COMPARATIVE ANALYSIS OF MACHINE LEARNING-DRIVEN PORE PRESSURE PREDICTION MODELS

200834M M.N.M.Naweed

200849M S.Suheerman

200811P S.M.D.K.R.Dilkushan

Research Project Thesis submitted in partial fulfillment of the requirements for the degree Bachelor of Science in Engineering

Department of Earth Resources Engineering

University of Moratuwa

Sri Lanka

July 2025

DECLARATION PAGE OF THE CANDIDATE

We declare that this is our own work and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any University or other institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

……. Date: ……

M.N.M.Naweed (200834M)

……. Date: ……

S.Suheerman (200849M)

……. Date: ……

S.M.D.K.R. Dilkushan (200811P)

DECLARATION PAGE OF THE SUPERVISOR

We have supervised and accepted this thesis for the submission of the degree.

……. Date: ……

Mrs. M.A.M.D.G. Wikrama

……. Date: ……

Dr. S. Thiruchittampalam

DEDICATION

We dedicate this research project report to,

Our supervisors, Mrs. M.A.M.D.G. Wikrama and Dr. S. Thiruchittampalam,

Head of the Department of Earth Resources Engineering, Dr. S.P. Chaminda,

Final year research coordinator, Mrs. M.A.M.D.G. Wikrama, lectures and staff members of the Department of the Earth Resources Engineering.

All the students, professionals, and researchers engaged in the field of petroleum engineering, who may derive value from the outcomes of this research.

ACKNOWLEDGMENT

We would like to acknowledge and extend our heart gratitude to the following personnel who have made the completion of this possible;

Project supervision panel: Mrs. M.A.M.D.G. Wikrama and Dr. S. Thiruchittampalam, lecturers in the Department of Earth Resources Engineering for the constant reminders, much needed motivation, vital encouragement and support provided throughout the period.

We would like to convey our sincere thanks to Dr. S.P. Chaminda, Head of the Department and Mrs. M.A.M.D.G. Wikrama the final year research project coordinator.

We are then thankful to officials of the Library, University of Moratuwa for their extra ordinary support given to purchase research papers regarding the subject.

We would also like to acknowledge and extend our appreciation to the contributor of the GitHub repository from which the dataset used in this research was obtained.

We would like to thank the non-academic staff of the department for kind support supplied at any time.

At last, but not least our grateful thank is expressed to everyone including our dearest

friends who helped us to make this project a success.

ABSTRACT

Accurate pore pressure prediction is essential for maintaining the safe mud weight window during drilling operations. Conventional methods, which rely on simplified empirical assumptions, often lack reliability in complex geologies because they fail to capture the multivariate and non-linear relationships inherent in subsurface data.

Machine learning (ML) provides a data-driven alternative capable of modeling these complexities. However, the practical application of ML is often inconsistent, as there is a lack of systematic understanding of how specific preprocessing choices, such as outlier treatment and feature selection, impact the performance of different algorithms. This study aims to resolve this ambiguity by identifying the optimal combination of preprocessing strategy and ML model for this task.

The methodology employed a rigorous four-scenario experimental design to systematically evaluate the effects of outlier capping and the removal of multicollinear features. Six machine learning algorithms, ranging from foundational models like K-Nearest Neighbors to advanced ensembles like XGBoost, were trained and optimized within each of the four distinct data configurations. Model efficacy was assessed using R-Squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

The results conclusively identify a Tuned XGBoost model as the top performer (R^2 = 0.9789), achieving this optimal accuracy on the raw, unprocessed dataset. The analysis further demonstrates that removing features based on linear correlation was detrimental to advanced model performance, and that the necessity of outlier treatment is highly algorithm-dependent, being critical for sensitive models like Support Vector Regression but unnecessary for robust ensembles.

This research concludes that while a specific champion model may not be universal, the findings reveal a universally applicable principle for model development: the optimal data preparation strategy is not a fixed routine but must be tailored to the chosen algorithm's inherent robustness. This study provides a valuable methodological framework for developing more reliable, context-aware predictive models in practical oilfield applications.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Full Form
ANN	Artificial Neural Network
API	American Petroleum Institute
EDA	Exploratory Data Analysis
FPWD	Formation Pressure While Drilling
GR	Gamma Ray
IQR	Interquartile Range
KNN	K-Nearest Neighbors
LIME	Local Interpretable Model-agnostic Explanations
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
NCT	Normal Compaction Trend
PP	Pore Pressure
R^2	R-Squared (Coefficient of Determination)
ReLU	Rectified Linear Unit
RF	Random Forest
RFE	Recursive Feature Elimination
RHOB	Bulk Density
RMSE	Root Mean Squared Error
SHAP	SHapley Additive exPlanations
SVR	Support Vector Regression
Vp	P-wave (Compressional) Velocity
Vsh	Volume of Shale
XGBoost	Extreme Gradient Boosting

CHAPTER 01: INTRODUCTION

1.1. Background and Motivation

In subsurface geomechanics, pore pressure exerted by fluids within the pore spaces of rock formations is a fundamental parameter that governs the mechanical behavior of reservoirs. Its relationship with the total overburden stress and the rock matrix stress is defined by Terzaghi's principle of effective stress, which is the foundation of modern rock mechanics (Zhang & Yin, 2017).

The primary application of pore pressure prediction is in determining the safe mud weight window. This window is the operational margin between the pore pressure gradient and the fracture pressure gradient. If the drilling mud weight is too low (below the pore pressure), formation fluids can enter the wellbore, leading to a "kick" and potentially a catastrophic blowout. Conversely, if the mud weight is too high (exceeding the fracture pressure), it can induce fractures in the formation, resulting in costly "lost circulation" events where drilling fluid is lost to the reservoir (Ramatullayev et al., 2019). Both scenarios pose significant risks to personnel, the environment, and the financial success of a drilling campaign.

For decades, industry has relied on conventional methods to predict pore pressure, such as those developed by Eaton (1975) and Bowers. These methods correlate petrophysical properties derived from well logs (e.g., sonic velocity or resistivity) with expected compaction trends. However, their reliability is often limited. These models are built on simplifying assumptions, such as a single, well-defined normal compaction trend, and require extensive local calibration to be effective. In geologically complex environments with multiple overpressure mechanisms, their predictive power diminishes significantly.

1.2. Machine Learning for Pore Pressure Prediction

The limitations of conventional models, particularly their difficulty in handling non-linear and multivariate relationships, have created a clear need for more adaptive, data-driven approaches. ML has emerged as a suitable alternative, offering the ability to learn these complex relationships directly from data without reliance on predefined empirical equations (Li et al., 2023a).

A primary capability of a data-driven ML model is its ability to approximate complex functions without explicit knowledge of the underlying physical laws. This is advantageous in geologically complex systems where physical relationships are too intricate to be accurately modeled by simplified equations. However, this data-driven nature also represents a trade-off; the lack of physical constraints can lead to "black-box" models that are difficult to interpret, a challenge that has given rise to the separate field of physics-informed machine learning. For the scope of this study, the focus remains on evaluating the full potential of the data-driven approach.

1.3. Problem Statement

While numerous studies have demonstrated the successful application of ML for pore pressure prediction (e.g., Feng et al., 2024; Sanei et al., 2024), a common theme in the existing literature is the adoption of a single, fixed data preprocessing pipeline without a systematic investigation into its necessity or its differential impact across various algorithm types (Li et al., 2023a). This creates a methodological ambiguity, as the optimal preprocessing strategy for a robust ensemble model may differ significantly from that for a sensitive kernel-based model. Therefore, the problem this study addresses is the need to move beyond a simple model-to-model comparison and to rigorously determine the optimal combination of data preparation techniques and machine learning algorithms for pore pressure prediction.

1.4.Research Objectives

This thesis aims to address the stated problem through the following specific objectives:

- To evaluate and compare the performance of six distinct machine learning algorithms (Decision Tree, K-Nearest Neighbors, Support Vector Regression, Random Forest, XGBoost, and Artificial Neural Networks) for pore pressure prediction.
- To investigate the impact of feature selection (based on multicollinearity) technique on the performance of each model.
- To systematically determine the optimal combination of preprocessing strategy and machine learning model for the given multi-well dataset.

- To interpret the champion model by identifying the most influential petrophysical parameters using feature importance analysis.

1.5. Research Questions

To achieve the above objectives, this research seeks to answer the following key questions:

- Which algorithm provides the highest predictive accuracy and robustness for pore pressure prediction on the given dataset?
- How do data preprocessing techniques affect model performance?
- What is the optimal end-to-end pipeline (i.e., the combination of preprocessing strategy and tuned algorithm) for this prediction task?
- According to the best model, which features are the most significant drivers of pore pressure variation?

1.6. Thesis Outline

This thesis is structured into five chapters. Chapter 1 introduces the research background, objectives, and questions. Chapter 2 provides a comprehensive review of the relevant literature, covering the fundamentals of pore pressure, conventional prediction methods, the theoretical basis of the selected machine learning algorithms, and identifies the key research gap. Chapter 3 details the complete methodology, including the dataset description, the four-scenario experimental design, and the model implementation and evaluation framework. Chapter 4 presents and discusses the results of the comparative analysis in detail. Finally, Chapter 5 concludes the thesis by summarizing the key findings, discussing the contributions and limitations of the study, and offering recommendations for future work.

CHAPTER 02: LITERATURE REVIEW

This chapter provides a comprehensive review of the theoretical principles and practical methodologies relevant to pore pressure prediction. It begins by establishing the fundamental concepts of geomechanics, followed by a critical evaluation of conventional empirical methods for pore pressure estimation. Subsequently, it delves into the theoretical underpinnings of the machine learning algorithms employed in this study. The chapter culminates in a review of existing applications of machine learning in geosciences, identifying a critical research gap that this thesis aims to address.

2.1. Fundamentals of Pore Pressure

In subsurface formations, the total downward pressure exerted by the weight of overlying rock and fluids is known as the overburden stress or vertical stress (σ_v). This stress is supported by two components: the rock matrix itself and the fluid contained within its pore spaces. The pressure of this pore fluid is termed the formation pore pressure (Pp). The portion of the overburden stress supported by the rock grain-to-grain framework is known as the effective stress (σ_e).

The relationship between these components was famously defined by Karl Terzaghi and is the foundational principle of modern geomechanics (Azadpour et al., 2015). Terzaghi's principle of effective stress is expressed as:

$$\boldsymbol{\sigma_e = \sigma_v - \alpha Pp}$$

where α is the Biot effective stress coefficient, often assumed to be unity in simplified contexts. This principle is crucial because it is the effective stress, not the total stress, that controls the mechanical behavior of rocks, including their compaction and strength.

