ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

SEMI-SUPERVISED LEARNING STRATEGY FOR IMPROVED FLASH POINT PREDICTION

M.Sc. THESIS

Mert SÜLÜK

Department of Computer Engineering

Computer Engineering Programme

AUGUST 2024

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

SEMI-SUPERVISED LEARNING STRATEGY FOR IMPROVED FLASH POINT PREDICTION

M.Sc. THESIS

Mert SÜLÜK
(504201527)

Department of Computer Engineering

Computer Engineering Programme

Thesis Advisor: Prof. Dr. Şule GÜNDÜZ ÖĞÜDÜCÜ

AUGUST 2024

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

PARLAMA NOKTASI TAHMİNİNİ İYİLEŞTİRMEK İÇİN YARI DENETİMLİ ÖĞRENME STRATEJİSİ

YÜKSEK LİSANS TEZİ

Mert SÜLÜK
(504201527)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Şule GÜNDÜZ ÖĞÜDÜCÜ

AĞUSTOS 2024

Mert SÜLÜK, a M.Sc. student of ITU Graduate School student ID 504201527 successfully defended the thesis entitled "SEMI-SUPERVISED LEARNING STRATEGY FOR IMPROVED FLASH POINT PREDICTION", which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Şule GÜNDÜZ ÖĞÜDÜCÜ** Istanbul Technical University	.....
Jury Members :	**Assist. Prof. Dr. Mehmet Tahir SANDIKKAYA** Istanbul Technical University	.....
	Assist. Prof. Dr. Gönül ULUDAĞ Fatih Sultan Mehmet Vakıf Üniversitesi	.....

Date of Submission : **24 May 2024**
Date of Defense : **20 August 2024**

To my.(self + family + friends) with deeply love.

FOREWORD

I extend my heartfelt gratitude to my advisor, Prof. Şule GÜNDÜZ ÖĞÜDÜCÜ, for her invaluable guidance. To my parents, Saniye and Ziya SÜLÜK, thank you for your unwavering support and love. To my friends, your encouragement and laughter kept me going. Lastly, I thank myself for persevering through this journey.

August 2024

Mert SÜLÜK
Research Assistant

TABLE OF CONTENTS

ABBREVIATIONS

SSL	**:** Semi-Supervised Learning
FP	**:** Flash Point
AI	**:** Artificial Intelligence
MM	**:** Min-Max Scaling
GPR	**:** Gaussian Process Regressor
RF	**:** Random Forest
MAE	**:** Mean Absolute Error
RMSE	**:** Root Mean Squared Error
BM	**:** Baseline Model
EW	**:** Expanding Window
ML	**:** Machine Learning

SYMBOLS

$\mathbf{X}$	: Total Data Set
$\mathbf{X_L}$	: Labeled Data
$\mathbf{X_U}$	: Unlabeled Data
$\mathbf{X_{Train}}$	: Training Data
$\mathbf{X_{Test}}$	: Test Data
$\mathbf{X_{LE}}$	: Enhanced labeled data set with semi-supervised labels
Thresh	: A predefined threshold value
winBounds	: Winsorization Boundaries
scaleBounds	: Scale Boundaries
L	: The number of labeled data
N	: Initial test size
E	: The number of pseudo-labeled data

LIST OF FIGURES

Page

SEMI-SUPERVISED LEARNING STRATEGY
FOR
IMPROVED FLASH POINT PREDICTION

SUMMARY

This thesis explores the application of semi-supervised learning techniques to enhance the prediction of flash points in the oil industry, which are critical for ensuring the safety of transporting and storing petroleum products. Flash points denote the lowest temperature at which a substance's vapors ignite in air, a crucial parameter that traditional methods ascertain through costly and time-consuming laboratory tests. This study proposes a data-driven approach to optimize these processes more efficiently and effectively.

Semi-supervised learning, which leverages both labeled and unlabeled data, provides a robust framework especially valuable in scenarios where data labeling is prohibitively expensive or logistically challenging. This research integrates sensor data such as pressure, temperature, and flow rates with sparse flash point measurements to develop a predictive model. The aim is to reduce dependency on extensive laboratory testing while enhancing operational efficiency and safety protocols.

The central research questions addressed are: How can flash points be accurately predicted in the oil industry when only a limited number of labeled data points are available? Given these constraint, could semi-supervised learning method be an effective solution? What are the specific advantages and limitations of these technique within the oil industry context? The study validates the effectiveness of semi-supervised learning method and develops a model that improves upon traditional approaches.

To address the research questions, particularly in the context of improving flash point predictions with limited labeled data, the study employs data preprocessing techniques and modeling processes that are essential for optimizing model performance. The methodology employs two principal data preprocessing techniques: Winsorization and Min-Max Scaling. Winsorization mitigates the effects of outliers by limiting extreme data points within a designated percentile range, ensuring the model is not skewed by anomalies. Min-Max Scaling normalizes the data, allowing for equitable evaluation of all features and preventing any single feature from dominating the model's output.

The modeling process involves the Gaussian Process Regressor and the Random Forest model. The Gaussian Process Regressor, suitable for continuous data, provides uncertainty estimates to gauge the reliability of predictions. The Random Forest model enhances stability and accuracy by aggregating predictions from multiple decision trees. Initially trained on labeled data, the Gaussian Process Regressor subsequently predicts labels for unlabeled data, incorporating those predictions within a specified confidence interval into the training set. This expanding dataset further trains the

Random Forest model, applying an expanding window approach to incrementally improve prediction capabilities.

Performance metrics such as Mean Absolute Error and Root Mean Squared Error assess model efficacy. The baseline model initially yielded an mean absolute error of 1.1 degrees in flash point predictions. With the application of the semi-supervised learning model, Mean Absolute Error improved to 1.01 and Root Mean Squared Error decreased to 1.63, demonstrating significant enhancements in accuracy through the inclusion of unlabeled data.

In conclusion, this thesis illustrates the potential of semi-supervised learning to bridge the gap caused by a scarcity of labeled data, particularly in critical industrial applications like oil processing. The findings suggest that semi-supervised learning not only reduces the financial and temporal expenditures associated with traditional testing methods but also offers a scalable, efficient alternative poised to transform industry practices. The methodologies developed here have broader implications, suggesting that semi-supervised learning could be similarly beneficial in other sectors where data labeling is a significant constraint and even small performance improvements are critical due to the importance of the parameters being predicted.

PARLAMA NOKTASI TAHMİNİNİ İYİLEŞTİRMEK İÇİN YARI DENETİMLİ ÖĞRENME STRATEJİSİ

ÖZET

Bu tez, petrol endüstrisinde parlama noktalarının tahminini iyileştirmek amacıyla yarı denetimli öğrenme tekniklerinin uygulanmasını araştırmaktadır. Parlama noktası, petrol ve petrol türevi ürünlerin buharlarının havayla karıştığında yanıcı hale geldiği en düşük sıcaklıktır. Bu sıcaklık, petrol ürünlerinin güvenli taşınması ve depolanması açısından kritik bir güvenlik parametresidir. Geleneksel yöntemlerle parlama noktası tahmini genellikle laboratuvar testleri gerektirir; bu testler hem zaman alıcı hem de maliyetlidir. Bu çalışmada, daha hızlı ve maliyet etkin bir alternatif olarak veri tabanlı bir yaklaşım kullanılarak bu süreci optimize etme amaçlanmaktadır. Yarı denetimli öğrenme, etiketli ve etiketlenmemiş verilerin birlikte kullanıldığı bir makine öğrenmesi yöntemidir. Bu yöntem, özellikle etiketlemenin zor veya maliyetli olduğu ve etiketsiz verinin mevcut olduğu durumlarda büyük avantajlar sağlar. Yarı denetimli öğrenme, sınırlı etiketli veri ile bile yüksek doğrulukta tahminler yapabilme, model performansını geliştirme potansiyeline sahiptir. Bu çalışmada, basınç, sıcaklık ve akış göstergeleri gibi sensör verileri, sınırlı parlama noktası laboratuvar ölçümleri ile entegre edilerek bir model geliştirilmiştir. Bu yaklaşım, geniş çaplı laboratuvar testlerine olan bağımlılığı azaltmayı ve operasyonel verimliliği ile güvenliği artırmayı hedeflemektedir.

Tezin ana araştırma soruları şunlardır: Petrol endüstrisinde, yalnızca sınırlı sayıda etiketlenmiş veri noktası mevcutken parlama noktaları nasıl doğru bir şekilde tahmin edilebilir? Bu veri kısıtlaması göz önüne alındığında, yarı denetimli öğrenme yöntemi etkili bir çözüm olabilir mi? Yarı denetimli öğrenme tekniği, sınırlı etiketli veri ile parlama noktalarını tahmin etmeye yönelik bir regresyon görevi için nasıl etkili bir şekilde uygulanabilir? Bu tekniğin petrol endüstrisi bağlamında özel avantajları ve sınırlamaları nelerdir? Bu soruları yanıtlamak amacıyla, bu tez, yarı denetimli öğrenme yöntemlerinin etkinliğini doğrulamayı ve geleneksel yöntemlere kıyasla iyileştirmeler sunan bir model geliştirmeyi hedeflemektedir.

Araştırmada kullanılan veri ön işleme teknikleri arasında Winsorization ve Min-Maks Ölçekleme bulunmaktadır. Winsorization yöntemi, veri setindeki uç değerlerin (çok yüksek veya çok düşük değerlerin) olumsuz etkilerini azaltmak için kullanılmıştır. Bu teknik, verilerin belirli bir yüzdelik dilim içerisinde sınırlanmasını sağlayarak, zaten kısıtlı olan verinin eksilmesini sağlamadan modelin aşırı değerlerden etkilenmesini önler. Min-Maks Ölçekleme ise, farklı ölçeklerdeki özniteliklerin (bağımsız değişkenlerin) model tarafından eşit şekilde değerlendirilmesini sağlamak amacıyla kullanılır. Bu iki teknik, veri setini model eğitimi için daha uygun hale getirir ve modelin daha doğru tahminler yapmasına olanak tanır. Modelleme sürecinde, Gaussian Süreç Regresörü ve Rastgele Orman modeli önemli rol oynamaktadır.

