1.6.1 Physical parameters	18
1.6.2 Chemical Parameters	21
1.7 Importance of BOD5	22

1.8 Limitations of the conventional measurement of BOD5.....	23
1.9 Need for a Software Sensor	24
1.10 Transition to a Framework	24
1.11 Conclusion.....	24
Chapter 2: Soft Sensors and Machine Learning Methods for Wastewater Quality	
Prediction.....	26
2.1 Introduction	27
2.2 Definition of a Software Sensor	27
2.3 Data Preprocessing	27
2.3.1 Cleaning.....	27
2.3.2 Normalization	28
2.3.3 Data Splitting.....	28
2.3.4 The Savitzky-Golay Filter	28
2.4 Software Sensors in Water Treatment.....	28
2.4.1 BOD5 Estimation	28
2.4.2 COD Estimation	28
2.4.3 Turbidity Prediction.....	29
2.4.4 Suspended Solids (SS) Estimation	29
2.4.5 Treated Water Quality Monitoring.....	29
2.5 Regression Models Used in the Framework.....	29
2.5.1 Random Forest Regression (RFR) Model	29
2.5.2 Linear Regression Model Prediction	31
2.5.3 CatBoost Model.....	32
2.5.4 Extra Trees Model	33
2.5.5 XGBoost Model.....	34
2.6 Evaluation Metrics.....	35
2.6.1 Coefficient of Determination (R^2)	35

2.6.2 Root Mean Square Error (RMSE)	36
2.6.3 Mean Absolute Error (MAE).....	36
2.7 Interpretation of Models	37
2.8 Model Optimization.....	37
2.8.1 Grid Search	38
2.9 Conclusion.....	38
Chapter 3: Design of the Configurable Soft Sensor Framework.....	39
3.1 Introduction	40
3.2 Definition of framework.....	40
3.3 Target users	41
3.4 General architecture of the framework.....	42
3.5 Description of the framework modules	43
3.6 Difference between a fixed application and a framework.....	44
3.7 Conclusion.....	45
Chapter 4: Implementation of the Framework Prototype.....	46
4.1 Development Environment.....	47
4.2 Software Architecture of the Prototype.....	48
4.3 Authentication and Role Management	49
4.4 Dataset Import Module.....	50
4.5 Preprocessing Module	51
4.6 Model Training Module	52
4.7 Evaluation and Comparison Module	53
4.8 BOD5 Prediction Module.....	54
4.9 Dashboard and Analytics Module	55
4.10 Conclusion.....	57
Chapter 5: Experimental Validation of the Framework: BOD5 Soft Sensor at the Sidi Merouane WWTP.....	58

5.1 Dataset Description	59
5.2 Experimental Protocol	61
5.2.1 Data cleaning	61
5.2.2 Data denoising	61
5.2.3 Data Split	62
5.2.4 Models Evaluated	62
5.2.5 Evaluation Metrics.....	63
5.2.6 Comparison Method	63
5.3 Scenario 1 — Local BOD5 Soft Sensor	63
5.4 Scenario 2 — Comparative Model Evaluation.....	65
5.5 Scenario 3 — Input Variable Combination Analysis	65
5.6 Best Soft Sensor Selection.....	66
5.7 Influential Variable Analysis.....	67
5.8 Discussion.....	68
GENERAL CONCLUSION.....	70
REFERENCES	74

List of Figures

Figure 1.1: Sidi Merouane Activated-Sludge Wastewater Treatment Plant (Chetti, 2019, modified)	8
Figure 1.2: Overview of the Sidi Merouane Wastewater Treatment Plant	9
Figure 1.3: Reception and Pumping Station.....	10
Figure 1.4: Automatic Coarse Screen (1) and Manual Screen (2).....	10
Figure 1.5: Automatic Fine Screen (1) and Manual Fine Screen (2)	11
Figure 1.6: Sand Separator and Oil Separator	12
Figure 1.7: Sand Pump	12
Figure 1.8: Sand Classifier	12
Figure 1.9: Biological Basin.....	13
Figure 1.10: Aeration System.....	14
Figure 1.11: Degasser	14
Figure 1.12: Clarifier	15
Figure 1.13: Sludge Recirculation	15
Figure 1.14: Drip Table (1) and Belt Filter (2).....	16
Figure 1.15: Drying Beds	16
Figure 1.16: Storage Areas	17
Figure 1.17: Control Room of the Sidi Merouane Wastewater Treatment Plant	17
Figure 1.18: Sidi Merouane Wastewater Treatment Plant Monitoring Laboratory	18
Figure 1.19: pH Measurement and Conductivity Measurement.....	19
Figure 1.20: Equipment for Measuring Suspended Solids: Vacuum Filtration and Centrifugation.....	20
Figure 1.21: Dissolved Oxygen (DO) Measurement by Oximetry	20
Figure 1.22: Equipment for Measuring Chemical Oxygen Demand: Thermoreactor and Spectrophotometer.....	21
Figure 1.23: BOD Meter and BOD5 Bottle for Measuring Biochemical Oxygen Demand over Five Days.....	22
Figure 2.1: Structure of the Random Forest Regression Model	31
Figure 2.2: Structure of the Linear Regression Model	32
Figure 2.3: Structure of the CatBoost Model	33
Figure 2.4: Structure of the Extra Trees Model.....	34
Figure 2.5: Structure of the XGBoost Model	35

Figure 3.1: Framework Architecture	42
Figure 4.1: Project Structure in Visual Studio Code	49
Figure 4.2: Application Login Interface	50
Figure 4.3: Dataset Management Page: Metadata, Validation Report, and Data Preview ..	51
Figure 4.4: Data Preprocessing Page: Configuration Panel and Preprocessing Log.....	52
Figure 4.5: Model Training Page: Algorithm Selection, Results Table, and Deploy Button	53
Figure 4.6: Prediction Page: Input Fields and Colour-Coded BOD5 Result Panel.....	54
Figure 4.7: Admin Dashboard: Model Registry and Status Overview.....	55
Figure 4.8: Operator Analytics Page: BOD5 Trend and Input-Parameter Trend Charts	56
Figure 4.9: Operator Analytics Page: Historical Table	56
Figure 5.1: Pearson Correlation Coefficients between Input Variables and BOD5.....	61
Figure 5.2: Savitzky-Golay Denoising of BOD5: Window-Size Comparison.....	62
Figure 5.3: Original BOD5 Values versus Filtered BOD5 Values	62
Figure 5.4: Predicted versus Actual BOD5 Values: CatBoost Model.....	64
Figure 5.5: SHAP Feature Contribution Analysis	67

List of Tables

Table 1.1: Technical Characteristics of the Sidi Merouane Wastewater Treatment Plant	8
Table 3.1: Framework Module Descriptions and Prototype Correspondence	43
Table 3.2: Comparison between a Fixed Application and the Proposed Framework.....	44
Table 4.1: Development Tools and Libraries Used in the Project.....	47
Table 4.2: Performance Comparison across All Models (80/20 Split, Full Feature Set)	53
Table 5.1: Experimental Dataset Summary	59
Table 5.2: Input Variable Information	60
Table 5.3: Model Performance on the Test Set with the Full Feature Set (80/20 Split).....	64
Table 5.4: Overfitting Analysis for the CatBoost Model	65
Table 5.5: CatBoost Performance across Nine Input-Variable Combinations	66

List of Equations

Equation 2.1: Random Forest Regression Prediction.....	30
Equation 2.2: Coefficient of Determination (R^2).....	36
Equation 2.3: Root Mean Square Error (RMSE).....	36
Equation 2.4: Mean Absolute Error (MAE)	36

List of Abbreviations

Wwtp : Wastewater Treatment Plant

BOD5 : Biochemical Oxygen Demand (5 days)

COD : Chemical Oxygen Demand

SS : Suspended Solids

TSS : Total Suspended Solids

DO : Dissolved Oxygen

pH : Potential of Hydrogen

EC : Electrical Conductivity

TDS : Total Dissolved Solids

RFR : Random Forest Regression

CART : Classification And Regression Trees

XGBoost : Extreme Gradient Boosting

CatBoost : Categorical Boosting

SHAP : SHapley Additive exPlanations

R² : Coefficient of Determination

RMSE : Root Mean Square Error

MAE : Mean Absolute Error

ML : machine learning

GENERAL INTRODUCTION

Wastewater treatment plants play a fundamental role in protecting public health, preserving aquatic ecosystems, and reducing the environmental impact of domestic, agricultural, and industrial discharges. Their efficiency depends on the continuous monitoring of physical, chemical, and biological parameters that describe the quality of influent and treated water. Among these parameters, the five-day biochemical oxygen demand, commonly denoted as BOD5, is one of the most important indicators of biodegradable organic pollution. It is widely used to assess the organic load of wastewater, evaluate treatment performance, and support the verification of effluent quality before discharge or reuse.

Despite its importance, BOD5 presents an important operational limitation. Its conventional laboratory measurement requires the incubation of a water sample for five days under controlled conditions. Consequently, the result becomes available long after the conditions that produced it have changed. This delay reduces the usefulness of BOD5 for rapid operational decisions, particularly when treatment-plant operators need to detect abnormal conditions, assess pollution variations, or adjust operating parameters within a short time. Laboratory measurement remains necessary as the reference method, but it cannot by itself provide immediate information about the current state of the treatment process.

This limitation has motivated the use of soft sensors. A soft sensor is a computational model that estimates a variable that is difficult, slow, or costly to measure from other variables that are more easily available. In the context of wastewater treatment, a BOD5 soft sensor can use physicochemical parameters such as flow rate, water temperature, pH, chemical oxygen demand, total suspended solids, and specific conductivity to estimate BOD5 rapidly. Machine learning regression techniques are particularly suitable for this purpose because they can model complex and non-linear relationships between input variables and the target parameter.

However, the development of a single predictive model is not sufficient to address the broader needs of wastewater-treatment applications. A model trained for one treatment plant, one dataset, and one group of input variables may not be directly transferable to another plant or another prediction task. Moreover, a fixed prediction application generally provides limited support for comparing algorithms, testing preprocessing choices, analysing

input-variable combinations, and documenting the experimental process. These limitations reduce the reproducibility and reusability of soft-sensor studies.

For this reason, the present work does not focus only on the development of a BOD5 prediction model. Its main objective is to design and implement a configurable framework for the development, evaluation, and use of soft sensors in wastewater treatment. The proposed framework is intended to provide a structured environment in which users can prepare data, define experimental configurations, train several regression models, compare their performance, select an appropriate model, interpret the prediction results, and use the retained model for new estimations. The BOD5 prediction problem at the Sidi Merouane wastewater treatment plant is used as a case study to validate the framework.

The work is guided by the following research question:

How can a configurable framework be designed and implemented to support the development, evaluation, and comparison of soft sensors for wastewater treatment, and how can this framework be validated through BOD5 prediction at the Sidi Merouane wastewater treatment plant?

To answer this question, the study pursues four main objectives. The first objective is to analyse the wastewater-treatment context and the operational importance of BOD5 monitoring. The second objective is to review the principles of soft sensors, data preprocessing, machine learning regression models, performance metrics, and model-interpretation methods. The third objective is to design and implement a configurable framework prototype that supports the main stages of soft-sensor development. The fourth objective is to validate this prototype through the generation and evaluation of a BOD5 soft sensor using historical data from the Sidi Merouane wastewater treatment plant.

The scope of this work is limited to the design and validation of a prototype framework and to an experimental case study based on historical station data. The application demonstrates the use of a selected trained model to generate BOD5 estimates from physicochemical inputs. However, direct integration with a live SCADA environment, continuous acquisition from real-time sensors, and operational validation under actual plant conditions remain outside the scope of the present study.

This dissertation is organised into five chapters. Chapter 1 presents the wastewater-treatment context, the Sidi Merouane wastewater treatment plant, and the BOD5 monitoring

challenge. Chapter 2 introduces soft sensors and the machine learning methods used for wastewater-quality prediction. Chapter 3 presents the design of the proposed configurable framework, including its objectives, architecture, and principal modules. Chapter 4 describes the implementation of the framework prototype and its main interfaces. Chapter 5 presents the experimental validation of the framework through the development and evaluation of a BOD5 soft sensor for the Sidi Merouane wastewater treatment plant. Finally, the general conclusion summarises the main contributions, results, limitations, and future research directions.

CHAPTER 1

WASTEWATER TREATMENT CONTEXT AND THE BOD5 MONITORING CHALLENGE

1.1 Introduction

Wastewater, regardless of its source, is a polluted effluent that can pose a serious threat to the environment and the health of users. It is therefore crucial not to discharge it directly into the environment. Instead, it must be treated in a wastewater treatment plant, where as many contaminants as possible will be removed to ensure compliance with current discharge standards.

Wastewater treatment plants are of vital importance for the preservation of the environment, public health, and aquatic ecosystems. They represent a fundamental pillar for sustainable development and the protection of our natural resources. Their main role is to treat wastewater from homes, industries, and communities before releasing it into the environment or reusing it for non-potable purposes such as irrigation, industrial cleaning, or groundwater recharge. This is a sustainable solution for preserving water resources in a context where access to water is becoming increasingly restricted in some regions.

In this context, part of this chapter is dedicated to the impact of wastewater treatment, highlighting the various aspects of this treatment process and the evaluation of the performance of the Sidi Merouane wastewater treatment plant by studying various pollution parameters such as: BOD5, COD, and TSS. Other parameters such as pH, temperature, conductivity, and oxygen concentration, in both raw and treated water, are also examined to ensure their compliance with discharge standards.

1.2 Wastewater

1.2.1 Definition of wastewater

According to Rejsek, wastewater is water laden with physical, chemical, or biological pollutants, primarily originating from human activity, that can contaminate the environment into which it is discharged. The term "residual water" refers to water from various sources that has lost its initial purity, in other words, its natural characteristics, due to contamination by pollutants following its use in human activities such as domestic, industrial, or agricultural activities.

1.2.2 Wastewater Classification

Wastewater comes from three main sources: Domestic wastewater, Industrial wastewater, and Runoff.

1.2.3 Domestic effluent

This consists of domestic wastewater: household wastewater (water from sinks, showers, bathtubs, household appliances, etc.). In addition to this highly polluted water, there is also (depending on the type of sewer system: separate, semi-separate, or combined) less polluted water that may come from roofs, drainage systems, yards, basements, and garages.

1.2.4 Effluents from industrial, municipal or craft establishments

which come from domestic activity (canteens, toilets, showers, etc.) or more specifically from waters linked to industrial activity, such as cooling water, or effluents resulting from the process used.

1.2.5 Agricultural effluents

Agriculture is the primary cause of diffuse pollution. Pollution from agricultural sources includes both that related to crops (pesticides and fertilizers) and livestock (manure and slurry).

1.2.6 Effluents of natural origin

The concentration of undesirable substances in water results from the activity of certain natural phenomena (volcanic eruptions, contact of water with mineral deposits, etc.).

1.3 Types of wastewater treatment

Wastewater treatment aims to reduce the pollutant load (organic and mineral matter) contained in wastewater before it is discharged into the natural environment or reused.

There are two types of wastewater treatment:

1.3.1 Centralized wastewater treatment

This is a municipal wastewater treatment system to which homes in an urban area are connected.

Wastewater is collected and then transported via a network of pipes to a treatment plant, which removes the pollutants. This type of wastewater treatment system primarily concerns densely populated areas.

1.3.2 On-site (or individual or autonomous) wastewater treatment

This applies to dwellings not connected to a public sewer system. On-site wastewater treatment systems consist of a pretreatment unit and a treatment unit that utilizes the soil's self-purifying capacity. This type of wastewater treatment is primarily used in areas with dispersed housing. According to Schriver Mazzuoli, an on-site wastewater treatment system must include a septic tank (Figure 4), a tank ventilation system, and a drain field.

1.4 Presentation of the Sidi Merouane wastewater treatment plant

The Sidi Merouane wastewater treatment plant is located 12 km northeast, south of the Beni Haroun Dam. The site covers an area of 16 hectares and is intended to reduce pollution in the lake located near the Beni Haroun Dam, which collects the waters of the Rhumel and Endja rivers. The treatment plant is one of three planned for this watershed. It was built by Sonatrach, a group specializing in civil engineering and construction (GCB), and equipped by OTV. It was commissioned in 2009 before being handed over to ONA in 2011.

The plant is designed to treat domestic wastewater from seven pumping stations, namely those of Mila, Grarem Gouga, Sidi Merouane, Ras El Bir, Annouche Ali, Sibari 1 and 2.

Once treated, this water is discharged into the catchment area of the Beni Haroune dam.

It is then conveyed to the drinking water treatment plant, which supplies the wilayas of Mila, Constantine, Batna, Jijel, Oum El Bouaghi, and Khenchela.

Furthermore, the Sidi Merouane wastewater treatment plant was established to purify the wastewater discharged upstream of the Beni Haroune dam and to address water pollution problems.

The purified water is intended for reuse in irrigation, while the sludge generated during the purification process is used for agricultural purposes. Thus, this wastewater treatment plant has a significant impact on agriculture.

1.4.1 Technical characteristics and basic sizing data for the WWTP

In general, the Sidi Merouane WWTP was designed to operate at a minimum load of 0.11 m³/day. The low-load activated sludge treatment process consists of a phase in which the raw water is brought into contact with the bacterial mass in the presence of oxygen (aeration), followed by the separation and settling of this mass (clarification).