Under normal conditions, as sediments are buried and compacted, pore fluids are expelled, and the pore pressure increases hydrostatically. This predictable relationship forms a **Normal Compaction Trend (NCT)**. However, various geological mechanisms can disrupt this process, preventing fluids from escaping and causing pore pressure to rise above the hydrostatic gradient, a condition known as **overpressure**. Key mechanisms for overpressure generation include disequilibrium compaction, where rapid sediment

deposition outpaces fluid expulsion, and fluid expansion mechanisms, such as aqua thermal pressuring or hydrocarbon generation (Kaiser, 2017; Zhang & Yin, 2017). Accurate prediction of these over pressured zones is paramount for safe drilling operations.

2.2. Conventional Pore Pressure Prediction Methods

For decades, the petroleum industry has relied on conventional predictive methods to estimate pore pressure profiles before and during drilling. These methods are primarily empirical, correlating measurable petrophysical properties from well logs or seismic data to changes in effective stress. They operate on the central assumption that any deviation from a normal compaction trend is attributable to an increase in pore pressure.

Among the most influential are Eaton's method (Eaton, 1975) and the D-exponent method. Eaton's method is particularly widespread and versatile, utilizing resistivity, sonic transit time, or conductivity logs. The method calculates pore pressure by comparing the observed log value at a given depth to the value expected from the NCT at that same depth. The mathematical basis is an extension of Terzaghi's principle, expressed as:

$$\boldsymbol{Pp} = \boldsymbol{\sigma}_{\mathrm{v}} - (\boldsymbol{\sigma}_{\mathrm{v}} - \boldsymbol{Ph}) * (\boldsymbol{X}_{\mathrm{n}} / \boldsymbol{X}_{\mathrm{o}})^{\mathrm{n}}$$

where Ph is the hydrostatic pressure, X_n is the log value from the NCT, X_o is the observed log value, and *n* is the Eaton exponent, an empirically derived constant specific to a basin or formation.

While foundational, these traditional methods possess significant limitations. Their primary weakness is the reliance on simplified assumptions, such as lithological homogeneity and the existence of a single, well-defined NCT, which often do not hold true in geologically complex settings (Ogbu et al., 2024). In regions with complex faulting, salt diapirism, or significant lateral variations in rock properties, these assumptions can lead to highly misleading predictions. Furthermore, the accuracy of these methods is heavily dependent on the quality and resolution of the input data, which is often a challenge (Li et al., 2023a). Most critically, these models often fail to account for overpressure mechanisms unrelated to compaction, such as fluid expansion or lateral pressure transfer, rendering them unsuitable for a wide range of geological environments (Tan et al., 2020). These

limitations underscore the need for more adaptive and data-driven models, creating a clear pathway for the application of machine learning.

2.3. Machine Learning Algorithms for Regression

ML offers a powerful alternative to conventional methods by learning complex, non-linear relationships directly from data without relying on predefined physical equations. This study evaluates six distinct regression algorithms, each with a unique theoretical basis.

2.3.1. Decision Tree

A Decision Tree is a non-parametric supervised learning method that predicts a target value by learning simple, hierarchical decision rules from data features. In petrophysics, its primary value lies in its high interpretability, where the learned rules (e.g., "if GR > 90 API and RHOB < 2.2 g/cm³...") can be directly compared to geological knowledge. However, due to its tendency to overfit complex well log data, it is rarely used as a primary predictive model and more often serves as a baseline for comparison or as the fundamental building block for more advanced ensemble methods (James et al., 2021).

2.3.2. K-Nearest Neighbors (KNN)

KNN is a distance-based, non-parametric algorithm that makes predictions based on the average value of the 'K' most similar data points in the training set. Its application in petrophysics is less common than other methods, primarily because its performance is highly sensitive to the scaling of input features with disparate units (e.g., sonic velocity in km/s vs. density in g/cm³). Furthermore, its "lazy learning" approach, which requires storing the entire training dataset for prediction, can be inefficient with the large, continuous datasets typical of multi-well studies (Hair et al., 2019).

2.3.3. Support Vector Regression (SVR)

SVR is a kernel-based method that has been successfully applied to predict various petrophysical properties, including shear wave velocity and porosity (Al-Anazi & Gates, 2010). Unlike models that minimize squared error, SVR fits a function within a specified margin (ε), making it robust to minor noise in well log data. Its strength lies

in the "kernel trick," which allows it to model highly non-linear relationships between logs without becoming computationally prohibitive. However, studies consistently note that its performance is critically dependent on the careful.

2.3.4. Random Forest

Random Forest is an ensemble method that has become a benchmark algorithm in many geoscientific applications due to its high accuracy and robustness. In pore pressure prediction, it is frequently used to model the complex, non-linear relationships between a full suite of well logs and pressure data (Li et al., 2023a). Its ensemble nature, which averages the predictions of many individual trees, makes it highly resistant to the overfitting and noise that can affect single Decision Trees, a common problem when dealing with real-world well log data.

2.3.5. XGBoost (Extreme Gradient Boosting)

XGBoost is an advanced gradient boosting algorithm that has recently emerged in numerous pore pressure prediction studies. Its sequential, error-correcting approach allows it to capture subtle and complex interactions between petrophysical parameters that other models may miss. For example, Feng et al. (2024) and Sanei et al. (2024) both demonstrated the performance of XGBoost over other models in complex geological settings, attributing its success to its built-in regularization features that effectively prevent overfitting, even with large and high-dimensional well log datasets.

2.3.6. Artificial Neural Networks (ANN)

ANNs have a long history of application in petrophysics for tasks such as lithology classification and permeability prediction. In the context of pore pressure, ANNs are valued for their ability to function as universal approximators, capable of modeling any arbitrarily complex, non-linear relationship between input logs and the target pressure (Abdelaal et al., 2021). Studies often employ multi-layer perceptron architectures to learn hierarchical patterns from the data. However, literature also consistently highlights their "black box" nature and their sensitivity to architectural choices and hyperparameter tuning as significant practical challenges.

2.4. Previous Applications for Machine Learning and Identification of the Research Gap

The application of machine learning to pore pressure prediction has gained significant attention, with numerous studies demonstrating its performance advantages over conventional methods, particularly in geologically complex settings (Ahmed et al., 2019; Li et al., 2023a). This body of research has explored a variety of algorithms, each with distinct characteristics. For instance, Artificial Neural Networks (ANNs) are frequently employed for their ability to model highly non-linear relationships, though their computational expense and sensitivity to hyperparameter tuning are well-documented challenges (Abdelaal et al., 2021). In parallel, ensemble methods like Random Forest are often utilized for their strong generalization capabilities and inherent robustness to variations in data quality (Li et al., 2023a). More recent studies have highlighted the performance of gradient boosting models like XGBoost, attributing their success to advanced regularization features that prevent overfitting on large, complex datasets (Sanei et al., 2024; Feng et al., 2024).

Despite these successful applications, the existing literature reveals several limitations and an important, unaddressed research gap. A primary challenge consistently cited is the dependency of ML models on the quality of the data input. For example, Li et al. (2023b) note that real-world well log datasets are often affected by noisy data—which can refer to measurement errors from logging tools, environmental effects, or small-scale geological heterogeneities that do not conform to the general trend. The presence of such data can significantly hinder the performance of predictive models. This dependency on data quality necessitates a robust data preprocessing strategy to ensure that the relationships learned by the model are valid and generalizable. Another common criticism is the "black box" nature of many ML models, which can reduce confidence in their predictions compared to transparent empirical equations (Feng et al., 2024).

While many studies acknowledge the need for preprocessing, they often treat it as a preliminary, fixed step. A typical study might apply outlier removal and feature selection and then proceed to compare several models on this single, cleaned dataset. However, there is a significant lack of research that systematically investigates the interaction between

different preprocessing strategies and the performance of various ML models. The implicit assumption in the literature is that a universal preprocessing pipeline is always beneficial. It is unclear whether robust algorithms like XGBoost derive the same benefit from outlier capping as sensitive models like SVR. Similarly, it is not well-documented whether removing linearly correlated features aids or harms complex models that might be capable of extracting unique non-linear information from them.

Therefore, the research gap this thesis addresses is the absence of a comprehensive, comparative study that evaluates machine learning models across a matrix of well-defined preprocessing scenarios. By testing all models on data that is (1) raw, (2) only outlier-capped, (3) only feature-selected, and (4) fully preprocessed, this study aims to move beyond merely identifying the best-performing model and instead explore how different combinations of data preparation techniques and machine learning algorithms affect model performance. This approach will lead to a refined and practically applicable framework for developing pore pressure prediction models.

CHAPTER 03: METHODOLOGY

The methodology employed in this study is illustrated in Figure 3.1. The workflow consists of two primary stages: first, an initial data processing phase to create four distinct experimental scenarios, and second, a standardized modeling pipeline through which each scenario is evaluated for a comprehensive comparative analysis.