Gaussian Süreç Regresörü, sürekli çıktılar üreten ve tahminlerle ilgili belirsizlik sağlayan güçlü bir regresyon modelidir. Bu model, öncelikle etiketli veriler üzerinde eğitilir ve ardından etiketlenmemiş veriler üzerinde tahminler yapar. Eğitim sürecinde, belirli bir güven sınırının altında kalan tahminler, modelin eğitim setine eklenir. Daha sonra, genişletilmiş veri seti üzerinde Rastgele Orman modeli eğitilir. Bu model, birden fazla karar ağacının tahminlerini birleştirerek, modelin genel tahmin yeteneğini artırır. Genişleyen pencere yaklaşımı sayesinde, bu iki model veri setinden maksimum faydayı sağlayarak daha doğru tahminler yapar.

Modelin performansı, ortalama mutlak hata ve kök ortalama kare hata metrikleri kullanılarak değerlendirilmiştir. Yarı denetimli öğrenme yaklaşımının katkısının direkt olarak gözlenebilmesi için yarı denetimli öğrenme modeli ile birlikte ayrıca bir de temel model kurulmuştur. Temel model Rastgele Orman metodunu genişleyen pencere yaklaşımıyla birlikte kullanmaktadır. Yarı denetimli öğrenmede ise yalnızca ek olarak Gaussian Süreç Regresörü ile veri setinin etiketsiz verisetinden faydalanarak genişletilmesi basamağı yer almaktadır. Temel model, parlama noktalarını tahmin ederken 1.1 derece ortalama mutlak hata ve 1.7 kök ortalama kare hata ile performans göstermiştir. Kök ortalama kare hata'nınn ortalama mutlak hatadan biraz daha yüksek olması, modelin genel olarak iyi performans göstermesine rağmen bazı durumlarda daha büyük hatalar sergileyebileceğini göstermektedir. Yarı denetimli öğrenme modeli kullanıldığında, modelin ortalama mutlak hata skoru 1.01'e düşmüş, kök ortalama kare hata skoru ise 1.63'e gerilemiştir. Bu iyileşme, etiketlenmemiş verilerin kullanılması sayesinde modelin doğruluk oranının artırılabileceğini göstermektedir. Ayrıca, model genel olarak daha az hata ile daha iyi tahminler yapabilmektedir. Bu da, hassas ölçümleme ve iyileştirmenin önemli olduğu ve dolayısıyla her türlü iyileştirmenin önem arz ettiği parlama noktası gibi parametreler adına kritik bir durum olabilmektedir.

Bu çalışma, yarı denetimli öğrenme tekniklerinin, veri etiketleme maliyetlerinin yüksek olduğu ve etiketli verinin sınırlı olduğu durumlarda nasıl etkili bir şekilde kullanılabileceğini göstermektedir. Petrol endüstrisindeki parlama noktası tahmini, bu tekniklerin etkinliğini ve uygulanabilirliğini kanıtlamaktadır. Yarı denetimli öğrenme yöntemleri, yalnızca petrol endüstrisi için değil, sağlık, finans ve çevre izleme gibi diğer sektörlerde de önemli uygulama potansiyeline sahiptir. Bu sektörlerde de veri etiketleme süreçleri genellikle maliyetli ve zaman alıcıdır, bu nedenle yarı denetimli öğrenme yöntemleri, bu tür problemleri aşmak için etkili bir çözüm sunar.

Sonuç olarak, bu tez, yarı denetimli öğrenmenin etiketli veri kıtlığı arasındaki boşluğu nasıl kapatabileceğini göstererek mevcut literatüre katkıda bulunmaktadır. Bulgular, yarı denetimli öğrenmenin endüstriyel uygulamalarda ve diğer alanlarda tahmin doğruluğunu artırmak için etkili bir strateji olabileceğini öne sürmektedir. Bu tez, doğruluk ve verimliliği dengeleyen sağlam bir metodoloji sunarak, hem akademik literatüre hem de endüstriyel uygulamalara önemli katkılar sağlamaktadır. Yarı denetimli öğrenme tekniklerinin, sınırlı veriyle çalışmak zorunda kalan çeşitli sektörlerde geniş bir uygulama potansiyeli bulunmaktadır. Petrol endüstrisinde parlama noktası tahmini, bu tekniklerin pratikte nasıl kullanılabileceğine dair somut bir örnek sunar. Bu yaklaşım, laboratuvar testlerinin yerini alabilecek hızlı ve maliyet etkin çözümler sunarak, hem güvenliği artırmakta hem de operasyonel süreçleri

optimize etmektedir. Bu tezde geliştirilen model ve yöntemler, gelecekte benzer zorluklarla karşılaşan diğer endüstri alanlarında da uygulanabilir ve bu sayede geniş bir etki alanı yaratabilir. Ayrıca, bu çalışma, yarı denetimli öğrenmenin sadece teorik bir kavram olmadığını, pratik uygulamalarda da önemli faydalar sağlayabileceğini kanıtlamaktadır. Bu nedenle, yarı denetimli öğrenme yöntemlerinin benimsenmesi ve geliştirilmesi, gelecekteki araştırmalar ve endüstriyel uygulamalar için önemli bir adım olarak değerlendirilmektedir.

Bu tezin ana katkıları, yarı denetimli öğrenme ve rastgele orman modellerinin petrol endüstrisindeki uygulamalarını keşfetmek, bu modellerin laboratuvar testlerine kıyasla daha hızlı ve maliyet etkin çözümler sunabileceğini göstermek ve literatürde mevcut olmayan yeni ve gerçek endüstri veriseti üzerinde keşif imkanı sunmasıdır. Modellerin geliştirilmesi sırasında, çeşitli veri ön işleme tekniklerinin ve iki güçlü makine öğrenmesi modelinin nasıl entegre edildiği ayrıntılı olarak incelenmiştir. Bu entegrasyon, yarı denetimli öğrenme sürecinin sadece veri tahmininde ve model optimizasyonunda nasıl faydalı olabileceğini ortaya koymaktadır.

1. INTRODUCTION

1.1 Purpose of Thesis

The primary aim of this thesis is to explore the application of semi-supervised learning (SSL) techniques to enhance the prediction of flash points (FP)s in the oil industry. Accurate FP prediction is critical for the safe transportation and processing of petroleum products. Traditional methods for determining FPs are often time-consuming and costly, making it difficult to perform real-time assessments. By leveraging SSL, which utilizes both labeled and unlabeled data, this study aims to improve prediction accuracy even with limited labeled data.

To narrow down this objective, this research focuses specifically on developing a model that can integrate sensor data (such as pressure, temperature, and flow indicators) with limited laboratory measurements to predict FPs. This approach not only seeks to reduce the reliance on extensive laboratory tests but also aims to provide timely predictions that can enhance operational efficiency and safety in the oil industry.

The research questions guiding this thesis include: *How can SSL techniques be effectively applied to predict FPs with limited labeled data? What are the specific advantages and limitations of using these techniques in the context of the oil industry?* The primary targets are to validate the efficacy of these techniques and to develop a model that offers significant improvements over traditional and purely supervised learning methods.

This thesis aims to make a substantial contribution to the existing literature by demonstrating how SSL can bridge the gap between the need for accurate FP predictions and the scarcity of labeled data. It also seeks to provide a framework that can be adapted to other industries facing similar challenges.

1.2 Problem Definition and Importance

The core problem addressed in this thesis is the challenge of accurately predicting the FPs of petroleum products using limited labeled data. FP, the lowest temperature at which a substance can vaporize to form an ignitable mixture in air, is a critical safety parameter in the oil industry. Ensuring accurate and timely FP predictions is essential for preventing accidents during the transportation and storage of petroleum products.

The importance of this problem is underscored by the potential hazards associated with inaccurate FP predictions. Incorrect assessments can lead to improper handling of flammable materials, increasing the risk of fires and explosions. Moreover, traditional laboratory methods for determining FPs are not only expensive but also slow, making them unsuitable for real-time process control.

Academically, this problem presents a significant challenge due to the complexities involved in modeling the physical and chemical properties of petroleum products. Industrially, the oil industry requires reliable and quick methods for predicting FPs to ensure safety and compliance with regulations. Existing methods, including experimental approaches and basic machine learning (ML) models, often fall short in terms of accuracy and efficiency when labeled data is scarce.

This thesis positions itself within the broader context of improving safety protocols in the oil industry by enhancing predictive models through advanced ML techniques. The SSL approach proposed here addresses the limitations of current methods by effectively utilizing both labeled and unlabeled data. This research aims to fill the gap in the literature by providing a robust methodology for FP prediction that balances accuracy with efficiency.

1.3 Structure of the Thesis

The structure of this thesis is designed to systematically address the research questions and objectives outlined above. It begins with a comprehensive review of related works in the field of ML and SSL, particularly focusing on their applications in industrial contexts.

The introductory chapter provides a detailed overview of the thesis's purpose, problem definition, and significance. The subsequent chapter on related works delves into existing methodologies and their limitations, setting the stage for the proposed approach.

The methodology chapter outlines the data preprocessing techniques and the implementation of the SSL model. This includes detailed descriptions of the dataset, the preprocessing steps such as handling outliers and feature scaling, and the specific models used (Gaussian Process Regressor (GPR) and Random Forest (RF)).

Following this, the experimental results chapter presents the findings from applying the proposed model to the dataset. This section compares the performance of the baseline model (BM) with the SSL approach, using metrics such as mean absolute error (MAE) and root mean square error (RMSE).

The conclusion chapter summarizes the results, discusses the effectiveness of the SSL techniques, and provides recommendations for future research. It highlights the contributions of the thesis to both academic literature and industrial practice.

By following this structured approach, the thesis aims to provide a clear and comprehensive examination of the use of SSL for FP prediction in the oil industry, demonstrating its potential to significantly improve safety and efficiency.

2. RELATED WORKS

The field of ML, particularly SSL, has seen significant advancements in recent years, providing innovative solutions to various complex problems. This section provides an overview of the key contributions and methodologies in ML and SSL, specifically focusing on their applications in predicting FPs in the oil industry. By examining these related works, this thesis aims to contextualize the proposed approach within the broader landscape of existing research and highlight the unique contributions of this study.

ML has become an integral part of various industrial applications, offering powerful tools to analyze and predict complex phenomena. Within ML, SSL stands out for its ability to leverage both labeled and unlabeled data, making it particularly useful in scenarios where obtaining labeled data is costly or time-consuming.

SSL is a hybrid approach that combines labeled and unlabeled data to improve learning accuracy. Unlike supervised learning, which relies solely on labeled data, SSL utilizes the inherent structure of unlabeled data to enhance model performance. The theoretical foundations of SSL are well-documented in the literature, with significant contributions from Zhu and Goldberg [1]. These studies provide comprehensive insights into various SSL algorithms and their potential to address the limitations of purely supervised methods.