Parameters	Quantity
Hydraulic load	–
Population equivalent	137,000 PE
Daily volume	20,650 m ³ /day
Average flow rate during dry weather	20,550 m ³ /day
Peak flow rate during dry weather	1,370 m ³ /day
Maximum admissible flow rate during rainy weather	25,340 m ³ /day
Mass load	0.11 m ³ /day
Pollution load	–
Daily COD load	19,180 kg/day
Daily BOD5 load	8,220 kg/day
Daily TSS load	12,330 kg/day

Table 1.1: Technical Characteristics of the Sidi Merouane Wastewater Treatment Plant

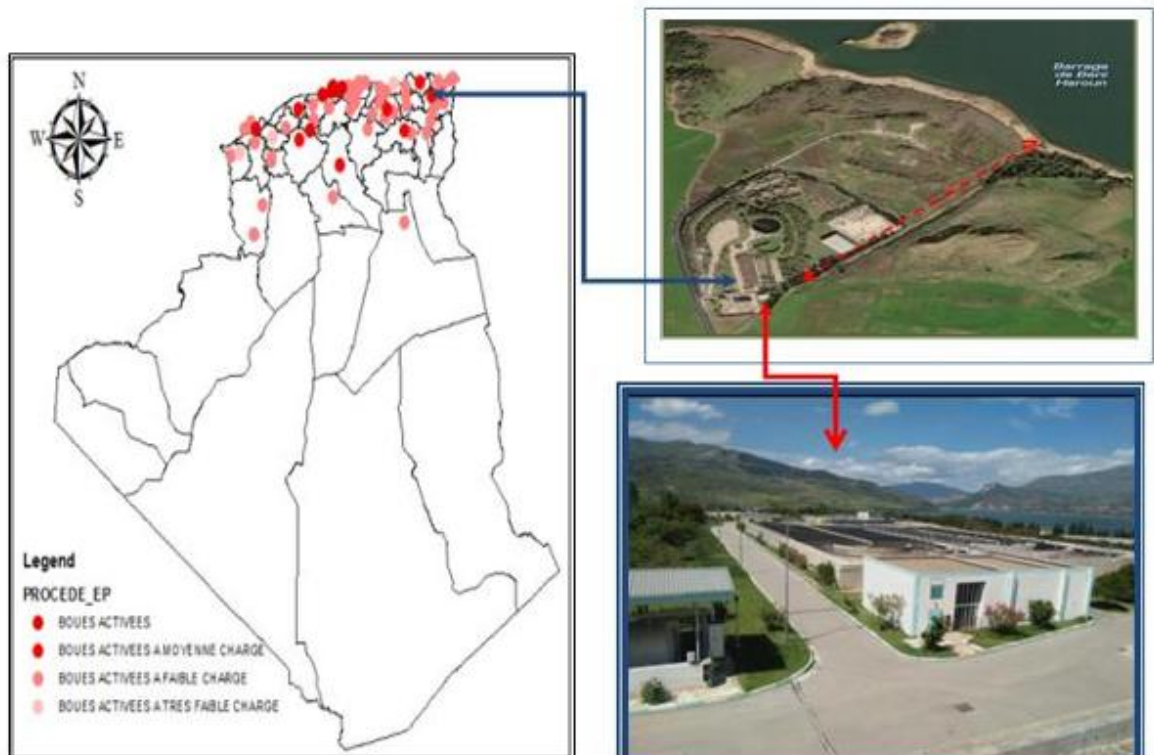


Figure 1.1: Sidi Merouane Activated-Sludge Wastewater Treatment Plant (Chetti, 2019, modified)

1.5 Treatment Processes at the Sidi Merouane Wastewater Treatment



Figure 1.2: Overview of the Sidi Merouane Wastewater Treatment Plant (Sidi Merouane Wastewater Treatment Plant, 2019)

1.5.1 Pumping Station

The pumping station is equipped with three centrifugal pumps, one of which is reserved for emergency use.

These pumps are equipped with variable frequency drives, ensuring smooth operation without water hammer.

In fact, the maximum flow rate recorded at this station was 1892 m³/h during the first phase,

which corresponds to 92% of the total rainfall. It was designed to allow for the installation of an additional pump during the second phase, in order to accommodate the projected flow rates for 2030. An overflow spillway is planned, allowing the bypass to directly discharge untreated raw water. The Viaduct/Ferdoua pumping station is responsible for pumping wastewater from the Mila, Grarem Gouga, and Ferdoua areas. It extends from the Viaduct to the eastern entrance of Sidi Merouane, towards the wastewater treatment plant and the sewer network. It connects to the wastewater treatment plant over a distance of 91 km.



Figure 1.3: Reception and Pumping Station (Sidi Merouane Wastewater Treatment Plant, 2019)

1.5.1.1 Coarse Screen

A coarse screen with a diameter of 1000 mm was installed in the raw water inlet channel before the pumping station. A second screening channel, equipped with a manual screen, was also added for use when needed. This coarse screen serves to protect the pumping pumps from clogging and to remove debris larger than 40 mm. In addition, an automatic screen is in place, linked to an upstream level detection system, which operates in synchronization with the pump. The debris collected by the coarse screen is then stored in a skip.

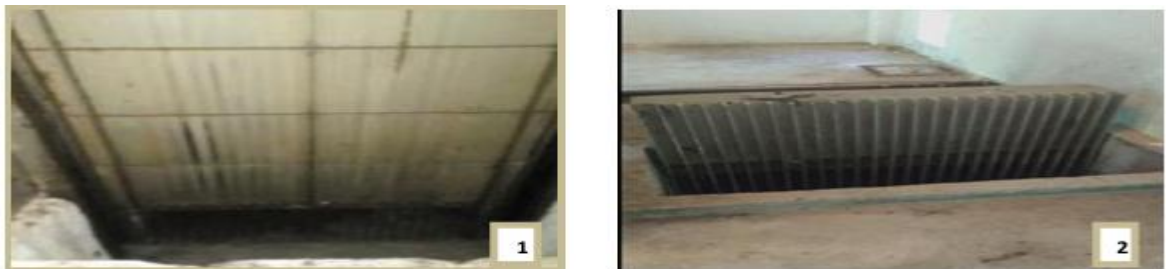


Figure 1.4: Automatic Coarse Screen (1) and Manual Screen (2) (Sidi Merouane Wastewater Treatment Plant, 2019)

1.5.1.2 Fine Screen

The raw water is first subjected to coarse screening and an elevator, before being mixed in the upstream distribution structure. It then passes through a fine screen, which retains fine solid debris; Two automatic fine screens inclined at 75° and arranged parallel.

Each screen is installed in a concrete channel, with a 15 mm spacing between the bars, designed to protect the entire installation against the accumulation of fiber and small suspended solids. A water level indicator, connected to a clock, is placed upstream, and a bypass of the wastewater treatment plant is provided before the screen.

The waste rejected by the screening machine is automatically collected by a screw conveyor, then conveyed to a high-pressure waste compactor.



Figure 1.5: Automatic Fine Screen (1) and Manual Fine Screen (2) (Sidi Merouane Wastewater Treatment Plant, 2019)

1.5.1.3 Effluent Counting

The treated effluent at the plant is collected downstream of the fine screening in a venturi channel.

Water levels are measured by ultrasound. Samples are automatically taken and kept chilled or frozen to monitor water quality at the inlet and outlet of the plant.

A bypass is also installed at the end of the grit removal process.

1.5.1.4 Grit and Oil Removal

The grit and oil removal unit is a rectangular-shaped installation designed for the sedimentation and separation of heavy residues (sand, suspended solids) from light residues (grease, fine waste). These units receive materials from external dumping.

Removing sand helps reduce wear on downstream equipment, while grease traps help transfer oxygen to the aeration basin. They are protected by cofferdams.

In fact, two grit removal units are in place, with a third planned as an emergency solution. The flat-bottomed structures are equipped with self-propelled bridges.

The primary function of these bridges is to scrape the bottom to remove sand. Next, they will scrape the surface to remove grease from the aerating pumps, which are installed on each structure. Finally, the surface scraping system moves the suspended solids to a common grease tank, which is equipped with an industrial water suction pump.



Figure 1.6: Sand Separator and Oil Separator (Sidi Merouane Wastewater Treatment Plant, 2019)

1.5.1.5 Sand Pump

Sand pumps transport the sand to the classifier, then store it in bins. To optimize the reliability of the installation, a sand extraction pump is provided for emergency situations.



Figure 1.7: Sand Pump (Sidi Merouane Wastewater Treatment Plant, 2019)

1.5.1.6 Sand Classifier

The sand classifier with a capacity of 51m³/h is installed to extract sand from the pumped residual water, with a sand pump, discharging it onto a conveyor belt common to the screening discharges.



Figure 1.8: Sand Classifier (Sidi Merouane Wastewater Treatment Plant, 2019)

1.5.1.7 Biological Treatment

The biological treatment system implemented is based on a low-load activated sludge system. This process combines nitrification and denitrification with biological dephosphorization.

To ensure optimal reliability and operational flexibility, an upstream distributor balances the distribution between these two lines by isolating them, using a set of valves or a bypass.

The plant has two reinforced concrete biological basins, each with a capacity of 13,000 m³, and consists of three distinct zones:

- Aeration zone in the channel: In the presence of oxygen, carbon pollution is oxidized and nitrified; phosphorus then accumulates in bacterial cells.
- Anoxic zone in the channel: Denitrification is achieved by circulating mixed liquid in the channel with a slow-moving agitator such as a "Banana Blade".

The alternation between aerated and non-aerated phases creates anoxic conditions favorable to denitrification.

- Anaerobic zone: Biological dephosphorization allows the transfer of phosphorus liquid from the wastewater to the purified biomass, progressively enriching it with phosphorus. This process takes place in an anaerobic environment, using only carbon without oxygen, making it easily assimilable.



Figure 1.9: Biological Basin (Sidi Merouane Wastewater Treatment Plant, 2019)

1.5.1.8 Aeration System

This involves injecting air into the bottom of the pool in the form of fine bubbles, using flexible and non-compact membranes. This method ensures efficient transfer while reducing energy costs.



Figure 1.10: Aeration System (Sidi Merouane Wastewater Treatment Plant, 2019)

1.5.1.9 Degassing

A circular structure with a diameter of 9 meters and a depth of 8 meters was installed to create a degassing pathway for the effluent in front of the clarifier. This allows the removal of gases (N_2 , CO_2) present in the water. Once the circular scrapers are installed on the surface of the structure, the suspended solids are collected by a chute, located below the degassing zone. These solids are then stored in tanks, then pumped and discharged to a pretreatment grease trap. The degassing system is designed to handle the projected flow rates by 2030.



Figure 1.11: Degasser (Sidi Merouane Wastewater Treatment Plant, 2019)

1.5.1.10 Clarifier (Settling Tank)

Water clarification takes place in two separate structures covering an area of 1521 m². The liquids from the biological treatment are purified and pumped by gravity to the clarification basins after undergoing degassing.

The sludge settles at the bottom of the structure and is then drawn up through a series of vertical pipes, progressively covering the surface of the bridge's rotation.

It is essential to emphasize that the suction pipe technology makes it possible to obtain a homogeneous and durable layer. The treated water is immediately returned to the Beni Haroune dam.



Figure 1.12: Clarifier (Sidi Merouane Wastewater Treatment Plant, 2019)

1.5.1.11 Sludge Recycling and Extraction

The generated sludge is drawn from the central shaft of each structure to the sludge tanks, where a volume of 551 m³/h is recovered in the biological basin. The excess sludge once recirculated, is measured and directed to the thickening station, with a maximum flow rate of 70 m³/h of sludge to be treated.



Figure 1.13: Sludge Recirculation (Sidi Merouane Wastewater Treatment Plant, 2019)

1.5.1.12 Sludge Treatment

The sludge treatment process goes through the following four stages:

1.5.1.13 Dewatering Table

Sludge is drawn from the tank using recirculation pumps. It undergoes a flocculation process by injecting a polymer at the beginning of each machine, which promotes the formation of flocs. Subsequently, gravity separation takes place in the dewatering zone, where almost all the liquid suspended between the materials is removed. Finally, the dewatering table is equipped with a comb, which directs the sludge to the thickening station.

1.5.1.14 Belt Filters

These filters, composed of two belts stretched between permeable fiber rollers, cover the sludge with a uniform layer and dewater it using grinding pressure. The dewatered sludge is then collected by a conveyor belt. A third belt filter is planned for 2030.



Figure 1.14: Drip Table (1) and Belt Filter (2) (Sidi Merouane Wastewater Treatment Plant, 2019)

1.5.1.15 Drying Beds

It should be mentioned that the dewatered sludge is conveyed to a collection skip by conveyor belts where it is spread onto the drying beds, which are formed by:

- *A sloping concrete lining;
- * An added sludge retention curb;
- * A drainage channel.

Thus, the management of drying beds encompasses the following steps:

- *Spreading of dewatered sludge on the beds;
- * Drying of the sludge;
- * Removal of the dried sludge.



Figure 1.15: Drying Beds (Sidi Merouane Wastewater Treatment Plant, 2019)

1.5.1.16 Storage Area

The dried sludge is produced in accordance with applicable standards. It is stored in a covered area for up to six months, after which it is placed near the beds and public roads.

It is important to emphasize that the choice of lining facilitates the treatment process. At this stage, the use of percolation sand is not required, as the sludge to be dried has already been mechanically dewatered.



Figure 1.16: Storage Areas (Sidi Merouane Wastewater Treatment Plant, 2019)

1.5.1.17 Control room

The control of the treatment operations and the operation of the wastewater treatment plant are carried out on a regular basis in the control room.

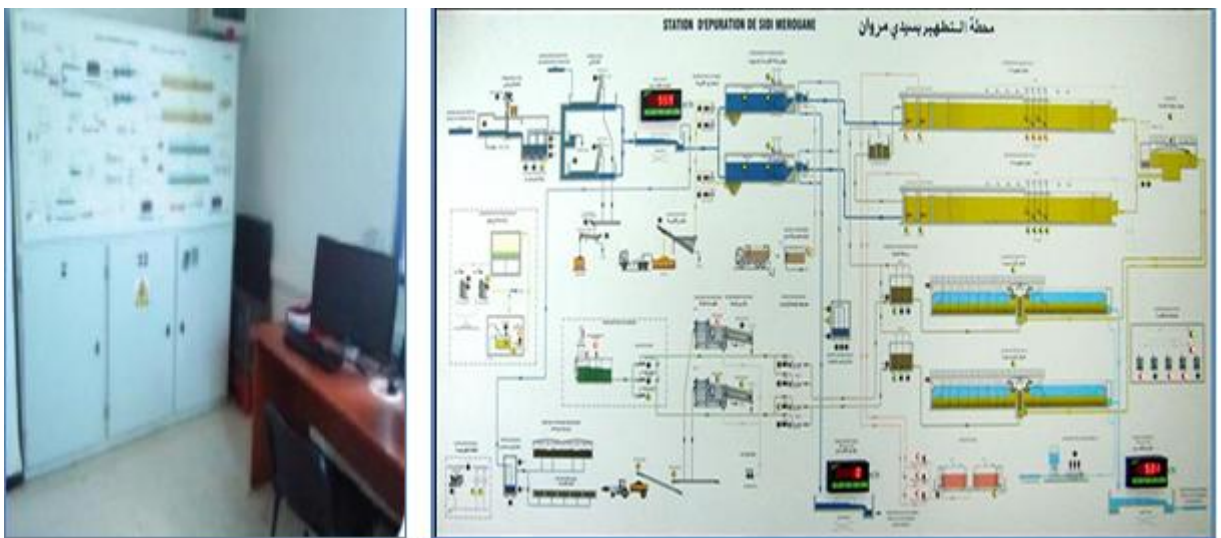


Figure 1.17: Control Room of the Sidi Merouane Wastewater Treatment Plant

1.5.1.18 Laboratory

In addition to its infrastructure, the station has a laboratory equipped with instruments and materials. This allows for essential physicochemical and biological measurements to be carried out, in order to guarantee the quality of the water entering and leaving the station. The objective is therefore to verify the effectiveness of the wastewater treatment process. [Sidi Merouane Wastewater Treatment Plant 2019].



Figure 1.18: Sidi Merouane Wastewater Treatment Plant Monitoring Laboratory (2019)

1.6 Parameters of treated water

1.6.1 Physical parameters

1.6.1.1 Temperature (T)

Temperature is an essential physical parameter of water. Its influence is significant on the solubility of gases in water, as well as on the rate of chemical and biochemical reactions, as Rejsek (2002) specified. Furthermore, it has an impact on biological nitrification and denitrification processes.

According to Bollag (1973) and Rodier et al. (2005), nitrification and denitrification occur optimally at temperatures ranging between 28 and 32 °C. It is important to note that the temperature of the effluents does not fall below 5 °C or exceed 45 °C. Outside this range, the reaction slows down considerably and the microorganisms begin to disappear.

Finally, these temperatures have a significant influence on the solubility of the salts and the pH, added (Rodier et al., 1996). Furthermore, the temperatures are recorded using a mercury thermometer and thermocouples.

1.6.1.2 The electrical potential (pH)

The potential of hydrogen (pH) is one of the key factors. It affects the biological activity of the aquatic microflora, because in the pH range of 4.5 to 8, microorganisms are scarce or absent. Optimal values are actually found between 5.5 and 7.5, as indicated (Batton et al. 1990, Merrill et al. 1977).

The protocol for measuring the pH of water is based on a potentiometric method. For this, a platinum electrode is connected to a probe. The measured temperatures range from 0 to 60°C, and the pH is between 3 and 10.

1.6.1.3 Electrical conductivity (EC)

Measuring electrical conductivity allows for the assessment of the overall mineralization of the water, mentioned (Rejsek, 2002). This assessment is crucial because treated wastewater is reused for irrigation. Indeed, salinity concentrations above 2500 $\mu\text{S}/\text{cm}$ limit the proliferation of microorganisms. This negatively impacts the productivity of the process of purification. The operating protocol describes the method for measuring conductivity using an electrical conductivity meter, expressed in microsiemens per centimeter ($\mu\text{S}/\text{cm}$).

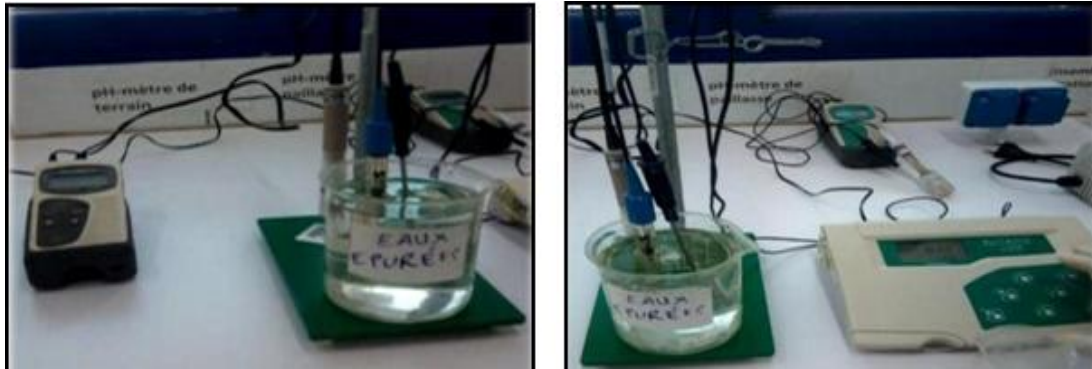


Figure 1.19: pH Measurement and Conductivity Measurement (Sidi Merouane Wastewater Treatment Plant, 2019)

1.6.1.4 Suspended solids (SS)

This is a portion of the pollutants consisting of undissolved organic and mineral particles. It is necessary to measure suspended solids in order to assess the distribution of the pollutant load, including both dissolved and sedimentary pollution.