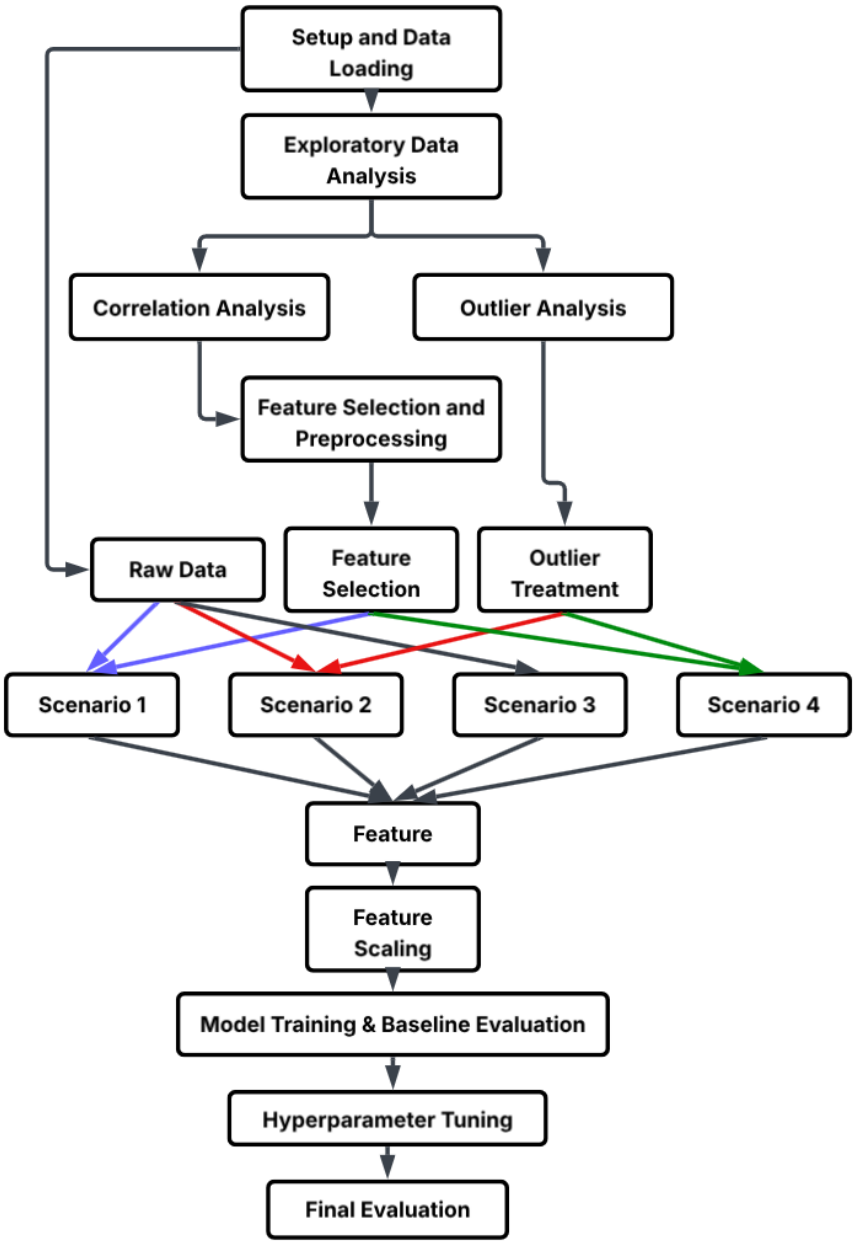


Figure 3. 1: Methodological Workflow for the Four-Scenario Comparative Analysis

3.1. Dataset Description

The foundation of this study is a publicly available dataset containing well-log measurements from eight distinct wells. The data was sourced via a GitHub repository (Tammy Reservoir, 2022) and is reported to originate from an oil field located in the United States. However, specific details regarding the geological basin, field name, or precise geographical coordinates were not provided in the accompanying documentation.

The absence of specific provenance is a known limitation of this study, which will be discussed further in Chapter 5. However, the dataset's value for this research lies in its comprehensive suite of standard petrophysical logs, which are directly relevant to pore pressure estimation. The primary objective of this thesis to systematically compare the performance and preprocessing sensitivities of various machine learning algorithms is therefore fully supported by the richness and internal consistency of the provided data. The data from all eight wells were aggregated into a single master dataset to provide a substantial basis for training and validating robust machine learning models.

The dataset includes nine predictor variables (features) and one target variable. A detailed description of each variable is provided in Table 3.1.

Table 3. 1: Description of Variables in the Dataset

Variable Name (Unit)	Description	Role in Model
Depth (m)	True vertical depth, representing the vertical distance from the surface.	Predictor Feature
GR (API)	Gamma ray log, measures natural radioactivity to identify lithology (e.g., shales vs. sands).	Predictor Feature
RHOB (g/cm³)	Bulk density log, measures the overall density of the formation, including rock and fluid.	Predictor Feature
Vp (km/s)	P-wave (compressional) sonic velocity measures the speed of sound through the formation.	Predictor Feature
Vsh (v/v)	Volume of shale, an interpreted log representing the fraction of shale in the formation.	Predictor Feature

Caliper (in)	Caliper log, measures the diameter of the wellbore, indicating potential washouts or swelling.	Predictor Feature
Porosity (%)	The percentage of void space within the rock, capable of holding fluids.	Predictor Feature
Resistivity (ohm.m)	Formation resistivity log, measures the rock's resistance to electrical current, sensitive to fluid type.	Predictor Feature
Stress (MPa)	Overburden stress, the pressure exerted by the weight of overlying formations.	Predictor Feature
PP (MPa)	Pore pressure, the pressure of the fluid within the rock's pore spaces.	Target Variable

A preliminary statistical summary of the raw, combined dataset is presented in Table 3.2. This summary reveals a wide range of values and differing scales across the various logs, highlighting the necessity for feature scaling prior to model training, which will be detailed in Section 3.4.3.

Table 3. 2: Descriptive Statistics of the Raw Dataset

Feature	Count	Mean	Std Dev	Min	25%	50% (Median)	75%	Max
DEPTH	11494	139.71	74.66	5.95	78.18	132.92	191.03	335.88
GR	11494	92.09	8.95	42.27	87.60	92.63	97.89	114.99
RHOB	11494	1.81	0.14	1.08	1.74	1.82	1.90	2.12
Vp	11494	-25.70	162.89	-999.25	1.51	1.54	1.58	1.72
Vsh	11494	0.66	0.45	-0.16	0.58	0.67	0.76	46.21
Caliper	11494	10.09	0.63	9.42	9.78	9.97	10.25	16.38
Porosity	11494	59.44	6.69	41.17	54.94	58.79	63.03	98.85
Resistivity	11494	0.99	0.27	0.36	0.84	0.95	1.09	2.87
Stress	11494	2.54E+06	1.45E+06	6.65E+04	1.38E+06	2.35E+06	3.42E+06	6.93E+06
PP (Pore Pressure)	11494	1840.08	219.43	1416.00	1668.00	1823.00	2006.00	2314.00

The descriptive statistics presented in Table 3.1.2 provide several critical insights that are foundational to the methodology of this study.

The table quantitatively confirms the wide variation in the scales and units of measurement across the predictor variables. For instance, the Stress feature has a mean value on the order of millions (2.54E+06), while RHOB has a mean of 1.81. This vast disparity underscores the methodological necessity of feature scaling, as detailed in Section 3.4. Without scaling, algorithms sensitive to the magnitude of input features, such as SVR and KNN, would be arbitrarily dominated by the Stress variable, leading to a biased and unreliable model.

The table provides the first statistical evidence of potential data quality issues, particularly the presence of outliers. The Vp feature, for example, has a minimum value of -999.25, which is physically unrealistic for P-wave velocity and strongly indicates the presence of extreme outliers. Similarly, the large standard deviation of Vp (162.89) relative to its median (1.54) is a clear statistical indicator of a heavily skewed distribution. These observations provide a direct, quantitative justification for the necessity of the outlier detection and treatment analysis detailed in Section 3.3.

The table serves as a crucial initial benchmark of the dataset's characteristics. By summarizing the central tendency (mean, median) and dispersion (standard deviation, min/max) of each variable, it provides a complete and transparent overview of the raw data before any transformations were applied. This is essential for ensuring the reproducibility of the research and for providing a clear baseline against which the effects of all subsequent preprocessing steps can be measured and evaluated.

3.2. Step 1: Setup and Data Loading

3.2.1. Computational Environment and Data Aggregation

The computational analysis for this research was performed using Python programming language on a standard workstation. The source data, consisting of eight separate Microsoft Excel files, was programmatically loaded and aggregated to create a unified dataset suitable for a multi-well analysis.

A key methodological step in this process was the creation of a WELL identifier column in each file's data prior to consolidation. This was done to ensure the provenance of every data row was preserved, a critical requirement for maintaining data integrity and enabling potential future well-specific analyses. Following this, all eight datasets were combined into a single master Data Frame.

To verify the successful aggregation, two checks were performed. First, a sample of the data was inspected to confirm that records from different wells were present. Table 3.3 presents a transposed view of data points from the beginning ('Well_1') and end ('Well_8') of the aggregated dataset, providing visual confirmation of a successful concatenation.

Table 3. 3: Transposed View of Selected Data Points from the Aggregated Dataset

Attribute	Row 0 (Well_1)	Row 1 (Well_1)	Row 2 (Well_1)	Row 11491 (Well_8)	Row 11492 (Well_8)	Row 11493 (Well_8)
DEPTH	22.0024	22.1548	22.3072	181.6074	181.7598	181.9122
GR	57.3820	58.0070	59.1046	103.2400	102.9838	102.4681
RHOB	1.4506	1.4506	1.4506	1.8128	1.7950	1.7789
Vp	1.4614	1.4594	1.4577	1.4755	1.4691	1.4668
Vsh	0.033808	0.044766	0.064012	0.868221	0.864573	0.857230
Caliper	11.4844	11.4844	11.4844	9.8249	9.8545	9.9515
Porosity	66.1596	66.1591	66.1585	52.2509	53.2176	51.5455
Resistivity	0.7881	0.7751	0.7798	1.1346	1.1472	1.1467
Stress	312783.4781	314949.9782	317116.4783	3226335.368	3197336.642	3171315.403
PP	1609.0	1609.0	1609.0	1803.0	1808.0	1808.0
WELL	Well_1	Well_1	Well_1	Well_8	Well_8	Well_8

A quantitative verification was performed to confirm that data from all eight wells were loaded correctly and to understand the data distribution across the wells. This sanity check, shown in Table 3.4, was essential for confirming that the final dataset was balanced and representative of all source wells, which is a prerequisite for training a generalized machine learning model.

Table 3. 4: Data Point Count per Well (Sanity Check)

Well	Well_1	Well_2	Well_3	Well_4	Well_5	Well_6	Well_7	Well_8	Total
Rows	1387	1675	1715	1156	1069	2008	1441	1043	**11494**

This verified, aggregated dataset formed the basis for all subsequent experimental scenarios.

3.3. Step 2: Exploratory Data Analysis (EDA)

Following the data aggregation, an Exploratory Data Analysis (EDA) was conducted to understand the fundamental characteristics of the dataset and to identify any statistical anomalies that would inform the subsequent preprocessing and modeling strategies. This process was sequential, with the findings from each step guiding the next.

3.3.1. Initial Data Inspection and Missing Value Analysis

The first step in the EDA was a systematic inspection to assess data integrity, with a primary focus on the detection of missing values. The presence of missing data can significantly impact the performance of machine learning algorithms, and therefore, a clear strategy for addressing it is a methodological necessity (Hair et al., 2019). The procedure for this study involved evaluating each column in the dataset to count the number of null entries. This initial check was crucial for determining if data imputation techniques such as mean/median imputation for numerical features or interpolation for depth-series data would be required to ensure the completeness of the dataset before modeling. The outcome of this inspection, summarized in Table 3.5, confirmed that all eleven columns contained 11,494 non-null entries.