The primary advantage of SSL lies in its ability to improve model accuracy with limited labeled data. This is particularly beneficial in fields like the oil industry, where obtaining labeled data can be challenging. In other contexts, SSL methods have shown to enhance model performance significantly by leveraging pseudo-labels generated from unlabeled data. For instance, these methods have been applied to seismic data classification and oil reservoir identification, showing promising results [2], and well overflow prediction in oil drilling [3]. Additionally, SSL techniques are known to

reduce the risk of overfitting, as they utilize a larger dataset for training, which helps in capturing the underlying data distribution more effectively.

FP prediction is a critical task in the oil industry, crucial for ensuring the safe handling and transportation of petroleum products. Traditional methods for determining FPs involve laboratory tests, which are often time-consuming and costly. Recent advancements in ML have opened new avenues for predicting FPs more efficiently.

Traditional methods for predicting FPs primarily involve experimental techniques and a variety of equation-based models, including empirical, ANN, and vapor-pressure-based models [4]. These methods have been extensively used due to their reliability and established protocols. However, they also come with significant limitations, particularly in terms of cost, time efficiency, and their suitability for real-time process control.

The most commonly used traditional method involves laboratory tests, where the FP is determined experimentally by gradually heating a sample and observing the temperature at which it ignites. This method, detailed by Alqaheem and Riazi [5], provides accurate results but is inherently time-consuming and requires specialized equipment. Additionally, these tests are not feasible for continuous monitoring or real-time applications.

Another traditional approach is the use of empirical correlations and equation-based methods. These methods, as discussed by Liu and Liu [6], involve using established equations to predict FPs based on the physical and chemical properties of the substance. While these methods can be quicker than laboratory tests, they often lack the accuracy required for critical safety assessments due to their reliance on approximations and assumptions.

Quantitative structure-property relationships (QSPR) represent another traditional method used for FP prediction. This approach, highlighted by Pan et al. [7], involves developing predictive models based on the molecular structure of the compounds. Although QSPR methods can provide reasonable estimates, they require extensive data on molecular structures and often involve complex calculations.

Furthermore, Gharagheizi et al. [8] introduced a group contribution method combined with neural networks for estimating FP temperatures of pure components. This method attempts to address some of the limitations of purely empirical approaches by incorporating ML techniques. However, it still depends heavily on the availability of detailed molecular data and may not be applicable to mixtures or complex substances.

Despite their widespread use, these traditional methods have several drawbacks. Laboratory tests, while accurate, are impractical for large-scale or real-time applications. Equation-based and QSPR methods, on the other hand, may lack the necessary precision for ensuring safety in all scenarios. These limitations underscore the need for more efficient and accurate methods for FP prediction.

In summary, while traditional methods have provided a foundation for FP prediction, they are often limited by their time, cost, and applicability constraints. This research aims to address these limitations by leveraging modern ML techniques, particularly SSL, to enhance the accuracy and efficiency of FP predictions in the oil industry.

Modern approaches to FP prediction leverage the advancements in ML to provide more accurate and efficient solutions. These methods have the potential to overcome the limitations of traditional techniques by utilizing large datasets and sophisticated algorithms to predict FPs in real-time.

Recent studies have demonstrated the effectiveness of various ML techniques in predicting FPs. For instance, Mirshahvalad et al. [9] used neural networks to predict the FPs of chemical compounds, achieving high accuracy and demonstrating the potential of ML in this field. Similarly, Mendia et al. [10] developed an adaptive soft sensor based on ML to infer FPs in real-time refinery processes. These studies highlight the capability of ML to handle complex datasets and provide reliable predictions.

One significant advantage of using ML approaches is their ability to continuously learn and improve from new data. This is particularly beneficial in industrial settings where conditions and data can change rapidly. For example, Ghorayeb et al. [11] integrated deep learning models with reservoir simulators to enhance the accuracy of

flash calculations. This integration allows for dynamic adjustments based on real-time data, improving the overall efficiency and safety of industrial operations.

Moreover, ML models can effectively handle the high-dimensional data typically associated with industrial applications. Koyanbayev et al. [12] employed ML techniques to assist in flash calculations for sour gas and crude oil, demonstrating significant improvements in computational efficiency and accuracy. These models can process large volumes of data and identify patterns that are not apparent through traditional methods, leading to more accurate predictions.

Despite the promising results, there are still challenges associated with the application of ML in FP prediction. One major issue is the need for large amounts of labeled data to train the models. However, SSL techniques offer a solution by utilizing both labeled and unlabeled data, as discussed in this thesis. By combining the strengths of ML and SSL, it is possible to achieve high accuracy even with limited labeled data.

In summary, modern ML approaches provide powerful tools for FP prediction, offering significant improvements over traditional methods. These techniques not only enhance accuracy but also enable real-time predictions, which are crucial for maintaining safety and efficiency in the oil industry. The integration of SSL further enhances these capabilities by effectively leveraging both labeled and unlabeled data, addressing one of the primary challenges in the field.

While significant advancements have been made in the field of FP prediction using ML, several gaps remain that need to be addressed to enhance the accuracy and applicability of these methods. Identifying these gaps is crucial for positioning the current research within the broader context of existing studies and highlighting its contributions.

One major gap in the existing literature is the limited focus on the integration of multiple ML approaches. Most studies tend to concentrate on individual algorithms, such as neural networks or regression models, without exploring the potential benefits of combining different techniques. For example, Mirshahvalad et al. [9] focused on neural networks for predicting chemical FPs, and Mendia et al. [10] utilized adaptive soft sensors for real-time inference. However, a comprehensive model that integrates various ML techniques to leverage their combined strengths is still underexplored.

Another significant gap is the lack of application of SSL methods in FP prediction. While SSL has shown promise in other fields by effectively utilizing both labeled and unlabeled data, its application in the context of FP prediction is yet to be explored. Studies like those mentioned have demonstrated the potential of SSL in enhancing model performance with limited labeled data, but similar approaches are not yet widespread in the oil industry.

Additionally, there is a need for more research addressing the real-time application of ML models in industrial settings. Many existing studies, such as those by Ghorayeb et al. [11] and Koyanbayev et al. [12], focus on offline predictions without considering the dynamic requirements of real-time process control. Real-time predictions are crucial for operational efficiency and safety, yet this aspect is often not fully addressed in current research.

Furthermore, the existing literature often does not account for the diversity and complexity of industrial datasets. Many models are developed and tested on relatively small and homogeneous datasets, which may not fully capture the variability present in real-world industrial data. This limitation highlights the necessity of working with new and diverse datasets to enhance model generalizability and applicability. For instance, Liu and Liu [6] and Alqaheem and Riazi [5] discuss predictive methods based on simplified datasets, which may not represent the full spectrum of industrial conditions.

This thesis aims to address these gaps by presenting a robust SSL model that integrates multiple ML techniques and applies them to real-time FP prediction using a comprehensive industrial dataset. By leveraging both labeled and unlabeled data, this research enhances prediction accuracy and operational efficiency, bridging the gap between academic research and industrial application.

In summary, while existing literature provides a foundation for FP prediction using ML, significant gaps remain in the integration of multiple approaches, the application of SSL, real-time implementation, and the handling of diverse industrial datasets. This thesis addresses these issues by developing a comprehensive SSL-based model tailored to the needs of the oil industry, thus advancing both the academic understanding and practical application of ML in this field.

3. METHOD AND EXPERIMENTAL RESULTS

The methodology of this thesis involves a systematic approach to data preprocessing, model development, and performance evaluation to enhance the prediction of FPs using SSL techniques. This section outlines the detailed steps taken to prepare the dataset, implement the SSL model, and evaluate its performance against BM. By adopting a structured methodology, this research aims to address the challenges associated with limited labeled data and demonstrate the effectiveness of SSL in an industrial context.

To begin with, the data preprocessing stage is crucial for ensuring the quality and consistency of the dataset used for model training and evaluation. This involves handling outliers, scaling features, and preparing the data for SSL. Following data preprocessing, the implementation of the SSL model, specifically the GPR combined with RF, is detailed. This model leverages both labeled and unlabeled data to improve prediction accuracy.

The next section focuses on the experimental settings and the performance evaluation of the proposed model. This includes the description of the dataset, the preprocessing techniques applied, the model training process, and the evaluation metrics used to assess the model's performance. By comparing the results of the SSL model with BM, this study aims to highlight the improvements achieved through the proposed approach.

This section is organized as follows: First, the dataset and its features are described, providing an overview of the data sources and key variables. Next, the data preprocessing steps, including handling outliers and feature scaling, are detailed. Following this, the implementation of the SSL model, including the use of GPR and RF, is explained. Finally, the experimental results are presented, comparing the

performance of the BM and the SSL approach using metrics such as mean absolute error (MAE) and root mean square error (RMSE).

3.1 Methodology

3.1.1 Data preprocessing

Data preprocessing is a crucial step in any ML project as it prepares the raw data for model training. This step involves cleaning, transforming, and organizing the data to enhance its quality and ensure that the ML models can learn effectively. Proper data preprocessing can significantly improve the accuracy and performance of the models. In this section, we will discuss two essential preprocessing techniques: handling outliers and feature scaling.

3.1.1.1 Handling outliers

Outliers are data points that differ significantly from other observations in the dataset. They can arise from various reasons, such as measurement errors, data entry mistakes, or genuine variability in the data. Outliers can skew and mislead the training process of ML models, leading to poor performance and inaccurate predictions.

There are several common methods for handling outliers. These include:

- **Removal**: Eliminating outliers from the dataset entirely. This method is straightforward but can result in the loss of valuable data, which is particularly problematic when data is limited.
- **Transformation**: Applying mathematical transformations to reduce the impact of outliers. Techniques such as logarithmic transformation or square root transformation are often used.
- **Capping**: Limiting the values of outliers to a specified range. This method retains all data points while reducing the influence of extreme values.

When data is limited, removing outliers might not be the best approach, as it reduces the amount of data available for training. Instead, capping methods like Winsorization

are more appropriate. Winsorization involves replacing the extreme values with the nearest values within a specified percentile range.

1. Calculate the lower and upper percentiles of the data distribution for each feature.
2. Replace values below the lower percentile with the lower percentile value.
3. Replace values above the upper percentile with the upper percentile value.

This technique minimizes the influence of extreme values while preserving the overall structure of the data, ensuring that the ML model can learn more effectively from the dataset.

