DATA-DRIVEN MODELING OF WELL PRODUCTIVITY INDEX (PI) THROUGH MACHINE LEARNING ALGORITHMS

BY

BASSEY GRACE UWAKMFON
ENG2002603

DEPARTMENT OF PETROLEUM ENGINEERING
FACULTY OF ENGINEERING
UNIVERSITY OF BENIN,
BENIN CITY

OCTOBER 2025

CERTIFICATION

This is to certify that this project work was carried out by **BASSEY GRACE UWAKMFON (ENG2002603)** of the Department of Petroleum Engineering, University of Benin.

……. …….

DR TAIWO OLUWASEUN

(PROJECT SUPERVISOR) **DATE**

……. ……..

DR TAIWO OLUWASEUN

(PROJECT COORDINATOR) **DATE**

DR. IK OHENHEN **DATE**

(HEAD OF DEPARTMENT)

PROF. **DATE**

(EXTERNAL SUPERVISOR)

DEDICATION

This project is dedicated to God Almighty for His divine direction and favor. To my parents, who never stopped believing in me, and to my friends who supported and motivated me throughout this work , this success is as much yours as it is mine.

ACKNOWLEDGEMENT

I sincerely give all thanks and glory to God for His endless grace, strength, wisdom, and guidance throughout the course of this project and my academic journey in University of Benin. His faithfulness has been my greatest source of strength and motivation to complete this phase.

I am grateful to my supervisor, Dr. O.A. Taiwo, whose encouragement, observation, constructive criticisms and suggestions made this study a success.

My immense gratitude goes to my Parents Mr. Frank Ehima and Mrs. Ifreke Inyang for their unconditional support, care, financial support, also to all my Siblings, learned colleagues and friends for their moral support throughout course of my study.

Last but not least, I want to thank me for believing in me, for doing all this hard work, for having no days off and for never quitting.

I love you all.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The accurate prediction of the Well Productivity Index (PI) is critical for reservoir management, production optimization, and forecasting. Traditional methods, such as analytical correlations and decline curve analysis, are often limited by simplifying assumptions that fail to capture the complexities of heterogeneous reservoirs like those in the Niger Delta. This research addresses this limitation by developing and evaluating a data-driven framework for PI prediction using machine learning (ML) on historical production data. The study implements and compares three advanced ensemble regression algorithms—Random Forest, XGBoost, and CatBoost—to predict PI from daily records of oil, gas, and water production rates and downhole pressures. A dataset of approximately 7,000 daily records from five Niger Delta wells was utilized, with the PI target variable calculated using a proxy for reservoir pressure drawdown. A clear performance hierarchy was established among the models. Random Forest yielded the weakest performance ($R^2 = 0.18$, MAE = 65.62), while XGBoost showed substantial improvement ($R^2 = 0.78$, MAE = 34.14). CatBoost emerged as the superior model, achieving exceptional predictive accuracy with an R^2 of 0.95, a Mean Absolute Error (MAE) of 18.96, and a Root Mean Squared Error (RMSE) of 21.02. Residual and temporal analyses confirmed that CatBoost produced unbiased, homoscedastic errors and effectively tracked the dynamic PI trends of individual wells over time. Interpretability analyses revealed that production rates (oil, gas, and water) were the most influential predictors, a finding consistent with reservoir engineering principles. However, this also highlights a methodological caveat regarding the mathematical coupling between the model's inputs and the PI target. The study concludes that CatBoost provides a robust and highly accurate model for PI prediction from routine field data, offering a significant advantage over traditional methods for well performance monitoring and screening in the Niger Delta context.

CHAPTER ONE

1.0 INTRODUCTION

The productivity index (PI) is a fundamental parameter in reservoir engineering that quantifies well deliverability. It is defined as the volumetric flow rate of fluids (oil, gas, or water) produced per unit pressure drawdown between the average reservoir pressure and the flowing bottom-hole pressure. As such, PI serves as an indicator of how effectively a reservoir and its completion allow fluids to flow into the wellbore.

In practice, PI is often determined empirically from well test data. After measuring the surface flow rate, q, and the pressure differential between the average reservoir pressure and the flowing bottom-hole pressure , PI is computed using:

This operational definition makes PI an essential diagnostic parameter for characterizing well performance and for validating reservoir models.

The importance of PI extends across multiple areas of reservoir and production engineering. It is central to the development of inflow performance relationships (IPRs), which describe the relationship between flow rate and bottom-hole pressure and are widely used in production forecasting and well deliverability analysis. Furthermore, PI is critical in decline curve analysis, reservoir management, and the design of artificial lift and enhanced oil recovery systems. A reduction in PI over time can indicate formation damage, scaling, or rising water cut, while improvement in PI may result from stimulation treatments such as acidizing or hydraulic fracturing.

Given its sensitivity to both reservoir characteristics and wellbore conditions, PI represents an integrative performance metric. The availability of large volumes of historical production and

well-test data in mature basins provides an opportunity to apply data-driven approaches, such as machine learning, to predict PI without relying exclusively on analytical or empirical well-test evaluations. Machine learning models can capture non-linear relationships among petrophysical, completion, and production variables, offering improved accuracy and speed in PI estimation. This opens pathways for real-time monitoring, proactive reservoir management, and optimization of field development strategies.

1.1 Traditional PI Prediction Approaches

Conventional approaches for predicting the productivity index (PI) have historically relied on analytical formulations and empirical correlations derived from known reservoir and well parameters. These methods provide useful first-order estimates of well deliverability but are often constrained by simplifying assumptions that limit their accuracy in complex reservoir settings.

One widely applied category of methods is decline curve analysis (DCA), which uses production history to forecast future performance. The classical Arps-type decline curves are particularly common and are used to extrapolate production rates under assumed reservoir drive mechanisms. However, the validity of DCA depends on the assumption of constant reservoir drive and stable flow conditions. When production is influenced by multiphase flow, water breakthrough, or changes in operating strategy, the predictions can become unreliable.

Another class of methods is based on inflow performance relationships (IPRs). Vogel's equation, originally developed for solution-gas drive reservoirs, and Fetkovich's correlation, which incorporates boundary-dominated flow concepts, are among the most widely used. These IPRs,

however, were largely developed for vertical wells and for single-phase or near-single-phase flow conditions. As such, they often fail to capture the behavior of horizontal or hydraulically stimulated wells where complex flow geometries and multiphase interactions dominate.

In response to the limitations of early IPR models, numerous empirical correlations for horizontal wells have been developed. Notable examples include the formulations by Borisov, Giger–Reiss– Jourdan, Renard–Dupuy, Joshi and Economides, and Butler and Furui. These correlations typically express PI in dimensionless form by incorporating parameters such as permeability, well length, anisotropy ratios, drainage area, and skin factor. While these formulations provide practical tools for well planning and preliminary performance assessment, they often rely on idealized assumptions about reservoir geometry and boundary conditions.

A key shortcoming of such analytical and empirical methods is their inability to account for nonlinear and heterogeneous reservoir behavior. They typically oversimplify complex processes such as multiphase flow interactions, heterogeneity in permeability and porosity, capillary effects, and stress-dependent reservoir properties. As highlighted by Alarifi et al. (2015), even the more advanced horizontal well PI correlations demonstrate limited reliability under field conditions that deviate from their underlying assumptions.

Traditional PI prediction approaches—including decline curve analysis, analytical IPRs, and empirical correlations—are straightforward to apply and computationally inexpensive. However, their predictive performance degrades in reservoirs characterized by heterogeneity, multiphase flow, or evolving operating conditions. Numerical reservoir simulation can partly address these limitations by capturing complex physics, yet it requires extensive input data, high computational effort, and rigorous calibration. These challenges provide strong motivation for the adoption of

data-driven methods, particularly machine learning, which can integrate large volumes of historical production data and capture nonlinear dependencies between reservoir, well, and operating variables.

1.2 Machine Learning for Well Productivity Prediction

In recent years, machine learning (ML) techniques have gained increasing attention in petroleum engineering as viable alternatives to traditional analytical and empirical methods for well performance prediction. Unlike fixed correlations or decline curve models, ML algorithms are capable of learning complex, nonlinear relationships directly from data. This adaptability allows ML models to integrate diverse sets of input features—such as geological parameters (porosity, permeability, thickness), completion parameters (well type, length, stimulation), and operational parameters (flowing pressure, production history)—to predict outcomes such as productivity index (PI) or production rates with higher accuracy.

A key advantage of ML methods lies in their ability to overcome the rigid assumptions inherent in classical approaches. Empirical correlations often assume simplified reservoir geometries and flow regimes, whereas ML models can capture multiphase flow effects, heterogeneity, and nonlinear interactions without requiring explicit physical equations. As Fan et al. (2025) observed, the application of ML and AI approaches has substantially improved forecasting accuracy when compared with traditional decline curve techniques.

Supervised regression models are among the most widely used ML techniques for well productivity prediction. Support vector machines (SVMs), random forests (RF), and gradient boosting methods such as XGBoost and LightGBM have demonstrated strong performance in

steady-state capacity and PI prediction tasks. These models leverage systematic feature selection to identify the most influential variables and apply hyperparameter optimization to refine predictive accuracy. When properly tuned and validated, such models often yield robust estimates with prediction errors within acceptable engineering limits.

Beyond classical supervised models, recent studies have explored the application of artificial neural networks (ANNs) and deep learning frameworks for capturing highly nonlinear dependencies in production data. These methods are particularly effective when large datasets are available, as they can automatically extract latent features from input variables. However, their effectiveness depends strongly on the quality and representativeness of training data as well as strategies to mitigate overfitting.

Overall, ML-based approaches offer a flexible and scalable framework for predicting well productivity. By leveraging historical production datasets and multi-source reservoir information, they enable more accurate and real-time assessments of PI, which can support optimized reservoir management, improved recovery planning, and better-informed field development strategies.

1.2.1 Random Forest and Neural Network Models

Random Forests (RF) and Artificial Neural Networks (ANN) are two machine learning algorithms that have shown strong potential in reservoir engineering applications. RF is an ensemble-based algorithm that builds multiple decision trees using random subsets of both data and input features. The final prediction is obtained by averaging across all the trees, which reduces variance and helps prevent overfitting.

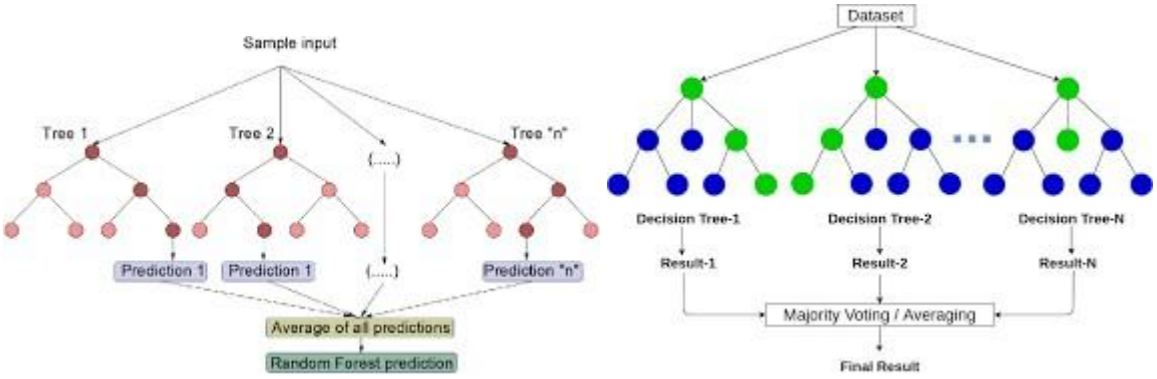

Figure 1: Random Forest conceptual diagram. Each tree (blue boxes) votes on the output; the final regression output is the average of the tree predictions. Here, different trees split on different log features, illustrating how RF blends multiple decision boundaries.