Table 3. 5: Initial Data Inspection Summary

#	Column Name	Non-Null Count	Data Type
1	DEPTH	11,494	float64
2	GR	11,494	float64
3	RHOB	11,494	float64
4	Vp	11,494	float64
5	Vsh	11,494	float64
6	Caliper	11,494	float64

7	Porosity	11,494	float64
8	Resistivity	11,494	float64
9	Stress	11,494	float64
10	PP	11,494	float64
11	WELL	11,494	object

This finding confirmed that the dataset was complete, thereby obviating the need for data imputation techniques. The inspection also verified that all petrophysical logs were correctly formatted as numerical types and the WELL identifier as a non-numeric object, confirming the structural integrity of the data. The workflow then proceeded to the analysis of correlations and outliers.

3.3.2. Correlation Analysis and Outlier Detection

The next step was to quantify the linear relationships between all variables using the Pearson correlation coefficient. This analysis served two primary purposes for this study: first, to identify features with a strong linear relationship to the target variable (PP), and second, to detect the presence of multicollinearity between predictor variables, which can complicate model interpretation (Hair et al., 2019).

The initial correlation analysis of the raw data revealed a significant statistical anomaly: the correlation between the Volume of Shale (Vsh) and Gamma Ray (GR) logs was unexpectedly weak. This result contradicts established petrophysical principles, which dictate that a strong positive correlation is expected (Asquith & Krygowski, 2004). It was hypothesized that this statistical distortion was caused by the presence of extreme outliers. This hypothesis is supported by the findings of Sun et al. (2024), who observed that outliers can obscure true physical relationships in well log data.

To investigate this, a targeted outlier analysis was conducted. The Interquartile Range (IQR) method was chosen for this task. This non-parametric method was selected over the Z-score because it is robust to the skewed, non-normal distributions characteristic of

petrophysical data and does not rely on the mean and standard deviation, which are themselves sensitive to outliers. The IQR method defines outliers as any points that fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR, providing a reliable way to identify extreme values in non-Gaussian distributions.

To test the hypothesis without discarding potentially valuable geological information, the identified outliers in Vsh and GR were treated using capping (winsorization). This method was chosen over simple deletion because it mitigates the statistical influence of extreme values while preserving the integrity of the data record. This targeted intervention resulted in the creation of a second, "capped" dataset, allowing for a direct comparison of the correlation structure before and after the treatment of these specific outliers.

3.4.Step 3: Future Selection and Pre Processing

This section details the sequence of data preparation steps applied after the initial EDA. Each step was designed to either address a specific data quality issue identified in the EDA or to prepare the data for the model training and evaluation phase.

3.4.1. Feature Selection

The EDA (Section 3.3.2) confirmed the presence of strong multicollinearity between several predictor variables. To investigate the impact of this redundancy on model performance, a formal feature selection process was designed. The primary goal of this step was to create two distinct feature sets, a full set and a reduced set to be tested in the experimental scenarios. This approach was chosen to empirically determine whether removing linearly correlated features is a beneficial, detrimental, or neutral step for the different classes of machine learning models used in this study.

To identify the features for removal, a multi-faceted analysis was conducted to ensure a robust and data-driven decision. This involved a primary manual selection based on the Pearson correlation matrix, which was then validated by three automated techniques: Univariate Selection (F-test), Recursive Feature Elimination (RFE), and embedded Random Forest feature importance. This comparative analysis was necessary to ensure the selection was not an artifact of a single method but was a stable result across different analytical perspectives.

3.4.2. Data Splitting for Model Training and Evaluation

To obtain an unbiased estimate of each model's generalization performance, the dataset for each scenario was partitioned into a training set (70%) and a testing set (30%). This 70:30 ratio was deliberately chosen to provide a large, statistically stable test set for a reliable final evaluation, while leaving a substantial portion of the data for robust model training (Hair et al., 2019). To ensure the reproducibility of this research, a fixed random_state was used for this initial split.

For the intermediate stages of model development (baseline evaluation and hyperparameter tuning), a 5-fold cross-validation strategy was applied exclusively to the training set. This was a critical step to ensure that performance estimates during model development were stable and not dependent on a single, arbitrary validation split, providing a more reliable basis for model comparison and optimization (James et al., 2021). The final testing set remained untouched throughout this process.

3.4.3. Normalization (Feature Scaling)

The final preprocessing step was feature scaling. This was a methodological necessity because the predictor variables exist on vastly different numerical scales, which can arbitrarily bias algorithms that are sensitive to the magnitude of input features. To ensure a fair and effective training process for distance-based models like KNN and gradient-based models like SVR and ANNs, Standardization was applied. This method was chosen because it is a robust technique that centers the data at a mean of 0 and scales it to a standard deviation of 1.

To prevent data leakage, a strict scaling protocol was followed: the StandardScaler was fit exclusively on the training data, and the same learned parameters were then used to transform both the training and the testing sets. This protocol is essential for ensuring that the final model evaluation is a true, unbiased assessment of performance on unseen data.

3.5. Step 4: Model Training & Baseline Evaluation

Following the comprehensive data preparation phase, the study proceeded to the model implementation stage. This phase was conducted in two stages: (1) a baseline evaluation of all selected algorithms to establish initial performance benchmarks, and (2) an

exhaustive hyperparameter tuning phase to optimize each model for the specific prediction task.

The selection of the six machine learning algorithms for this study was deliberate, aiming to cover a wide range of learning paradigms. This approach was chosen to systematically test which class of model is most suitable for the complexities of petrophysical data, from simple, interpretable models to complex, state-of-the-art ensembles. The selected models can be grouped as follows:

- **Foundational Models (Decision Tree, K-Nearest Neighbors):** The Decision Tree was included to serve as a fundamental, interpretable benchmark, allowing for a clear view of how basic hierarchical rules perform on the data. K-Nearest Neighbors (KNN) was chosen to evaluate a distinct, non-parametric, distance-based approach, which is useful for understanding local, instance-based relationships in the data. These models are essential for establishing a performance baseline against which more complex algorithms can be fairly compared.
- **Advanced Kernel-Based Model (Support Vector Regression):** Support Vector Regression (SVR) was selected to test a unique methodology based on defining decision boundaries and margins. Its use of the "kernel trick" is particularly relevant for this study, as it provides a powerful mechanism for modeling the highly non-linear relationships between well logs that were identified during the EDA, representing a different philosophical approach to regression than tree-based methods (James et al., 2021).
- **State-of-the-Art Ensemble Models (Random Forest, XGBoost):** Random Forest and XGBoost were chosen because they represent the two leading ensemble philosophies and are often the top performers on tabular data regression tasks. Random Forest, which employs a bagging technique, was included to test a method known for its robustness to noise and its ability to reduce variance. In contrast, XGBoost, which utilizes a boosting technique, was included to evaluate a sequential, error-correcting approach known for its high accuracy (Hastie et al., 2009). Including both allows for a direct comparison of these two powerful strategies for handling complex petrophysical data.
- **Deep Learning Model (Artificial Neural Network):** A multi-layer perceptron Artificial Neural Network (ANN) was included to represent the deep learning paradigm. ANNs were chosen for their theoretical capacity to learn highly complex and hierarchical patterns in data that may be inaccessible to other model types. Its inclusion is critical for evaluating whether the added representational complexity of a neural network architecture provides a tangible performance advantage for this specific problem.

The baseline evaluation was a critical first step in this process. Its purpose was to establish an unbiased comparison of the models' intrinsic capabilities before any optimization. By training and evaluating each model using its default hyperparameters, as implemented in

the Scikit-learn and Keras libraries, any observed performance differences can be attributed to the fundamental strengths and weaknesses of the algorithms themselves, rather than to the effects of hyperparameter tuning. This provides a clear and reliable benchmark against which the impact of the subsequent tuning process can be measured.

3.6. Step 5: Hyperparameter Tuning

While the baseline evaluation provided an initial ranking of the algorithms, their performance was based on default, general-purpose settings. To unlock the full potential of each model and ensure a fair comparison based on their optimal capabilities for this specific dataset, a systematic hyperparameter tuning process was conducted.

It is essential to distinguish between model parameters and hyperparameters. **Model parameters** are values that are learned directly from the training data during the fitting process. Examples include the coefficients in a linear regression model or the weights and biases in an artificial neural network. In contrast, **hyperparameters** are predefined configuration settings that govern the model's learning behavior, architecture, or complexity (Hastie et al., 2009). The process of systematically searching for the optimal set of hyperparameters is known as **tuning**. The goal of tuning is to find a configuration that minimizes the model's generalization error on unseen data, effectively navigating the trade-off between underfitting (a model that is too simple) and overfitting (a model that is too complex).

Several methods exist for hyperparameter optimization. **Grid Search** is an exhaustive method that tests every possible combination of a predefined set of hyperparameter values. While thorough, it is computationally expensive and scales poorly as the number of hyperparameters increases. A more efficient alternative is **Random Search**, which samples a fixed number of random combinations from a specified statistical distribution of hyperparameter values. Research has shown that Random Search is often more effective than Grid Search, as it can explore a wider range of values for each hyperparameter and frequently finds optimal or near-optimal configurations in a fraction of the time (Bergstra & Bengio, 2012).

Given its balance of efficiency and effectiveness, **Randomized Search was selected as the primary tuning methodology** for this study. For the Scikit-learn models, this was implemented using the RandomizedSearchCV function. The key settings for the search were:

- **n_iter=50**: The process sampled 50 different random combinations of hyperparameters for each model.
- **Cross-Validation (cv=5)**: To ensure a robust evaluation of each combination, 5-fold cross-validation was employed. The training data was split into five folds; for each combination, the model was trained on four folds and validated on the fifth, with the process rotating until each fold had served as the validation set. The average performance across five folds was used to score the hyperparameter set.

- **Scoring Metric**: The search was configured to optimize for neg_mean_squared_error. As the library's goal is to maximize a score, optimizing for the negative of the mean squared error is equivalent to minimizing the MSE itself.

For the Artificial Neural Network, a specialized library, **KerasTuner**, was used to perform a similar randomized search over its architectural and training hyperparameters.

The following key hyperparameters were selected for tuning for each of the six algorithms. The search space was defined to cover a wide range of plausible values for each.