3.1.1.2 Feature scaling

Feature scaling is another vital preprocessing step that ensures all features contribute equally to the model training process. Raw data often contains features with varying units and scales, which can adversely affect the performance of the model. Scaling the features brings them to a common scale, making the training process more efficient and improving model performance.

Several common scaling techniques are used in data preprocessing, including:

- **Standardization**: This method scales the data so that it has a mean of zero and a standard deviation of one. It is useful when the features follow a Gaussian distribution.
- **Min-Max Scaling**: This method scales the data to a fixed range, usually between 0 and 1. It is widely used because of its simplicity and effectiveness.

In this study, Min-Max scaling is applied to normalize the features. The Min-Max scaling formula is:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3.1}$$

where x is the original value, $\min(x)$ is the minimum value of the feature, and $\max(x)$ is the maximum value of the feature. This transformation scales the features to a range between 0 and 1, ensuring that all features are on the same scale.

The steps to apply Min-Max scaling are:

1. Calculate the minimum and maximum values for each feature.
2. Apply the Min-Max scaling formula to transform the feature values to the range [0, 1].

By scaling the features, we ensure that no single feature dominates the learning process due to its scale, leading to a more balanced and effective model training.

Through these preprocessing steps—handling outliers and feature scaling—the dataset is prepared for optimal performance of the ML models. These steps are crucial for improving the quality of the data, thereby enhancing the accuracy and reliability of the predictions made by the models in this study.

3.1.2 Data split

Data splitting is a fundamental step in developing robust and reliable ML models. It involves dividing the available dataset into distinct subsets to evaluate the model's performance accurately. Commonly, data is split into training and testing sets, where the training set is used to build the model and the testing set is used to validate its predictive capabilities. This approach ensures that the model can generalize well to new, unseen data, which is crucial for its application in real-world scenarios.

Traditional data splitting methods are commonly used for static datasets where the temporal order of observations is not a primary concern. These methods include *holdout validation*, *k-fold cross-validation*, and *stratified sampling*, each with its theoretical foundation and specific applications.

Traditional data splitting methods, such as holdout validation, involve partitioning the dataset X into two subsets: the training set X_{Train} and the testing set X_{Test}. Mathematically, if X contains N data points, then X_{Train} and D_{Test} are disjoint subsets

such that $X_{\text{Train}} \cup D_{\text{Test}} = X$ and $D_{\text{Train}} \cap X_{\text{Test}} = \emptyset$. Typically, X_{Train} constitutes $70\% - 80\%$ of the data, and X_{Test} constitutes the remaining $20\% - 30\%$.

Another widely used method is *k-fold cross-validation*, where the dataset X is divided into k approximately equal-sized folds. The model is trained k times, each time using $k-1$ folds for training and the remaining fold for testing. Formally, let $X = \{X_1, X_2, \ldots, X_k\}$ where X_i represents the i-th fold. The training set for the i-th iteration is given by $X_{\text{train}}^{(i)} = \bigcup_{j \neq i} X_j$, and the validation set is $X_{\text{val}}^{(i)} = X_i$. The overall performance is then averaged across all k iterations, providing a robust estimate of the model's generalization ability.

Stratified sampling is another method, particularly useful when dealing with imbalanced datasets. It ensures that each subset maintains the same class distribution as the original dataset. Let $\mathscr{C} = \{C_1, C_2, \ldots, C_m\}$ be the set of classes in D. Stratified sampling ensures that for each class C_i, the proportion $\frac{|C_i \cap X_{\text{Train}}|}{|X_{\text{Train}}|} = \frac{|C_i \cap X|}{|X|}$.

Mathematically, the holdout method can be represented as:

$$X_{\text{Train}}, X_{\text{Test}} \sim \text{split}(X, \text{ratio})$$

For k-fold cross-validation, the process can be represented as shown in Equation (3.2):

$$X_{\text{Train}}^{(i)} = \bigcup_{j \neq i} X_j, \quad X_{\text{val}}^{(i)} = X_i$$

$$\text{Performance} = \frac{1}{k} \sum_{i=1}^{k} \text{evaluate}(X_{\text{Train}}^{(i)}, X_{\text{val}}^{(i)}) \tag{3.2}$$

And for stratified sampling, the process can be represented as shown in Equation (3.3):

$$X_{\text{Train}}, X_{\text{Test}} \sim \text{stratified\_split}(X, \text{ratio})$$

$$\frac{|C_i \cap X_{\text{Train}}|}{|X_{\text{Train}}|} = \frac{|C_i \cap X|}{|X|}, \quad \forall C_i \in \mathscr{C} \tag{3.3}$$

These traditional methods, while effective for many applications, may fall short when dealing with time-series data due to their disregard for temporal dependencies. In contrast, the Expanding Window method preserves the chronological order even though the , making it more suitable for time-series analysis and providing a more realistic evaluation of the model's performance over time.

In our study, the Expanding Window approach is adopted due to the limited availability of labeled data, dynamic environmental conditions, and the data's ability to exhibit temporal characteristics, albeit weakly.

Expanding Window (EW) is a technique used in time series analysis and ML to incrementally increase the size of the training dataset over time. This method is particularly useful when dealing with sequentially ordered data, such as time series data, where past observations are used to predict future values. The fundamental principle of EW is to start with an initial window of data and progressively expand this window by including new data points as they become available. This approach ensures that the model is continually updated with the most recent data, allowing it to adapt to new patterns and trends.

Mathematically, let $\{x_1, x_2, \ldots, x_t\}$ represent a time series of data points. The initial window of size w_0 can be defined as follows (see Equation 3.4):

$$W_0 = \{x_1, x_2, \ldots, x_{w_0}\} \tag{3.4}$$

As new data points $\{x_{w_0+1}, x_{w_0+2}, \ldots, x_{w_0+s}\}$ become available, the window expands to include these points. The window at step k can be expressed as follows (see Equation 3.5):

$$W_k = \{x_1, x_2, \ldots, x_{w_0+ks}\} \tag{3.5}$$

where s is the step size, representing the number of new data points added to the window at each step. This incremental expansion continues as more data points are collected, resulting in a dynamically growing training set (see Equation 3.6):

$$W_{k+1} = W_k \cup \{x_{w_0+ks+1}, \ldots, x_{w_0+(k+1)s}\} \tag{3.6}$$

The key parameters in an expanding window approach include the initial window size (w_0) and the step size (s). The initial window size determines the starting point of the training data, while the step size defines the number of new data points added to the window at each step. Proper selection of these parameters is crucial for balancing the trade-off between model accuracy and computational efficiency.

In this study, the expanding window technique is utilized to enhance the predictive performance of the model by incorporating the latest available data in the training

process. This approach is particularly effective in scenarios where the underlying data distribution may change over time, necessitating continuous model updates. By doing so, the model is able to better capture evolving patterns and trends, thereby improving its predictive accuracy and robustness.

3.1.3 Random forest

The RF model is a widely-used, versatile, and powerful supervised learning algorithm in ML. Developed by Breiman, this model consists of multiple decision trees combined to form a forest, aiming to improve the overall performance of the predictive model [13]. RF can be effectively applied to both classification and regression problems, making it a valuable tool for various data science applications. Its ability to handle a large number of input variables without overfitting and to provide estimates of feature importance makes it particularly useful in complex datasets.

The fundamental working principles of RF are rooted in the concept of ensemble learning, where multiple models are trained and their predictions are aggregated to achieve better performance than individual models. Specifically, RF constructs a multitude of decision trees during training time and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Formally, given a training set $X = \{x_1, x_2, \ldots, x_n\}$ with corresponding labels $Y = \{y_1, y_2, \ldots, y_n\}$, RF builds B decision trees $\{T_1, T_2, \ldots, T_B\}$. The prediction of the RF for a new instance x' is obtained by aggregating the predictions from all individual trees:

- For regression, $\hat{y}$ averages B trees (Eq. 3.7):

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x') \tag{3.7}$$

- For classification, $\hat{y}$ is the majority vote of B trees (Eq. 3.8):

$$\hat{y} = \text{mode}\{T_1(x'), T_2(x'), \ldots, T_B(x')\} \tag{3.8}$$

The key advantage of the RF model lies in its robustness and accuracy. By averaging the results from numerous decision trees, RF reduces the risk of overfitting, which is

a common problem in decision tree models. This ensemble approach also enhances the model's generalization capabilities. Furthermore, RF can handle large datasets with higher dimensionality and missing values efficiently. It provides insights into feature importance, aiding in feature selection and understanding the underlying data structure. RF models are extensively used in various domains, including finance for credit scoring, healthcare for disease prediction, and environmental science for predicting climate change impacts.

3.1.4 Gaussian process regressor

GPR is a powerful and flexible regression model that leverages the properties of Gaussian processes to provide probabilistic predictions. GPR is particularly useful in scenarios where quantifying uncertainty is crucial. It models the distribution of possible functions that fit the data, offering a measure of confidence in the predictions.

The fundamental working principle of GPR is based on the assumption that the data can be represented as a sample from a multivariate Gaussian distribution. For a given set of training data (X,Y), where $X = \{x_1, x_2, \ldots, x_n\}$ are the inputs and $Y = \{y_1, y_2, \ldots, y_n\}$ are the corresponding outputs, the goal is to predict the output y' for a new input x'.

The prediction in GPR involves calculating the posterior distribution over the possible values of y', given the training data and the new input. This is achieved by defining a kernel function $k(x,x')$, which measures the similarity between different inputs. Commonly used kernels include the Radial Basis Function (RBF) and the Matern kernel.

The predictive distribution is given by Eq. 3.9:

$$\begin{aligned} \hat{y}(x') &= k(x',X)[K(X,X) + \sigma_n^2 I]^{-1} Y, \\ \text{Var}(x') &= k(x',x') - k(x',X)[K(X,X) + \sigma_n^2 I]^{-1} k(X,x'), \end{aligned} \tag{3.9}$$

where $K(X,X)$ is the covariance matrix computed using the kernel function, and σ_n^2 is the noise variance. This provides both the predicted mean $\hat{y}(x')$ and the variance $\text{Var}(x')$, offering a complete probabilistic view of the predictions.

Key hyperparameters in GPR

Optimizing the performance of a Gaussian Process Regressor (GPR) involves careful selection and tuning of several key hyperparameters. These hyperparameters control various aspects of the model, such as its flexibility, smoothness, and ability to generalize from the training data. By understanding and properly tuning these hyperparameters, we can significantly enhance the model's predictive accuracy and reliability.