The evolution of these two aspects varies depending on the natural environment or septic tanks, as highlighted in Tarmoul, 2007.

Furthermore, suspended solids affect the balance of aquatic environments by increasing turbidity. They hinder the development of aquatic organisms by decreasing the amount of dissolved oxygen, which creates an imbalance between the different species. Furthermore, measuring suspended solids is also crucial for guiding the choice of irrigation system when reusing treated wastewater.

According to the FAO (2003), excess suspended solids can clog irrigation systems, particularly drip irrigation systems. Suspended solids are assessed using two methods:

- The filtration method: for slightly contaminated water, the gravimetric method is used to determine the mass of dry matter (mg/l) after having separated the suspended solids from the water, followed by dehydration at temperatures reaching 105°C.

- The centrifugation method: this measurement approach is applicable to wastewater with a high concentration of suspended solids.

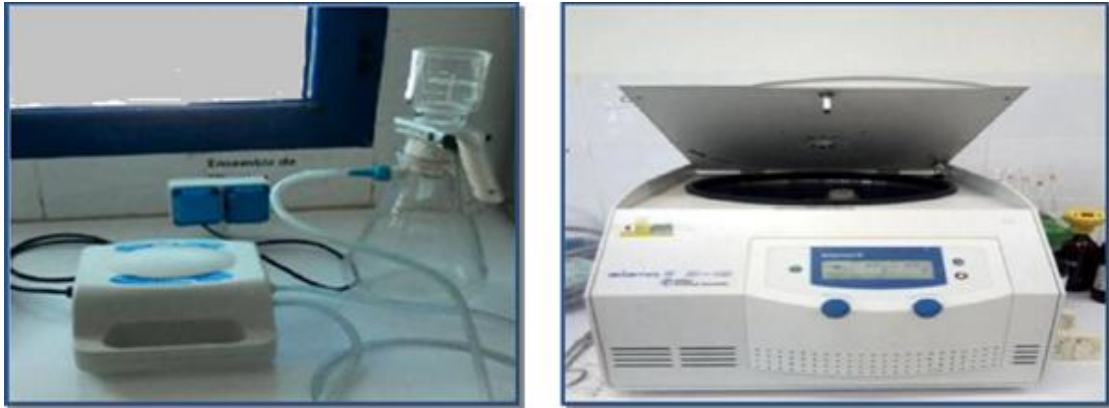


Figure 1.20: Equipment for Measuring Suspended Solids: Vacuum Filtration and Centrifugation (Sidi Merouane Wastewater Treatment Plant, 2019)

1.6.1.5 Dissolved Oxygen (DO)

Dissolved oxygen is necessary for the composition of water, as it is essential for the biological degradation of pollutants. This phenomenon occurs mainly under aerobic conditions. In raw wastewater, the oxygen concentration is very low, almost zero, due to the high presence of reducing compounds and the intense activity of microorganisms.

The dissolution of oxygen in water depends on several factors, such as temperature, pressure, and the ionic strength of the medium.

Finally, the most commonly used technique for assessing the concentration of dissolved oxygen in water is oximetry.



Figure 1.21: Dissolved Oxygen (DO) Measurement by Oximetry (Sidi Merouane Wastewater Treatment Plant, 2019)

1.6.2 Chemical Parameters

1.6.2.1 Chemical Oxygen Demand (COD)

According to Rodier et al. (2009), chemical oxygen demand allows for the assessment of the concentration of organic or mineral matter. Whether dissolved or suspended in the water, based on the amount of oxygen required for their total chemical oxidation.

This refers to the total amount of oxidizable organic matter. It is measured as a function of the amount of oxygen supplied by potassium dichromate and required for the oxidation of the organic compounds (proteins, carbohydrates, lipids, etc.) present in the wastewater.

The difference in results obtained between BOD and COD values indicates the importance of the polluting substances that are less biodegradable or not biodegradable.

Indeed, COD is measured by spectrophotometry (standard NFT 90-101).

This measurement reflects the correlative estimate of oxidizable substances present in the water, whether they are of organic or mineral origin. The units are in mg O₂/l. (Bliefert and Perraud, 2001).



Figure 1.22: Equipment for Measuring Chemical Oxygen Demand: Thermoreactor and Spectrophotometer (Sidi Merouane Wastewater Treatment Plant, 2019)

1.6.2.2 Biochemical Oxygen Demand (BOD5)

According to Berne and Cordonnier (1991), BOD₅ corresponds to the oxygen consumption by bacteria during the biological oxidation of a portion of carbonaceous organic matter. samples are previously inoculated and incubated for five (05) days in the dark at a temperature of 20°C is used to prevent parasitic photosynthesis.

It is important to note that this temperature is conducive to the activity of oxygen-consuming microorganisms. BOD₅ tests are also performed before and after treatment to assess the effectiveness of the treatment and predict the impact of the effluent on receiving waters, added (Gaujous, 1995).

Their oxygen requirement (BOD₅) is calculated using a manometric method (standard NFT 90-720). The incubator is then kept closed for five days. The results of this test depend on the temperature and the incubation period.

If the standard temperature is set at 20°C, the oxidation rate may vary. During this period, oxidation is not complete (60-723% to 70%); the reaction is finished. It is estimated at 90% to 95% after incubation.



Figure 1.23: BOD Meter and BOD5 Bottle for Measuring Biochemical Oxygen Demand over Five Days (BOD₅)

1.7 Importance of BOD₅

Biochemical Oxygen Demand (BOD) is an important parameter used to assess the biodegradable fraction of organic pollution in water. It represents the amount of dissolved oxygen required by microorganisms, mainly bacteria, to biologically degrade organic matter present in a water sample over a specified period.

BOD₅ corresponds to the amount of oxygen consumed after five days of incubation at 20 °C in the dark. It is generally determined according to the NF EN 1899-1 and NF EN 1899-2 standards.

The experimental protocol consists of measuring the initial dissolved oxygen concentration, denoted by C_0 , in the sample at the beginning of the test, and the final dissolved oxygen concentration, denoted by C_5 , after five days of incubation in the dark at 20 °C. The BOD₅ value is calculated as follows:

$$\text{BOD}_5 = C_0 - C_5$$

The result is usually expressed in mg O₂/L.

During incubation, a nutrient solution containing phosphates may be added to ensure that nitrogen and phosphorus do not become limiting factors for microbial growth. Allylthiourea may also be used to inhibit nitrification and prevent ammonium oxidation from affecting the BOD₅ measurement.

This method is relatively simple and provides information about the natural biodegradability of organic matter. However, it requires a long incubation period of five days and is generally less reproducible than Chemical Oxygen Demand (COD) analysis.

During the BOD test, microorganisms transform organic matter through metabolic processes. Catabolism involves the degradation of organic compounds to produce energy, whereas anabolism involves the synthesis of new cellular material. Part of the organic matter is oxidized and consumes oxygen, while another part is converted into microbial biomass, including proteins, lipids, and carbohydrates.

Not all organic matter present in wastewater is biodegradable. Therefore, the BOD₅ value is generally lower than the COD value, since COD measures both biodegradable and non-biodegradable oxidizable organic matter.

BOD₅ measurements may be influenced by high ammonium concentrations and the presence of toxic substances. Toxic compounds may inhibit or reduce the activity of microorganisms involved in biological degradation. Similarly, ammonium oxidation may lead to additional oxygen consumption and affect the interpretation of the results.

When BOD₅ values are high, particularly in domestic wastewater or agricultural effluents, the sample may need to be diluted using water saturated with dissolved oxygen. This prevents complete depletion of oxygen during the incubation period.

Unlike COD, BOD₅ specifically reflects the biodegradable fraction of organic matter and is therefore more representative of the natural biological decomposition processes that occur in aquatic environments. Nevertheless, BOD₅ analysis is more time-consuming and usually less reproducible than COD analysis.

BOD₅ is thus an essential parameter for evaluating the organic pollution load of wastewater, monitoring treatment efficiency, and assessing the potential impact of discharges on receiving water bodies.

1.8 Limitations of the conventional measurement of BOD5

The measurement process takes five days, and this relatively long timeframe poses a significant obstacle for operators of the processing stations, requiring ad hoc decisions regarding operational procedures. Furthermore, the laboratory analyses, which require specific equipment, qualified personnel, and operational costs, cannot be overlooked.

Consequently, the resulting results reflect an outdated system condition, limiting its use for online control and immediate process optimization.

1.9 Need for a Software Sensor

Faced with the limitations of traditional methods, the development of software sensors (Soft Sensors) appears as a promising solution. A software sensor is a mathematical or artificial intelligence model capable of estimating a variable that is difficult or costly to measure based on other variables that are readily available in real time.

1.10 Transition to a Framework

Data characteristics, measured parameters, and operating conditions can vary considerably from one station to another. In this context, creating a single model does not necessarily guarantee good generalization capabilities.

Therefore, this thesis is not limited to the design of a simple predictive model for BOD5. Instead, it proposes the development of a generic framework for building software sensors. This framework aims to provide a flexible architecture that allows for the integration of different stages of data preprocessing, variable selection, model training, and performance evaluation.

In the case of BOD5, software sensors utilize continuously measured physicochemical parameters, such as pH, temperature, electrical conductivity, total dissolved solids (TDS), turbidity, and chemical oxygen demand (COD). Thanks to advanced modeling techniques, including machine learning and artificial intelligence, it becomes possible to quickly predict BOD5 with satisfactory accuracy.

The use of such sensors offers several advantages: reduced waiting time, lower analysis costs, continuous water quality monitoring, and improved operator responsiveness to changes in operating conditions.

1.11 Conclusion

Wastewater, of diverse compositions and origins, poses a challenge/risk to the environment when discharged without prior treatment. Understanding the characteristics of wastewater is essential for the proper planning, management, and use of these resources. In

this chapter, we presented wastewater and study the Sidi Merouane wastewater treatment plant Identifying the obstacles facing the stations.

CHAPTER 2

SOFT SENSORS AND MACHINE LEARNING FOR WATER TREATMENT

2.1 2.1 Introduction

This is a revolution. Almost every field of scientific research has been dominated by artificial intelligence. Every day, it becomes more and more complex and offers us extraordinary possibilities. We will first present the definition of a software sensor, models based on artificial intelligence, and evaluation metrics.

2.2 2.2 Definition of a Software Sensor

The concept of a software sensor consists of associating available or easily obtainable measurements, representative of the evolution of the process under study, with mathematical models linking the measured quantities and the quantities to be determined [Baya Hadid].

A software sensor is a "software" which, using physical measurements, is capable of predicting the behavior of the physical quantity, detecting faults in physical sensors, reconstructing missing data, estimating the value of a quantity, and improving the accuracy of the measurands. This software uses physical sensors to predict/estimate a specific quantity.

2.3 2.3 Data Preprocessing

Data preprocessing involves transforming raw data into a more suitable and relevant format for model analysis and training. Data preprocessing plays a crucial role in improving the quality and effectiveness of machine learning models by addressing issues such as missing values, noise, inconsistencies, and outliers in the data.

2.3.1 2.3.1 Cleaning

Data cleaning is a critical component of the data preprocessing pipeline in machine learning. It involves identifying and correcting errors or inconsistencies in the dataset to ensure that the data is high-quality and suitable for analysis or model training.

Data cleaning typically includes:

Outliers: are data points that differ significantly from other observations in the dataset and can skew statistical analysis or machine learning models.

Handling missing or incomplete values

These values can result from a data entry error, a non-response in surveys, or data lost during collection. But don't worry, you have several solutions for handling this type of situation. You can exclude the relevant observations or replace missing values using the mean, median, or linear regression. As is often the case in data analysis,

2.3.2 2.3.2 Normalization

Normalization is a data preprocessing technique used to scale and standardize the values of features in a dataset. The main goal of normalization is to bring all variable values into a similar range without distorting differences in value ranges. This is important because many machine learning algorithms perform better or converge faster when input features are on a similar scale and have a similar distribution.

2.3.3 2.3.3 Data Splitting

Data splitting is the separation of your initial dataset into distinct subsets to build and evaluate a model. The main sets involved in data splitting are: Training set and test set.

2.3.4 2.3.4 The Savitzky-Golay Filter

The Savitzky-Golay filter is a digital filter that smooths data points by fitting successive sub-sets of adjacent data points with a low-degree polynomial using the method of linear least squares. Its primary advantage lies in its ability to smooth data while preserving the features of the original signal, such as peaks and valleys, which might be lost with other smoothing techniques

2.4 2.4 Software Sensors in Water Treatment

Software sensors are increasingly used in water treatment to estimate parameters that are difficult, costly, or time-consuming to measure directly, such as BOD5. By leveraging readily available real-time data and machine learning techniques, they enable improved monitoring, control, and optimization of treatment processes.

2.4.1 2.4.1 BOD5 Estimation

Traditional BOD5 measurement requires five days of incubation, which is incompatible with real-time monitoring needs. Software sensor models typically use variables such as COD, pH, temperature, electrical conductivity, turbidity, and suspended solids to rapidly predict BOD5. This approach allows wastewater treatment plant operators to obtain a near-instantaneous estimate of the organic load and adjust treatment parameters accordingly.

2.4.2 2.4.2 COD Estimation

Although COD measurement is faster than BOD5 measurement, it still requires laboratory analysis and the use of chemical reagents. Software sensors allow for COD prediction based on online measurements such as conductivity, pH, temperature, and turbidity.

This continuous estimation facilitates the rapid detection of pollutant load variations and improves the management of treatment processes.

2.4.3 2.4.3 Turbidity Prediction

Software sensors can be used to predict turbidity changes based on hydraulic, meteorological, or physicochemical parameters. These predictions allow for anticipating variations in raw water quality and adapting coagulation, flocculation, and filtration processes to maintain optimal performance.

2.4.4 2.4.4 Suspended Solids (SS) Estimation

Direct SS measurements generally require laboratory analyses involving filtration and weighing. Software sensors allow for the estimation of SS concentrations based on parameters such as turbidity, conductivity, or hydraulic flow rates. This real-time estimation helps improve the control of treatment plants and reduce analytical costs.

2.4.5 2.4.5 Treated Water Quality Monitoring

Software sensors are also used to ensure comprehensive monitoring of treated water quality. They allow for the simultaneous estimation of several critical parameters, such as BOD₅, COD, SS, nitrogen, phosphorus, and dissolved oxygen. Thanks to these continuous estimates, operators have a real-time view of the process status and can quickly detect any anomalies or system deviations. This advanced monitoring capability improves regulatory compliance, reduces environmental risks, and promotes more efficient water resource management.

2.5 2.5 Regression Models Used in the Framework

2.5.1 2.5.1 Random Forest Regression (RFR) Model

Random Forest Regression (RFR) was developed by Breiman (2001). It is one of the most accurate general-purpose learning algorithms. RFR has proven to be extremely efficient, using minimal computing resources. RFR performance is significantly improved compared to single-tree algorithms. Its speed is high, and its error rates are similar to those of more traditional and resource-intensive algorithms (Desai, 2019). Random Forest Regression (RFR) consists of a regression tree (RT) and aggregation (bootstrap, bagging). It is a perfect model for solving real-time problems in various geophysical fields, including the environment (Alizamir et al., 2021; Kim et al., 2020; Melesse, 2020), hydrometeorology (Alizamir et al., 2020), and hydrology (Ghorbani et al., 2021; Yaseen, 2020).

In random forest regression, the predictors of the tree $h(x, \theta_k)$, $k = 1 \dots K$, take numerical values based on random vectors $\{\theta_k\}$ (Breiman, 2001). It is important to note that $\{\theta_k\}$ are identically distributed and independent random vectors. The training data are randomly and independently drawn from a joint distribution of (X, Y) , where the random vector X is the observed input and the random vector Y is the expected numerical output. The individual trees are grown using the CART (Classification and Regression Trees) algorithm. The algorithm for random forest regression is presented in Trevor et al. (2009).

➤ For $b = 1$ to B : (a) Draw a bootstrap sample Z^* of size N from the training data.

(b) Create a random forest tree T_b for the data (bootstrap) by recursively repeating the following steps for each terminal node of the tree, until the minimum size of the node $nmin$ is reached.

(i) Select m variables randomly from p variables.

(ii) Choose the best variable/best split point from the m .

(iii) Split the node into two child nodes.

➤ Outputs: Produce the set of trees $\{T_b\}_{1B}$

Make a prediction at a new point x :

$$\bar{f}_{rf}^B = \frac{1}{B} \sum_{b=1}^B T_b(x)$$

Equation 2.1: Random Forest Regression Prediction

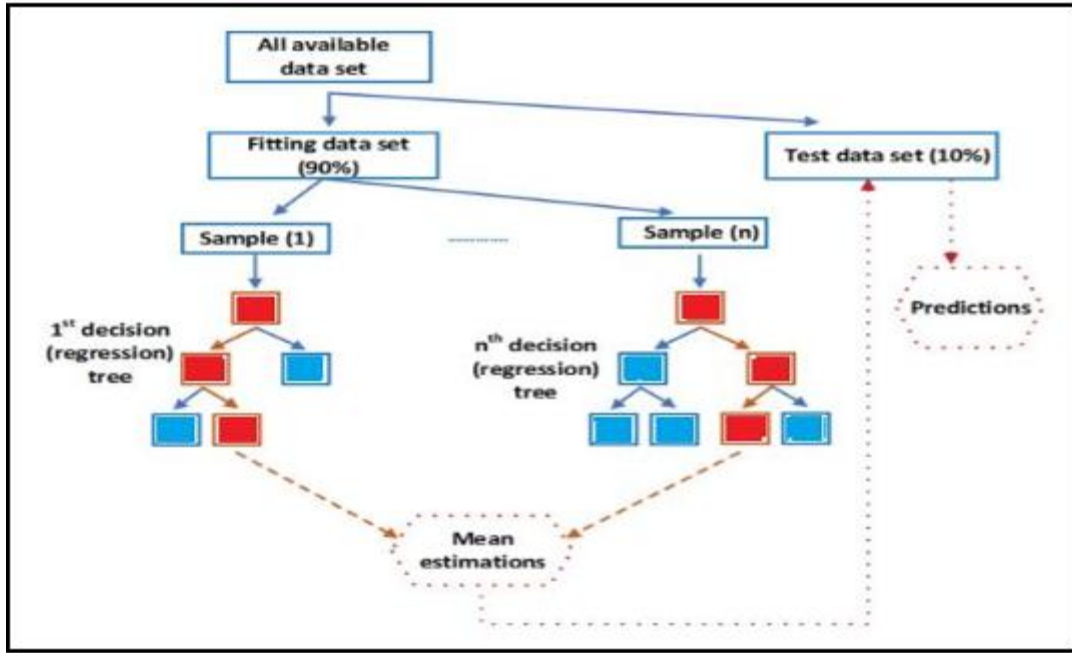


Figure 2.1: Structure of the Random Forest Regression Model (Diamantopoulou, 2022)

Three parameters are used in the RFR modeling procedure: n tree (the number of trees to be developed), $mtry$ (randomly selected predictive variables), and $nodesize$ (the minimum number of observations). Furthermore, if predictors are ignored individually in the RFR, the importance of each predictor is calculated using a higher MSE.