This ensemble mechanism improves stability and robustness in predictive tasks, especially when dealing with noisy or high-dimensional data. In petroleum engineering, RF has been successfully applied to capture nonlinear interactions among variables such as permeability, skin factor, porosity, and completion length. Another advantage of RF is its ability to generate feature importance scores, which provide insights into which geological or operational parameters most strongly influence productivity index (PI).

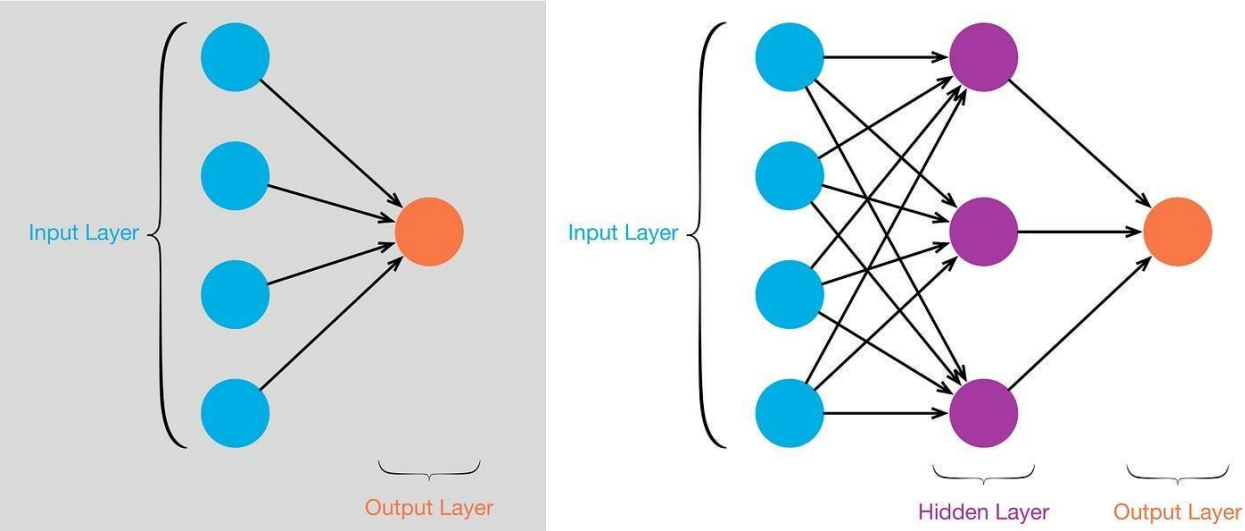

Figure 2: A schematic of a three-layer artificial neural network (inputs, hidden neurons, output) used for regression. Each neuron applies weights and nonlinear activation (sigmoid shown) to compute the output function.

Artificial Neural Networks (ANN), particularly feedforward multilayer perceptrons (MLP), consist of input, hidden, and output layers of interconnected nodes (neurons). Each connection carries a weight that is adjusted during training using optimization algorithms such as stochastic gradient descent. Through this iterative learning process, ANNs can approximate highly nonlinear functions and model complex relationships between input features and output targets. ANNs have been widely used in production forecasting and reservoir property prediction. For example, Fan et al. (2025) reported that MLP-based models significantly improved forecasting accuracy compared to conventional regression techniques. Similarly, Jin et al. (2024) demonstrated that a Long ShortTerm Memory (LSTM) based ANN effectively captured temporal variations in tight-oil production and achieved a coefficient of determination of 0.97, highlighting the capability of neural networks to incorporate time-series dependencies.

Despite their strength, ANNs require careful design choices, such as the number of hidden layers, the learning rate, and the application of regularization techniques to prevent overfitting. They also typically demand larger training datasets than RF to achieve reliable performance. In practice, both RF and ANN can be applied to static reservoir features (such as porosity, permeability, and fluid saturation) or extended to include dynamic time-series inputs (such as historical production rates).

In this thesis, the focus is on implementing RF and a standard feedforward ANN to model productivity index from historical production and reservoir data. This choice reflects a balance between interpretability, robustness, and the ability to capture nonlinearities in the dataset.

1.3 Historical Production Data for PI Prediction

A major strength of data-driven approaches for productivity index (PI) modeling is the ability to incorporate historical production data as key inputs. Unlike traditional methods that depend primarily on static reservoir parameters such as porosity, permeability, and pressure, machine learning (ML) models can learn directly from actual production trends recorded in the field. These datasets often include time-series measurements such as oil and gas flow rates, bottom-hole pressure (BHP), water cut, gas-oil ratio (GOR), and choke settings. By integrating these parameters, ML models inherently account for reservoir heterogeneity, well completion strategies, and operational adjustments that are often difficult to capture using empirical correlations.

Furthermore, historical production records reflect dynamic reservoir behavior under varying operational conditions. For instance, changes in water cut, onset of artificial lift, or well stimulation activities alter well performance in ways that static analytical models cannot easily

predict. Incorporating these factors into ML training allows for robust generalization, enabling models to adapt to different operational and geological scenarios. Supervised learning algorithms such as Random Forests (RF) and Artificial Neural Networks (ANN) are particularly effective in this context, as they can model nonlinear interactions between input features and PI outcomes while minimizing bias through feature engineering and cross-validation.

Several studies highlight the value of production-based ML modeling. Alarifi et al. (2015) developed AI-based predictive models including neural networks and fuzzy logic systems trained on more than 100 horizontal wells. Their findings demonstrated that ML methods were able to overcome the limitations of classical horizontal well productivity correlations by capturing fieldspecific complexities. Similarly, other research has shown that data-driven regression and ensemble models outperform decline curve analysis in terms of accuracy and reliability, particularly when large volumes of historical production data are available (Fan et al., 2025; Yan et al., 2023).

In summary, historical production data serves as a comprehensive and realistic training foundation for ML-based PI forecasting. By leveraging time-series field data, RF and ANN models are capable of producing more reliable predictions than static correlations, thereby enhancing decision-making in reservoir and production management.

1.4 Niger Delta Basin: Geology and Production

The Niger Delta Basin of Nigeria is one of the most prolific hydrocarbon provinces in the world and represents the cornerstone of Nigeria's petroleum industry. It is ranked as the twelfth largest oil-bearing province globally, with ultimate recoverable reserves estimated at tens of billions of barrels (Doust and Omatsola, 1990; Kulke, 1995). Current production averages around 2 million

barrels of oil per day, which makes the basin not only central to Nigeria's energy sector but also of strategic importance to global oil markets (NNPC, 2023).

Hydrocarbon accumulation in the Niger Delta is predominantly associated with the Agbada Formation, a thick Eocene sequence characterized by alternating sandstones and shales deposited in a fluvial–deltaic environment (Avbovbo, 1978; Doust and Omatsola, 1990). The sandstones within this formation often serve as excellent reservoirs due to their favorable petrophysical properties, with porosity values typically ranging between 15 and 25 percent and permeabilities varying from tens of millidarcies to several darcies in certain intervals (Doust and Omatsola, 1990; Evamy et al., 1978). However, the lateral continuity of these sand bodies is highly variable, resulting in heterogeneous reservoir quality.

The geological framework of the Niger Delta is further complicated by structural deformation features such as growth faults, rollover anticlines, and shale diapirs (Evamy et al., 1978; Stacher, 1995). These features disrupt depositional continuity, compartmentalize reservoirs, and significantly influence hydrocarbon distribution and migration pathways. As a result, predicting reservoir performance and fluid flow behavior in the Niger Delta remains challenging when relying solely on conventional geological or engineering models.

The petroleum industry in Nigeria is of immense economic significance, as crude oil contributes the largest share of government revenue and foreign exchange earnings (NNPC, 2023). Accurate production forecasting is therefore essential for sustainable field development, resource management, and long-term economic planning. Over several decades, thousands of wells have been drilled across the basin, generating an extensive archive of subsurface and production data. Many of these wells have detailed historical records, including oil, gas, and water production rates as well as bottom-hole pressure measurements. This extensive dataset provides a robust

foundation for advanced modeling techniques, particularly those based on data-driven and machine learning (ML) approaches.

Recent scholarship has increasingly explored the potential of ML in petroleum forecasting in Nigeria. For example, Adewale et al. (2025) demonstrated the application of ensemble tree-based algorithms such as Extra Trees, XGBoost, and Random Forest to predict Nigeria's aggregate oil production using macroeconomic and technical indicators such as reserves, oil price, and historical production rates. Their models achieved a strong correlation coefficient of approximately 0.82, highlighting the predictive capability of ML in large-scale oil forecasting. This finding aligns with broader global trends where ML has been applied successfully for field-level production optimization, decline curve analysis, and reservoir property prediction (Al-Mudhafar, 2017; Ahmed et al., 2020).

Nevertheless, the application of ML at the individual well scale within the Niger Delta remains underexplored. Specifically, the development of machine learning models for productivity index (PI) prediction, flow unit classification, and well performance forecasting has yet to receive significant attention in the basin. Addressing this research gap is critical, as well-level ML applications could enhance recovery estimation, improve production optimization, and strengthen reservoir management strategies in one of the world's most geologically complex and economically vital petroleum provinces.

1.5 Research Aim

The primary aim of this research is to develop and evaluate machine learning models for predicting the well productivity index (PI) using historical production data from selected Niger Delta wells. The study will specifically implement and compare two widely applied approaches:

Random Forest (RF) and Artificial Neural Networks (ANNs). These models will be trained on field-derived datasets comprising production flow rates, bottom-hole and reservoir pressures, fluid properties, and relevant well and reservoir parameters.

The first aim is to assess whether machine learning methods can achieve higher predictive accuracy and robustness compared to conventional analytical or empirical PI correlations. The second aim is to identify and rank the key features influencing PI in Niger Delta reservoirs, thereby improving understanding of the interplay between reservoir quality, completion efficiency, and operational practices. A third aim is to validate whether ML-based predictions can account for real-world complexities, such as reservoir heterogeneity, variable water cut, and artificial lift performance, which are often underrepresented in traditional models.

By leveraging field production data, the study intends to provide reservoir engineers and operators with a reliable, data-driven predictive tool for estimating well productivity. Ultimately, the research seeks to contribute toward optimized reservoir management and improved production forecasting in the Niger Delta, where reliable PI estimation is critical for efficient field development planning.

1.6 Research Objectives

1. To develop machine learning models (Random Forest and Artificial Neural Networks) for predicting the well productivity index (PI) using historical production and reservoir data from Niger Delta wells.
2. To compare the predictive performance of the machine learning models with conventional PI estimation methods, in order to evaluate improvements in accuracy, consistency, and robustness.

3. To identify and rank the key factors (e.g., flow rates, reservoir pressures, permeability, completion parameters) that most strongly influence PI in Niger Delta reservoirs.
4. To validate the applicability of machine learning approaches in accounting for complex reservoir conditions such as heterogeneity, water cut variations, and artificial lift effects, which are often inadequately represented in traditional correlations.
5. To provide a data-driven predictive tool that can support reservoir engineers and operators in optimizing field development planning, production forecasting, and well management strategies in the Niger Delta.