- **Length scale** (*LS*): The length scale parameter determines the smoothness of the function that the Gaussian process models. A small length scale means that the function can change rapidly, while a large length scale implies a smoother function. In the RBF kernel, the length scale is a critical parameter that controls the distance over which the correlations between points are significant.

- **Length scale boundaries (LSB)**: The length scale boundaries define the range within which the length scale parameter can vary. These boundaries are essential for setting reasonable limits during the hyperparameter optimization process, ensuring the model does not overfit or underfit the data by choosing an excessively small or large length scale.

- **Signal variance** (σ_f^2): The signal variance controls the vertical variation of the function. It determines the amplitude of the variations in the function values. A higher signal variance allows the function to vary more significantly, capturing more of the underlying patterns in the data.

- **Noise level** (σ_n^2): The noise level represents the variance of the Gaussian noise added to the observations. It captures the inherent noise in the data and helps in regularizing the model to avoid overfitting. By correctly estimating the noise level, the model can better distinguish between the underlying signal and the noise in the data.

- **Alpha** (α): In some implementations, alpha represents a value added to the diagonal of the kernel matrix during fitting. It acts as a regularization term to ensure numerical stability and to control the smoothness of the model. A larger alpha can

lead to a smoother model, while a smaller alpha can make the model more sensitive to the training data.

- **Smoothness parameter** (ν): In the Matern kernel, the smoothness parameter ν controls the smoothness of the resulting function. Different values of ν correspond to different degrees of differentiability of the function. For instance, $\nu = 1/2$ results in an exponential kernel, while higher values of ν produce smoother functions. Choosing an appropriate ν is crucial for capturing the desired level of smoothness in the data.

3.1.5 Semi-supervised learning

SSL is an advanced ML technique that leverages both labeled and unlabeled data to improve model performance. This approach is particularly beneficial when labeled data is scarce or expensive to obtain, a common scenario in many real-world applications. SSL aims to utilize the vast amount of unlabeled data available to enhance learning accuracy and model generalization, bridging the gap between supervised and unsupervised learning methods.

The core idea behind SSL is to use the structure of the unlabeled data to inform and refine the learning process, effectively combining the strengths of supervised and unsupervised learning. SSL can be particularly advantageous in scenarios where acquiring labeled data is costly or time-consuming, but a large amount of unlabeled data is readily available.

In SSL, the learning process can be formally described as follows: Given a small set of labeled data $L = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$ and a large set of unlabeled data $U = \{x_{l+1}, x_{l+2}, \ldots, x_{l+u}\}$, the goal is to build a model f that can predict the labels for new instances.

The SSL algorithm typically involves the following steps:

1. Initialization: Train an initial model using the labeled data L.

2. Label Propagation: Use the model to predict labels for the unlabeled data U. These predictions can be treated as pseudo-labels.

3. Refinement: Retrain the model using both the original labeled data L and the pseudo-labeled data U.

4. Iteration: Repeat the label propagation and refinement steps until convergence or a predefined stopping criterion is met.

A common SSL approach is self-training, where the model iteratively improves by incorporating its own predictions as additional training data. Another popular method is co-training, which involves training multiple models on different subsets of features and using each model's predictions to iteratively refine the other models.

Mathematically, the optimization objective in SSL can be represented as a combination of supervised and unsupervised learning objectives:

$$L_{SSL} = L_{supervised}(L) + \lambda L_{unsupervised}(U)$$

where $L_{supervised}(L)$ is the loss function computed on the labeled data, $L_{unsupervised}(U)$ is the loss function computed on the unlabeled data. The unsupervised loss function is often designed to enforce smoothness or consistency across the data, meaning that data points that are close to each other in the input space should have similar predictions. This is typically achieved through techniques such as consistency regularization, where the model's predictions on unlabeled data are encouraged to be invariant to small perturbations, or through cluster assumption-based methods, where the model is encouraged to place decision boundaries in low-density regions. The parameter λ is a regularization factor that balances the contribution of the supervised and unsupervised objectives.

The advantages of SSL include improved learning accuracy with limited labeled data, better generalization by leveraging the structure of the data, and cost efficiency by reducing the need for extensive labeling. SSL is widely used in applications such as natural language processing, image recognition, and, as discussed in this thesis, predicting FPs in the oil industry.

To illustrate the SSL process, Figure 3.1 depicts a typical workflow of a SSL model:

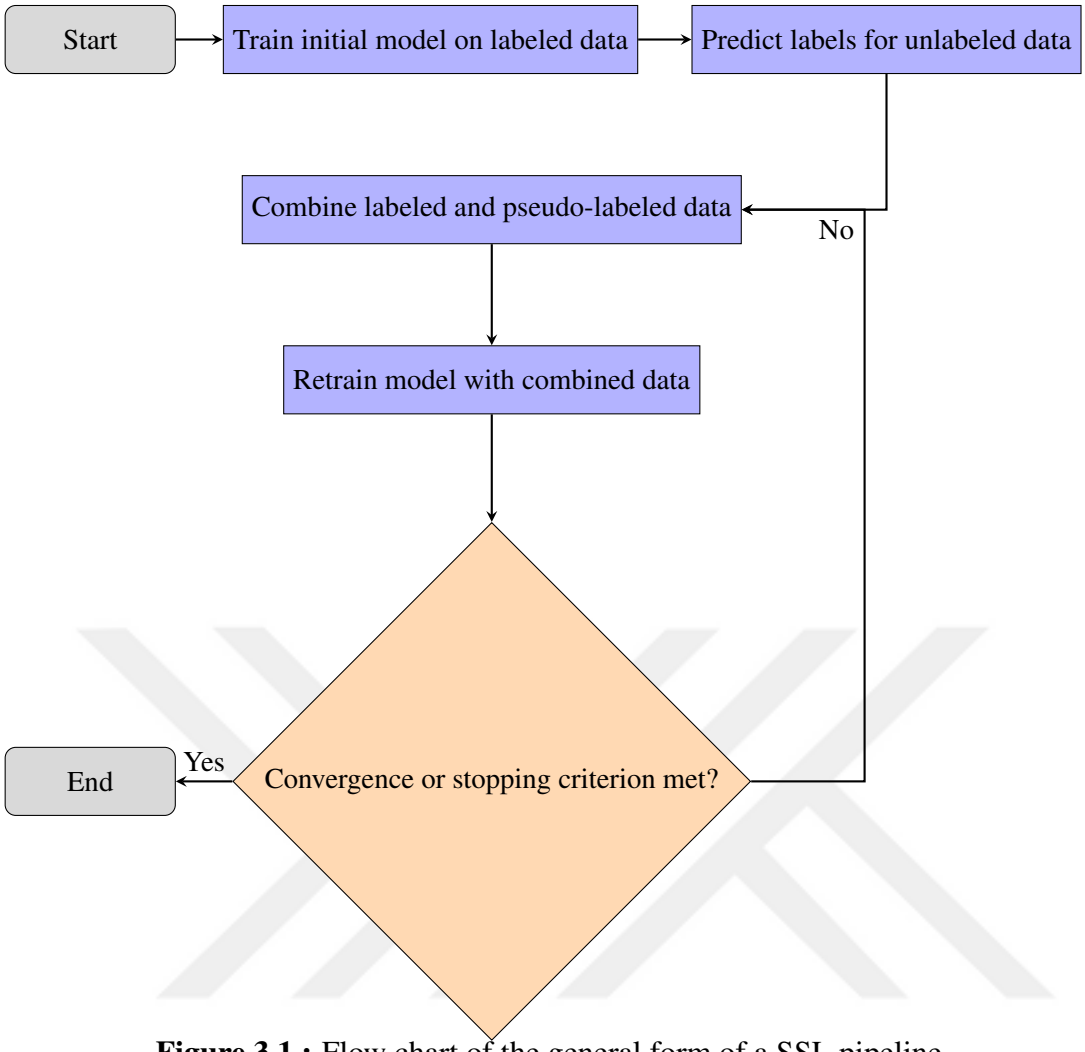

Figure 3.1 : Flow chart of the general form of a SSL pipeline.

In the context of FP prediction, SSL is used to enhance the predictive accuracy by utilizing both the limited labeled sensor data and the abundant unlabeled sensor data. This approach allows for more accurate and efficient predictions, contributing to improved safety and operational efficiency in the oil industry.

3.1.6 Baseline model

In this section, the methodology for the BM is detailed, including the methods used and the training process of the model. The BM is established to serve as a reference for evaluating the effectiveness of the SSL model. As illustrated in Figure 3.2, the process begins with data collection and preprocessing. The BM model is first trained using this preprocessed data, followed by an expanding window technique to refine predictions. Finally, the model's performance is evaluated using a score evaluation method.

The BM utilizes the RF Regressor algorithm. RF is chosen due to its robustness and ability to handle complex relationships in the data without requiring extensive parameter tuning.

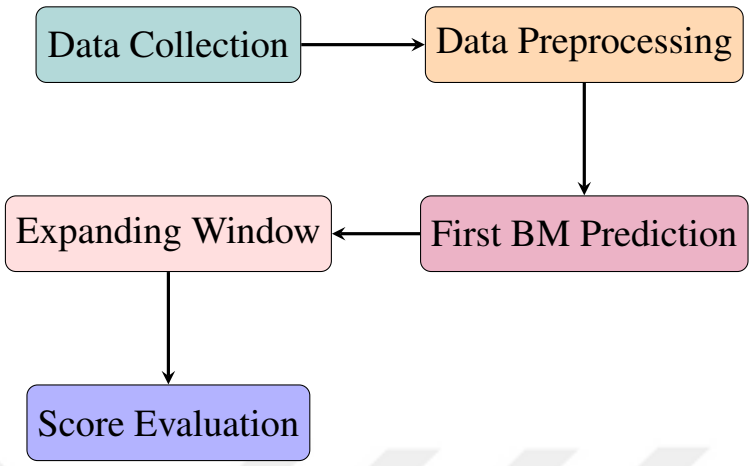

Figure 3.2 : Flow chart of the data processing pipeline for the BM.

The entire dataset, denoted as X, contains both labeled and unlabeled data. The labeled portion, L, consists of l data points $(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)$, and the test portion consists of N data points separated from L. The splitting of the data is done as follows:

- **Training data (Initial):** The initial training set is $X_{\text{train}} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{l-N}, y_{l-N})\}$.
- **Test data:** The test set is $X_{\text{test}} = \{(x_{l-N+1}, y_{l-N+1}), (x_{l-N+2}, y_{l-N+2}), \ldots, (x_l, y_l)\}$.