The relative importance of each predictor can be chosen according to the specific functioning of the RFR (Alizamir et al., 2021). Random forest regression (RFR), which consists of a regression tree (RT) and aggregation (bootstrap, Bagging), is a perfect model for solving real-time problems in various geophysical fields, including the environment (Alizamir et al, 2021; Kim et al, 2020; Melesse, 2020), hydrometeorology (Alizamir et al, 2021), and hydrology (Wang et al., 2015; Ghorbani et al, 2020; Yaseen 2020).

2.5.2 Linear Regression Model Prediction

We looked at a simple regression model of life satisfaction:

$$\text{life_satisfaction} = \theta_0 + \theta_1 \times \text{GDP_per_capita}$$

This model is simply a linear function of the input feature GDP_per_capita.

The parameters θ_0 and θ_1 are the model's parameters.

More generally, a linear model makes a prediction by computing a weighted sum of the input features and adding a constant called the bias term, also known as the intercept term.

The Linear Regression model prediction is:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

Where:

- $\hat{y}$ is the predicted value.
- n is the number of features.
- x_i is the i -th feature value.
- θ_j is the j -th model parameter, including the bias term θ_0 and the feature weights $\theta_1, \theta_2, \dots, \theta_n$.

In vectorized form, the Linear Regression model prediction can be written as:

$$\hat{y} = h\theta(x) = \theta \cdot x$$

Where:

- θ is the model's parameter vector, containing the bias term θ_0 and the feature weights θ_1 to θ_n .
- x is the instance's feature vector, containing x_0 to x_n , where x_0 is always equal to 1.
- $\theta \cdot x$ is the dot product of the vectors θ and x , which is equal to:
 $\theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$
- $h\theta$ is the hypothesis function that uses the model parameters θ .

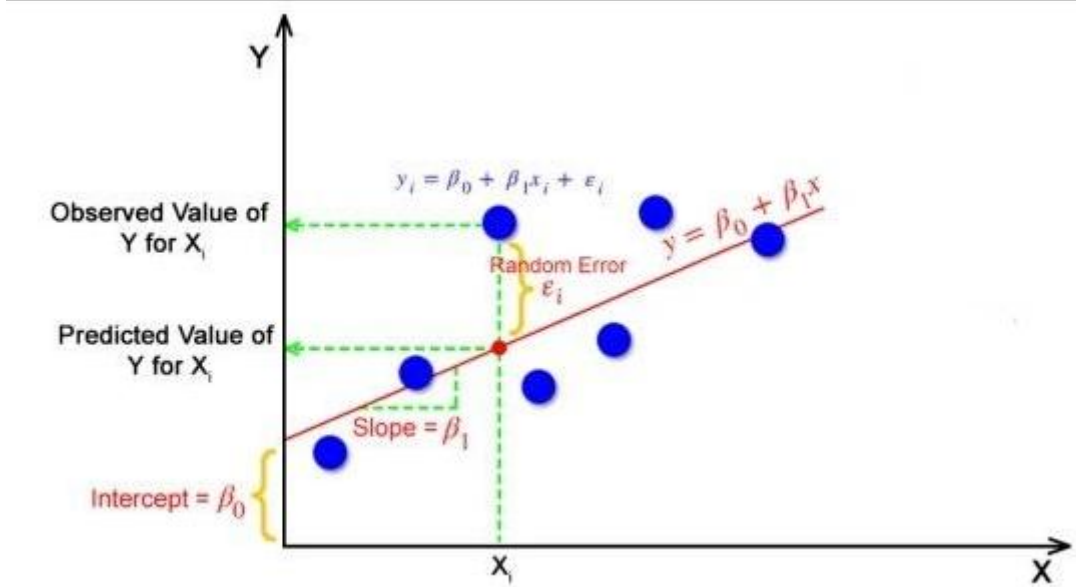


Figure 2.2: Structure of the Linear Regression Model

2.5.3 CatBoost Model

CatBoost is an implementation of gradient boosting, which uses binary decision trees as base predictors. A decision tree [4, 10, 27] is a model built by a recursive partition of the feature space R_m into several disjoint regions (tree nodes) according to the values of some splitting attributes a . Attributes are usually binary variables that identify that some feature

x_k exceeds some threshold t , that is, $a = 1\{x_k > t\}$, where x_k is either numerical or binary feature, in the latter case $t = 0.5$.² Each final region (leaf of the tree) is assigned to a value, which is an estimate of the response y in the region for the regression task or the predicted class label in the case of classification problem.³ In this way, a decision tree h can be written as

$$h(x) = \sum_{j=1}^J b_j 1\{x \in R_j\},$$

where R_j are the disjoint regions corresponding to the leaves of the tree.

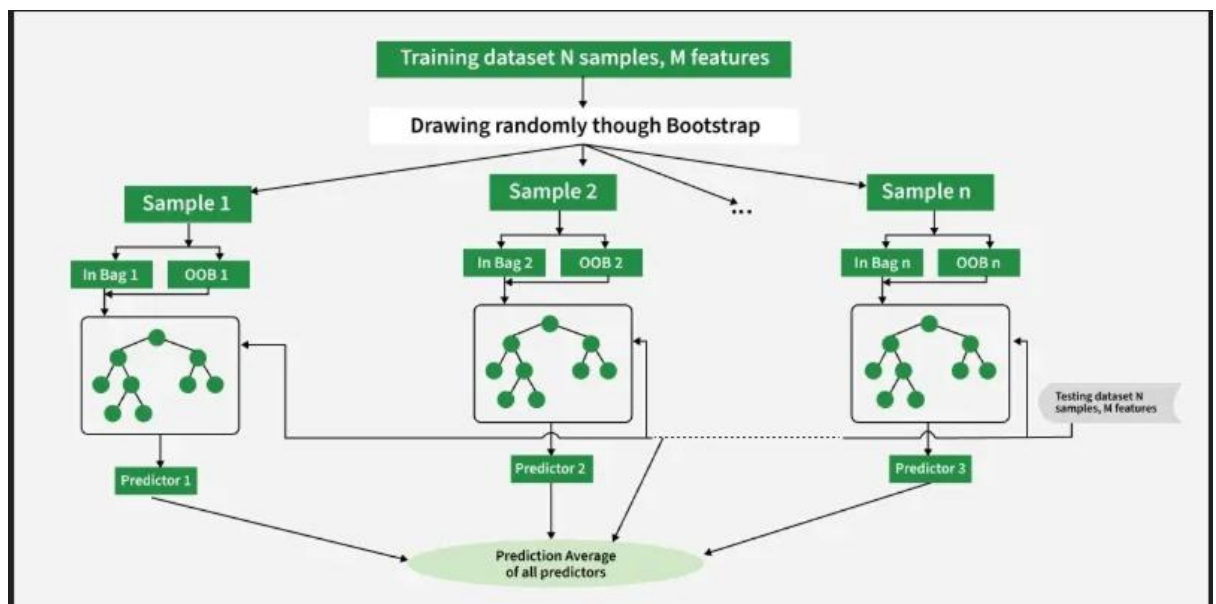


Figure 2.3: Structure of the CatBoost Model

2.5.4 Extra Trees Model

When you are growing a tree in a Random Forest, at each node only a random subset of the features is considered for splitting (as discussed earlier).

It is possible to make trees even more random by also using random thresholds for each feature rather than searching for the best possible thresholds (like regular Decision Trees do). A forest of such extremely random trees is simply called an Extremely Randomized Trees ensemble¹² (or Extra-Trees for short).

Once again, this trades more bias for a lower variance. It also makes Extra-Trees much faster to train than regular Random Forests since finding the best possible threshold for each feature at every node is one of the most time-consuming tasks of growing a tree.

You can create an Extra-Trees classifier using Scikit-Learn's `ExtraTreesClassifier` class. Its API is identical to the `RandomForestClassifier` class. Similarly, the `Extra TreesRegressor` class has the same API as the `RandomForestRegressor` class.

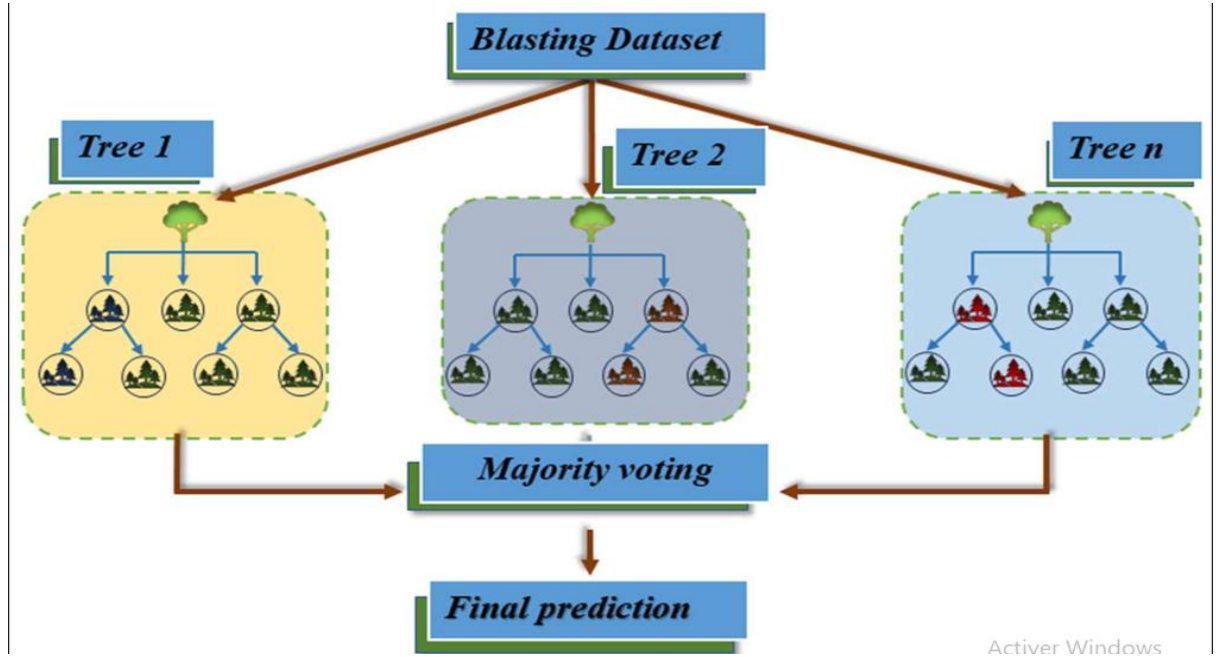


Figure 2.4: Structure of the Extra Trees Model

2.5.5 XGBoost Model

XGBoost (eXtreme Gradient Boosting) is a highly optimized machine learning algorithm based on gradient-boosted decision trees. It is widely used for structured/tabular data and has won many machine learning competitions.

We review gradient tree boosting algorithms in this section. The derivation follows from the same idea in existing literatures in gradient boosting. Specially the second order method is originated from Friedman et al. We make minor improvements in the regularized objective, which were found helpful in practice.

For a given data set with n examples and m features $D = \{(x_i, y_i)\}$ ($|D| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}$),

a tree ensemble model uses K additive functions to predict the output.

$$y_i = \varphi(x_i) = \sum_{k=1}^K f_k(x_i), f_k \in F,$$

- where $F = \{f(x) = wq(x)\} (q : \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^T)$ is the space of regression trees (also known as CART). Here q represents the structure of each tree that maps an example to the corresponding leaf index. T is the number of leaves in the tree. Each f_k

corresponds to an independent tree structure q and leaf weights w . Unlike decision trees, each regression tree contains a continuous score on each of the leaf, we use w_i to represent score on i -th leaf. For a given example, we will use the decision rules in the trees (given by q) to classify it into the leaves and calculate the final prediction by summing up the score in the corresponding leaves (given by w). To learn the set of functions used in the model, we minimize the following regularized objective.

$$L(\phi) = l(\hat{y}_i, y_i) + \lambda \sum_{k=1}^K \Omega(f_k)$$

- where $\Omega(f) = \gamma T + \frac{1}{2} \lambda w^2$. Here l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i . The second term Ω penalizes the complexity of the model (i.e., the regression tree functions). The additional regularization term helps to smooth the final learnt weights to avoid over-fitting. Intuitively, the regularized objective will tend to select a model employing simple and predictive functions. A similar regularization technique has been used in Regularized greedy forest (RGF) [25] model. Our objective and the corresponding learning algorithm is simpler than RGF and easier to parallelize. When the regularization parameter is set to zero, the objective falls back to the traditional gradient tree boosting.

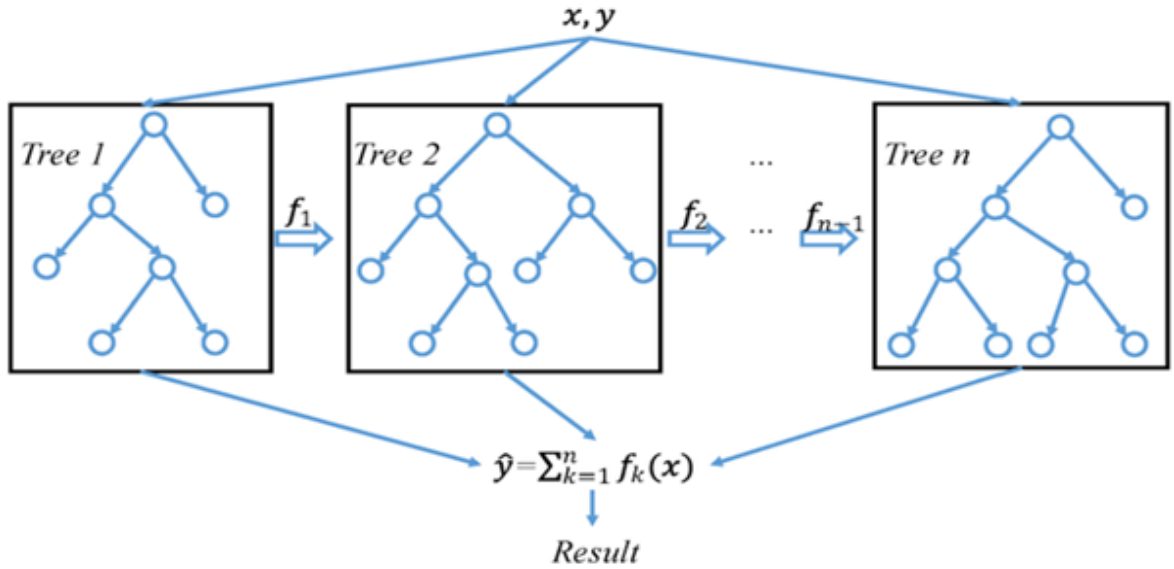


Figure 2.5: Structure of the XGBoost Model

2.6 Evaluation Metrics

In this section, we first introduce the mathematical background of the analyzed rates (“Mathematical Background”).

2.6.1 2.6.1 Coefficient of Determination (R^2)

(worst value = $-\infty$; best value = $+1$)

The coefficient of determination (Wright, 1921) can be interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^m (X_i - Y_i)^2}{\sum_{i=1}^m (Y - Y_i)^2}$$

Equation 2.2: Coefficient of Determination (R^2)

- $R^2 = 1$: Parfitt prediction.
- R^2 close to 1: A very good pattern.
- $R^2 = 0$: The pattern is not as good as it seems.
- $R^2 < 0$: A very bad pattern.

2.6.2 2.6.2 Root Mean Square Error (RMSE)

(best value = 0; worst value = $+\infty$)

The two quantities MSE and RMSE are monotonically related (through the square root). An ordering of regression models based on MSE will be identical to an ordering of models based on RMSE.

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (X_i - Y_i)^2}$$

Equation 2.3: Root Mean Square Error (RMSE)

2.6.3 2.6.3 Mean Absolute Error (MAE)

(best value = 0; worst value = $+\infty$)

MAE can be used if outliers represent corrupted parts of the data. In fact, MAE is not penalizing too much the training outliers (the L1 norm somehow smooths out all the errors of possible outliers), thus providing a generic and bounded performance measure for the model. On the other hand, if the test set also has many outliers, the model performance will be mediocre.

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^m |X_i - Y_i|$$

Equation 2.4: Mean Absolute Error (MAE)

2.7 Interpretation of Models

Its purpose is to measure the contribution of each variable to the prediction of the swell.

In the case of BOD5 estimation, explanatory variables can include pH, temperature, electrical conductivity, turbidity, suspended solids (MES), or DCO2. Analyzing their significance allows us to identify those measurements that influence the model's predictions.

For example, the model might show that DCO and turbidity contribute to the BOD5 prediction of temperature or pH. This information is particularly useful for:

- Understanding the model's behavior;
- Identifying the most relevant variables,
- Reducing model complexity;
- Reducing data acquisition costs and improving overall performance.

SHAP is a modern interpretation method based on cooperative game theory. It attributes to each change an individual contribution to the model's realistic prediction. SHAP allows you to explain each individual prediction. For each observation, each measurement indicates that each change increases or decreases the preset value. For example, to achieve a high BOD5 estimate, SHAP depended a strong DCO value contributes to the prediction increase, a significant turbidity also enhances this increase;

The SHAP approach offers additional advantages: intuitive visualization of variable contributions, understanding the details of prediction mechanisms, revealing dominant variables and Improving user confidence in the results. Within this framework, SHAP can be used to compare the behavior of different algorithms and identify the most influential variables under diverse exploitation conditions.

2.8 2.8 Model Optimization

Therefore, we offer a model optimization module designed to improve the accuracy of machine learning algorithms' predictions by adjusting their operating parameters.

This module automatically searches for optimal parameter combinations to improve prediction accuracy and minimize modeling errors.

2.8.1 2.8.1 Grid Search

Grid search is a method to optimize a model pipeline's tuning parameters. It creates a pre-defined set of candidate values and computes performance for each. From there, the numerically best candidate could be chosen, or the relationship between the tuning parameter(s) and model performance can be inspected to see if the model might benefit from additional optimization.