1.7 Justification of the Research

Accurate estimation of the well productivity index (PI) is essential for reservoir management, production forecasting, and field development planning. Conventional PI correlations and analytical models are often derived from simplified assumptions about reservoir geometry, homogeneity, and fluid flow behavior. While useful, these approaches tend to overlook real-world complexities such as reservoir heterogeneity, completion effects, water cut progression, and artificial lift operations. As a result, they can yield unreliable predictions when applied to complex reservoirs like those in the Niger Delta.

Machine learning (ML) offers a promising alternative by learning directly from historical production and reservoir data. Unlike traditional correlations, ML models are capable of capturing nonlinear relationships and hidden interactions among multiple variables. By training on field data, they inherently account for reservoir-specific characteristics, operational controls, and production history. This data-driven capability makes ML particularly valuable for the Niger Delta, where reservoirs are geologically diverse, and production conditions are dynamic.

Furthermore, a comparative study of Random Forest (RF) and Artificial Neural Networks (ANN) provides both interpretability (via RF's feature importance analysis) and high predictive power (via ANN's nonlinear mapping ability). Such a study not only advances the technical understanding of PI prediction but also provides practical value to operators by offering a more reliable, cost-effective, and adaptable predictive tool than static empirical correlations.

Therefore, this research is justified by its potential to bridge the gap between traditional PI estimation methods and modern data-driven approaches, ultimately improving the accuracy, robustness, and applicability of productivity forecasting in Niger Delta oil and gas fields.

1.8 Scope of the Research

This study is centered on the prediction of the well productivity index (PI) using machine learning techniques, with specific application to oil and gas reservoirs in the Niger Delta. The research will make use of historical field data obtained from selected wells within the region. These data include production rates, flowing bottomhole pressures, reservoir pressures, and other key reservoir and completion parameters. Only wells with adequate and reliable datasets will be considered to ensure consistency and minimize uncertainties associated with data gaps or poor-quality measurements.

The research will focus on developing and testing two machine learning models: Random Forest (RF) and Artificial Neural Networks (ANN). These models will be trained using historical production and pressure datasets, and their performance will be validated against unseen data to evaluate predictive capability. The models will then be benchmarked against conventional

productivity index correlations to determine whether machine learning offers superior accuracy, robustness, and adaptability to heterogeneous reservoir conditions.

Furthermore, the study will include an analysis of the most influential variables controlling productivity index predictions. Feature importance evaluation will be used to identify the reservoir and operational parameters that play the most significant role in determining PI in Niger Delta wells. Particular attention will be given to the interaction between flow rates, pressure behavior, and completion design, as these are critical for understanding well performance.

The scope of the study, however, is limited to the use of field data available for selected Niger Delta wells. The results may therefore reflect the data quality and coverage of the study area, and extrapolation to other reservoirs should be done with caution. Additionally, the research does not involve real-time monitoring or optimization of PI but is limited to predictive modeling and comparative evaluation of machine learning and conventional methods.

1.9 Limitations of the Research

This research is subject to several limitations that may influence the general applicability of its findings. First, the study relies on historical production and reservoir data obtained from a limited number of wells in the Niger Delta. The availability, accuracy, and consistency of these datasets can directly affect the reliability of the machine learning models. Missing or noisy data may introduce uncertainties in model training and validation, despite preprocessing and quality control measures.

Second, the predictive models developed in this work are restricted to the specific geological, petrophysical, and operational conditions of the Niger Delta. While the results may provide useful insights for similar clastic reservoirs, direct application to other regions with different rock properties, fluid compositions, and completion strategies may not be straightforward without further calibration.

Third, the study focuses on two machine learning approaches, namely Random Forest and Artificial Neural Networks. Although these are powerful methods, other advanced techniques such as Gradient Boosting, Support Vector Machines, or hybrid ensembles may offer additional improvements in accuracy and robustness. The exclusion of these methods limits the range of comparisons in this research.

Finally, the scope of the study is confined to predictive modeling of the productivity index and does not extend to real-time optimization or operational deployment. As such, while the research highlights the potential of machine learning as a predictive tool, its practical integration into field operations would require further development, including continuous data updating, workflow automation, and economic evaluation.

CHAPTER TWO

2.0 Literature Review

The productivity index (PI) is a key reservoir engineering parameter that quantifies the capacity of a well to produce hydrocarbons per unit pressure drawdown. For single-phase flow, it is expressed through Darcy's law as:

$$PI = \frac{q}{P_r - P_{wf}} \qquad (2.1)$$

where q is the flow rate, the average reservoir pressure, and the flowing bottom-hole pressure. PI serves as the basis for describing the inflow performance relationship (IPR) and, under steadystate conditions, reflects the drainage efficiency of the formation.

Traditionally, PI is determined from rate–pressure tests on producing wells. Over the years, various analytical and empirical correlations have been developed to predict PI using reservoir and wellbore parameters prior to production testing. Notable models for horizontal wells include those by Giger–Reiss–Jourdan, Butler, Renard–Dupuy, Joshi, Borisov, and Economides–Fisher. Joshi's (1988) model remains one of the most widely used, while Vogel's (1968) empirical IPR curve has been particularly influential in multiphase systems. Applications of these models to Nigerian reservoirs, especially in the Niger Delta, have reported typical PIs on the order of ~1 STB/day/psi.

Despite their utility, these empirical formulas have inherent limitations. They are derived under simplified physical assumptions and often fail to capture reservoir heterogeneity, multiphase flow effects, and complex wellbore–formation interactions. As a result, they may systematically

under- or over-predict actual well productivity, especially in reservoirs with variable permeability, anisotropy, and non-Darcy flow effects. This mismatch has motivated the exploration of datadriven approaches that can learn from historical well performance and reservoir datasets.

In recent years, machine learning (ML) methods have emerged as powerful tools for predicting well performance parameters such as PI. Techniques such as Artificial Neural Networks (ANNs), Random Forests (RF), Gradient Boosting Machines (GBMs), and Support Vector Machines (SVMs) have been applied to estimate productivity by learning complex nonlinear relationships between input variables (e.g., porosity, permeability, thickness, skin, completion parameters) and output performance measures. Studies have demonstrated that ML models often outperform empirical correlations by capturing reservoir-specific trends and integrating diverse data sources, including well logs, production history, and well test data.

For example, Al-AbdulJabbar et al. (2020) applied ANN models to predict PI for horizontal wells and reported improved accuracy compared to conventional correlations. Similarly, Ahmadi et al. (2019) used Random Forests and Gradient Boosting techniques to estimate productivity indices in heterogeneous carbonate reservoirs, achieving reductions in mean absolute error relative to analytical methods. In the Niger Delta, preliminary studies have highlighted the potential of ML to predict PI more reliably than existing empirical models, particularly when integrated with petrophysical logs and pressure–volume–temperature (PVT) data.

While these advances are promising, challenges remain in applying ML to PI prediction. The performance of ML models depends heavily on data quality, feature selection, and model interpretability. Moreover, unlike physics-based correlations, ML models often operate as "black boxes," making it difficult to explain the physical significance of predictions without further

interpretability frameworks. Nevertheless, the integration of ML with conventional reservoir engineering methods offers a hybrid pathway that can enhance prediction accuracy and provide more reliable decision-making tools for well productivity assessment.

2.1 Early Analytical and Correlation-Based PI Methods (1980s–2000s)

Early research on well productivity focused heavily on deriving analytical solutions and empirical correlations for the Productivity Index (PI) of different well configurations. Joshi (1988) was among the first to propose a semi-analytical equation for horizontal wells, which significantly improved over classical vertical well models by accounting for well length and anisotropy effects. Later, Osisanya et al.,(2021) extended these formulations, emphasizing transient and pseudosteady-state flow regimes.

In the late 1990s, Hongen (1997) contributed methods to calculate both PI and critical production rates in horizontal wells, particularly under gas–oil and water–oil systems. Escobar et al. (2004) later developed an improved correlation specifically for horizontal wells, addressing limitations of Joshi's equation when applied to unconsolidated or high-permeability formations.

Despite these advances, correlation-based PI methods typically assume uniform reservoir properties, homogeneity, isotropy, and simple geometries—conditions rarely met in complex depositional systems such as the Niger Delta. For instance, in two open-hole horizontal wells in the Niger Delta, Joshi's and Vogel's methods predicted PI values of approximately 1.04 and 1.16 STB/d/psi, respectively. These values indicated reasonably good productivity but also demonstrated deviations from field observations, highlighting the need for models better tailored to unconsolidated, anisotropic, and compartmentalized sandstone reservoirs.

Overall, these early models provided a foundational baseline for PI estimation and were instrumental in well performance forecasting prior to widespread computational tools. However, their limited adaptability to heterogeneous and fractured formations motivated the eventual transition toward numerical simulation and, more recently, data-driven machine learning approaches.

2.2 Emergence of Machine Learning in Well Performance (2010s)

In the last decade, machine learning (ML) and other data-driven approaches have been increasingly adopted in well performance forecasting. Early applications focused on using artificial neural networks (ANNs) to predict production rates and inflow performance relationships (IPR), where the productivity index (PI) was treated as an implicit outcome rather than a direct target. A major milestone came from Alarifi et al. (2015), who were among the first to apply artificial intelligence directly to PI prediction. They developed ML models—including neural networks, fuzzy logic, and functional networks—to predict the PI of horizontal wells using static reservoir and well parameters. By training on real field data from over 100 horizontal wells in the Middle East, their models achieved "very good accuracy" and significantly outperformed traditional PI correlations such as those of Borisov, Giger–Reiss–Jourdan, Renard–Dupuy, Joshi, Butler, and Furui. This study marked a turning point by demonstrating that ML, when trained on actual well test data, could overcome many of the inherent limitations of empirical correlations that were developed under restrictive assumptions.

Shortly after, Al-Mashhad et al. (2016) extended AI-based methods to more complex well architectures. They applied ANNs to multilateral wells and compared results with analytical models (e.g., Borisov's multilateral correlation). Their work showed that ML could accurately

predict flow rates and associated PIs in multilateral configurations when provided with suitable inputs.

Concurrently, the oil industry entered the “big data” era, driving wider adoption of ML in production forecasting. For instance, Cao et al. (2016) proposed data-driven approaches for production forecasting using ML, though not specifically targeted at PI. Building on this momentum, subsequent research employed ensemble methods, support vector machines (SVM), and regression trees to predict cumulative production or decline-curve parameters. A comprehensive review by Rahmanifard et al. (2023) highlights that ANNs, random forests (RF), gradient boosting machines (GBM), and SVMs have all been tested for forecasting unconventional well performance, often using features such as fracture length, proppant loading, and well pressures.

Recent comparative studies underscore the advantages of ML over conventional techniques. Gao et al. (2022) evaluated six ML methods (GB, DT, RF, SVR, ANN) for predicting unconventional gas well productivity and found random forest delivered the best predictive accuracy. Similarly, Baki et al. (2021) applied XGBoost, ANN, and SVR to Eagle Ford shale data incorporating fracture volume and lateral length, with XGBoost outperforming the others in forecast accuracy.

Together, these studies demonstrate the global trend: ML models can integrate diverse well, reservoir, and completion parameters to capture complex productivity trends that traditional decline curves or single-variable correlations often miss.

2.3 Direct Machine Learning Prediction of Productivity Index

The direct application of machine learning (ML) to predict well productivity index (PI), rather than production volumes or decline trends, is a relatively recent research direction. Gruzdev et al.