The choice of N is critical and is based on several considerations:

- **Dataset size:** The total number of labeled data points l determines how much data can be allocated for training and testing. A typical choice might be to use around 20-30% of the labeled data for testing, depending on the dataset size and the specific problem domain.
- **Model performance:** To ensure that the model is adequately tested, N should be large enough to provide a reliable estimate of the model's performance but small enough to leave sufficient data for training.

- **Cross-validation considerations:** If cross-validation is used, N can be chosen based on the number of folds, ensuring each fold has a representative sample of the data.

To address the limited labeled data, the Expanding Window approach is employed, which iteratively incorporates new data points into the training set as follows:

1. **Initial training:**
    - Train the RFR model on the initial training set X_{train}.
    - Predict the label for x_{l-N+1}.
2. **Iterative training:**
    - Add the predicted data point (x_{l-N+1}, y_{l-N+1}) to the training set.
    - Retrain the RFR model on the updated training set.
    - Predict the label for the next data point x_{l-N+2}.
    - Repeat this process until all N test data points are predicted.

This method allows the model to continuously learn and adapt as new data points are incorporated, simulating a real-time learning scenario.

3.1.7 SSL model

In this study, we utilize a combined approach involving both GPR and RF models to enhance prediction accuracy. The modeling process begins with the GPR model, which provides initial predictions along with uncertainty estimates. These predictions, particularly those with high confidence (low uncertainty), are then used to augment the training dataset for the RF model. As illustrated in Figure 3.3, the data processing pipeline starts with data collection and preprocessing, followed by GPR-based predictions for labeled data. High-confidence predictions are then used to expand the dataset, which is subsequently employed to train the RF model.

The process can be summarized in the following steps:

1. *Initial Training with GPR:* Train the GPR model using the available labeled data X_L. The GPR model predicts labels for the unlabeled data X_U, providing both the predicted values and their associated uncertainties.

2. *Selection of High-Confidence Predictions:* Select the predictions from the GPR model that have uncertainties below a predefined threshold. These high-confidence predictions are treated as pseudo-labels.

3. *Augmenting the Training Set:* Combine the original labeled data X_L with the high-confidence pseudo-labeled data from X_U to create an augmented training set.

4. *Training the RF Model:* Train the RF model using the augmented training set. The RF model benefits from the increased volume of training data, which now includes both the original labels and the confident predictions from the GPR model.

5. *Iterative Refinement:* Optionally, the process can be iterated by using the RF model's predictions to further refine the training set, similar to a self-training approach.

By combining GPR and RF models, we leverage the strengths of both methods: the probabilistic predictions and uncertainty quantification of GPR, and the robustness and flexibility of RF. This hybrid approach aims to improve the overall prediction accuracy, particularly in scenarios with limited labeled data. The iterative refinement process ensures that the model continuously improves as more data becomes available, making it well-suited for dynamic environments such as FP prediction in the oil industry.

The entire dataset, denoted as X, contains both labeled and unlabeled data. The labeled portion, X_L, consists of L data points $(x_1, y_1), (x_2, y_2), \ldots, (x_L, y_L)$, and the test portion consists of N data points separated from X_L.

- **Training data (Initial):** The initial training set is $X_{\text{Train}} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{L-N}, y_{L-N})\}$.
- **Test data:** The test set is $X_{\text{Test}} = \{(x_{L-N+1}, y_{L-N+1}), (x_{L-N+2}, y_{L-N+2}), \ldots, (x_L, y_L)\}$.

To address the limited labeled data, the Expanding Window approach is employed, which iteratively incorporates new data points into the training set as follows:

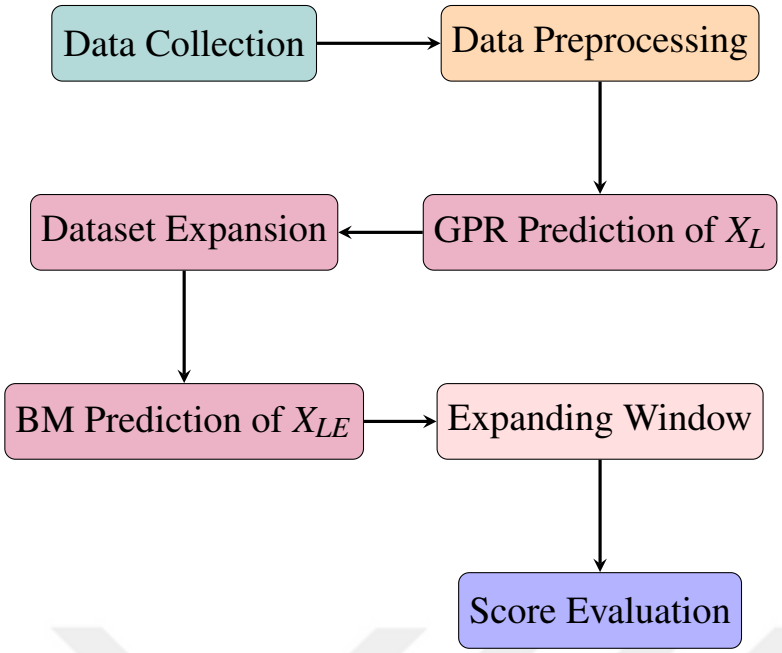

Figure 3.3 : Flow chart of the data processing pipeline for the SSL model.

1. **Initial training:**
   - Train the RFR model on the initial training set X_{Train}.
   - Predict the label for x_{L-N+1}.
2. **Iterative training:**
   - Add the predicted data point (x_{L-N+1}, y_{L-N+1}) to the training set.
   - Retrain the RFR model on the updated training set.
   - Predict the label for the next data point.
   - Repeat this process until all N test data points are predicted.

This method allows the model to continuously learn and adapt as new data points are incorporated, simulating a real-time learning scenario.

3.1.7.1 SSL-pseudocode

This section elaborates on the SSL model, which employs the expanded window technique to enhance predictive accuracy by effectively utilizing both labeled and unlabeled data. The algorithm is structured to adapt and refine its predictive capabilities.

Algorithm overview:

The SSL model described in Algorithm 1 initiates with the division of data into labeled and unlabeled datasets. These datasets undergo a series of preprocessing steps to ensure robust model training and prediction accuracy. The following is a step-by-step breakdown of the process:

Data initialization:

- Total Data Set (X): Represents the entire collection of data points.
- Labeled Data (X_L): Consists of data points that have associated labels, with L representing the number of labeled instances.
- Unlabeled Data (X_U): Comprises data points without labels, with U denoting the number of unlabeled instances.
- Initial Test Set Size (N): Specifies the number of instances in the initial test set.

Data partitioning:

- Training Data (X_{Train}): Obtained by excluding the last N instances from X_L, used initially for training the model.
- Test Data (X_{Test}): Defined as the last N instances from X_L, used for evaluating the model.

Data preprocessing:

- Winsorization: Applied to reduce the influence of outliers. The bounds for Winsorization (winBounds) are determined based on X_{Train}.
- Min-Max Scaling: Normalizes the data to a specific range, typically [0, 1]. Scaling bounds (scaleBounds) are calculated from X_{Train} to ensure that the scaling is appropriately calibrated.

Algorithm 1 SSL Model Using Expanded Window Technique

$X \leftarrow$ All data
$X_L \leftarrow$ Labelled data, $L \leftarrow$ COUNT(X_L)
$X_U \leftarrow$ Unlabelled data, $U \leftarrow$ COUNT(X_U)
$N \leftarrow$ Initial size of test set
$X_{Train} \leftarrow X_L[1 : L-N]$
$X_{Test} \leftarrow X_L[L-N+1 : L]$
$X_{Expanded} \leftarrow$ copy of X_{Train}
winBounds $\leftarrow$ WINSORIZATION(X_{Train})
$X_{Train} \leftarrow$ APPLYWIN$(X_{Train}, \text{winBounds})$
$X_{Test} \leftarrow$ APPLYWIN$(X_{Test}, \text{winBounds})$
$X_U \leftarrow$ APPLYWIN$(X_U, \text{winBounds})$
$X_{Expanded} \leftarrow$ APPLYWIN$(X_{Expanded}, \text{winBounds})$
scaleBounds $\leftarrow$ MINMAXSCALE(X_{Train})
$X_{Train} \leftarrow$ APPLYMM$(X_{Train}, \text{scaleBounds})$
$X_{Test} \leftarrow$ APPLYMM$(X_{Test}, \text{scaleBounds})$
$X_U \leftarrow$ APPLYMM$(X_U, \text{scaleBounds})$
$X_{Expanded} \leftarrow$ APPLYMM$(X_{Expanded}, \text{scaleBounds})$
GPR $\leftarrow$ TRAINGPR(X_{Train})
$X_E \leftarrow \emptyset, E \leftarrow 0$
for all $x_i \in X_U$ **do**
    label, $\sigma \leftarrow$ GPR.PREDICT(x_i)
    if $\sigma <$ Thresh **then**
        Add (x_i, label) to X_E
        $E \leftarrow E + 1$
    end if
end for
$X_{LE} \leftarrow X_{Train} \cup X_E$
$X_{Expanded} \leftarrow X_{LE}$
predictions $\leftarrow$ []
for $i \leftarrow 1$ to LEN(X_{Test}) **do**
    if $i > 1$ **then**
        Add $X_{Test}[i-1]$ to $X_{Expanded}$
    end if
    RF $\leftarrow$ TRAINRF$(X_{Expanded})$
    $p_i \leftarrow$ RF.PREDICT$(X_{Test}[i])$
    predictions.add(p_i)
end for
return predictions

Model training and prediction:

- Gaussian Process Regression (GPR): The GPR model is trained using the preprocessed X_{Train}. It is then used to predict labels and associated uncertainties (sigma) for each instance in X_U.
- Uncertainty Thresholding: If the predicted uncertainty (sigma) for an unlabeled instance is below a predefined threshold (Thresh), the instance, along with its predicted label, is added to the set of confidently labeled data (X_E).

Iterative model refinement:

- Expanded Training Data (X_{LE}): Combines X_{Train} with X_E, forming an augmented training set that incorporates both originally labeled data and newly labeled instances from X_U.
- RF Training: For each test instance in X_{Test}, a RF model is trained on the continuously updated X_{Expanded}, which includes X_{LE} and potentially previous test instances.
- Prediction Collection: Predictions for each test instance are stored in an array or list, progressively building the set of output predictions.