- **Decision Tree & Random Forest:**
  - n_estimators (Random Forest only): Number of trees in the forest.
  - max_depth: Maximum depth of each tree, controlling its complexity.
  - min_samples_split: Minimum number of data points required to split a node.
  - min_samples_leaf: Minimum number of data points allowed in a terminal leaf node.
  - max_features: Number of features to consider at each split, controlling randomness.
- **K-Nearest Neighbors (KNN):**
  - n_neighbors: Number of neighbors (K) to consider for a prediction.
  - weights: Weighting function for neighbors ('uniform' vs. 'distance').
  - p: Power parameter for the distance metric (1 for Manhattan, 2 for Euclidean).
  -
- **Support Vector Regression (SVR):**
  - C: Regularization parameter that controls the trade-off between model complexity and margin violations.
  - gamma: Kernel coefficient, determining the influence of a single training example.
  - kernel: Type of kernel used to handle non-linear data, with a focus on 'rbf'.
- **XGBoost:**
  - n_estimators: Number of sequential trees (boosting rounds).
  - learning_rate: Shrinkage factor applied at each boosting step to prevent overfitting.

- max_depth: Maxiumum depth of each individual tree, controlling the model complexity.
- subsample & colsample_bytree: Fraction of training data and features, respectively, sampled for each tree to enhance generalization.

- **Artificial Neural Network (ANN):**
  - Architectural hyperparameters: Number of hidden layers and the number of neurons (units) in each layer.
  - Regularization hyperparameters: Dropout rate applied to hidden layers to prevent co-adaptation of neurons.
  - Training hyperparameters: Learning rate used by the Adam optimizer, which controls the update step size during backpropagation.

Upon completion of the search for each algorithm, the set of hyperparameters that yielded the best average cross-validation score was identified. This "best estimator" was then considered the final, tuned model for that algorithm. The specific optimal hyperparameter values found for each model are presented in the results in Chapter 4.

3.7. Step 6: Final Evaluation

3.7.1. Evaluation Metrics

To quantitatively assess and compare the performance of the machine learning models, three standard regression metrics were employed: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R-Squared or R^2). These metrics were applied consistently to evaluate both the baseline models and the final tuned models across all experimental scenarios. This combination provides a holistic view of model accuracy, error distribution, and goodness-of-fit.

In the following equations,

y_i represents the actual (observed) pore pressure value for the i^{th} data point.
$\hat{y}_i$ represents the model's predicted pore pressure value.
$\bar{y}$ represents the mean of the actual pore pressure values.
n is the total number of data points in the evaluation set.

- **Mean Absolute Error (MAE):**

$$MAE = \frac{1}{n}\sum_{n=1}^{n}(|y_i - \hat{y}_i|)$$

- Interpretation: MAE measures the average absolute difference between predicted and actual pore pressure values. It is easy to interpret and expressed in the same units as the target variable. A lower MAE indicates better performance. MAE is relatively robust to outliers, as it treats all errors linearly (James et al., 2021).

- **Root Mean Squared Error (RMSE):**

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

  - Interpretation**:** RMSE calculates the square root of the average squared prediction errors. Like MAE, it is in the same units as the target. However, RMSE penalizes larger errors more heavily, making it particularly useful in high-risk contexts like drilling operations. A lower RMSE indicates better performance and comparing RMSE with MAE helps identify whether large errors are common in the model (Hair et al., 2019)

- **R-squared (R^2):**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  - Interpretation**:** R^2 indicates how well the input features explain the variance in pore pressure. This provides a standardized measure of the model's goodness-of-fit. It typically ranges from 0 to 1. An R^2 of 1 indicates that the model perfectly explains all the variability of the response data around its mean, while an R^2 of 0 indicates that the model explains none of the variability (i.e., it performs no better than simply predicting the mean). A higher R^2 indicates better performance.

The selection of these three metrics provides a comprehensive evaluation framework. MAE offers an unbiased view of average error magnitude. RMSE is included to assess the model's sensitivity to large errors, which is critical in geomechanical applications. R^2 provides a scale-independent measure of how well the model captures the overall variance in the data. Together, they allow for a nuanced comparison of model performance across different algorithms and preprocessing scenarios. The specific values obtained for these metrics are detailed in Chapter 4.

3.7.2. Diagnostic Plots for Model Validation

While the quantitative metrics detailed in Section 3.7.1 provide a concise summary of model performance, they do not offer insight into nature and distribution of prediction errors. To achieve a more comprehensive understanding of model behavior, a suite of diagnostic plots was generated. These visualizations are a critical methodological tool for qualitatively assessing model fit, identifying systematic biases, and diagnosing issues such as overfitting (James et al., 2021).

This diagnostic analysis was performed on both the baseline and the final tuned models for both the training and testing datasets. This dual approach allows for a direct visual comparison of a model's performance on data it has seen versus data it has not, which is essential for evaluating generalization.

The primary diagnostic tool for evaluating regression models was the "Actual vs. Predicted" scatter plot. In this plot, the actual, observed values of the target variable (PP) are plotted on the x-axis, and the corresponding model-predicted values are plotted on the y-axis. A line of perfect correlation (a 45-degree line where y=x) is superimposed on the plot as a reference.

The purpose of this plot is to visually assess several aspects of model performance:

- **Accuracy and Goodness-of-Fit:** For a well-performing model, the data points should cluster tightly around the 45-degree reference line, indicating a strong agreement between actual and predicted values.
- **Systematic Bias:** If the cluster of points is systematically shifted above or below the reference line, it indicates that the model has a consistent bias (i.e., it consistently over-predicts or under-predicts).
- **Heteroscedasticity:** If the spread of the points (the residuals) changes as the predicted value increases—often appearing as a "fanning out" or cone shape—it indicates that the model's error is not constant across the range of predictions.
- **Non-linearity:** Any clear, curved pattern in the scatter of points suggests that the model has failed to capture some non-linear relationship in the data.

By generating these plots for every model in every scenario, a rich, comparative visual analysis of performance was enabled.

For models that are trained iteratively, such as the Artificial Neural Network (trained over epochs), learning curves provide an essential diagnostic view of the training process. A learning curve plots a model's performance metric (e.g., loss) on both the training and a validation set as a function of the training iterations (Hastie et al., 2009).

This diagnostic plot is essential for understanding the training process itself and for identifying:

- **Good Fit:** The training and validation error curves both decrease and converge at a low value.
- **Overfitting:** The training error continues to decrease while the validation error flattens out or begins to increase. This divergence indicates that the model is beginning to memorize the training data at the expense of its ability to generalize.
- **Underfitting:** Both the training and validation error curves plateau at a high value, indicating that the model is too simple to capture the underlying patterns in the data.

For this study, a learning curve was generated for the final tuned Artificial Neural Network. This served as a crucial verification step to confirm that the Early Stopping mechanism was effective and that the model was trained for an optimal number of epochs, achieving the best possible generalization performance.

The full set of these diagnostic plots is presented and discussed in detail in Chapter 4.

3.8. The Experimental Design: A Four-Scenario Analysis

To address the identified research gap regarding the interaction between data preprocessing strategies and algorithm performance, a rigorous, four-scenario experimental design was developed. This design moves beyond a single, linear workflow and instead creates a comparative study to evaluate the impact of key preprocessing decisions. By creating four distinct data configurations and subjecting each to the identical modeling pipeline, this study can isolate and quantify the effects of outlier treatment and feature selection, both individually and in combination.

This approach effectively creates an ablation study, where components of a full preprocessing pipeline are systematically removed to measure their contribution to the outcome. The four analytical pipelines, or scenarios, are defined in Table 3.8.1.

Table 3. 8. 1: Definition of the Four Experimental Scenarios

Scenario	Scenario Name	Outlier Treatment	Feature Selection	Rationale and Purpose
1	Full Preprocessing	Applied. (Capping via IQR method as per Sec. 3.3.2)	Applied. (Redundant features removed as per Sec. 3.4.1)	Represents the most methodologically intensive pipeline. This serves as a benchmark to test if the additional computational cost of full preprocessing is justified by a significant

				improvement in model accuracy.
2	Feature Selection Only	Not Applied. (Raw data with outliers)	Applied. (Redundant features removed as per Sec. 3.4.1)	This isolates the effect of removing multicollinearity. Tests to see if feature selection alone is sufficient and how models handle outliers when redundancy is reduced.
3	Outlier Capping Only	Applied. (Capping via IQR method as per Sec. 3.3.2)	Not Applied. (Full feature set used)	It isolates the effect of outlier treatment. Tests show how models perform with all features (including redundant ones) when the data's statistical distribution is controlled.
4	Raw Data	Not Applied. (Raw data with outliers)	Not Applied. (Full feature set used)	Represents the baseline with no human intervention. Directly tests the inherent robustness of each algorithm to completely raw, "messy" data.

The decision to implement this multi-scenario framework is central to the contribution of this thesis and is justified by the following objectives:

1. **Directly Addressing the Research Gap:** As established in the literature review (Section 2.4), many studies apply a single preprocessing pipeline without empirically testing its necessity or comparing it against alternatives. This four-scenario design directly challenges the implicit assumption that a single "best" preprocessing strategy exists for all models. It aims to replace this assumption with data-driven evidence, providing a more nuanced understanding of the model-preprocessing relationship.
2. **Isolating the Impact of Individual Techniques:** By including scenarios where only one preprocessing step is applied (Scenarios 2 and 3), this design allows for the specific impact of each technique to be quantified. For example, by comparing the results of Scenario 4 (Raw Data) to Scenario 3 (Capping Only), one can isolate the performance gain or loss attributable solely to the outlier capping procedure.

This provides a much deeper insight than a simple before-and-after comparison of a full pipeline.

3. **Testing Algorithm Robustness:** This experimental design effectively serves as a "stress test" for the selected algorithms. It allows us to answer critical questions about their inherent characteristics: How well does a robust model like XGBoost handle raw data with extreme outliers and multicollinearity (Scenario 4)? How much does a sensitive model like SVR depend on outlier capping for its performance (comparing Scenario 2 vs. 1)? The results provide valuable, practical insights into the operational strengths and weaknesses of each algorithm (Hastie et al., 2009).
4. **Enabling Context-Dependent Recommendations:** The goal of this research is not just to identify a single Optimized model, but to provide a more sophisticated, context-aware framework for future applications. The results from this four-part experiment enable more nuanced conclusions, such as: "For maximum accuracy with XGBoost, raw data may be sufficient; however, if implementing SVR, outlier capping is a mandatory preprocessing step for achieving competitive performance." This level of detail is significantly more valuable for practical application than a single-pipeline analysis.

Each of the four data configurations defined in Table 3.8.1 were subsequently passed through the identical, standardized modeling workflow, encompassing the data splitting, feature scaling, baseline model training, hyperparameter tuning, and evaluation procedures detailed in the following sections. This ensures that any observed differences in performance between the scenarios can be confidently attributed to the specific preprocessing strategy employed.