2.9 2.9 Conclusion

it is essential to have a robust, reliable, accurate, and uncertainty-resistant predictive model. Such models must have high predictive capacity to be used in characterizing and forecasting changes in wastewater quality.

This chapter aims to present different methods for modeling wastewater quality in order to select the most appropriate model to achieve the objectives. By leveraging the power of artificial intelligence models, we can move towards a future where wastewater is managed intelligently, thus ensuring its sustainable reuse.

CHAPTER 3

DESIGN OF THE CONFIGURABLE SOFT SENSOR FRAMEWORK

3.1. Introduction

The previous chapters established the problem setting and theoretical background, namely the operational difficulty of routine BOD5 determination in wastewater treatment plants and the principles of soft sensing and machine learning regression. The present chapter shifts the focus to the main contribution of this work: the design of a configurable framework for data-driven soft sensor generation. In contrast to a single-purpose prediction tool that embeds one fixed model for one fixed dataset, the proposed framework provides a structured environment in which users can define, run, and compare multiple experimental configurations for soft sensors. The chapter specifies the objectives and intended users of the framework, outlines its architectural organisation, details the internal modules, and clarifies how it departs from conventional fixed applications. The chapter concludes by delineating the scope and current limits of the prototype that is validated experimentally in this dissertation.

3.2. Definition of framework

Here is a definition by Ralph Johnson [Joh97a]: “A framework is a reusable design of all or part of a system that is represented by a set of abstract classes and the way their instances interact.” This can be complemented by the following definition: “A framework is the skeleton of an application that can be customized by an application developer” [Joh93]. By offering modularity, extensibility, and minimal behaviour, a framework aims to facilitate the construction of an application by providing a small set of known and proven components (classes) to compose or extend.

It goes beyond simple code reuse by promoting the reuse of analysis and design. We will primarily retain from this set of classes its adaptability and extensibility, and the partial, albeit incomplete, representation it offers of an application.

A framework therefore has both a structural aspect (a class architecture) and a behavioural aspect (the functionalities provided by these classes). They cover areas as varied as graphical interfaces in the broadest sense (HotDraw, ET++, AWT, X11), the construction of development environments (Eclipse), enterprise management (SanFrancisco), software buses (DCOM, CORBA), component models (COM, EJB, JavaBeans), etc.

3.3. Target users

The proposed framework is designed to meet the needs of several user categories involved in water treatment and artificial intelligence. It is aimed at water treatment experts, who can use it to analyze water quality data. Researchers also constitute an important user category, as the framework offers them a flexible environment for conducting experiments, comparing different modeling approaches, and evaluating new machine learning algorithms.

Furthermore, engineers can leverage the framework to test and deploy predictive models tailored to the needs of water treatment plants.

In addition, users specializing in Machine Learning can use the framework to train different models, compare their performance and analyze their explainability using methods such as variable importance.

3.4. General architecture of the framework

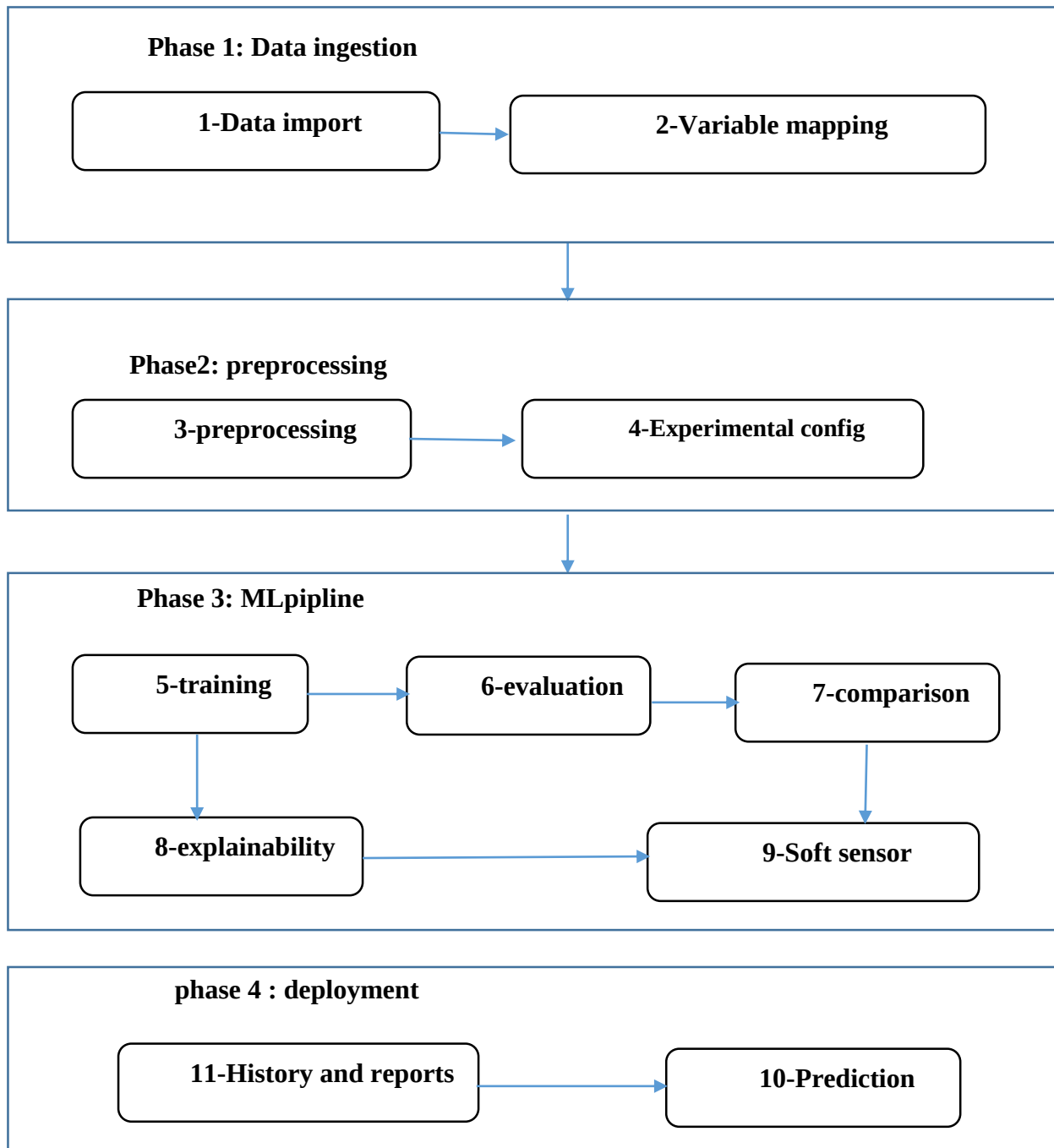


Figure 3.1: Framework Architecture

3.5. Description of the framework modules

Module	Role	Prototype component
1. Data import	Load raw data from a station file (CSV or Excel); validate completeness and quality	dataset_page.py — Dataset Management
2. Variable mapping	Assign dataset columns to the physicochemical roles required by the framework (COD, pH, TW, TSS, SC, Q, BOD5)	Integrated in dataset loading — automatic column recognition
3. Preprocessing	Apply missing value treatment, outlier removal (IQR), signal filtering (Savitzky-Golay), and optional feature scaling	preprocessing_page.py — Data Preprocessing
4. Experimental config	Select regression models, set train/test split ratio, and choose the target variable	training_page.py — Model Training configuration panel
5. Training	Fit each selected model on the training partition using Scikit-Learn, CatBoost, or XGBoost	training_page.py — training execution
6. Evaluation	Compute R^2 , RMSE, and MAE on the held-out test set for each trained model	training_page.py — results table
7. Comparison	Rank all trained models by test R^2 ; identify and highlight the best model	training_page.py — Best Model field + Deploy button
8. Explainability	Analyse the contribution of each input variable using feature importance scores and SHAP analysis	Conducted in Jupyter Notebooks (CatBoost importance + SHAP)
9. Soft sensor generation	Serialise the best model via Joblib, register it as active in the SQLite database	training_page.py — Deploy action
10. Prediction	Load the active model and estimate BOD5 from six real-time sensor inputs; classify result (Good / Moderate / Polluted)	prediction_page.py — Prediction Page (Operator)
11. History and reports	Persist all predictions with timestamps and inputs in SQLite; display aggregated trends and statistics	analytics_page.py + dashboard_page.py

Table 3.1: Framework Module Descriptions and Prototype Correspondence

3.6. Difference between a fixed application and a framework

A fixed application represents a specialized solution dedicated to a specific task, while a framework offers a configurable approach allowing the user to select variables, learning algorithms, training parameters, and evaluation methods according to the study's objectives.

The framework facilitates the comparison of multiple machine learning models.

The framework's modular architecture improves its reusability and extensibility, because it allows for the addition of new algorithms.

Table 3.2 summarizes the key differences between a fixed application and the proposed framework.

Dimension	Fixed application	Proposed framework
Input data	Hardcoded or fixed format	Any CSV or Excel file
Variable selection	Fixed	Configurable mapping
Preprocessing	Fixed pipeline	Configurable (outliers, filter, scaler)
Algorithm	Single predefined model	User selects from 5 algorithms
Train/test split	Fixed ratio	Configurable (e.g. 80/20)
Comparison	Not available	Automatic multi-model comparison
Explainability	Not available	Feature importance + SHAP
Reproducibility	Limited	Full — config saved to JSON + SQLite
Target user	Station operator	Researcher, engineer, analyst, operator

Table 3.2: Comparison between a Fixed Application and the Proposed Framework

3.7. Conclusion

This is a working prototype of a framework predicting the five-day Biological Oxygen Demand (BOD5) using data from the Sidi Merouane wastewater treatment plant. This validation has enabled the import and preprocessing of data up to the training, evaluation, comparison, and interpretation of machine learning models.

CHAPTER 4

IMPLEMENTATION OF FRAMEWORK PROTOTYPE

4.1. Development Environment

We implemented and tested the prototype using Visual Studio Code as our main integrated development environment, which we found suitable for managing a multi-module Python project and integrating version control and virtual environments. The desktop application itself is written entirely in Python, relying on a set of scientific and engineering libraries that together cover all stages of the pipeline, from reading raw plant data to deploying the BOD5 soft sensor. The principal tools and their roles are summarised in Table 4.1.

Tool / Library	Role in the project
Python 3.x	Main programming language
Visual Studio Code	Primary development environment (IDE)
Jupyter Notebook	Exploratory experiments, model development and validation notebooks
PyQt6	Desktop graphical user interface (GUI) framework
Pandas	Data loading, manipulation, and tabular operations
NumPy	Numerical computations and array operations
Scikit-Learn	Classical ML models (Linear Regression, Random Forest, Extra Trees) and evaluation metrics
CatBoost	Gradient-boosted decision tree regressor (selected soft sensor)
XGBoost	Extreme gradient boosting regressor
Matplotlib	In-application data and result visualisation
Seaborn	Statistical plots used during exploratory analysis in notebooks
SHAP	Model interpretability — feature contribution analysis (notebooks)
Joblib	Model serialisation and loading (.pkl files)
SQLite (sqlite3)	Embedded relational database for result and prediction persistence

Table 4.1: Development Tools and Libraries Used in the Project

Some of these tools are restricted to the experimental phase and do not appear in the final deployed application. In particular, we relied on Seaborn and SHAP only within Jupyter Notebooks to explore distributions, correlations, and feature importance, whereas the running application uses Matplotlib for all interactive plots and Joblib for model persistence. We chose Joblib instead of the standard pickle module to ensure robust serialisation of both scikit-learn and CatBoost models with consistent loading behaviour across sessions.

4.2. Software Architecture of the Prototype

We organised the prototype as a layered architecture that separates the main responsibilities into seven interacting layers. This organisation helped us maintain a clean separation between data management, learning logic, user interface, and persistence, while still ensuring that the layers work together in a coherent pipeline.

- **Data layer:** manages the loading of CSV and Excel datasets and coordinates interactions with the SQLite database.
- **Preprocessing layer:** implements data cleaning operations such as missing value handling, IQR-based outlier detection and removal, signal filtering, and optional feature scaling.
- **Machine learning layer:** integrates the regression models and coordinates the training workflow, including the train/test split and model configuration.
- **Evaluation layer:** computes the regression metrics (R^2 , RMSE, MAE) for each trained model and aggregates results for comparison.
- **Prediction layer:** loads the active model from disk and generates BOD5 estimates from new user inputs.
- **Interface layer:** provides the PyQt6 graphical interface and role-based navigation between the different pages.
- **Storage layer:** handles the persistence of datasets, trained models, experimental results, and prediction logs in the SQLite database.

We implemented these layers within a single Python package named **SMART_WASTEWATER_FRAMEWORK**, which groups the code into dedicated modules for each GUI page and a central Database.py module responsible for database operations. The trained model artefacts are stored in a Models/ subdirectory, allowing the application to retrieve the current soft sensor and related metadata as needed.

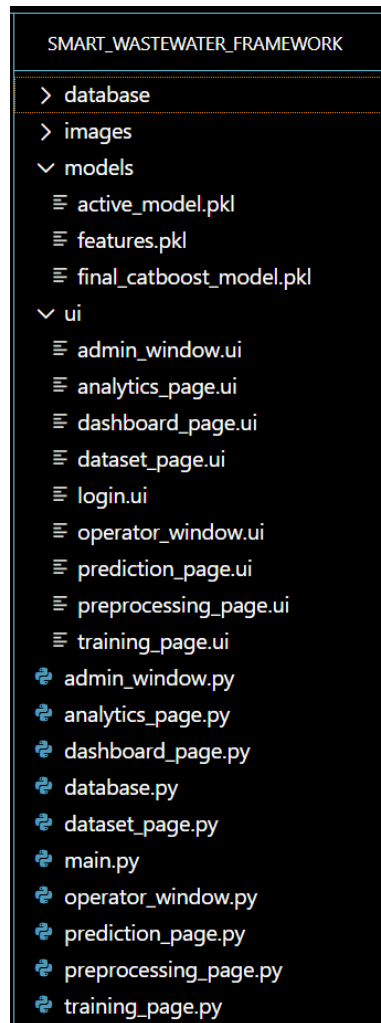


Figure 4.1: Project Structure in Visual Studio Code

4.3. Authentication and Role Management

The application starts from `main.py`, which displays a login window requesting a username, password, and role selection (Admin or Operator).

User credentials and roles are validated against the `users` table in the SQLite database, and successful authentication triggers loading of the corresponding window via `QStackedWidget`.

This role-based mechanism ensures that operational users cannot adjust or retrain models, while administrators retain full control over the experimental configuration and deployment process.

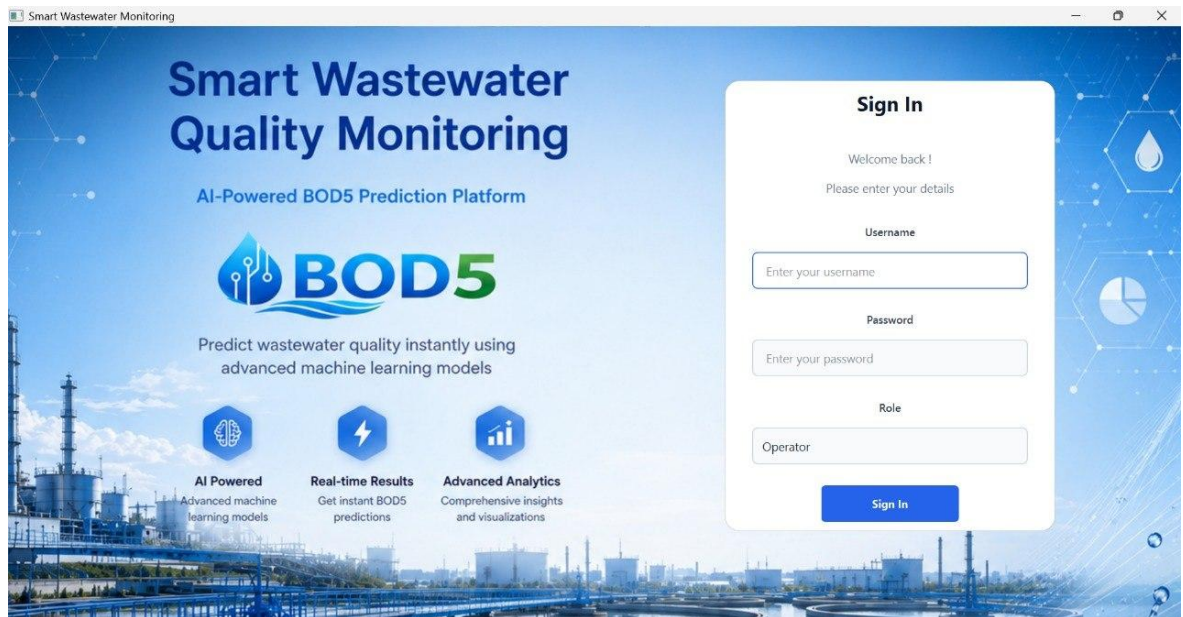


Figure 4.2: Application Login Interface

4.4. Dataset Import Module

The Dataset Management page (`dataset_page.py`), accessible only to Admin users, implements the data import functionality. Through the Import Dataset button, a local CSV or Excel file is loaded using pandas, and the interface then displays key metadata, including filename, station name, total number of rows, and number of features.

A Data Validation card summarises three quality indicators—duplicate count, missing values, and IQR-based outlier count—while a Data Preview table shows the first rows of the dataset for visual inspection.

For the Sidi Merouane case study, the imported file (`sidi_merouane_2025.xlsx`) contains 1,235 rows, with six input features and one target variable. The validation step reported zero duplicate rows, no missing values, and 15 outliers, which are handled in the preprocessing phase.