Output: The model returns a list of predictions corresponding to each instance in X_{Test}, offering insights into the model's performance and accuracy.

3.2 Experimental Settings and Performance Evaluation

In this section, we discuss the experimental setup and the methodologies employed to evaluate the performance of the proposed models. We focus on the *BM* and the *SSL* approach, detailing the techniques used to optimize and assess their performance. The evaluation metrics used in this study provide a comprehensive understanding of the model's predictive accuracy and adaptability over time. The following subsections describe the *BM*, the *EW* approach, and the process of model optimization and parameter settings, highlighting their respective roles in enhancing prediction accuracy.

3.2.1 Dataset description

This subsection provides an overview of the dataset used in this study. The dataset consists of FP values obtained through laboratory measurements, alongside pressure, temperature, and flow indicator data collected from sensors. It comprises 52 continuous features over a one-year observation period, totaling 10178 observations. However, FP values measured in the laboratory are available in only 275 of these observations. This dataset serves as the foundation for training and evaluating the BM and SSL models.

3.2.2 Data preprocessing

In this subsection, we discuss the essential data preprocessing techniques applied to prepare the dataset for the BM and SSL models. Key preprocessing steps include handling outliers and feature scaling, ensuring the data's integrity and suitability for model training.

3.2.2.1 Handling outliers

In this study, Winsorization is used to handle outliers for BM and SSL models. The steps are as follows:

1. Calculate the 2nd and 98th percentiles of the data distribution for each feature.
2. Replace values below the 2nd percentile with the 2nd percentile value.
3. Replace values above the 98th percentile with the 98th percentile value.

3.2.2.2 Feature scaling

The features were normalized to a range of 0 to 1 in order to give equal importance to features of varying scales during model training for BM and SSL models.

3.2.3 Baseline model (BM)

The BM serves as a fundamental benchmark to evaluate the performance improvements gained through SSL. In this study, the *RF* algorithm is used for the BM,

leveraging its robustness and effectiveness in handling complex, high-dimensional data.

The BM is built using labeled data, specifically using 235 (L-40) labeled data points. This quantity as "40" was chosen to provide a relatively robust training set while managing computational resources effectively. The RF algorithm was configured to use a maximum of 15 features, which were selected based on their relevance and contribution to the predictive accuracy for FPs. The (*EW*) approach is employed to iteratively update the model with new labeled data. This method allows the model to incorporate both existing test data and any new incoming data. Initially, the model is trained with the available $L-N$ labeled data to predict the 40th data point (*L-40*). After predicting *L-40*, this data point, along with its true label, is added to the training set. The model is then retrained with the updated training set, which includes *L-39* data points, and the 39th data point (*L-39*) is predicted next. This process continues iteratively, with each newly predicted data point being added to the training set until the model is trained on *L-1* data points, where *L* is set to 40, and the final prediction is made for the original *L*th data point.

Formally, the EW approach was implemented with a step size of 1 data point and an initial window size of 40 data points. These parameters were chosen to balance model performance and computational efficiency, ensuring that the model remains updated with the most recent data trends without overfitting.

3.2.4 SSL model

This section details the parameter settings of the semi-supervised learning (SSL) model. Proper parameter tuning is crucial to optimize the performance of the model. Based on the methodology outlined in the paper, the parameter settings for the SSL model are as follows:

Kernel selection and parameters:

In the GPR model, the Radial Basis Function (RBF) kernel is primarily used to measure the similarity between data points, enhancing the model's flexibility with a length scale set to 3.0. Additionally, the Matern kernel, which provides a broader function space, is

employed with a length scale of 10.0 and a smoothness parameter (ν) of 1.5. The alpha parameter (α), controlling the level of observation noise, is set to 0.3. These kernel selections and parameter settings are critical for optimizing the model's performance in this study.

3.2.5 Model optimization, parameter settings and system spesifications

Model optimization is crucial for enhancing the performance of ML models. In this study, several optimization techniques were employed to fine-tune the parameters of both the *RF* and GPR models. For the *RF* model, parameters such as the number of trees, maximum depth, and the number of features considered for splitting were optimized using grid search and cross-validation techniques. For the *GPR* model, kernel parameters like the length scale and noise level were adjusted to improve prediction accuracy.

The optimization process involved iteratively testing different parameter combinations and evaluating the model's performance using metrics such as *Mean Absolute Error (MAE)* and *Root Mean Squared Error (RMSE)* . The optimal parameters were selected based on the lowest error rates achieved during cross-validation. This systematic approach to parameter tuning ensured that the models were well-calibrated to handle the specific characteristics of the dataset, thereby maximizing their predictive accuracy and reliability.

This thesis work was carried out on an Asus N550JV computer with an Intel Core i7-4700HQ processor, 16 GB of RAM, and an NVIDIA GeForce GT 750M graphics card. Python 3.9 was used as the programming language on the Windows 10 operating system. Throughout the study, pandas, numpy, matplotlib, and seaborn were utilized for data processing and analysis, while sklearn and the *sklearn.gaussian_process.kernels* library for kernel functions used in Gaussian Process Regression (GPR) models were employed for machine learning, model evaluation, and optimization. Coding and data analysis were performed using the Spyder development environment. Additionally, the argparse library was used for

managing command-line arguments, and joblib was utilized for parallel processing and model saving.

3.2.6 Experimental results

In this section, we present the results of the experiments conducted to evaluate the performance of the proposed models. The focus is on comparing the *BM* with the *SSL* approach, assessing their effectiveness using various metrics. The results highlight the improvements in prediction accuracy and model adaptability achieved through the implementation of the *SSL* technique.

The following subsections provide a detailed analysis of the performance comparison between the baseline and *SSL* approaches, the evaluation of *MAE* and *RMSE* metrics, and the assessment of the model's adaptability and time performance.

3.2.6.1 MAE and RMSE evaluation

The *MAE* and *RMSE* are crucial metrics for evaluating the accuracy of regression models. This subsection discusses the *MAE* and *RMSE* values obtained from the experiments, providing insights into the models' predictive performance.

The *MAE* and *RMSE* are calculated as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3.10}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{3.11}$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and N is the total number of samples. As it was given before, N referes to 40.

Performance of BM In the evaluation using the BM, the model achieved an MAE score of 1.1 and an RMSE score of 1.7. These results indicate that the model predicts the FPs of petroleum products with an average error of 1.1 degrees. The RMSE being higher than the MAE suggests that while the model generally performs well, it does

encounter larger errors in some cases, indicating occasional significant deviations from the actual values.

Performance of SSL Using the SSL approach, the model's MAE score decreased to 1.01 and the RMSE score dropped to 1.63. This enhancement highlights the potential of leveraging unlabeled data to boost the model's accuracy. Consequently, the model consistently produces more accurate predictions with reduced errors. As illustrated in Figure 3.4, the semi-supervised model effectively captures the test data's pattern. Although the RMSE and MAE scores decreased, the RMSE remained higher compared to MAE score. This pattern is particularly evident in the SSL model, as illustrated by the 24th observation shown in Figure 3.4.

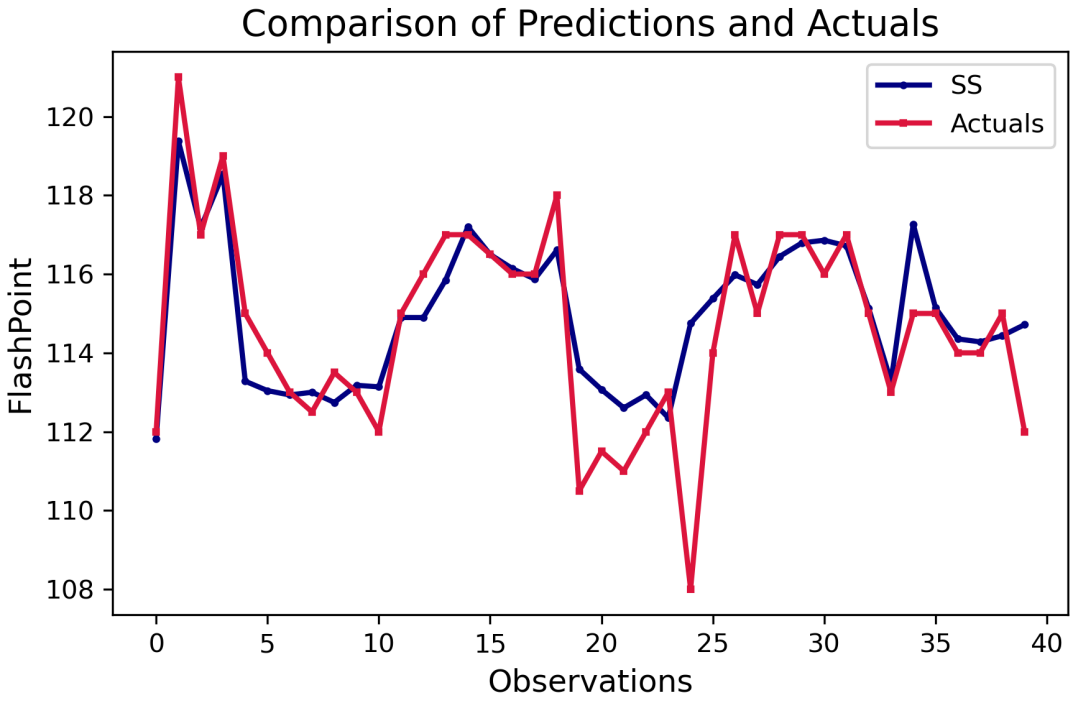

Figure 3.4 : Comparison of the actual values with the predictions for SSL of FP.

Figure 3.4 illustrates the comparison between the actual flash point values and the predictions made by the SSL model. The blue line represents the predictions, while the red line shows the actual values. The close alignment of the two lines indicates the model's accuracy in predicting flash points across various observations.

Figure 3.5 illustrates a histogram showing the distribution of differences between the predicted and actual values, categorized into specific ranges. The x-axis represents the difference ranges, from greater than 2 degrees below to greater than 2 degrees above, while the y-axis shows the count of predictions that fall within each range. The majority of predictions fall within the range of 0-1 degrees below, indicating a general accuracy in the predictions, while fewer outliers exist in the extreme ranges (both below and above 2 degrees). This visualization highlights the precision of the model in predicting values, with a tendency toward smaller errors.