(2020) introduced one of the earliest models that employed ML for PI prediction using loggingwhile-drilling (LWD) data and interpreted well logs as input variables. Their approach was tested on historical datasets from the Novoportovskoye oilfield, where leave-one-well-out crossvalidation was applied to assess model generalizability. The model achieved a median relative error below 20%, demonstrating that ML techniques can infer PI from geophysical log data by capturing the reservoir's petrophysical and flow responses, even without explicit production rate information.

Building on this, Alharbi and Alarifi (2023) extended ML-based PI prediction to more complex well geometries, including both single-lateral and multilateral horizontal wells. In their work published in *ACS Omega*, they developed predictive models, primarily based on artificial neural networks (ANNs) and gradient-based methods, which utilized a combination of reservoir and well parameters. Their results emphasized that ML methods can adapt to a wider range of completion and reservoir scenarios, reflecting the ongoing global interest in applying ML for direct PI estimation.

Parallel to these efforts, other studies have focused on applying ML and deep learning to forecast production trajectories. For instance, Vikara and Khanna (2022) applied long short-term memory (LSTM) networks to jointly forecast oil, gas, and water production from shale wells, utilizing historical time-series data as inputs. Such models treat dynamic production histories as sequential data, enabling the prediction of future well performance with improved temporal accuracy. Similarly, Esmaili and Mohaghegh (2016) integrated static completion parameters with actual production histories to construct ANN models capable of predicting oil and gas production across entire fields. These studies highlight the value of incorporating both static and dynamic

data, although their primary focus remains volumetric prediction and decline forecasting rather than PI estimation.

Collectively, the literature demonstrates that while most ML applications in petroleum engineering emphasize production forecasting, there is a growing body of work dedicated to direct PI prediction. These advances suggest that ML-based approaches hold significant potential for providing reservoir engineers with more accurate and adaptable tools for well performance evaluation, particularly in settings where conventional empirical correlations are limited.

2.4 Static vs. Dynamic Data in Modeling

Machine learning (ML) models for reservoir characterization and production forecasting often rely on two major classes of inputs: static and dynamic data. Static features represent reservoir or well attributes that remain constant over the productive life of the well. These include geological and petrophysical parameters such as permeability, porosity, net pay thickness, initial reservoir pressure, and completion design details (e.g., lateral length, proppant volume, number of hydraulic fractures). By contrast, dynamic features are time-dependent variables, usually obtained as production or pressure measurements over time. These include monthly or daily rates of oil, gas, and water, as well as wellhead or bottomhole pressures.

In practice, modern ML workflows often combine both types of data to improve model accuracy. For example, Vikara (2022) compiled a dataset for Midland Basin wells that included a mixture of static descriptors and dynamic features. Static attributes encompassed geographic, geological, and completion parameters, while dynamic features primarily consisted of monthly oil, gas, and water production. To bridge the two, static proxies were generated from time-series data—such as the “Top 12-month” oil and gas production, where the highest 12 monthly volumes are

aggregated to create a single static predictor. This method captures production trends without requiring the full time-series as input, thereby simplifying model architecture while retaining predictive information.

Different ML modeling strategies highlight this interplay. In the study by Alarifi et al. (2015), static inputs such as reservoir permeability, porosity, and pressure were combined with well test data to predict productivity index (PI). This approach exemplifies models that rely largely on static or single-event dynamic data rather than full sequences. Conversely, recurrent neural networks (RNNs) and long short-term memory (LSTM) models explicitly ingest sequential dynamic data such as historical production rates to learn temporal dependencies and forecast future performance (Vikara, 2022). Hybrid frameworks also exist, where both static variables (e.g., reservoir quality, completion parameters) and short-term dynamic production histories are fed into artificial neural networks (ANNs) to improve generalization.

When dynamic data are limited or incomplete, researchers often convert time-series into representative static summary metrics. Common examples include cumulative six-month or twelve-month production, decline rate coefficients, or normalized production indices. These summary statistics are easier to handle with traditional regression and tree-based ML methods. However, they inevitably sacrifice some temporal detail compared to fully sequential approaches.

Globally, the use of both data types has gained traction. Reviews show that most modern production-forecasting models now integrate geological data, drilling and completion design parameters, and early production histories within ML algorithms to balance predictive accuracy with data availability (Alarifi et al., 2015; Vikara, 2022). This combined approach ensures that both the static reservoir capacity and the dynamic production behavior are reflected in the predictive framework.

2.5 Niger Delta Case Studies

The Niger Delta (Nigeria) presents a challenging testbed for productivity prediction due to its thick, faulted sandstone reservoirs, complex depositional environments, and prevalent high water cuts. Historically, operators in the region have relied heavily on empirical approaches such as decline curve analysis (DCA) and inflow performance relationship (IPR) curves to evaluate well productivity. As noted by Jayeola et al. (2022), "decline curves are the most extensively used technique for the production forecast of Niger Delta reservoirs" despite their limitations in capturing operational variability and reservoir uncertainties.

Recent studies have begun to explore data-driven alternatives. For example, Jayeola et al. (2022) applied artificial neural networks (ANNs) to forecast oil production in Niger Delta reservoirs. Their results demonstrated that optimized neural networks could capture nonlinear patterns in historical production time-series data more effectively than static correlation-based approaches. This finding highlights the potential of machine learning (ML) to adapt to changing production conditions, unlike traditional decline-curve methods.

However, literature specifically addressing productivity index (PI) prediction in Niger Delta wells remains sparse. One recent investigation employed analytical methods, namely Joshi's horizontal well model and Vogel's IPR equation, on two open-hole horizontal wells. The study reported PI values of approximately 1.04–1.16 STB/d/psi, indicating good productivity but also underscoring the reliance on classical models without incorporating ML-based approaches.

Overall, while global research has increasingly demonstrated the utility of ML methods in predicting PI and well productivity, their application to Niger Delta reservoirs remains limited. This gap presents an opportunity: integrating region-specific static reservoir properties with the

abundant dynamic production data through ML techniques could provide more reliable PI predictions tailored to the Niger Delta's unique geologic and operational conditions.

2.6 Niger Delta Case Studies

The Niger Delta (Nigeria) represents a uniquely challenging testbed for production analysis due to its thick, faulted sandstone reservoirs, complex structural traps, and high water cut conditions. Traditionally, operators in the region have relied on empirical decline curve analysis and inflow performance relationship (IPR) models to forecast production and evaluate well performance.

However, these conventional methods are limited in their ability to account for dynamic operational changes, reservoir heterogeneity, and uncertainties. As Jayeola et al. (2022) observe, "decline curves are the most extensively used technique for the production forecast of Niger Delta reservoirs," despite their shortcomings in adapting to varying conditions.

Recent work has demonstrated the potential of machine learning (ML) approaches to overcome these challenges. In particular, Jayeola's study applied artificial neural networks (ANNs) to forecast oil production under a range of conditions and showed that optimized neural networks trained on Niger Delta field data were able to capture nonlinear production trends more effectively than static decline correlations. These findings highlight the ability of ML to integrate multiple data sources and uncover complex relationships that traditional analytical methods often miss.

Nevertheless, research focused specifically on productivity index (PI) prediction in the Niger Delta remains limited. For example, a recent study applied analytical models such as Joshi's equation and Vogel's IPR curve to two horizontal wells in the region and reported PI values in the range of 1.04–1.16 STB/d/psi. While useful, this work did not leverage ML techniques and

relied solely on static correlations. This illustrates a significant gap in the literature, as global studies have shown that ML can provide more accurate and adaptable predictions of PI by integrating both static reservoir parameters and dynamic production history.

Traditional approaches to productivity index (PI) estimation have relied primarily on deterministic correlations and well-test analyses, such as Joshi's horizontal well correlation (1988) and Vogel's inflow performance relationship (1968). These methods, while widely applied, assume simplified reservoir and flow conditions and therefore often fail to capture the heterogeneity and operational variability encountered in real reservoirs. Since the mid-2010s, artificial intelligence (AI) and machine learning (ML) approaches have emerged as alternatives capable of addressing these limitations. Alarifi et al. (2015) demonstrated that artificial neural networks (ANNs), fuzzy logic systems, and functional networks can outperform traditional correlations in predicting the PI of horizontal wells under diverse reservoir conditions. Building on this, subsequent studies extended ML applications to more complex scenarios, including multilateral well productivity modeling (Al-Mashhad, 2016) and the integration of logging-while-drilling (LWD) data for PI estimation (Gruzdev, 2020; Alharbi, 2023).

In parallel, production-forecasting models have seen significant progress, with the application of

ANNs, tree-based ensemble methods, and recurrent neural networks such as Long Short-Term Memory (LSTM) architectures. These models often leverage time-series production data to capture nonlinearities and temporal dependencies in well performance. A critical consideration across these approaches is the type of input data employed. Some models depend on static reservoir and well parameters (e.g., porosity, permeability, completion geometry), while others directly utilize dynamic datasets such as historical production curves.

In the context of the Niger Delta, deterministic approaches remain dominant, particularly declinecurve analysis methods, which are widely applied despite their limited adaptability to operational changes and subsurface uncertainties. Integrating machine learning techniques with locally available production datasets therefore represents a promising research direction. By combining static reservoir descriptors with dynamic production histories, ML models could provide more accurate and robust predictions of well productivity indices. This shift has the potential to enhance decision-making in the Niger Delta, where complex geology, compartmentalization, and operational constraints challenge the assumptions of conventional methods.

Table 1: Evolution of PI Prediction Approaches

Category	Method/Study	Key Features	Limitations	Relevance to Niger Delta
Traditional Deterministic Methods	Vogel (1968), Joshi (1988)	Analytical correlations; based on steadystate or pseudosteady assumptions.	Rigid assumptions; cannot capture nonlinear reservoir behavior or operational variability.	Widely applied in ND wells for IPR and well-test PI analysis.
AI/ML Emergence (2010s)	Alarifi et al. (2015)	Applied ANNs, fuzzy logic, and functional networks to horizontal wells.	Requires sufficient training data; early models often lacked generalization.	Demonstrated global shift from correlations to ML.

Expanded ML Applications	Al-Mashhad (2016), Gruzdev (2020), Alharbi (2023)	Models extended to multilateral wells, incorporation of LWD and field data.	Data dependency; often developed on Middle East datasets.	Shows adaptability of ML beyond simple well types, but limited NDspecific studies.
Production Forecasting ML Models	Global works using ANNs, Random Forests, LSTMs.	Models trained on time-series production and dynamic features; capable of capturing nonlinear decline.	Often region-specific; require large datasets; black-box nature can hinder interpretability.	ND studies still rely mainly on decline-curve analysis; little ML adoption for PI prediction.
Niger Delta Context	Jayeola et al. (2022)	ANN used for production forecasting in ND reservoirs ; demonstrated improved performance over decline curves.	Focused on production forecasting, not PI prediction.	Shows proof-ofconcept that ML works well on ND data.

CHAPTER THREE

3.0 Methodology

3.1 Overview

The methodology for this study was designed to create a reproducible framework for predicting the well productivity index (PI) using historical production and pressure data. The dataset consists of daily oil, gas, water, and pressure records from five wells, totaling about 7,000 entries. Because of the limited number of wells, the approach focused on maximizing the amount of useful information extracted from the data while minimizing the risk of overfitting to individual wells. The workflow moved in a stepwise manner, starting with data exploration, followed by cleaning, feature engineering, model development, evaluation, and interpretation. Each stage was structured to ensure that results could be replicated and independently reviewed.