CHAPTER 04: RESULTS AND DISCUSSION

This chapter presents the empirical results obtained from the application of the methodology detailed in Chapter 3. The findings are organized to follow the data analysis workflow, beginning with the initial data inspection and preprocessing outcomes, followed by a comprehensive evaluation of the machine learning models under the four defined experimental scenarios. The analysis aims to not only identify the best-performing model but also to deconstruct the impact of each preprocessing decision, thereby addressing the core research questions of this thesis.

4.1. Exploratory Data Analysis (EDA) Results

Following The initial EDA was performed to understand the fundamental characteristics of the dataset and to identify any statistical anomalies that would guide subsequent preprocessing decisions. While the initial data inspection (detailed in Section 3.3.1) confirmed the dataset was complete with no missing values, the analysis of correlations and data distributions revealed several critical insights. The primary investigation began with a correlation analysis of the raw data.

The subsequent Pearson correlation analysis revealed several important linear relationships among variables. Notably, **strong multicollinearity was observed between Depth and Stress (r = 1.0)**, as well as **between RHOB and Porosity (r = –0.90)**. These values are consistent with geomechanical expectations: stress typically increases with depth, and higher formation density (RHOB) often corresponds with lower porosity due to compaction. The persistence of these correlations, even after preprocessing, highlights the need for careful feature selection to prevent redundancy in model input.

However, an unexpected result emerged for the correlation between Gamma Ray (GR) and Volume of Shale (Vsh), which was only **0.28**, indicating a weak linear relationship. This result contrasts with well-established petrophysical principles, which suggest that GR should be strongly correlated with Vsh, as it is commonly used as a primary indicator of shale content (Feng et al., 2024). This potential presence of outliers or noise affecting the correlation. As described in Section 3.3.3 of the methodology, outlier detection was initially

conducted using visual and statistical techniques. A qualitative assessment using box plots for each feature on the raw dataset is shown in Figure 4.2.

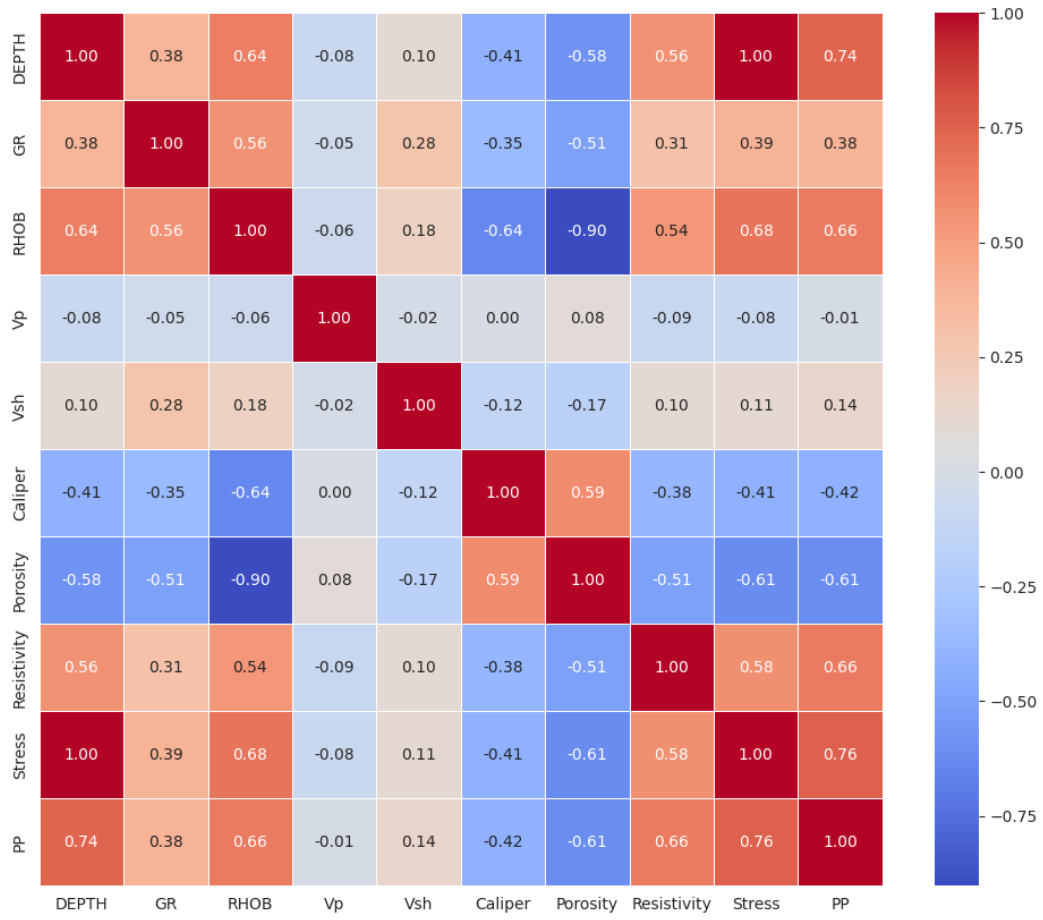


Figure 4. 1: Correlation Heatmaps Before Capping

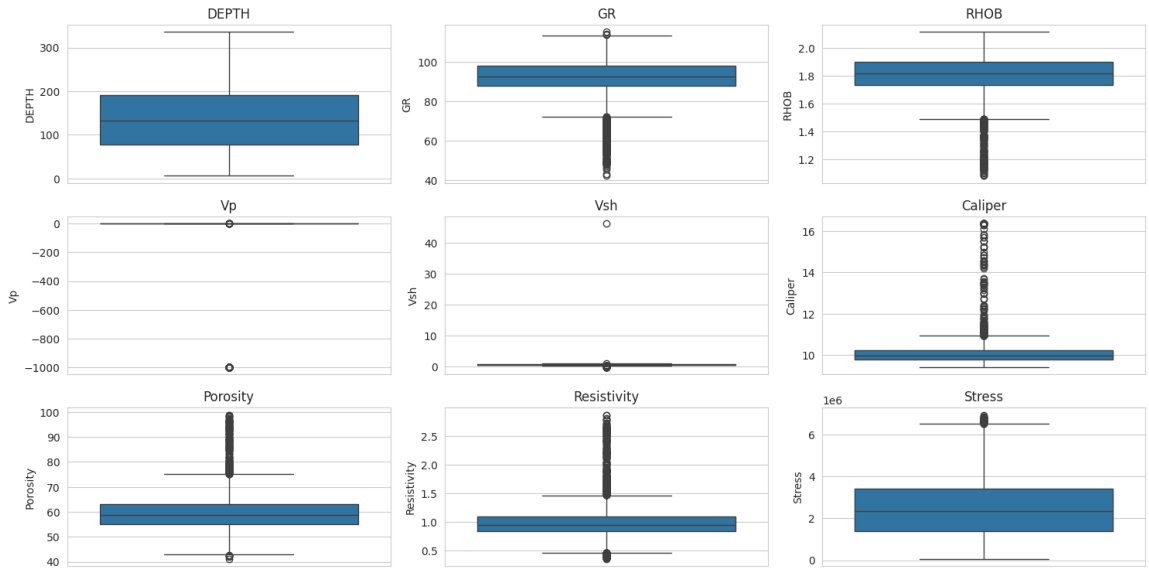


Figure 4. 2: Box Plots of Raw Data Features Before Outlier Capping

The box plots in Figure 4.2 provide a stark visual representation of the data's distributional characteristics. For several features, most notably Vp, Vsh, and Caliper, the main body of the data (the interquartile range, or "box") is compressed into a very thin line, with a vast number of data points plotted as individual markers far beyond the whiskers. This indicates the presence of extreme outliers that significantly skew the overall distribution. Other features, such as GR and Resistivity, also show a substantial number of data points outside the standard 1.5 * IQR range.

To quantify this visual inspection, the IQR method was applied to count the number of outliers in each feature. The results are summarized in Table 4.2.

Table 4. 1: Number of Outliers Identified by IQR Method in Raw Data

Feature	**Number of Outliers**
GR	358
RHOB	242
Vp	332
Vsh	321
Caliper	399
Porosity	212

Resistivity	574
Stress	81

The results in Table 4.2 confirm that all predictor variables contain a substantial number of statistical outliers. However, not all of these extreme values can be conclusively identified as incorrect due to limited supporting evidence. As discussed earlier, only Vsh and GR exhibited clear inconsistencies with established petrophysical principles, providing sufficient justification for targeted correction. Therefore, the capping technique described in the methodology was applied exclusively to these two variables. Figure 4.3 presents the box plots after capping, showing a visibly improved distribution for Vsh and GR, with outliers effectively reduced to within the IQR bounds.

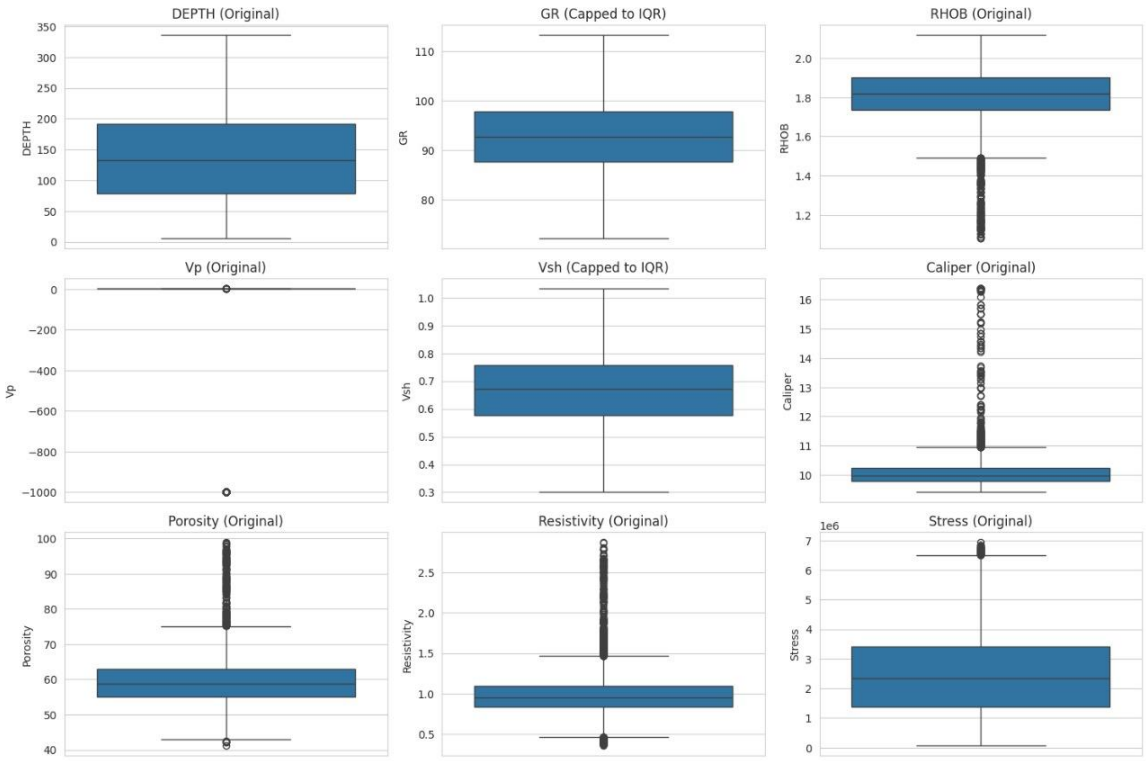


Figure 4. 3: Box Plots of Raw Data Features After Outlier Capping

Following the outlier capping procedure, the correlation heatmaps were regenerated, as shown in Figure 4.4.