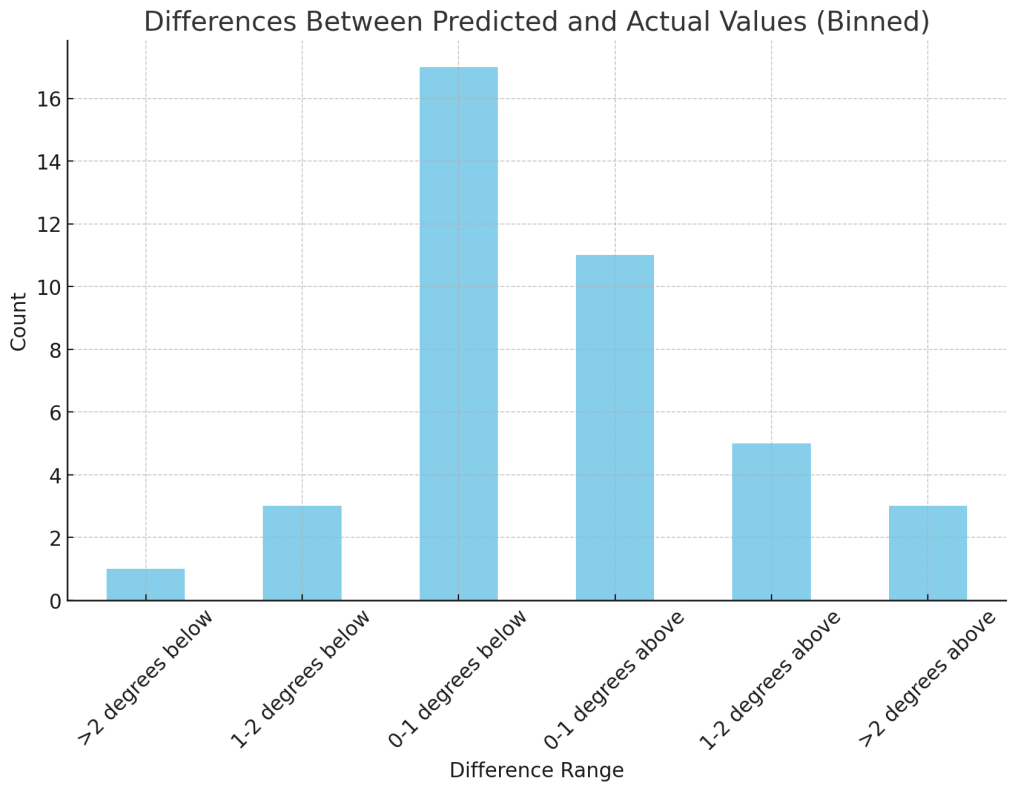

Figure 3.5 : Histogram of Difference Ranges Between Predicted and Actual Values

The results indicate that the *SSL* approach achieves lower *MAE* and *RMSE* values compared to the *BM*, confirming its superior predictive accuracy.

3.2.6.2 Performance comparison of BM and SSL approach

The performance of the *BM* and the *SSL* approach are compared to highlight the advantages of incorporating unlabeled data into the training process. This subsection presents the results of this comparison, emphasizing the improvements in predictive accuracy achieved through the *SSL* approach. The comparison reveals that the *SSL*

approach outperforms the *BM* in terms of prediction accuracy. Specifically, while the improvement in *MAE* and *RMSE* is slight, the *SSL* model still achieves a lower error compared to the *BM*. Given that the topic is as sensitive as flash point prediction, these slight improvements can be significant. These results demonstrate the effectiveness of leveraging both labeled and unlabeled data to enhance model performance.

4. CONCLUSION AND RECOMMENDATIONS

In this section, we summarize the key findings of our study, discuss the effectiveness of the SSL techniques employed, and provide recommendations for future research. This section aims to encapsulate the overall contributions of the research and suggest potential areas for further investigation.

The following subsections provide a detailed summary of the results, evaluate the effectiveness of the SSL techniques used in the study, and offer recommendations for future work.

4.1 Summary of Results

The experimental results demonstrate performance improvements when employing the SSL approach compared to the BM. The SSL model exhibited lower MAE and RMSE values, indicating higher predictive accuracy. These findings highlight the potential of SSL techniques to enhance model performance, particularly in scenarios with limited labeled data.

Moreover, the iterative process of incorporating newly predicted data points back into the training set proved effective in continuously improving the model's accuracy. This approach ensured that the model remained relevant and adaptable to new data, enhancing its overall robustness and reliability.

4.2 Effectiveness of SSL Techniques

The SSL techniques employed in this study have shown to be effective in improving prediction accuracy for flash point prediction in the oil industry. By leveraging both labeled and unlabeled data, the SSL approach effectively mitigates the limitations posed by the scarcity of labeled data.

Compared to traditional methods, the SSL approach provides a more flexible and efficient solution, capable of adapting to new data and improving over time. The use

of techniques such as the Expanding Window (EW) approach further enhances the model's adaptability and ensures continuous learning.

However, the SSL approach also has limitations, such as the need for careful tuning of parameters and the potential computational overhead associated with iterative training. Despite these challenges, the benefits of improved accuracy and adaptability make SSL a valuable technique for predictive modeling in data-scarce environments.

4.3 Recommendations for Future Work

Future research should focus on addressing the limitations identified in this study and exploring additional avenues for enhancing the application of SSL techniques. Specific recommendations include:

- Investigating alternative SSL algorithms and comparing their performance to the approaches used in this study.
- Exploring advanced parameter tuning and optimization techniques to further improve model accuracy and efficiency.
- Expanding the dataset to include a broader range of variables and conditions, which could provide more comprehensive training data and enhance the model's generalizability.
- Implementing real-time data acquisition and model updating mechanisms to fully leverage the benefits of the SSL approach in dynamic industrial environments.
- Conducting cross-industry studies to evaluate the applicability of SSL techniques in other domains with similar data limitations, such as healthcare, finance, and environmental monitoring.

By addressing these areas, future research can build on the findings of this study and contribute to the ongoing development and refinement of SSL techniques for predictive modeling.

REFERENCES

[1] **Zhu, X. and Goldberg, A.B.** (2009). *Introduction to Semi-Supervised Learning*, Synthesis Lectures on Artificial Intelligence and Machine Learning, Springer International Publishing, Cham, https://link.springer.com/10.1007/978-3-031-01548-9.

[2] **Nishitsuji, Y.**, **Exley, R. and Nasseri, J.** (2018). Semi-Supervised Deep-Learning Applied To UK North Sea Well And Seismic Data, https://doi.org/10.3997/2214-4609.201803013.

[3] **Liu, W.**, **Fu, J.**, **Liang, Y.**, **Cao, M. and Han, X.** (2022). A Well-Overflow Prediction Algorithm Based on Semi-Supervised Learning, *Energies*, https://consensus.app/papers/w.

[4] **Phoon, L.Y.**, **Mustaffa, A.A.**, **Hashim, H. and Mat, R.** (2014). A Review of Flash Point Prediction Models for Flammable Liquid Mixtures, *Industrial & Engineering Chemistry Research*, *53*(32), 12553–12565, https://pubs.acs.org/doi/10.1021/ie501233g.

[5] **Alqaheem, S.S. and Riazi, M.R.** Flash Points of Hydrocarbons and Petroleum Products: Prediction and Evaluation of Methods, *31*(4), 3578–3584, https://pubs.acs.org/doi/10.1021/acs.energyfuels.6b02669.

[6] **Liu, X. and Liu, Z.** Research Progress on Flash Point Prediction, *55*(9), 2943–2950, https://pubs.acs.org/doi/10.1021/je1003143.

[7] **Pan, Y.**, **Jiang, J. and Wang, Z.** Quantitative structure–property relationship studies for predicting flash points of alkanes using group bond contribution method with back-propagation neural network, *147*(1), 424–430, https://linkinghub.elsevier.com/retrieve/pii/S0304389407000775.

[8] **Gharagheizi, F.**, **Alamdari, R.F. and Angaji, M.T.** A New Neural NetworkGroup Contribution Method for Estimation of Flash Point Temperature of Pure Components, *22*(3), 1628–1635, https://pubs.acs.org/doi/10.1021/ef700753t.

[9] **Mirshahvalad, H.R.**, **Ghasemiasl, R.**, **Raufi, N. and Malekzadeh Dirin, M.** A Neural Networks Model for Accurate Prediction of the Flash Point of Chemical Compounds, *39*(4), https://doi.org/10.30492/ijcce.2019.35001.

[10] **Mendia, I.**, **Gil-López, S.**, **Landa-Torres, I.**, **Orbe, L. and Maqueda, E.** Machine learning based adaptive soft sensor for flash point inference in a refinery realtime process, *13*, 100362, https://linkinghub.elsevier.com/retrieve/pii/S2590123022000329.

[11] **Ghorayeb, K.**, **Mogensen, K.**, **El Droubi, N.**, **Kloucha, C.K.**, **Ramatullayev, S. and Mustapha, H.** Integration of Deep-Learning-Based Flash Calculation Model to Reservoir Simulator, *Day 3 Wed, November 02, 2022*, SPE, p.D031S073R001, https://onepetro.org/SPEADIP/proceedings/22ADIP/3-22ADIP/D031S073R001/513811.

[12] **Koyanbayev, M.**, **Wang, L. and Zhumatai, A.** Machine Learning Assisted Flash Calculation for Sour Gas and Crude Oil, *Day 1 Tue, January 24, 2023*, SPE, p.D011S001R005, https://onepetro.org/SPERCSC/proceedings/22RCSC/1-22RCSC/D011S001R005/515777.

[13] **Breiman, L.** (2001). Random Forests, *Machine Learning*, *45*(1), 5–32, http://link.springer.com/10.1023/A:1010933404324.

CURRICULUM VITAE

Mert SÜLÜK

EDUCATION:

- **B.Sc.:** 2019, Karadeniz Technical University, Faculty of Engineering
- **M.Sc.:** 2024, Istanbul Technical University, Faculty of Computer and Informatics Engineering, Department of Computer Engineering

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2022-2024 Turkcell Research Assistant in Computer Engineering at Istanbul Technical University
- 2024-Continues Research Assistant in Computer Engineering at Istanbul University

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Sülük M.**, Oguducu SG. (2024). Semi-Supervised Learning for Sensor-Based Flash Point Prediction in Oil Industry. *9th International Conference on Computer Science and Engineering, UBMK 2024*, October 26-28, 2024 Antalya, Turkey.