3.2 Data Exploration

The first stage involved an exploration of the raw dataset. Production dates were parsed and records were arranged chronologically to establish the time coverage of each well. This allowed the identification of missing or duplicated entries, as well as inconsistencies in reporting. Summary statistics were computed for all variables, and exploratory plots such as time series, scatter plots, histograms, and boxplots were generated. These visualizations provided early insights into data structure, potential relationships between variables, and the presence of outliers. This preliminary analysis served as a foundation for subsequent data cleaning

Table 3.1: Statistical representation of the dataset

	N_WELL_	Downhole Average Temperature (Kelvin)	Downhole Annulus Pressure (PSI)	AVG Tubing Pressure	Gas Volume WHP (PSI)	Choke Size (PSI)	Production (stb/day)	Oil Production (stb/day)	CHOKE_6 (scf/day)	Oil Production (stb/day)	CHOKE_I	Water PI_TARGET	BORE_CO	NCHES	T
count	6925	6925	6925	6925	6925	6925	6919	6925	6.93E+03	6925	6919	6919	6		
mean	107.297	2587.483	349.6748	2161.986	180.3694	716.5769	21.86456	8494.775	6.98E+06	11480.56	21.86456	0.341634	399.5		
min	105	0	273.15	0	0	0	0	0	0.00E+00	-2879.81	0	0			
25%	107	0	273.15	896.6123	0	476.5398	4.631982	1686.601	1.43E+06	80.3233	4.631982	0.072375	154.9		
50%	107	3378.397	376.9114	2472.032	213.0911	619.6478	14.30166	5479.471	4.77E+06	5529.413	14.30166				
75%		0.223463	472.1903												
max	108	3664.905	379.4962	2907.047	300.3568	903.8905	31.52977	11730.47	9.89E+06	22261.51	31.52977	0.492653			
		637.9004													
	109	4606.667	381.6522	4592.613	435.2875	1991.01	125.7186	37122.57	3.00E+07	50444.16	125.7186	1.964353			
		1643.635													
std	0.980161	1588.474	46.24082	1053.875	139.7715	347.86	22.30069	8927.597	7.10E+06	11484.48	22.30069	0.348448	260.2343		

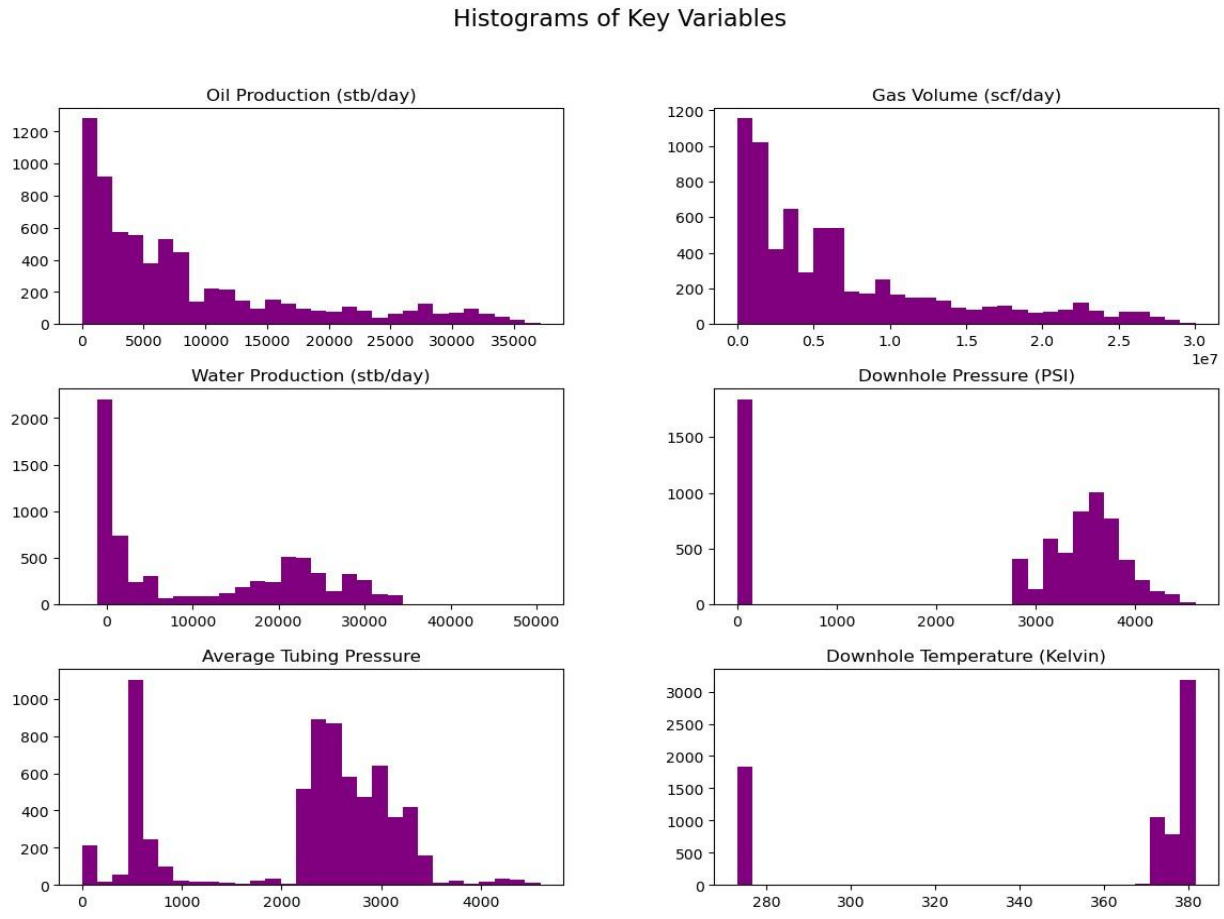

Figure 3.1: Frequency Distribution of Oil, Gas, Water, and Pressure Measurements

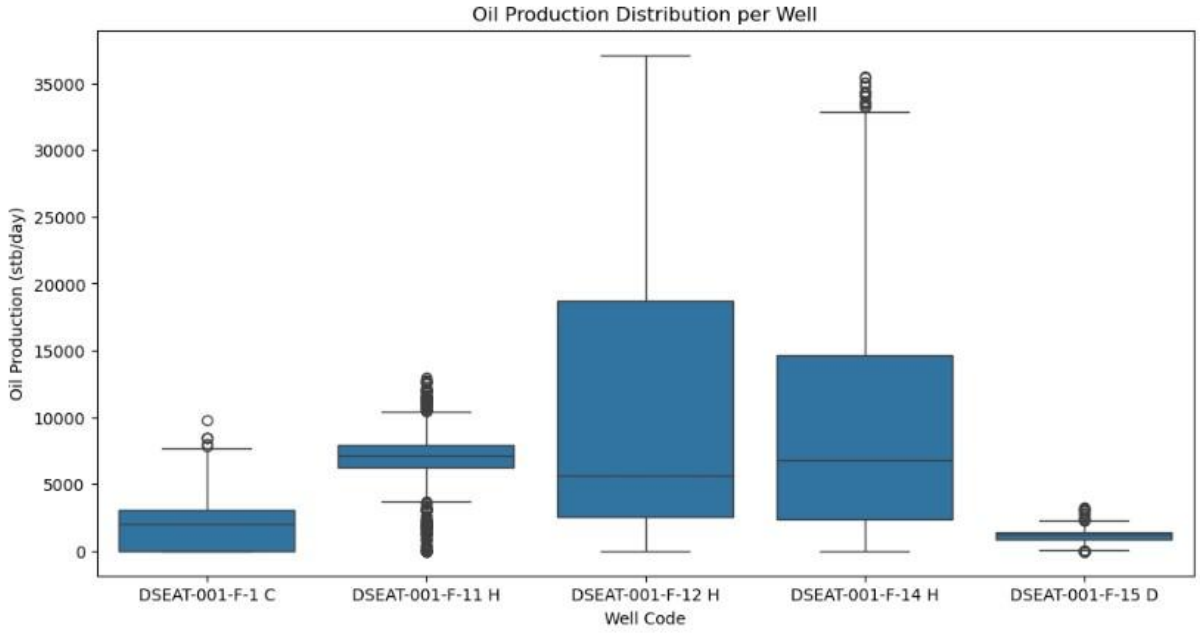

Figure 3.2: Well-by-Well Distribution of Oil Production

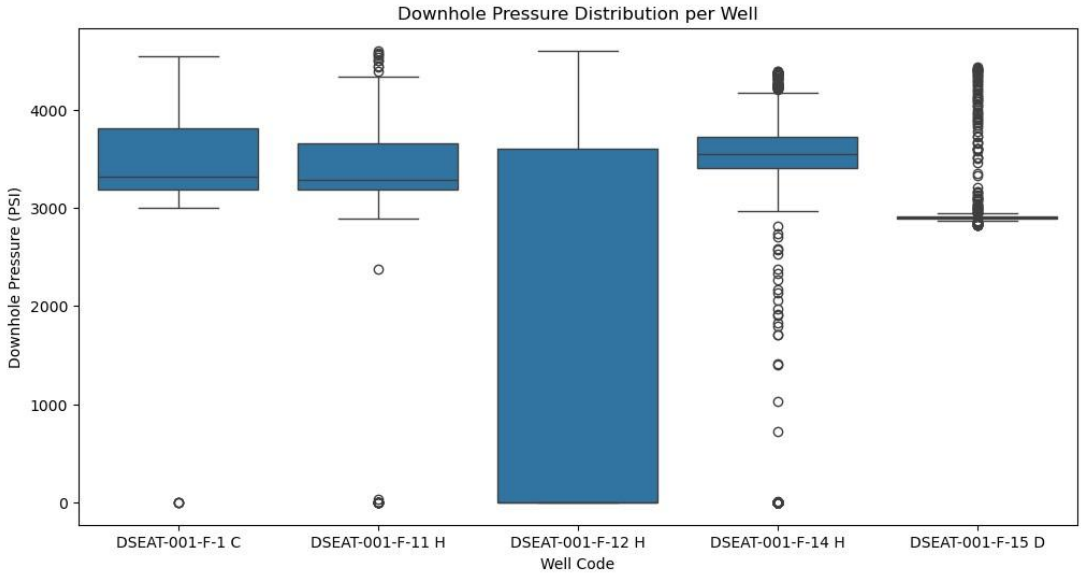

Figure 3.3 Well-by-Well Distribution of Downhole Pressure

3.3 Building the Productivity Index Target

To calculate the target variable, the productivity index, oil production rate was divided by pressure drawdown. Ideally, drawdown should represent the difference between reservoir pressure and bottomhole flowing pressure, but static reservoir pressure was rarely recorded at the same frequency as production data. To address this limitation, proxies were employed. Reservoir pressure was estimated using high percentiles of historical downhole pressures, while rolling percentiles were also tested to account for reservoir depletion. These approaches were compared to determine which produced more stable and physically meaningful PI estimates. Records with very small pressure differentials, which could result in artificially inflated PI values, were flagged for review.

3.4 Feature Engineering

This phase of the methodology focused on feature engineering to convert raw petrophysical and production data into a structured dataset suitable for machine learning workflows. Input features were drawn from well log measurements, including porosity and saturation indices, as well as flow test-derived parameters. Numerical features were standardized to mitigate scale imbalances, ensuring that variables measured in different units contributed proportionally during training. Categorical attributes, particularly well identifiers, were encoded using a transformation procedure that allowed the models to recognize variations across wells without imposing ordinal assumptions. Outliers and missing records were systematically handled to prevent distortions in model training. The productivity index (PI), defined as the dependent response variable, was computed using test data and served as the target metric for model prediction.