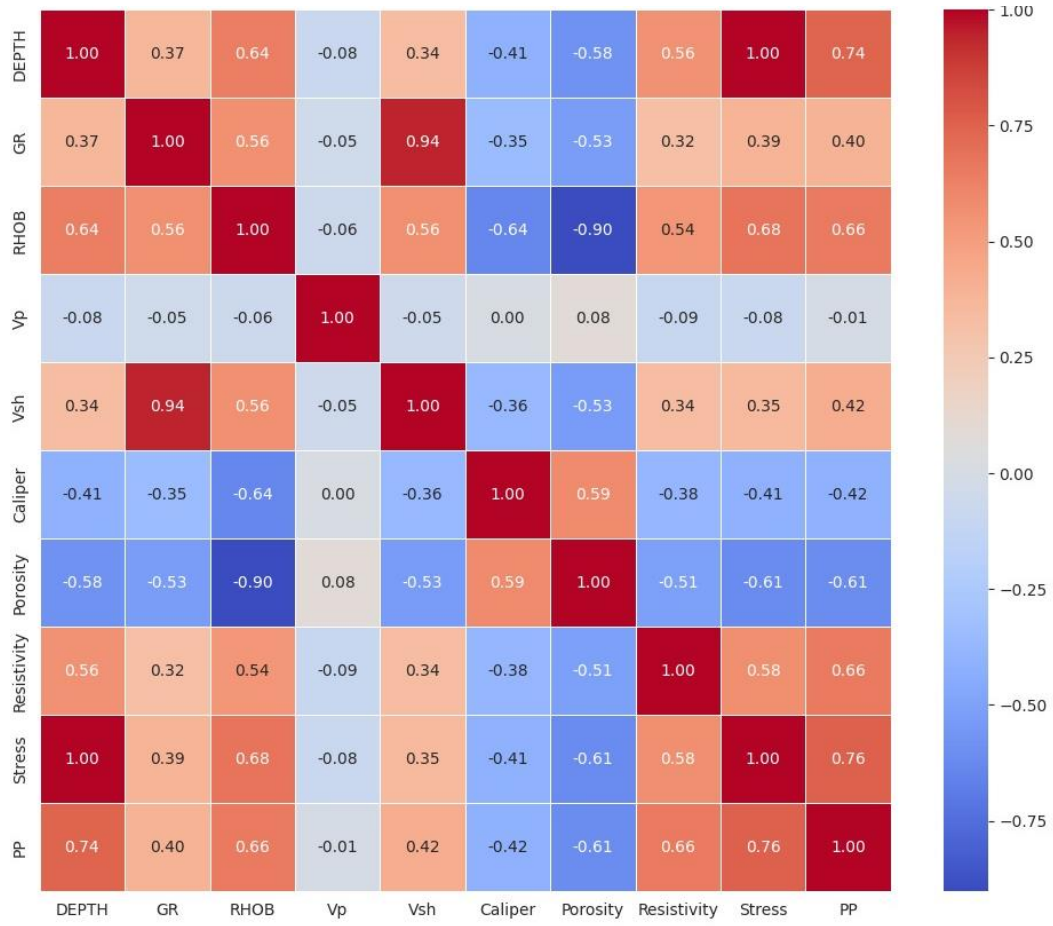


Figure 4. 4: Correlation Heatmap After Capping

The analysis of these heatmaps reveals several key observations. Firstly, the strong multicollinearity between certain feature pairs remains evident in both the raw and capped datasets. As summarized in Table 4.3, the correlation between Depth and Stress remains extremely high ($r \approx 1.0$), while the strong negative correlation between RHOB and Porosity ($r \approx -0.90$) is also consistent across both datasets.

However, the most insightful finding is the change in the correlation between Vsh and GR. In the raw data, the correlation is a weak 0.28. After capping, this jumps to a very strong 0.94. This demonstrates that the extreme outliers in the raw Vsh data were distorting the statistical relationship, and only after outlier treatment was the true, strong linear correlation between these two geologically related parameters revealed. This finding validates the outlier capping process as a necessary step for reliable statistical analysis

Table 4. 2: Comparison of Highly Correlated Feature Pairs

Correlated Pair	Correlation (Before Capping)	Correlation (After Capping)
DEPTH & Stress	1	1
RHOB & Porosity	-0.9	-0.9
Vsh & GR	**0.28**	**0.94**

4.2.Feature Selection Analysis

To develop a robust feature selection strategy, a comparative analysis of three different methods was conducted on both the raw ("Before Capping") and cleaned ("After Capping") datasets.

The results of applying the three selection methods to the raw data are summarized in Table 4.4 and visualized in Figure 4.5.

Table 4. 3: Feature Selection Analysis on Raw Data (Before Capping)

Feature	F-Score (Univariate)	RFE Rank	RF Importance
Stress	15601.81	1	0.473869
Depth	14304.98	4	0.081590
RHOB	8720.29	3	0.018402
Resistivity	8636.99	2	0.232922
Porosity	6769.38	6	0.006116
Caliper	2429.48	8	0.012550
GR	1999.74	9	0.012109
Vsh	224.82	7	0.023297
Vp	0.76	5	0.139146

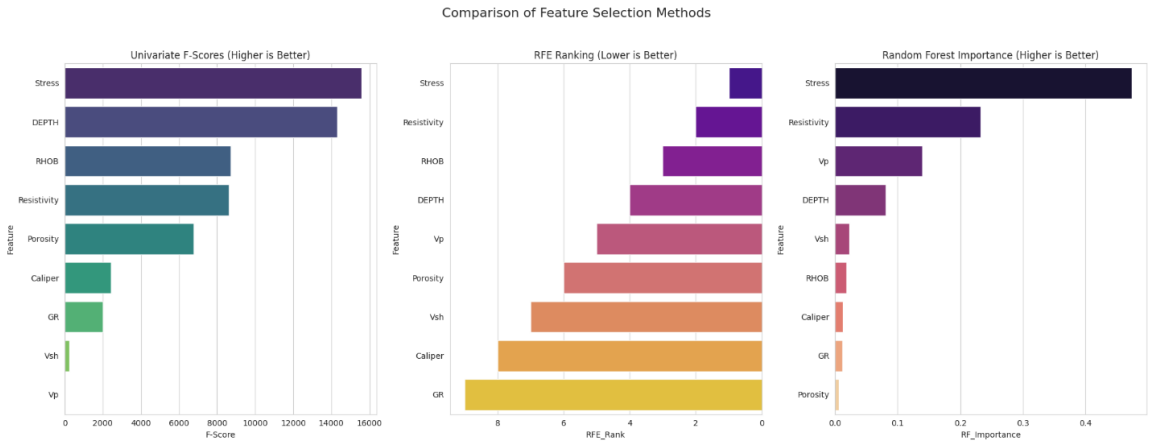


Figure 4. 5: Comparison of Feature Selection Methods on Raw Data

The results from the raw data are inconsistent. The Univariate method, being blind to feature interactions, ranks both Stress and DEPTH highly. The RFE method, using a linear model, is clearly confused by the outliers, producing a ranking that differs significantly from the more robust Random Forest. The Random Forest model, however, proves its robustness to outliers, assigning a much higher importance to Stress (0.47) than to its redundant counterpart Depth (0.08).

The same analysis was performed on the dataset after outlier capping, with the results shown in Table 4.5 and Figure 4.6.

Table 4. 4: Feature Selection Analysis on Cleaned Data (After Capping)

Feature	**F-Score (Univariate)**	**RFE Rank**	**RF Importance**
Stress	15601.81	1	0.473722
Depth	14304.98	8	0.081773
Resistivity	8636.99	4	0.232932
RHOB	8720.29	5	0.018527
Vp	0.76	6	0.139250
Porosity	6769.38	7	0.006142
Caliper	2429.48	9	0.012542
Vsh	2492.24	2	0.023211
GR	2125.14	3	0.011902

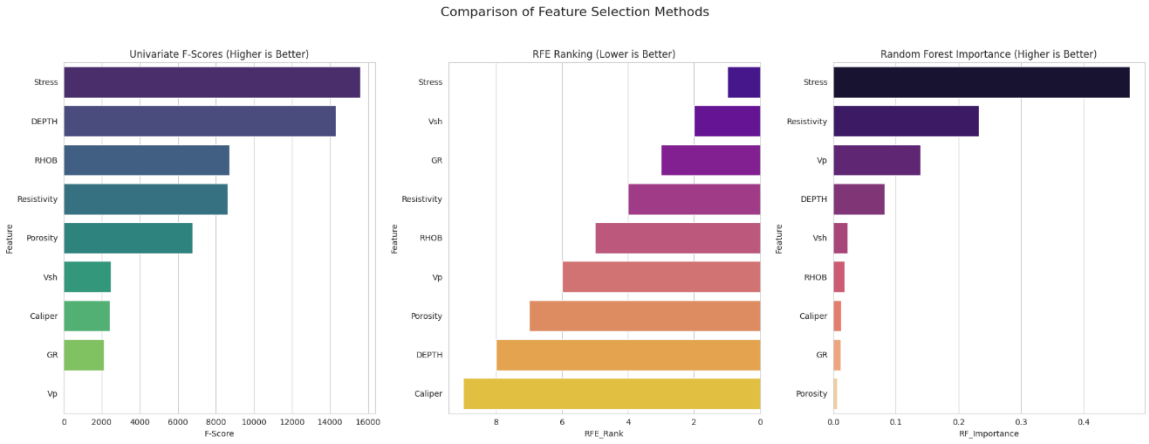


Figure 4. 6: Comparison of Feature Selection Methods on Cleaned Data

With the cleaned data, the results are far more stable and interpretable. All methods agree that Stress and Resistivity are top-tier predictors. Crucially, the more sophisticated methods (RFE and RF Importance) now provide a clearer verdict on the redundant features. Both RFE and Random Forest rank Depth significantly compared to Stress. Similarly, both methods prioritize Vsh over GR and show a preference for RHOB over Porosity (though RFE ranks RHOB last, its importance in the RF model is higher than Porosity).