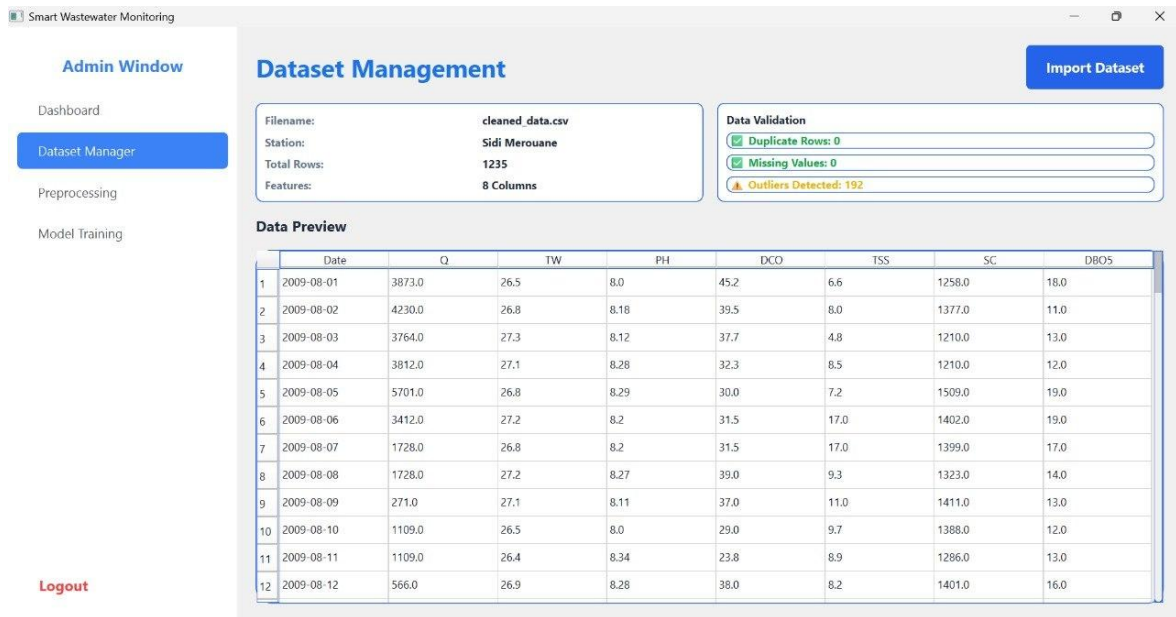


Figure 4.3: Dataset Management Page: Metadata, Validation Report, and Data Preview

4.5. Preprocessing Module

The Data Preprocessing page (`preprocessing_page.py`) provides the Admin with a configurable pipeline for data preparation. The screen is divided into a left-hand Configuration panel, where preprocessing options are defined, and a right-hand area dedicated to visualisation and logging. The Cleaning section enables the selection of a missing value strategy (row removal or mean/median imputation) and activation of IQR-based outlier removal. The Signal Filtering section offers Savitzky–Golay or Moving Average smoothing, with adjustable window size and polynomial degree, while the Feature Scaling section allows optional normalisation via StandardScaler or MinMaxScaler. Once the configuration is set, the Process Data button executes the preprocessing pipeline.

In the present study, rows with missing values were removed, IQR filtering was applied to outliers, and Savitzky–Golay filtering with a window of 5 and polynomial degree of 2 was used. The resulting variables—`Q_filtered`, `DCO_filtered`, `TSS_filtered`, `SC_filtered`, `PH_filtered`, and `TW_filtered`—constituted the final input set for model training. The preprocessing settings are stored as a JSON file and logged in SQLite for traceability, while the right panel provides before/after plots for selected variables and a Processing Log reporting real-time execution message.

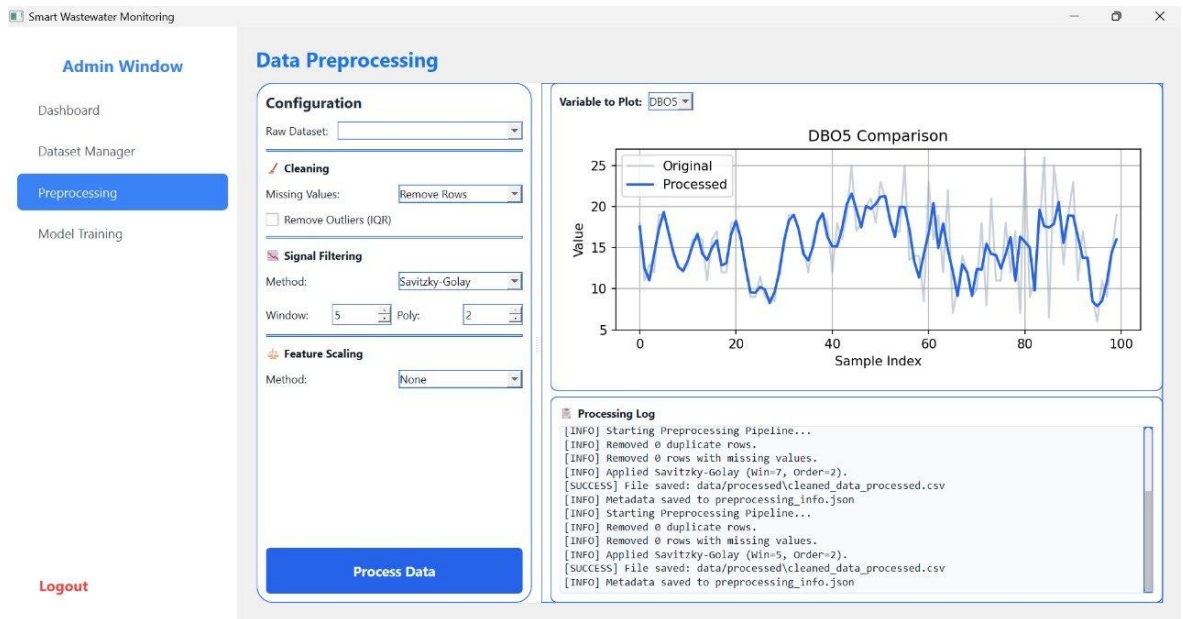


Figure 4.4: Data Preprocessing Page: Configuration Panel and Preprocessing Log

4.6. Model Training Module

The Model Training page (`training_page.py`) embodies the machine learning core of the framework. Through its configuration panel, the user selects a processed dataset, one or more algorithms (Linear Regression, Random Forest, Extra Trees, XGBoost, CatBoost), the target variable, and the test set ratio. For the experiments reported here, an 80/20 split (20% test set) with random seed 42 was used.

When the Start Training button is pressed, each selected algorithm is fitted on the training subset and evaluated on the test subset. The Current Training Results table is updated in real time with R^2 , RMSE, and MAE for each trained model, and the best model by R^2 is highlighted in the Best Model field. The Deploy button serialises this model using Joblib, marks it as active in the SQLite database, and archives any previously active model.

During the research phase, GridSearchCV was applied to CatBoost to explore combinations of depth, iterations, learning rate, and L2 regularisation. The tuned configuration improved test R^2 by only 0.003 and at the same time increased the gap between training and test performance, indicating a slight overfitting tendency. Given that the default CatBoost settings already integrate overfitting control, the default configuration was retained for deployment, while automated hyperparameter optimisation is left as a future extension of the framework.

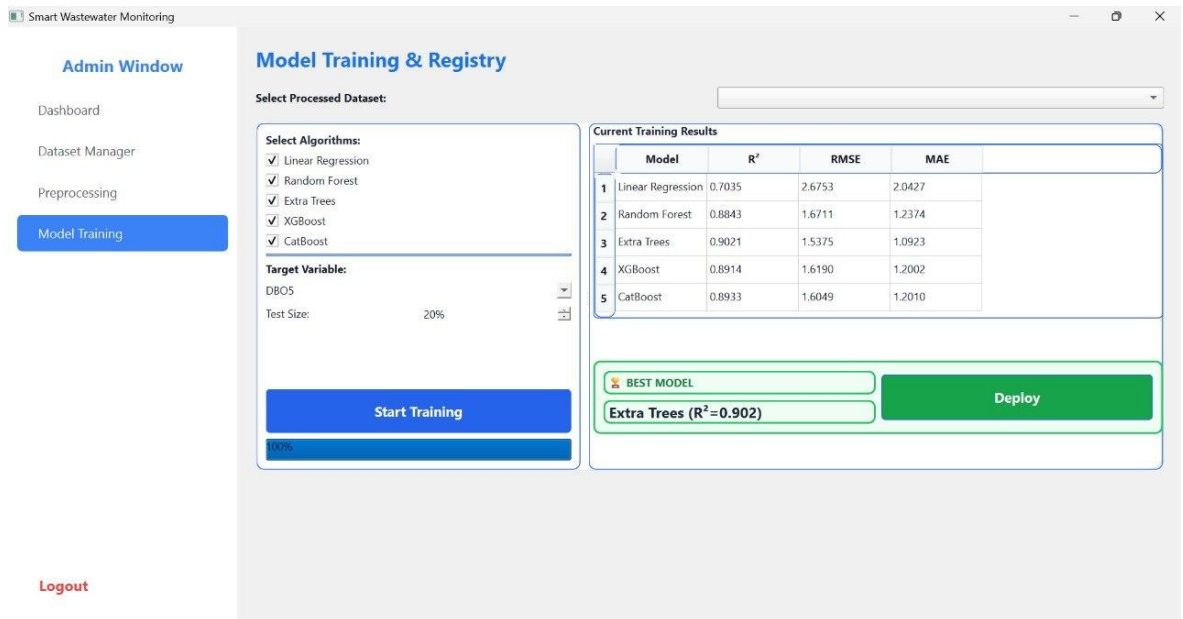


Figure 4.5: Model Training Page: Algorithm Selection, Results Table, and Deploy Button

4.7. Evaluation and Comparison Module

Model evaluation is tightly coupled to the training process. For each trained model, R^2 , RMSE (mg/L), and MAE (mg/L) are computed on the held-out test set and displayed in a consolidated table, allowing immediate side-by-side comparison without leaving the training session. Table 4.2 reproduces the performance obtained for the Sidi Merouane case study.

Model	R ²	RMSE (mg/L)	MAE (mg/L)
Linear Regression	0.7035	2.6753	2.0427
Random Forest	0.8829	1.6810	1.2458
Extra Trees	0.8977	1.5711	1.1067
CatBoost	0.9024	1.5351	1.1404
XGBoost	0.8894	1.6342	1.2084

Table 4.2: Performance Comparison across All Models (80/20 Split, Full Feature Set)

4.8. BOD5 Prediction Module

The Prediction Page (prediction_page.py) serves as the primary operational interface for the Station Operator. It provides six input fields corresponding to key physicochemical

measurements: water temperature (TW), flow rate (Q), chemical oxygen demand (COD), pH, specific conductance (SC), and total suspended solids (TSS).

when clicking the 'Predict' button, the application dynamically loads the active machine learning model via Joblib. The model processes the inputs to generate an instantaneous BOD5 estimate, displayed alongside a color-coded quality status:

- Good (< 15 mg/L, green)
- Moderate (15–30 mg/L, orange)
- Polluted (> 30 mg/L, red).

To ensure model transparency and foster operator trust, the interface integrates SHAP (SHapley Additive exPlanations) for real-time interpretability. Rather than presenting the prediction as a black-box output, the system generates a detailed breakdown of feature contributions. Using a Tree Explainer, the application calculates the marginal impact of each parameter on the final BOD5 value. These impacts are visualized through intuitive, color-coded indicators: red markers highlight parameters that actively drive the pollution level upward (e.g., elevated COD), while green markers indicate parameters that mitigate it (e.g., optimal pH levels). This explainability component allows operators to quickly identify the root physical causes behind a specific prediction.

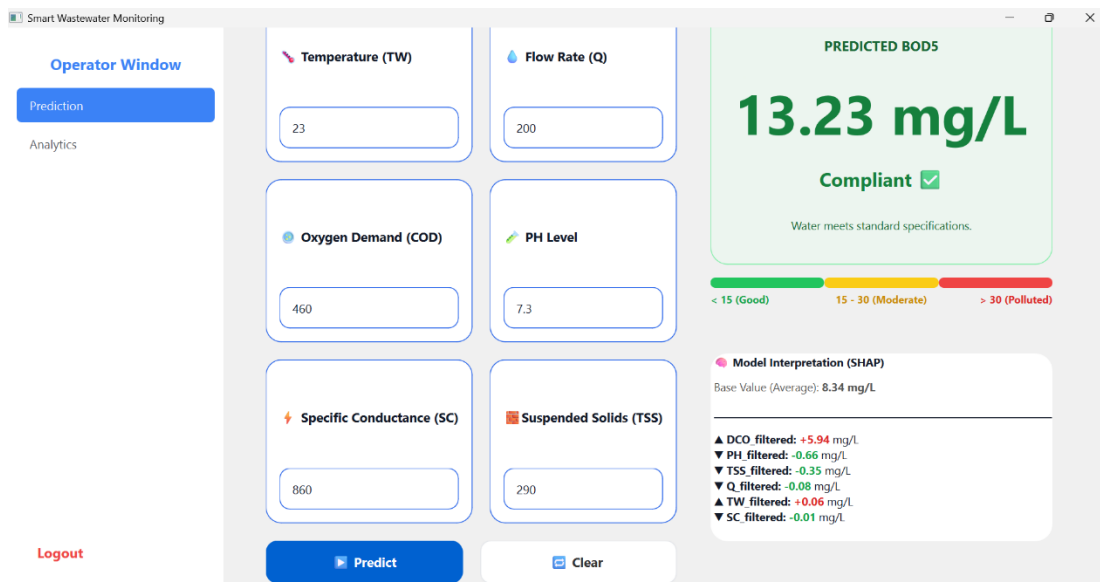


Figure 4.6: Prediction Page: Input Fields and Colour-Coded BOD5 Result Panel

4.9. Dashboard and Analytics Module

The Admin Dashboard provides a high-level overview of the framework's usage and the models produced. It includes four summary cards (total datasets, processed datasets,

number of trained models, and current active model status) and a registry table listing all models with their names, origin dataset, R^2 score, and deployment status (active or archived) in addition to the action column, which allows activate or archive any of the models saved in the database. This persistent registry allows administrators to track the evolution of soft sensor experiments over time and easy control.

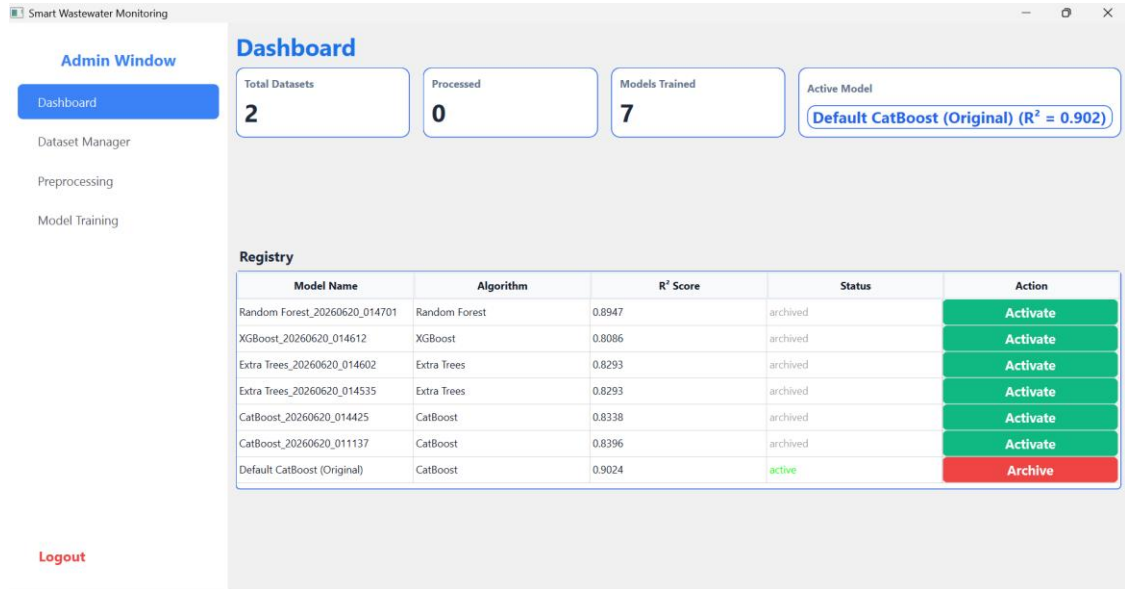


Figure 4.7: Admin Dashboard: Model Registry and Status Overview

For operators, the Analytics page summarises the behaviour of the deployed soft sensor. It features three key indicators (average BOD5, total number of predictions, weekly alerts). In addition to *BOD5 trend and input parameters trend charts*.

These visualisations support operational surveillance and anomaly detection.

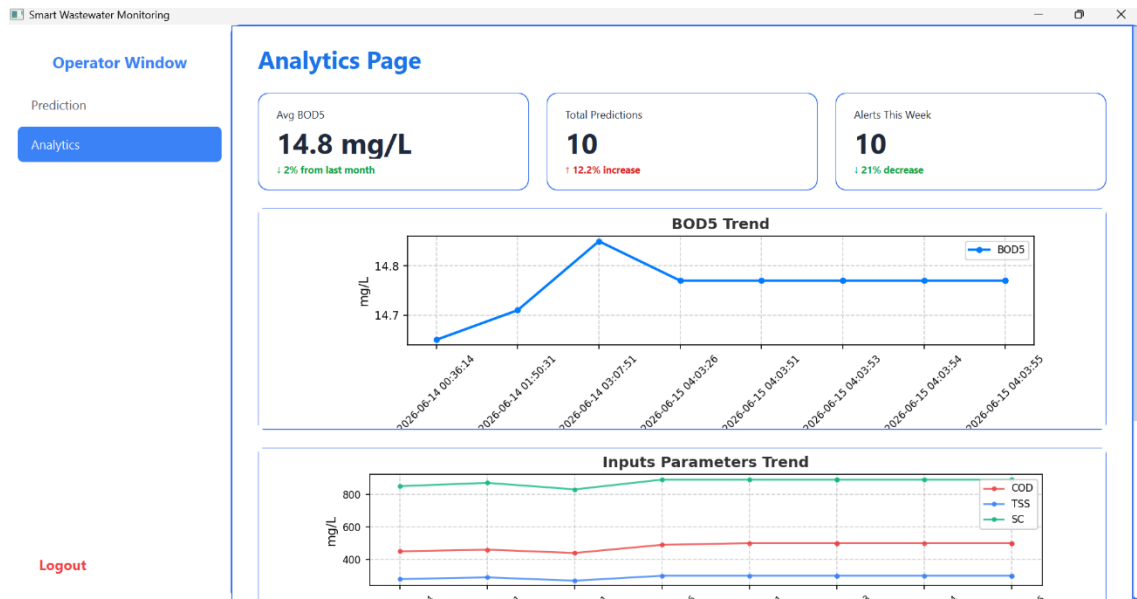


Figure 4.8: Operator Analytics Page: BOD5 Trend and Input-Parameter Trend Charts

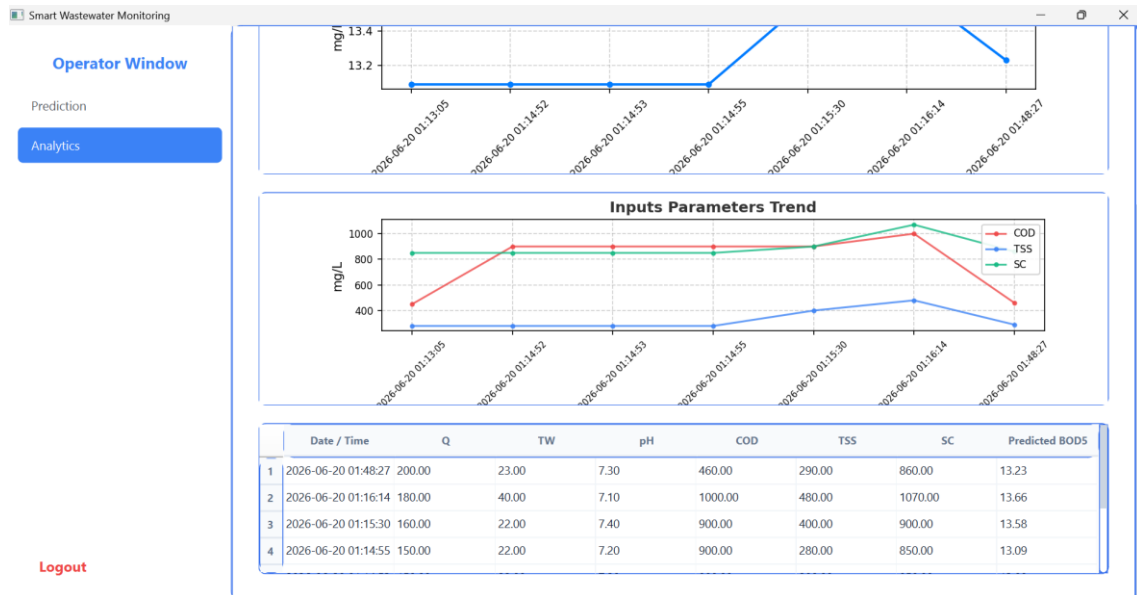


Figure 4.9: Operator Analytics Page: Historical Table

The current version does not generate PDF reports automatically; instead, SQLite acts as the central persistence mechanism, and historical information remains accessible via the Dashboard and Analytics interfaces. Export of preprocessed datasets as CSV files is supported from the preprocessing page.