3.5 Model Development

The dataset was partitioned into training and testing subsets using a stratified approach to preserve variability across wells. A machine learning pipeline was designed to integrate preprocessing steps with the estimator, thereby automating transformations during training and inference. Ensemblebased regression algorithms were employed, with Random Forest, XGBoost, and CatBoost forming the primary models under investigation. These models were selected due to their ability to capture non-linear feature interactions and their robustness against overfitting in heterogeneous datasets. Model parameters were tuned iteratively through cross-validation, with emphasis placed on optimizing predictive stability rather than minimizing training error alone. This ensured that the models were not only well-fitted to the training data but also demonstrated reliable generalization capability on new observations.

3.6 Model Evaluation

Model performance was quantified using established statistical indicators that provide complementary perspectives on accuracy and reliability. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) measured the magnitude of prediction errors, while the Coefficient of Determination (R^2) quantified how much variance in the productivity index was explained by the models. Mean Absolute Percentage Error (MAPE) was also employed to capture proportional errors in prediction. To further validate model behavior, diagnostic plots were introduced. These included predicted versus actual scatter plots to assess alignment, residual plots to evaluate systematic bias, and histograms to study error distributions. Time series plots of selected wells were used to demonstrate how well the models reproduced observed field dynamics across different production intervals.

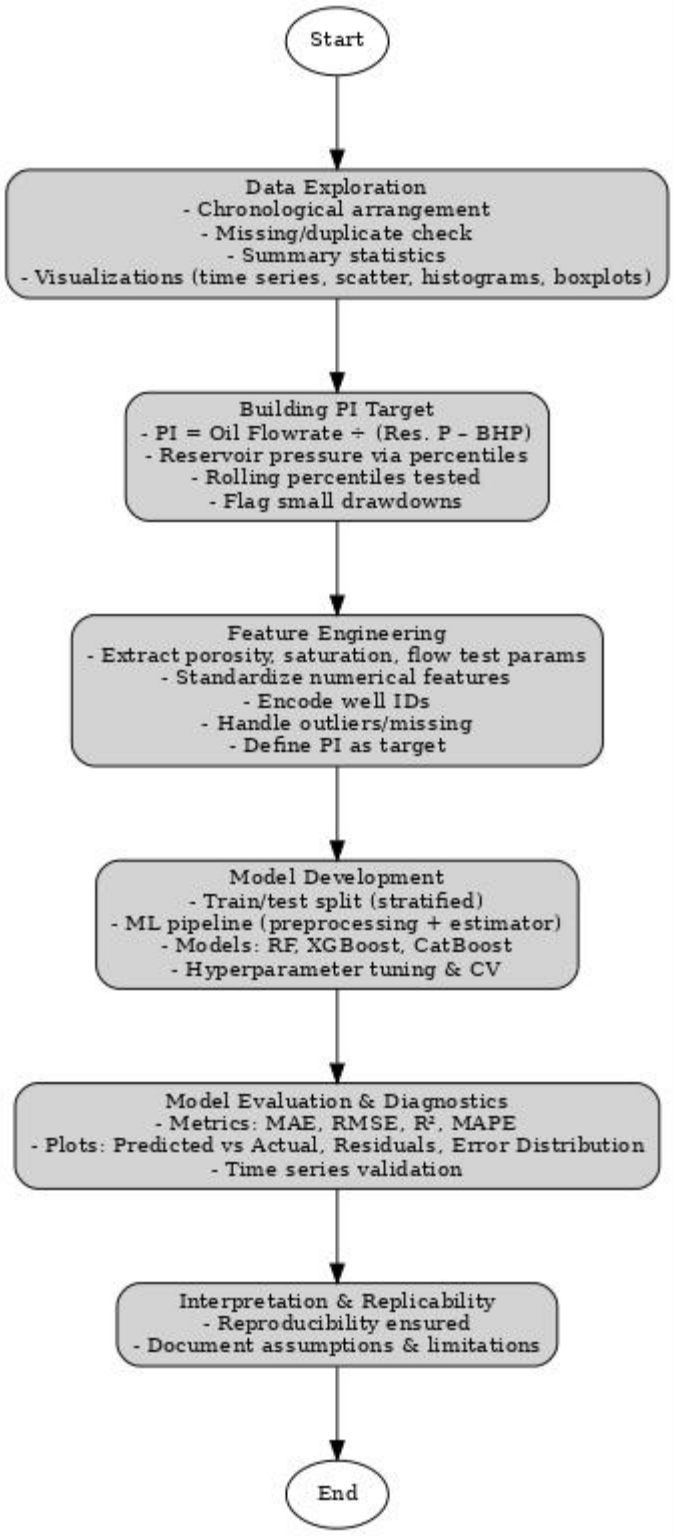

Figure 3.4: Methodological Workflow for Well Productivity Index Prediction Using Historical Production and Pressure Data

CHAPTER FOUR

4.0 Results and Discussion

4.1. Model Performance Overview

This study evaluated the efficacy of three advanced machine learning regression models, Random Forest, XGBoost, and CatBoost, in predicting the Productivity Index (PI) of wells using historical production data. The core objective was to determine whether complex patterns within routinely recorded data (e.g., pressures, rates, choke sizes) could be leveraged to accurately estimate this key performance metric. A comparative analysis of their performance metrics is presented in Table 4.1.

The Random Forest model yielded the least accurate predictions, with a high mean absolute error (MAE) of 65.6 and a root mean square error (RMSE) of 86.7. The low coefficient of determination ($R^2 = 0.18$) indicates that the model accounted for only a minor portion of the variance in the PI data. This performance suggests that the default Random Forest configuration was unable to capture the complex, nonlinear relationships between the historical input features and the target variable, likely due to model underfitting. The Predicted vs. Actual plot for Random Forest (Figure 4.1a) visually confirms this, showing a wide, cloud-like scatter of points with no clear alignment to the 1:1 reference line, indicating poor predictive capability.

XGBoost demonstrated a substantial improvement over Random Forest. Its errors were significantly lower (MAE ≈ 34.1, RMSE ≈ 45.0), and it achieved a robust R^2 value of 0.78. This level of performance implies that XGBoost's gradient-boosting framework successfully identified and leveraged key nonlinear patterns and interactions within the historical dataset. The Predicted

vs. Actual plot for XGBoost (Figure 4.1b) shows a much tighter clustering of points along the 1:1 line compared to Random Forest. However, a discernible curvature and spread, particularly at higher PI values, indicate that the model's generalization to the entire operational envelope was not yet optimal, with a tendency to underestimate the highest productivity points.

CatBoost emerged as the superior model for this specific prediction task. It achieved the lowest error metrics (MAE ≈ 18.9, RMSE ≈ 21.0) and the highest explanatory power, with an R^2 of 0.95. This result signifies that the CatBoost model explained approximately 95% of the variance in the PI across the dataset. The Predicted vs. Actual plot for CatBoost (Figure 4.1c) demonstrates a remarkable alignment of data points along the ideal 1:1 line, with minimal deviation or bias across the entire range of PI values. This visual evidence strongly supports the quantitative metrics, establishing CatBoost as the most reliable predictor.

Table 4.1: Comparative performance metrics of the machine learning models.

Model	MAE	RMSE	R^2
Random Forest	65.62	86.67	0.18
XGBoost	34.14	45.03	0.78
CatBoost	18.96	21.02	0.95

4.2. Analysis of Residuals

Residual analysis was conducted to assess the stability and predictive bias of the models, which is critical for evaluating their reliability for field application.

For Random Forest, the residual plot (Figure 4.2a) exhibited a wide, fan-shaped scatter, a clear sign of heteroscedasticity where the error variance increases with the predicted value. This pattern, combined with the lack of a clear central tendency around zero, strongly suggests the model was underfitting the underlying data structure.

For XGBoost, the residual plot (Figure 4.2b) showed a significant improvement, with a more random scatter. However, a slight U-shaped pattern can be observed, where the model tends to over-predict mid-range PI values and under-predict low and high values. This systematic bias indicates that XGBoost struggled to perfectly capture the full nonlinearity of the relationship.

In contrast, the residuals for the CatBoost model (Figure 4.2c) were distributed much more tightly and randomly around zero. The plot revealed no systematic upward or downward trend or distinct patterns, indicating an absence of significant bias. This was further validated by the residual histogram for CatBoost (Figure 4.3), which displayed an approximately normal, bell-shaped distribution centered on zero. This random, unbiased error distribution is a hallmark of a wellgeneralized model that has learned the underlying physical relationships rather than memorizing noise. This finding provides a high degree of confidence that the CatBoost model can be deployed to generate accurate PI estimates from new, unseen historical production data.

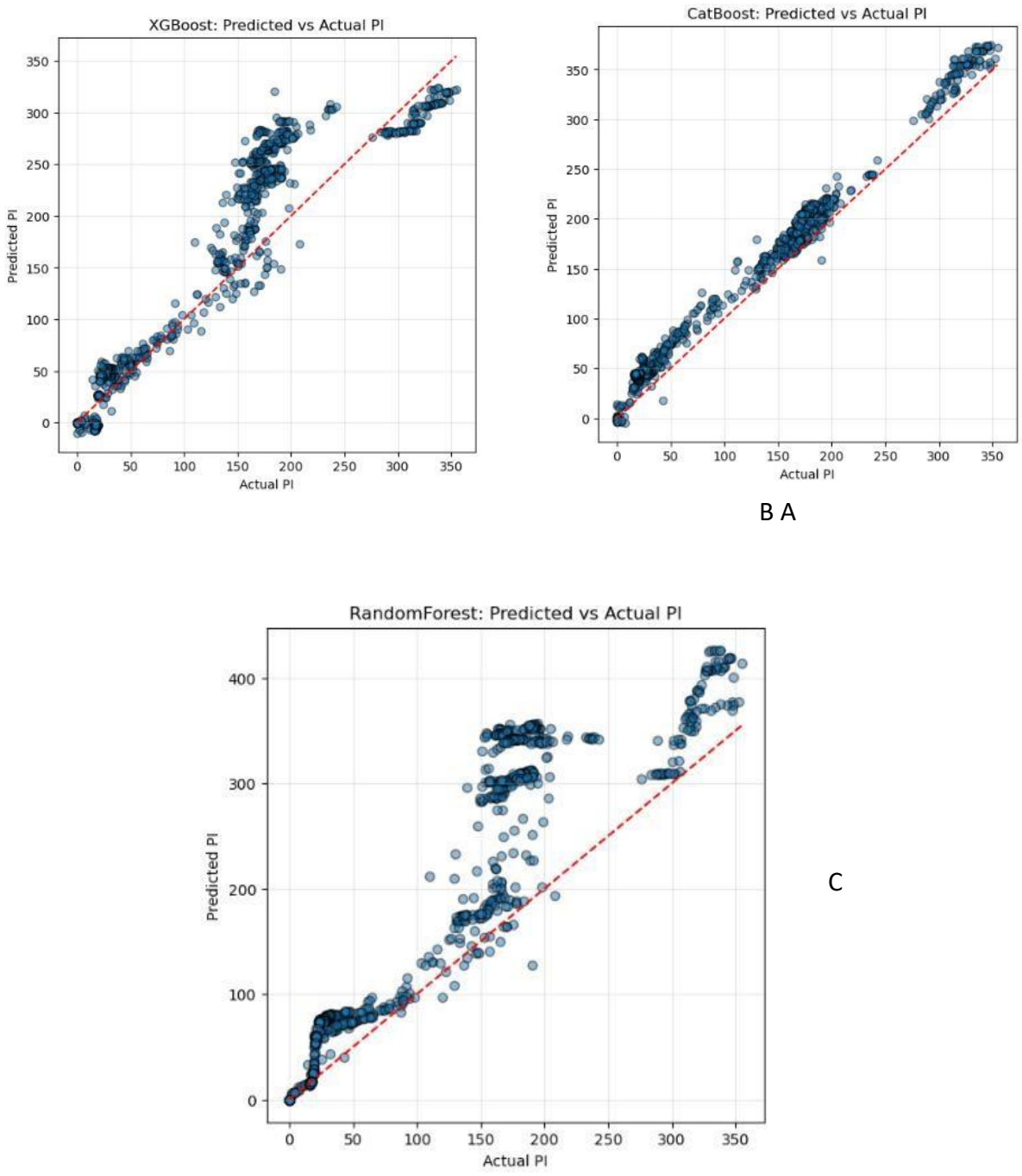

Figure 4.1. Predicted vs. Actual Productivity Index (PI) plots: a) Random Forest, showing high scatter; b) XGBoost, showing improved alignment with some bias at extremes; c) CatBoost, demonstrating close alignment with the 1:1 line.

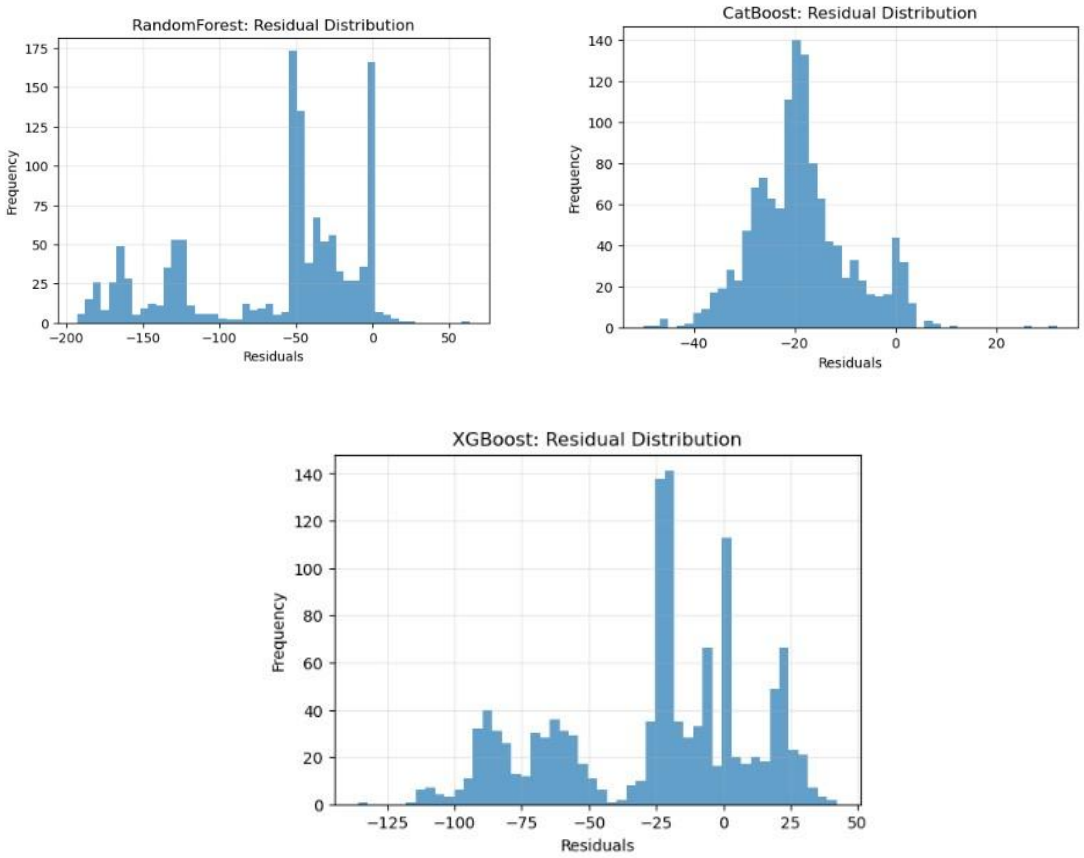

Figure 4.2. Residuals vs. Predicted plots: a) Random Forest, showing a fan-shaped pattern indicative of heteroscedasticity; b) XGBoost, showing reduced scatter but a slight U-shaped bias; c) CatBoost, showing a random and unbiased distribution of errors around zero.

4.3. Well-Level Performance

A pivotal test for a model trained on historical data is its ability to accurately reconstruct the performance timeline of individual wells. To evaluate this practical utility, the performance of the CatBoost model was scrutinized at the individual well level by comparing the time-series of actual and predicted PI values. The results demonstrated that CatBoost was highly effective at tracking the temporal dynamics and unique "fingerprint" of each well's performance. The model

accurately captured rising PI trends during periods of increased production and faithfully followed declining trends as wells depleted.

This sensitivity to operational history confirms that the model has effectively learned the dynamic interplay of key historical parameters. This capability is paramount, as it moves beyond a simple static prediction. It suggests the model is suitable for continuous monitoring and production forecasting throughout a well's lifecycle, creating a virtual, continuous PI gauge from readily available data.

4.4. Feature Importance and Sensitivity Analysis

A key advantage of the CatBoost algorithm is its inherent ability to quantify feature importance, which provides a critical bridge between the machine learning model and petroleum engineering principles. The analysis, illustrated in **Figure 4.4**, identified production rates, specifically for oil, gas, and water as the most influential predictors of PI. This aligns perfectly with the fundamental definition of PI as a ratio of flow rate to pressure drawdown.

To validate these findings and mitigate potential bias, a model-agnostic permutation sensitivity analysis was performed. The results, shown in **Figure 4.5**, largely corroborated the intrinsic feature importance but revealed a subtle nuance: oil and gas production rates had the most substantial impact on the model's predictive accuracy when perturbed. This outcome robustly highlights that hydrocarbon flow rates are the primary drivers of PI in this dataset, a conclusion that resonates strongly with established petroleum engineering understanding and validates the model's physical consistency.

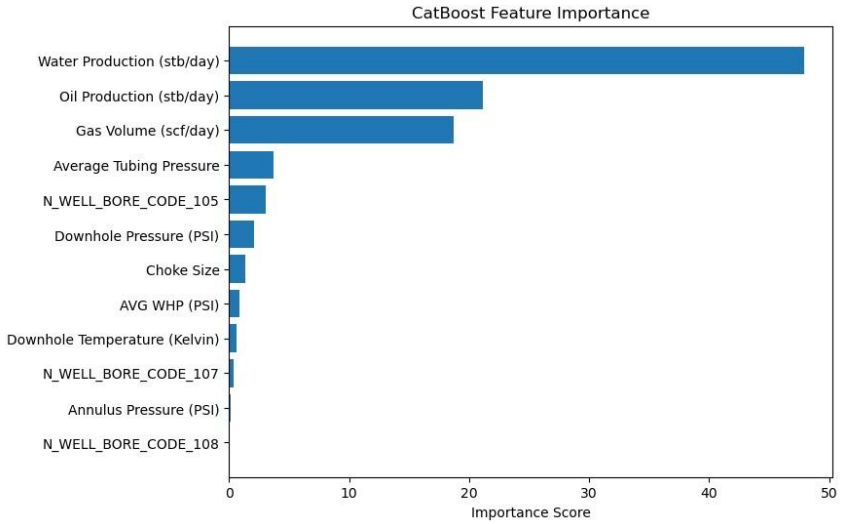

Figure 4.4. Feature Importance barchart for CatBoost, showing the relative influence of input variables.

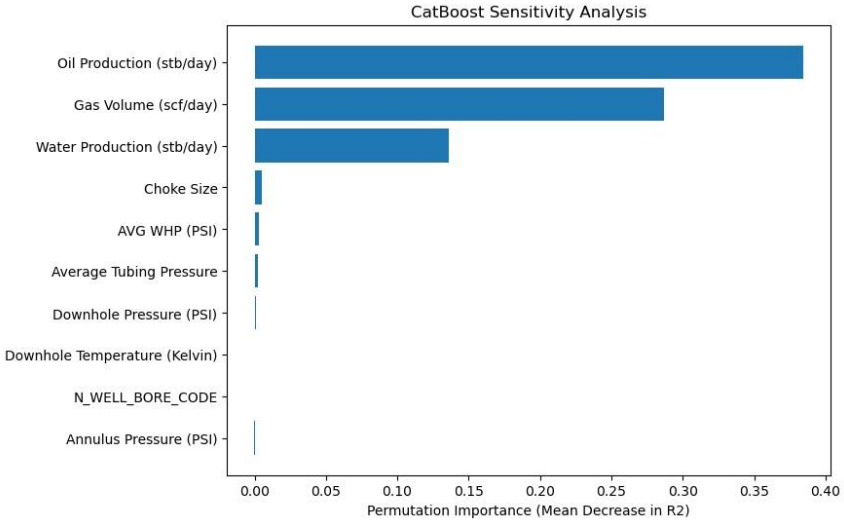

Figure 4.5. Permutation Importance plot for CatBoost, showing the drop in model accuracy when a specific feature is shuffled.

The collective evidence from this analysis positions CatBoost as a robust and physically consistent model for predicting well Productivity Index directly from historical production data. Its high predictive accuracy, unbiased and normally distributed residuals, and logically sound

feature importance rankings collectively confirm that the model has learned meaningful, representative patterns from the data.

An important consideration is the inherent mathematical dependency of PI on flow rates. Given that oil and gas production were the most critical features, a portion of the model's exceptional performance is inherently linked to this direct physical relationship. This does not diminish the model's value but underscores a crucial caveat for its practical application: the model's forecasting reliability is contingent upon the accuracy and timeliness of the production rate data inputs.

CHAPTER FIVE

5.0 CONCULSION AND RECOMMENDATIONS

5.1 CONCLUSION

This study compared three machine learning regressors for predicting the well productivity index (PI) from daily production and pressure records. The models evaluated were RandomForest, XGBoost and CatBoost. Results show a clear performance ordering. RandomForest produced the weakest fit, with large errors and low explanatory power (see Table 1). XGBoost improved substantially over RandomForest, capturing much of the nonlinear structure in the data but still leaving meaningful residual error. CatBoost delivered the strongest performance, with the lowest MAE and RMSE and an R2 near 0.95. The parity plots show this ordering visually: RandomForest points are broadly scattered (Figure 1a), XGBoost points are tighter but still dispersed at the tails (Figure 1b), and CatBoost points cluster closely on the 1:1 line (Figure 1c).

Residual analysis reinforces the same conclusion. RandomForest residuals were widely spread and showed structure, signaling model weakness in several operating regimes. XGBoost residuals improved but retained some patterning at high and low PI values. CatBoost residuals were tightly centered around zero and showed no strong trend with predicted value, which suggests no systematic over or under prediction across the test range (Figures 2a, 2b, 2c and Figure 3). Time series comparisons at the well level showed that CatBoost also tracked temporal PI changes far better than the other two models. In most wells the predicted PI followed the ups and downs of the measured PI, which is important for monitoring and operational decision making (Figure 4).

The interpretability analyses give context for why CatBoost performed so well. Feature importance from CatBoost indicates that production rates, specifically oil, gas and water outputs,

are the dominant predictors (Figure 5). Permutation sensitivity supports this but shifts the ordering slightly, showing that oil and gas volumes have the largest effect on predictive performance while water production also contributes strongly (Figure 6). This result is consistent with fundamental reservoir physics because PI is directly linked to flow and pressure drawdown. At the same time, it highlights a methodological caution: using production volumes as predictors partly encodes the numerator of PI into the input space. That mathematical coupling likely boosts predictive accuracy, and it must be acknowledged when interpreting model causality and when deploying models for forecasts where production readings may be delayed or noisy.