4.10. Conclusion

This chapter detailed the implementation of the framework prototype, from the development stack and layered software architecture to the design of each functional module. The conceptual model of Chapter 3 has been instantiated as a Python/PyQt6 desktop application with an embedded SQLite database, ensuring full traceability from raw data import to model deployment and prediction history. The role-based separation between Admin and Operator interfaces mirrors the distinction between users who configure and evaluate models and those who exploit the resulting soft sensor. With the prototype operational and validated on the Sidi Merouane dataset, the next chapter turns to a systematic evaluation of the framework through an in-depth case study on BOD5 soft sensor generation.

CHAPTER 5

EXPERIMENTAL VALIDATION OF THE FRAMEWORK: BOD5 SOFT SENSOR AT THE SIDI MEROUANE WWTP

5.1. Dataset Description

For our experiments, we used data from the wastewater treatment plant (WWTP) of Sidi Merouane in Algeria. The measurements span the period from 2009 to 2013 and therefore cover several seasons and operating conditions. After removing incomplete records and unnamed columns introduced during export, the cleaned dataset contains 1,235 observations across seven physicochemical variables.

The response variable is BOD5, the five-day biochemical oxygen demand expressed in mg/L, which we aim to estimate using the soft sensor. The six input variables are COD/DCO, pH, water temperature (TW), total suspended solids (TSS), specific conductivity (SC), and volumetric flow rate (Q). Table 5.1 summarises the main characteristics of this dataset, including the split into 988 training and 247 test observations under the 80/20 scheme and table 5.2 contained parameters informations.

Parameter	Value
Station	STEP Sidi Merouane, Algeria
Monitoring period	2009 – 2013
Total observations (after cleaning)	1,235
Training set (80%)	988 observations
Test set (20%)	247 observations
Target variable	BOD5 (mg/L)
Input variables	COD, pH, TW, TSS, SC, Q (Savitzky-Golay filtered)

Table 5.1: Experimental Dataset Summary

BOD5 Soft Sensor at the Sidi Merouane WWTP

Variable	Description	Unit	Role
BOD5 / BOD5	Biochemical Oxygen Demand (5-day)	mg/L	Target variable
DCO / COD	Chemical Oxygen Demand	mg/L	Input feature
MES / TSS	Total Suspended Solids	mg/L	Input feature
PH	Acidity / alkalinity index	—	Input feature
TW	Water temperature	°C	Input feature
SC	Specific conductivity	μS/cm	Input feature
Q	Wastewater flow rate	m ³ /h	Input feature

Table 5.2: Input Variable Information

A preliminary correlation analysis confirmed that COD is the variable most strongly correlated with BOD5 (Pearson $r = 0.647$), consistent with the known biochemical relationship between these two oxygen demand indicators. The remaining variables showed moderate to weak linear correlations, motivating the use of non-linear ensemble methods capable of capturing more complex interactions.

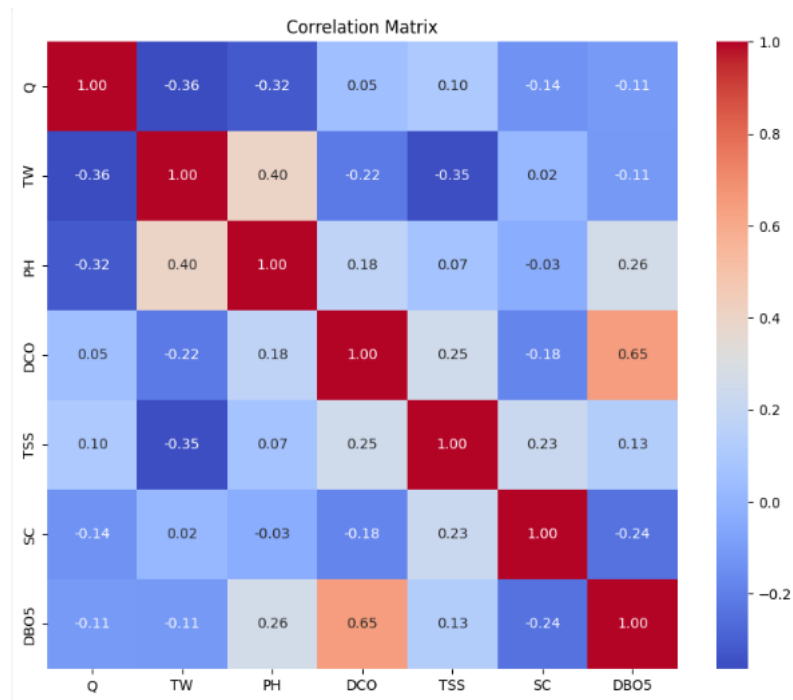


Figure 5.1: Pearson Correlation Coefficients between Input Variables and BOD5

5.2. Experimental Protocol

The experimental protocol was designed to test not only model performance but also the capacity of the application to manage the complete workflow in a systematic way. The key choices are summarized below.

5.2.1. Data cleaning

5.2.2. Data denoising

To reduce noise from wastewater sensor signals while preserving vital peaks representing system events, we applied the Savitzky-Golay filter. We tested three window sizes (5, 7, and 11) to balance noise reduction and peak preservation. Our observations showed that a window of 5 was too narrow, allowing too much noise to pass, while a window of 11 caused over-smoothing, obscuring important peaks. Ultimately, we selected a window of 7, which effectively denoised the data without losing significant detail. As shown in the filtered signal curve of BOD5 appeared in figure 5.2, this configuration successfully isolated high-frequency electronic noise and maintained the true hydraulic peaks, resulting in a cleaner dataset for subsequent analysis.

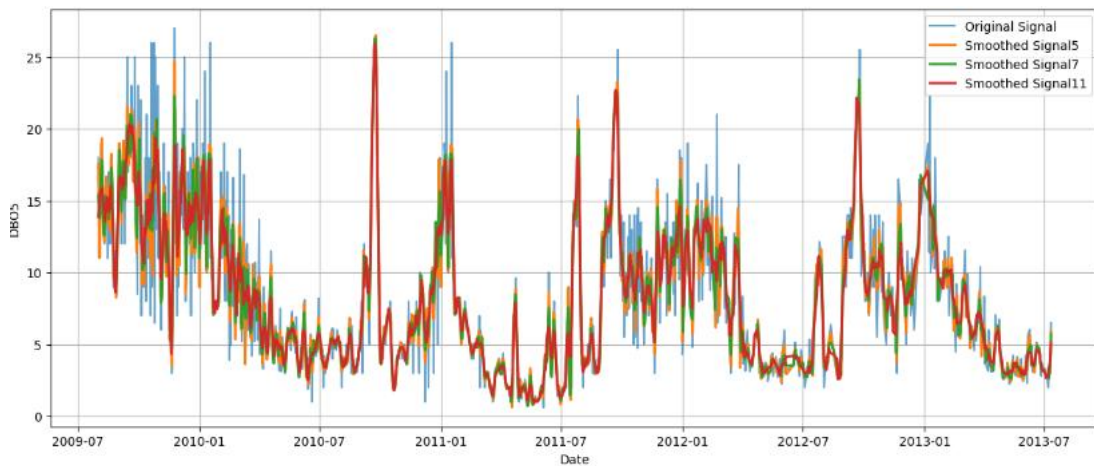


Figure 5.2: Savitzky-Golay Denoising of BOD5: Window-Size Comparison

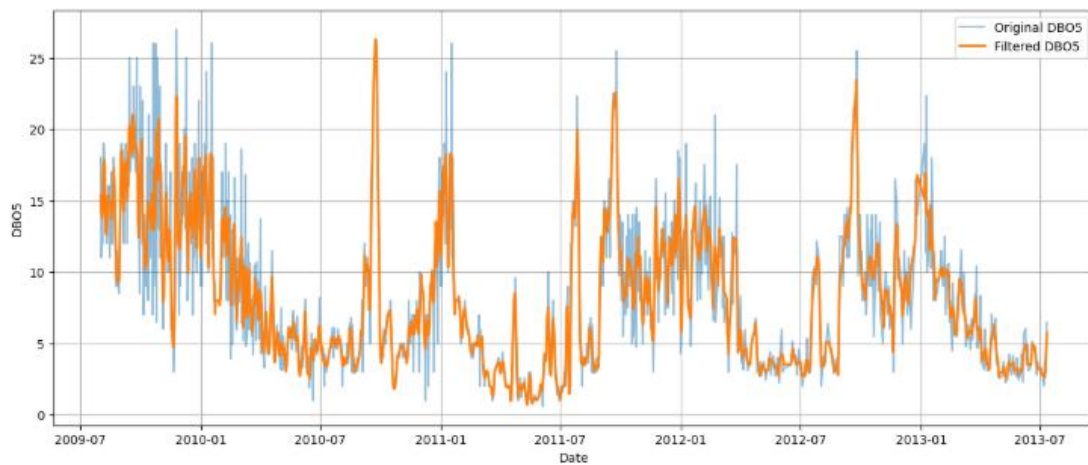


Figure 5.3: Original BOD5 Values versus Filtered BOD5 Values

5.2.3. Data Split

A 80/20 train-test split was applied to the Sidi Merouane dataset, resulting in 988 training points and 247 test points. A fixed random seed (42) was used in all experiments to keep the split consistent and make comparisons between models directly meaningful.

5.2.4. Models Evaluated

Five regression algorithms were evaluated in this study: Linear Regression, Random Forest Regressor, Extra Trees Regressor, XGBoost Regressor, and CatBoost Regressor. The tree-based ensembles were configured with default hyperparameters and a fixed random seed. No extensive hyperparameter tuning was carried out during the initial comparison, so that the models could be evaluated primarily under their default configurations and compared as a first baseline.

5.2.5. Evaluation Metrics

As in the software, model performance was assessed using three metrics: R^2 , RMSE, and MAE.

- R^2 (coefficient of determination) measures the proportion of variance in BOD5 explained by the model; values approaching 1.0 indicate better predictive performance, and values above 0.85 are often considered satisfactory for soft sensor applications.
- RMSE (root mean squared error) is expressed in mg/L and gives more weight to larger errors.
- MAE (mean absolute error) is also in mg/L and summarizes the average magnitude of prediction errors in a more interpretable way.

5.2.6. Comparison Method

All five models were trained and evaluated within a single application session so that their results appeared together in the training comparison table. The model with the highest R^2 on the test set was automatically flagged as the best candidate, reducing the risk of misinterpretation by users unfamiliar with the metrics. To deepen the analysis, a secondary series of experiments was conducted with nine different input variable combinations, in order to examine how performance depends on the set of features made available to the model.

5.3. Scenario 1 — Local BOD5 Soft Sensor

In the first scenario, we evaluated the framework's ability to generate a local soft sensor for BOD5 at a single plant, STEP Sidi Merouane. We applied the full preprocessing pipeline — removal of missing values, IQR-based outlier filtering, and Savitzky–Golay signal smoothing — and used the six filtered variables as the input feature set for all models.

Table 5.2 presents the performance of the five regression algorithms on the held-out test set. CatBoost achieved the best results with an R^2 of 0.9024, an RMSE of 1.5351 mg/L, and an MAE of 1.1404 mg/L, indicating that the model can estimate BOD5 with high accuracy despite the conventional five-day delay required for laboratory measurements. The ensemble models in general outperformed Linear Regression, which reached an R^2 of 0.7035 and higher error values.

BOD5 Soft Sensor at the Sidi Merouane WWTP

Model	R^2	RMSE (mg/L)	MAE (mg/L)
Linear Regression	0.7035	2.6753	2.0427
Random Forest	0.8829	1.6810	1.2458
Extra Trees	0.8977	1.5711	1.1067
CatBoost	0.9024	1.5351	1.1404
XGBoost	0.8894	1.6342	1.2084

Table 5.3: Model Performance on the Test Set with the Full Feature Set (80/20 Split)

We further inspected the CatBoost model by plotting predicted versus actual BOD5 values on the test set. The scatter plot shows points clustering around the diagonal line, confirming that the model tracks the ground truth reasonably well over the range of observed values. A residual analysis, either as a residual scatter plot or a time-series comparison, indicates that the residuals remain mostly centred and do not display strong systematic patterns, supporting the reliability of the model under these conditions.

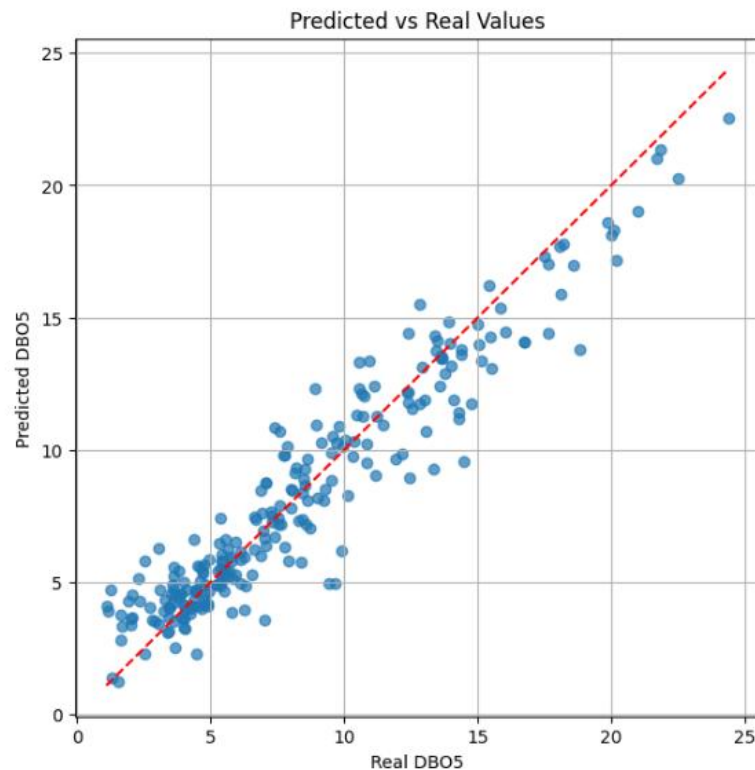


Figure 5.4: Predicted versus Actual BOD5 Values: CatBoost Model

5.4. Scenario 2 — Comparative Model Evaluation

The second scenario focuses on comparing the behaviour of the five algorithms under the same data and preprocessing conditions, with particular attention to generalisation and overfitting. For CatBoost, we examined both training and test performance; the corresponding metrics are reported in Table 5.3.

Partition	R ²	RMSE (mg/L)
Training set	0.9851	0.6057
Test set	0.9024	1.5351

Table 5.4: Overfitting Analysis for the CatBoost Model

On the training set, CatBoost reached an R² of 0.9851 and an RMSE of 0.6057 mg/L, whereas on the test set it achieved an R² of 0.9024 and an RMSE of 1.5351 mg/L. The difference between training and test R² indicates a moderate degree of overfitting, which is expected for a flexible ensemble applied to a dataset of this size. Nonetheless, the test scores remain high, and CatBoost’s internal overfitting detection mechanisms keep the performance gap at an acceptable level for practical use.

5.5. Scenario 3 — Input Variable Combination Analysis

In the third scenario, we investigated how the choice and number of input variables influence the predictive performance of the CatBoost model. We defined nine distinct input combinations, ranging from single-variable configurations to the full six-variable set, and evaluated each configuration under the same training protocol. This analysis reflects realistic situations where some sensors may be unavailable or where reducing the number of measurements is desirable for cost or maintenance reasons.

Table 5.4 lists the nine combinations, sorted by descending test R².

The full feature set (C6) — including COD, TSS, pH, TW, SC, and Q — achieves the highest performance with an R² of 0.9042, an RMSE of 1.5205 mg/L, and an MAE of 1.1270 mg/L. Combinations C5 and C4, which comprise COD, TSS, pH, and TW (with or without SC), still reach R² values above 0.87 and maintain relatively low errors, indicating that a

BOD5 Soft Sensor at the Sidi Merouane WWTP

configuration with five sensory inputs remains highly effective when some measurements are not available.

At the other end of the ranking, combination C9 (TSS + TW + Q) yields an R^2 of 0.4687 and markedly higher errors, reflecting the difficulty of predicting BOD5 accurately without COD. COD alone (C1) already explains around 67% of the variance ($R^2 = 0.6679$) but performs less well than combinations that include additional variables. These results confirm the dominant predictive role of COD and underline the added value of including other physicochemical variables, particularly TSS, pH, and TW.