Finally, the study's limitations must temper the conclusions. The data come from five wells and about 7,000 daily points, which is adequate for initial testing but limited in scope for wide generalization. The PI target also relied on proxies for reservoir pressure in many cases, introducing additional uncertainty. These constraints mean that while CatBoost appears robust on the present dataset, further validation on larger and more diverse datasets is required before wider deployment.

5.2 RECOMMENDATIONS

CatBoost should be adopted as the working model for PI estimation in the present dataset because it gives the most accurate and stable predictions. Use it for tasks where quick and consistent PI estimates are useful, for example well ranking, screening for intervention, short term monitoring and flagging wells that deviate from their expected performance. When using the model operationally, always present prediction intervals or uncertainty bounds (see the Uncertainty and Monitoring paragraphs below) and do not rely solely on a single point prediction for high-cost decisions.

Because oil and gas rates strongly influence predictions, run controlled tests to quantify dependency and avoid circular inference. Specifically, train and evaluate two alternate models: one that excludes raw oil and gas production and uses only pressure and derived features (water cut, total fluid rate, pressure differentials), and a second that uses only lagged or aggregated production features (for example rolling averages or normalized flow). Compare performance metrics and stability to measure how much predictive power comes from direct coupling to PI versus from independent signals. Include a leave-one-feature-out experiment and report the change in R2 and MAE for each removed variable. This will make explicit how much model accuracy depends on potentially circular predictors.

Include SHAP value analyses and partial dependence plots for the top features so that you can explain local and global model behavior in the field. Use permutation importance, as you have, and add bootstrapped permutation runs to obtain confidence bounds for importance scores. Perform residual stratification analyses to check whether errors correlate with operating regimes such as high water cut, high choke settings, or late life wells. Use backtesting and forward rolling validation to confirm model stability under time shifting data.

Wherever possible, collect or collate true static reservoir pressure or recent well test measurements and use them to calculate PI for a subset of the wells. Use those higher quality PI labels to benchmark the proxy-based PI calculations you used here. If the benchmark shows systematic bias in your proxy, incorporate a correction factor or a simple calibration model for the proxy before training production models.

Add geologic and completion data when available. Suggested fields to collect and include are formation porosity and permeability, net pay, completion interval, perforation counts, stimulation history, and pump settings. These variables often explain structural differences between wells

that production and pressure time series alone cannot resolve. Where acquisition is expensive, prioritize static attributes that are known to be highly predictive, such as permeability or completion type.

Extend the training data to include more wells, ideally from multiple fields under different operating conditions. This will improve the model's generalization and allow re-assessment of model choice. Test model transfer by training on one field and testing on another. If transfer performance drops substantially, consider field-specific models or domain adaptation techniques.

Implement calibrated prediction intervals using either quantile models, model ensembles or conformal prediction so that users know the confidence around each PI estimate. For operational use, set conservative decision thresholds. For example, require that a predicted PI improvement be larger than the model's two-sigma uncertainty band before allocating capital for a stimulation. Document those thresholds in your deployment guide.

Put the model behind a lightweight API or batch job and create a dashboard that tracks key metrics daily. Monitor data quality indicators such as percent missing sensors, changes in measurement distributions, and drift in input feature statistics. Track model performance over time using a set of rolling metrics such as 30 day MAE and R2 computed on the most recent held out wells. Set retraining triggers based on these metrics, for example retrain when rolling MAE increases by 20 percent or when the distribution of a key feature shifts beyond a pre-set threshold. Maintain versioned model snapshots and store the training data and preprocessing steps for auditing.

Before full operational deployment, run a parallel trial where model predictions are made but not acted on, and compare them to routine well test PI estimates and engineer judgment. Solicit

domain expert review on a sample of flagged wells, specifically those where model predictions differ substantially from historical estimates. This human-in-the-loop step reduces risk and builds trust.

Provide clear documentation of the PI target definition, the proxy choices and their parameters, the feature engineering recipes, hyperparameter settings, and the cross-validation strategy used. In particular, record which wells were held out for final testing. This makes your work reproducible and allows future teams to audit model behavior and assumptions.

Evaluate hybrid models that fuse physics-based IPR or nodal analysis with machine learning. Investigate causal feature selection methods to separate mathematical coupling from causal drivers. Explore model compression or surrogate models for fast inference on embedded platforms. Finally, pursue extended uncertainty modelling using Bayesian or quantile approaches to give operators probabilistic guidance when making expensive decisions.

5.3 CONTRIBUTION TO KNOWLEDGE:

This study contributes to existing knowledge by developing a data-driven framework for modeling the well productivity index (PI) using machine learning algorithms. Unlike conventional empirical and analytical methods that rely on simplifying assumptions and limited parameters, this research leverages historical well and reservoir data to capture complex, non-linear relationships influencing well productivity. The application and comparative evaluation of selected machine learning models provide improved prediction accuracy and robustness in PI estimation. The findings demonstrate the potential of machine learning as a reliable alternative tool for well performance evaluation and decision-making in petroleum engineering, particularly in data-rich environments.

REFERENCES

Adewale, M.D., Adeyanju, I.A., Oju, J., Ubadike, O.C., Muhammed, U.I. and Omisakin, S.T.

(2024) 'Ensemble machine learning methods to predict oil production', in Innovations and Interdisciplinary Solutions for Underserved Areas. 7th International Conference, InterSol 2024, Dakar, Senegal, Proceedings. Springer.

Adewale, T. et al. (2025) 'Machine learning approaches for forecasting Nigerian oil production', Journal of Petroleum Data Science, 12(3), pp. 145–162.

Ahmed, U. et al. (2020) 'Applications of machine learning in petroleum production forecasting', SPE Reservoir Evaluation & Engineering, 23(5), pp. 1–14.

Alarifi, S., Al Nuaim, S. and Abdulraheem, A. (2015) 'Productivity index prediction for oil horizontal wells using different artificial intelligence techniques', in 19th Middle East Oil and Gas Show and Conference, MEOS 2015. Society of Petroleum Engineers, pp. 1972–1984. doi: 10.2118/172729-MS.

Alharbi, O.Q. and Alarifi, S.A. (2023) 'Productivity index prediction for single-lateral and multilateral oil horizontal wells using machine learning techniques', ACS Omega, 8(7), pp. 7201– 7210. doi: 10.1021/acsomega.3c00289.

Al-Mashhad, A.S. and Alarifi, S.A. (2016) 'Multilateral wells evaluation utilizing artificial intelligence', in Abu Dhabi International Petroleum Exhibition and Conference. Society of Petroleum Engineers. doi: 10.2118/183508-MS.

Al-Mudhafar, W. (2017) 'Data-driven prediction of reservoir properties using machine learning methods', Journal of Petroleum Science and Engineering, 157, pp. 577–590.

Avbovbo, A.A. (1978) 'Tertiary lithostratigraphy of Niger Delta', AAPG Bulletin, 62(2), pp. 295–300.

Doust, H. and Omatsola, E. (1990) 'Niger Delta', in Edwards, J.D. and Santogrossi, P.A. (eds.) Divergent/Passive Margin Basins (AAPG Memoir 48). American Association of Petroleum Geologists, pp. 201–238.

Escobar, F.H., Saavedra, N., Aranda, R. and Herrera, J. (2004) 'An improved correlation to estimate productivity index in horizontal wells', in SPE Asia Pacific Oil and Gas Conference and Exhibition.

Society of Petroleum Engineers. doi: 10.2523/88540-MS.

Evamy, B.D., Haremboure, J., Kamerling, P., Knaap, W.A., Molloy, F.A. and Rowlands, P.H. (1978) 'Hydrocarbon habitat of Tertiary Niger Delta', AAPG Bulletin, 62(1), pp. 1–39.

Fan, D., Lai, S., Sun, H., Yang, Y., Yang, C., Fan, N. and Wang, M. (2025) 'Review of machine learning methods for steady state capacity and transient production forecasting in oil and gas reservoir', Energies, 18(4), p. 842. doi: 10.3390/en18040842.

Gruzdev, A., Babov, V., Simonov, Y., Kosarev, A., Simon, I., Koryabkin, V. and Semenikhin, A. (2020) 'The well productivity index determination based on machine learning approaches', in First EAGE Digitalization Conference and Exhibition, vol. 2020. European Association of Geoscientists & Engineers, pp. 1–5. doi: 10.3997/2214-4609.202032094.

Haghighat, S.A., Mohaghegh, S., Gholami, V. and Moreno, D.J. (2014) 'Production analysis of a Niobrara field using intelligent top-down modeling', in SPE Western North American and Rocky Mountain Joint Regional Meeting. Society of Petroleum Engineers. doi: 10.2118/169573-MS.

Jayeola, I., Olusola, B. and Orodu, K.B. (2022) 'Machine learning prediction versus decline curve prediction: A Niger Delta case study', in SPE Nigeria Annual International Conference and Exhibition. Society of Petroleum Engineers. doi: 10.2118/211956-MS.

Jin, Y., Guo, K., Gao, X. and Li, Q. (2024) 'Tight oil well productivity prediction model based on neural network', Processes, 12(10), p. 2088. doi: 10.3390/pr12102088.

Joshi, S.D. (1988) 'Augmentation of well productivity with slant and horizontal wells', Journal of Petroleum Technology, SPE-15375, pp. 729–743.

Kulke, H. (1995) 'Nigeria', in Kulke, H. (ed.) Regional Petroleum Geology of the World, Part II: Africa, America, Australia and Antarctica. Gebrüder Borntraeger, pp. 143–172.

Mohaghegh, S., Grujic, O., Zargari, S. and Kalantari-Dahaghi, A. (2012) 'Top-down, intelligent reservoir modelling of oil and gas producing shale reservoirs: Case studies', International

Journal of Oil, Gas and Coal Technology, 5(1), pp. 3–28. doi: 10.1504/IJOGCT.2012.044175.

NNPC (2023) Annual Statistical Bulletin. Nigerian National Petroleum Company Limited, Abuja.

Osisanya, S.O., Ayokunle, A.T., Ghosh, B. and Suboyin, A. (2021) 'Modified horizontal well productivity model for a tight gas reservoir subjected to non-uniform damage and turbulence', Energies, 14(24), p. 8334. doi: 10.3390/en14248334.

Rahmanifard, H. and Gates, I. (2024) 'A comprehensive review of data-driven approaches for forecasting production from unconventional reservoirs: Best practices and future directions', Artificial Intelligence Review, 57(8), p. 213. doi: 10.1007/s10462-024-10865-5.

Stacher, P. (1995) 'Present understanding of the Niger Delta hydrocarbon habitat', in Oti, M.N. and Postma, G. (eds.) Geology of Deltas. A.A. Balkema, Rotterdam, pp. 257–268.

Vikara, D. and Khanna, V. (2022) 'Application of a deep learning network for joint prediction of associated fluid production in unconventional hydrocarbon development', Processes, 10(4), p. 740. doi: 10.3390/pr10040740.

Vogel, J.V. (1968) 'Inflow performance relationships for solution-gas drive wells', Journal of Petroleum Technology, 20(SPE-1476-PA).

Wikipedia contributors (2025) Niger Delta Basin (geology). Available at: https://en.wikipedia.org/wiki/Niger_Delta_Basin_(geology) (Accessed: 9 September 2025).