Rank	Combination	Variables	R^2	RMSE (mg/L)	MAE (mg/L)
1	C6 — Full Set	COD, TSS, pH, TW, SC, Q	0.9042	1.5205	1.1270
2	C5	COD + TSS + pH + TW + SC	0.8848	1.6678	1.2554
3	C4	COD + TSS + pH + TW	0.8738	1.7457	1.3286
4	C3	COD + TSS + pH	0.7771	2.3195	1.7548
5	C8	COD + pH + SC	0.7759	2.3256	1.7645
6	C2	COD + TSS	0.7220	2.5903	2.0131
7	C7	COD + Q	0.7152	2.6218	1.9320
8	C1	COD only	0.6679	2.8313	2.1637
9	C9	TSS + TW + Q	0.4687	3.5813	2.6028

Table 5.5: CatBoost Performance across Nine Input-Variable Combinations

5.6. Best Soft Sensor Selection

By jointly considering the three experimental scenarios, we identified the configuration that offers the best compromise between accuracy and practical applicability. The selected soft sensor is the CatBoost Regressor trained on the full set of six filtered variables (combination C6). This model attains an R^2 of 0.9042, an RMSE of 1.5205 mg/L, and an

BOD5 Soft Sensor at the Sidi Merouane WWTP

MAE of 1.1270 mg/L on the test set, and therefore provides the most accurate BOD5 estimates among the evaluated alternatives.

We saved this model in the application as `Final_catboost_model.pkl` using Joblib, along with the corresponding input feature specification. The prediction module invokes this artefact to generate BOD5 estimates from new real-time sensor measurements, relying solely on physicochemical variables that are typically available at wastewater treatment plants, without any need for contemporaneous BOD5 laboratory results.

5.7. Influential Variable Analysis

To gain insight into how the soft sensor uses the different inputs, we examined the feature importance scores provided by CatBoost and complemented this with a SHAP-based interpretability study conducted in Jupyter Notebooks. Both analyses consistently indicate that COD (DCO) is the most influential variable for predicting BOD5, followed by specific conductivity (SC) and pH.

Water temperature (TW) contributes by capturing seasonal and diurnal variations that affect the biological processes underlying BOD5. Flow rate (Q) and TSS appear with lower individual importance scores, yet the combination analysis from Scenario 3 shows that including them still leads to improved overall performance compared to configurations that omit them. In this way, the feature importance results align with the quantitative findings from the variable combination experiments.

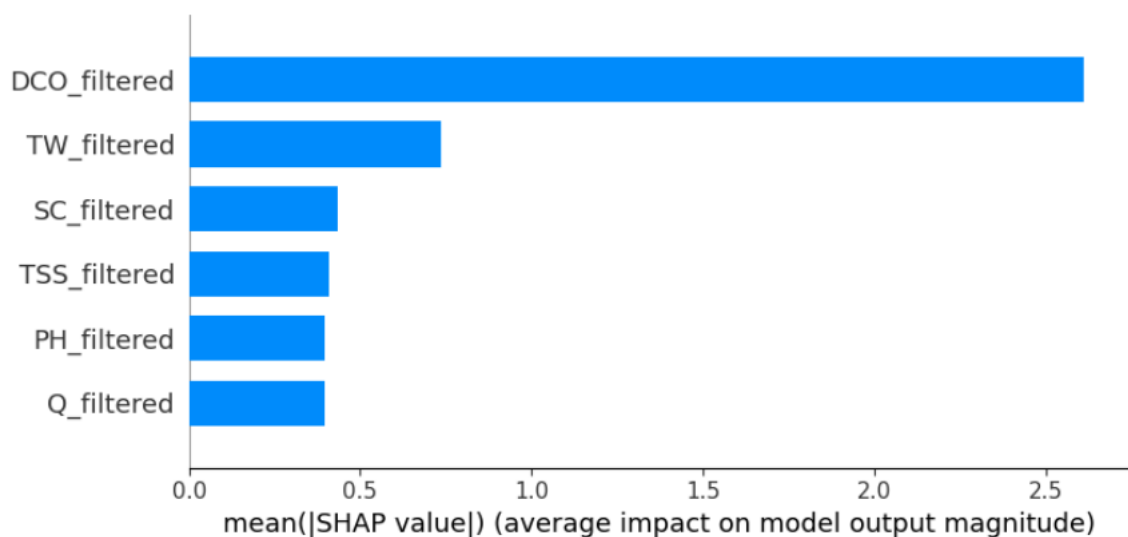


Figure 5.5: SHAP Feature Contribution Analysis

5.8. Discussion

The results obtained across the three scenarios demonstrate that the proposed framework is capable of generating a reliable soft sensor for BOD5 estimation at STEP Sidi Merouane. An R^2 of approximately 0.90 is considered a strong result for this type of application, given the modest dataset size and the inherent measurement noise in real treatment plant data. From a practical standpoint, the soft sensor delivers a BOD5 estimate in seconds from online sensor readings, replacing an analysis that conventionally requires five days of laboratory incubation — a difference with direct implications for process control and regulatory compliance monitoring.

The combination analysis carried out in Scenario 3 adds an important dimension that a fixed-model application could not provide. By systematically evaluating nine input configurations, we quantified the individual contribution of each variable and identified reduced-input alternatives that retain acceptable accuracy. This type of sensitivity analysis is a direct benefit of the framework's configurable design and would be impractical to reproduce without the structured experimental pipeline we implemented.

Several limitations must nonetheless be acknowledged. The dataset covers a single station over a four-year period, restricting the diversity of operating conditions represented in the training data. Seasonal non-stationarities, changes in industrial effluent load, and equipment maintenance events can shift the underlying distribution in ways the model cannot anticipate. The random train/test split used in this study also does not fully simulate a chronological deployment scenario; a complementary chronological-split experiment confirmed that performance degrades somewhat under this more realistic condition, underscoring the importance of periodic model retraining in production environments.

With respect to hyperparameter optimisation, GridSearchCV applied to CatBoost during the research phase produced only a marginal test improvement (+0.003 in R^2) while slightly widening the train/test performance gap. The default parameters already delivered stable and competitive results, so we retained them for the deployed model. Hyperparameter tuning is planned as a future module of the framework once it is validated on a broader set of stations and datasets.

More broadly, the generalisation of the resulting soft sensor to other stations remains the primary open challenge, and is identified as the main direction for future work, including multi-station training, transfer learning, and automated model drift detection.

General Conclusion

This dissertation addressed the problem of delayed BOD5 measurement in wastewater treatment plants and investigated the use of machine learning-based soft sensors as a complementary solution. BOD5 is a major indicator of biodegradable organic pollution and is essential for assessing wastewater-treatment performance. However, the conventional laboratory method requires five days of incubation, which limits its usefulness for rapid monitoring and operational decision-making. This delay creates a gap between the time at which process conditions change and the time at which the corresponding BOD5 result becomes available.

The main contribution of this work is not limited to the construction of a single BOD5 prediction model. Instead, the study proposed the design and implementation of a configurable framework for soft-sensor development in wastewater treatment. The framework provides a structured approach for preparing data, applying preprocessing operations, configuring experiments, training regression models, evaluating their performance, comparing alternatives, interpreting prediction results, and using a selected model for BOD5 estimation. The framework therefore supports experimentation and model selection rather than imposing a single fixed modelling solution.

A functional prototype of the proposed framework was implemented. Its interfaces provide support for user authentication, dataset management, preprocessing, model training, performance comparison, BOD5 prediction, and visualisation of prediction-related information. The implemented prototype demonstrates how the proposed framework can organise the major stages of a soft-sensor development workflow within one software environment. Nevertheless, the prototype should be considered an experimental and decision-support tool, rather than a fully deployed industrial monitoring system.

The validation of the framework was carried out through a case study involving BOD5 prediction at the Sidi Merouane wastewater treatment plant. Historical physicochemical data were used to train and compare five regression models: Linear Regression, Random Forest Regression, Extra Trees Regression, XGBoost, and CatBoost. The study considered three complementary scenarios: the development of a local BOD5 soft sensor, the comparison of several candidate models, and the analysis of different combinations of input variables.

The obtained results showed that CatBoost was the best-performing model among the evaluated alternatives. It achieved an R^2 of 0.9042, an RMSE of 1.5205 mg/L, and an MAE of 1.1270 mg/L on the test set. These results indicate that the selected model can provide useful BOD5 estimates from routinely available physicochemical variables. The interpretation analysis based on feature importance and SHAP values showed that COD was the most influential variable, followed by specific conductivity and pH. Flow rate, temperature, and total suspended solids also contributed to the prediction quality, especially when considered jointly with the other variables.

The study also confirmed the value of the configurable framework approach. The comparison of several algorithms and multiple input-variable combinations would be more difficult to conduct and document within a fixed application designed around a single model. The framework made it possible to identify the most suitable model for the available dataset and to evaluate the effect of different input configurations. In this sense, the BOD5 case study validates the feasibility of the proposed framework at the prototype level.

Several limitations should nevertheless be acknowledged. The experimental dataset originates from a single wastewater treatment plant, which limits the generalisation of the findings to other stations and operating conditions. The data may not capture all possible seasonal variations, maintenance events, sensor failures, or changes in influent composition. In addition, the random train-test split used for the main evaluation does not fully reproduce a real chronological deployment situation. Finally, the prototype was not connected to a live SCADA system and was not tested under continuous operation at the wastewater treatment plant.

Future work should therefore focus on extending the framework in several directions. First, the framework should be validated with datasets from multiple wastewater treatment plants in order to evaluate its ability to support different operational contexts. Second, chronological validation strategies, periodic retraining procedures, and model-drift detection mechanisms should be introduced to improve long-term reliability. Third, the framework may be extended to other target variables, such as COD, turbidity, suspended solids, dissolved oxygen, or treated-water quality indicators. Finally, future work may consider the integration of the framework with online monitoring systems and SCADA platforms, after appropriate technical validation in a real operating environment.

In conclusion, this work demonstrates that a configurable framework can support the systematic development of machine learning-based soft sensors for wastewater treatment. The BOD5 case study at the Sidi Merouane wastewater treatment plant confirms the feasibility of the approach and provides a basis for future extension toward broader, multi-station, and operationally validated applications.

References

-
- Abba, S. I., & Elkiran, G. (2017). Effluent prediction of chemical oxygen demand from the astewater treatment plant using artificial neural network application. *Procedia Computer Science*, 120, 156–163. <https://doi.org/10.1016/j.procs.2017.11.223>
 - Abd, S. A., & Al-Sulttani, A. O. (2024). Predicting biochemical oxygen demand at the inlet of Al-Rustumiya wastewater treatment plant using different mathematical techniques. *Journal of Engineering*, 30(2), 16–29. <https://doi.org/10.31026/j.eng.2024.02.02>
 - Abou-Elela, S. I., Kamel, M. M., & Fawzy, M. E. (2010). Biological treatment of saline wastewater using a salt-tolerant microorganism. *Desalination*, 250(1), 1–5. <https://doi.org/10.1016/j.desal.2009.03.022>
 - Aghdam, E., Mohandes, S. R., Manu, P., Cheung, C. M., Yunusa-Kaltungo, A., & Zayed, T. (2023). Predicting quality parameters of wastewater treatment plants using artificial intelligence techniques. *Journal of Cleaner Production*, 405, Article 137019. <https://doi.org/10.1016/j.jclepro.2023.137019>
 - Alam, G., Ihsanullah, I., Naushad, M., & Sillanpää, M. (2022). Applications of artificial intelligence in water treatment for optimization and automation of adsorption processes: Recent advances and prospects. *Chemical Engineering Journal*, 427, Article 130011. <https://doi.org/10.1016/j.cej.2021.130011>
 - Alizamir, M., Kim, S., Zounemat-Kermani, M., Heddami, S., Shahrabadi, A. H., & Gharabaghi, B. (2021). Modelling daily soil temperature by hydro-meteorological data at different depths using a novel data-intelligence model: Deep echo state network model. *Artificial Intelligence Review*, 54(4), 2863–2890. <https://doi.org/10.1007/s10462-020-09915-5>
 - Alsulaili, A., & Refaie, A. (2021). Artificial neural network modeling approach for the prediction of five-day biological oxygen demand and wastewater treatment plant performance. *Water Supply*, 21(5), 1861–1877. <https://doi.org/10.2166/ws.2020.199>
 - Arévalo, G., & Borne, I. (2001). Architectural description of object-oriented frameworks: An approach. *Langages et Modèles à Objets – L’Objet*, 7, 183–198.
 - Baki, O. T., Aras, E., Özel Akdemir, U., & Yılmaz, B. (2019). Biochemical oxygen demand prediction in wastewater treatment plant by using different regression

- analysis models. *Desalination and Water Treatment*, 157, 79–89. <https://doi.org/10.5004/dwt.2019.24158>
- Bahramian, M., Dereli, R. K., Zhao, W., Giberti, M., & Casey, E. (2023). Data to intelligence: The role of data-driven models in wastewater treatment. *Expert Systems with Applications*, 217, Article 119453. <https://doi.org/10.1016/j.eswa.2022.119453>
 - Bali, M. (2024). *Traitement des eaux usées* [Mémoire de licence, Institut supérieur des sciences et techniques des eaux, Université de Gabès]. HAL. <https://hal.science/hal-04838747>
 - Baumont, S., Camard, J.-P., Lefranc, A., & Franconie, A. (2004). *Réutilisation des eaux usées épurées : Risques sanitaires et faisabilité en Île-de-France*. Observatoire régional de santé d'Île-de-France.
 - Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
 - Butler, G., & Dénommée, P. (1999). Documenting frameworks. In *Building application frameworks* (Chap. 21, pp. 495–503). Wiley.
 - Cardot, C. (1999). *Les traitements de l'eau : Procédés physico-chimiques et biologiques : Cours et problèmes résolus*. Ellipses.
 - Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
 - Ching, P. M. L., Zou, X., Wu, D., So, R. H. Y., & Chen, G. H. (2022). Development of a wide-range soft sensor for predicting wastewater BOD5 using an eXtreme gradient boosting (XGBoost) machine. *Environmental Research*, 210, Article 112953. <https://doi.org/10.1016/j.envres.2022.112953>
 - Desai, S. (2019). *Application du modèle Random Forest Regression dans l'analyse fréquentielle régionale des événements hydrologiques extrêmes* [Mémoire de maîtrise, Institut national de la recherche scientifique, Centre Eau Terre Environnement].

- Ding, G. K. C. (2017). Wastewater treatment and reuse—The future source of water supply. In *Encyclopedia of Sustainable Technologies* (pp. 43–52). Elsevier. <https://doi.org/10.1016/B978-0-12-409548-9.10170-8>
- Frat, L. (2020). *Caractérisation des eaux usées d'un réseau d'assainissement pour un diagnostic des pollutions en amont des stations d'épuration* [Thèse de doctorat, AgroParisTech / Sorbonne Université].
- Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (3e éd.). O'Reilly Media.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774).
- Paepae, T., Bokoro, P. N., & Kyamakya, K. (2021). From fully physical to virtual sensing for water quality assessment: A comprehensive review of the relevant state-of-the-art. *Sensors*, 21(21), Article 6971. <https://doi.org/10.3390/s21216971>
- Pattnaik, B. S., Pattanayak, A. S., Udgata, S. K., & Panda, A. K. (2021). Machine learning based soft sensor model for BOD estimation using intelligence at edge. *Complex & Intelligent Systems*, 7, 961–976. <https://doi.org/10.1007/s40747-020-00259-9>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems* (Vol. 31, pp. 6638–6648).
- Rapicault, P. (2002). *Modèles et techniques pour spécifier, développer et utiliser un framework : Une approche par méta-modélisation* [Thèse de doctorat, Université Nice Sophia Antipolis]. TEL. <https://tel.archives-ouvertes.fr/tel-00505470>

- Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8), 1627–1639. <https://doi.org/10.1021/ac60214a047>

الجمهورية الجزائرية الديمقراطية الشعبية

Republique Algérienne Démocratique Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur de la Recherche Scientifique



جامعة 20 août 1955 Skikda

جامعة 20 أوت 1955 - شكندة

Faculté des Sciences

Département d'informatique

كلية العلوم

قسم الإعلام الآلي

رقم الملف: 13 / 0 / 2016

Autorisation de Dépôt de Mémoire de Master

Je soussigné **TDUIL Ghassen**

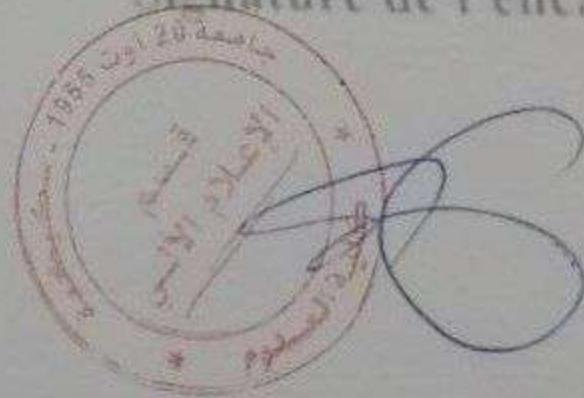
Certifie que l'étudiant(e) **BEN LOUCIF Nour El Houda**

Spécialité **IA**

Ayant soutenu le projet intitulé **Design and development of a configurable framework for generating software sensors in wastewater treatment**

A apporté les corrections nécessaires sur son manuscrit de Master

Signature de l'encadreur



الجمهورية الجزائرية الديمقراطية الشعبية

Republique Algérienne Démocratique Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur de la Recherche Scientifique



Université 20 août 1955 Skikda

الجامعة 20 أوت 1955 - سكيكدة

Faculté des Sciences

Département d'informatique

الكلية العلمية

قسم الإعلام الآلي

البريد الإلكتروني: 2005@univ-ski.dz

Autorisation de Dépôt de Mémoire de Master

Je soussigné

TOUL Ghassen

Certifie que l'étudiant(e)

TEGHDOUDI Achouak

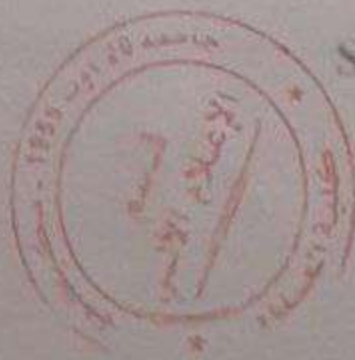
Spécialité

IA

Ayant soutenu le projet intitulé:

Design and development
of a configurable framework
for generating software sensors

A apporté les corrections nécessaires sur son manuscrit de Master



Signature de l'encadreur

[Handwritten signature]