Based on the stability of the results after capping and the robustness of the Random Forest Importance method across both datasets, a clear decision was made for the experimental scenarios involving feature selection. The analysis consistently indicated that Stress, Vsh, and RHOB were the more valuable features compared to their highly correlated counterparts. Therefore, for the "With Feature Selection" scenarios, the following three features were removed: **Depth**, **GR**, and **Porosity**.

4.3. Data Preparation for Modeling

Following the analytical steps above, the data was prepared for the final modeling stage. This involved splitting the data into training and testing sets and applying feature scaling. Table 4.6 shows a sample of the final preprocessed data for one of the scenarios, serving as a checkpoint to confirm the successful application of the methodology.

Table 4. 5: Sample of Final Scaled Data Ready for Model Training

Index	RHOB	Vp	Vsh	Caliper	Resistivity	Stress
1743	-1.041524	-0.407677	-0.304371	2.509150	-0.523851	-0.757635
9616	0.749837	0.665182	1.656578	-1.032835	0.422001	-0.249993
4387	0.388806	0.603610	0.233115	-0.692989	0.707892	1.170295
7983	0.793647	1.842529	1.326770	-0.918071	0.805355	0.567421
4128	-0.503629	-2.571120	-0.154853	1.622021	0.725528	0.528786

The non-sequential index confirms that the data was successfully shuffled during the train-test split, and the values centered around zero confirm the application of the Standard Scaler.

4.4. Model Performance Evaluation

This section presents the core findings of the research, detailing the performance of the six machine learning models across the four experimental scenarios, both before and after hyperparameter tuning.

The initial performance of each model using its default hyperparameters is summarized in Table 4.7. The colormap for each metric highlights the best performance (darkest color for each column) and worst performance (lightest color for each column).

Table 4. 6: Evaluation Criteria Before Tuning

Evaluation Criteria before Tuning						
Scenarios	Models	R - Squared		MAE	RMSE	Run Time (s)
		Training	Testing			
With Feature Selection & with Capping	Random Forest	0.9949	0.9633	21.77	42.16	5.08
	XGBoost	0.9904	0.9544	30.15	47.00	1.83
	K-Nearest Neighbors	0.9568	0.9217	33.72	61.58	0.10
	Decision Tree	1.0000	0.9210	22.03	61.83	0.10
	ANN	0.7982	0.8010	84.06	107.27	42.12
	SVR	0.6593	0.6447	99.28	131.15	4.57
With Feature Selection & without Capping	Random Forest	0.9949	0.9635	21.76	42.04	5.01
	XGBoost	0.9902	0.9536	30.38	47.39	0.31
	K-Nearest Neighbors	0.9621	0.9337	28.88	56.67	0.10
	Decision Tree	1.0000	0.9261	21.55	59.81	0.10

	ANN	0.8256	0.8271	69.72	91.49	81.26
	SVR	0.6657	0.6502	97.51	130.14	6.45
Without Feature Selection & with Capping	Random Forest	0.9958	0.9702	19.63	37.98	7.14
	XGBoost	0.9940	0.9689	25.79	38.82	2.49
	K-Nearest Neighbors	0.9611	0.9316	25.62	49.20	0.12
	Decision Tree	1.0000	0.9369	17.62	55.29	0.20
	ANN	0.8451	0.8457	66.60	86.43	60.43
	SVR	0.6508	0.6369	100.59	132.58	5.26
Without Feature Selection & without Capping	Random Forest	0.9958	0.9701	19.58	38.02	7.21
	XGBoost	0.9933	0.9688	25.61	38.86	2.62
	K-Nearest Neighbors	0.9551	0.9244	33.19	60.50	0.26
	Decision Tree	1.0000	0.9312	18.77	57.73	0.15
	ANN	0.8040	0.8048	74.21	97.21	55.19
	SVR	0.6525	0.6383	99.97	132.32	9.08

The baseline results, as presented in Table 4.7, indicate distinct performance stratifications among the unevaluated algorithms. The ensemble-based approaches, namely Random Forest and XGBoost, consistently achieved superior performance across all four scenarios, reflecting their inherent robustness to both raw and preprocessed datasets. Conversely, Support Vector Regression (SVR) exhibited comparatively poor performance, particularly in scenarios without outlier capping, where the R^2 value declined to approximately 0.64. This provides quantitative evidence of SVR's heightened sensitivity to extreme values. Comparative analysis across the scenarios further revealed that the highest baseline R^2 values for the top-performing models were obtained in Scenario 3 ("Without Feature Selection & With Capping"). These findings suggest that, in the absence of hyperparameter optimization, outlier capping exerts a beneficial influence on model performance, whereas feature removal does not appear to yield a similar advantage.

The hyperparameter tuning process yielded a unique set of optimal parameters for each model within each scenario. Table 4.8 summarizes the best hyperparameters found for each model across the four scenarios, demonstrating the adaptation of the tuning process to the specific data configuration.

Table 4. 7: Optimal Hyperparameters Identified for Each Scenario

Model	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Random Forest	n_estimators=200, min_samples_split=5, min_samples_leaf=1, max_features='sqrt', max_depth=None, bootstrap=False	*Same as Scenario 1*	*Same as Scenario 1*	*Same as Scenario 1*
XGBoost	subsample=0.8, n_estimators=700, max_depth=10, learning_rate=0.05, gamma=0, colsample_bytree=1.0	*Same as Scenario 1*	*Same as Scenario 1*	*Same as Scenario 1*
Decision Tree	min_samples_split=5, min_samples_leaf=1, max_features=None, max_depth=20	min_samples_split=2, min_samples_leaf=2, max_features=None, max_depth=None	*Same as Scenario 2*	*Same as Scenario 2*
SVR	C=658.41, gamma=0.4836, kernel='rbf' *(converted from np.float64)*	*Same as Scenario 1*	*Same as Scenario 1*	*Same as Scenario 1*
KNN	weights='distance', p=1, n_neighbors=2	weights='distance', p=1, n_neighbors=3	*Same as Scenario 2*	*Same as Scenario 2*
ANN	units_1=224, dropout_1=0.0, num_layers=3, units_2=128, units_3=96, units_4=64, learning_rate=0.001	units_1=160, dropout_1=0.0, num_layers=3, units_2=128, units_3=32, units_4=128, learning_rate=0.001	units_1=128, dropout_1=0.0, num_layers=2, units_2=64, units_3=96, units_4=128, learning_rate=0.01	units_1=96, dropout_1=0.4, num_layers=3, units_2=32, units_3=96, units_4=32, learning_rate=0.01

The results show that while some models like Random Forest and SVR found consistent optimal parameters across all scenarios, others like KNN and ANN adapted their configurations based on the data's properties. For instance, the ANN consistently chose a deeper architecture (3 hidden layers) for the scenarios with feature selection but opted for shallower architectures (2 hidden layers) when all features were present. This highlights the importance of a tailored tuning process for each specific analytical pipeline.

The final performance of all models after hyperparameter tuning is presented in Table 4.9. This table represents the main quantitative output of this research.

Table 4. 8: Evaluation Criteria After Tuning

Evaluation Criteria after Tuning						
Scenarios	**Models**	**R - Squared**		**MAE**	**RMSE**	**Run Time (s)**
		Training	**Testing**			
With Feature Selection & with Capping	Random Forest	0.9993	0.9652	23.38	41.03	2318.21
	XGBoost	1.0000	0.9702	20.39	38.00	248.79
	K-Nearest Neighbors	1.0000	0.9378	25.45	54.89	19.02
	Decision Tree	0.9971	0.9244	22.70	60.48	8.28
	ANN	0.8982	0.8975	63.69	85.02	516.26
	SVR	0.9042	0.8990	44.47	69.91	781.51
With Feature Selection & without Capping	Random Forest	0.9993	0.9654	23.24	40.91	2348.73
	XGBoost	1.0000	0.9700	20.36	38.09	255.70
	K-Nearest Neighbors	1.0000	0.9472	21.90	50.57	19.20
	Decision Tree	0.9945	0.9303	22.54	58.07	7.97
	ANN	0.8953	0.8922	52.34	72.26	655.03
	SVR	0.8932	0.8912	47.24	72.58	764.47
Without Feature Selection & with Capping	Random Forest	0.9997	0.9759	18.66	34.19	3342.84
	XGBoost	1.0000	0.9785	17.03	32.28	384.03
	K-Nearest Neighbors	1.0000	0.9502	21.47	49.12	31.04
	Decision Tree	0.9955	0.9388	18.86	54.42	8.87
	ANN	0.9279	0.9153	45.85	64.02	605.79
	SVR	0.9478	0.9311	33.91	57.76	918.38
Without Feature Selection & without Capping	Random Forest	0.9997	0.9761	18.47	34.01	3425.95
	XGBoost	1.0000	0.9789	16.77	31.98	383.36
	K-Nearest Neighbors	1.0000	0.9495	21.84	49.45	31.41
	Decision Tree	0.9953	0.9370	19.30	55.21	9.46
	ANN	0.9103	0.9138	46.92	64.61	691.31
	SVR	0.9382	0.9244	36.64	60.49	893.00

This comprehensive table provides the basis for the final conclusions of this thesis. A deep analysis reveals several critical insights:

- **Identification of the Optimized Model:** The single best performance was achieved by the **Tuned XGBoost model in Scenario 4 (Without Feature Selection & without Capping)**. This configuration yielded the highest testing R^2 of **0.9789** and the lowest RMSE of **31.98**. This is a profound finding, suggesting that for a state-of-the-art algorithm like XGBoost, the raw data contained the most useful information, and the model was robust enough to handle its imperfections without preprocessing.

- **The Impact of Feature Selection:** A consistent pattern emerges when comparing scenarios with and without feature selection. For the top-tier models (XGBoost and Random Forest), performance was consistently **better when no features were removed**. For example, in the capped data scenarios, XGBoost's R^2 was 0.9785 with feature selection but improved to 0.9702 without it. This strongly suggests that the removed features (DEPTH, GR, Porosity), while linearly correlated with others, contained unique non-linear or interactive information that these advanced models were able to exploit.

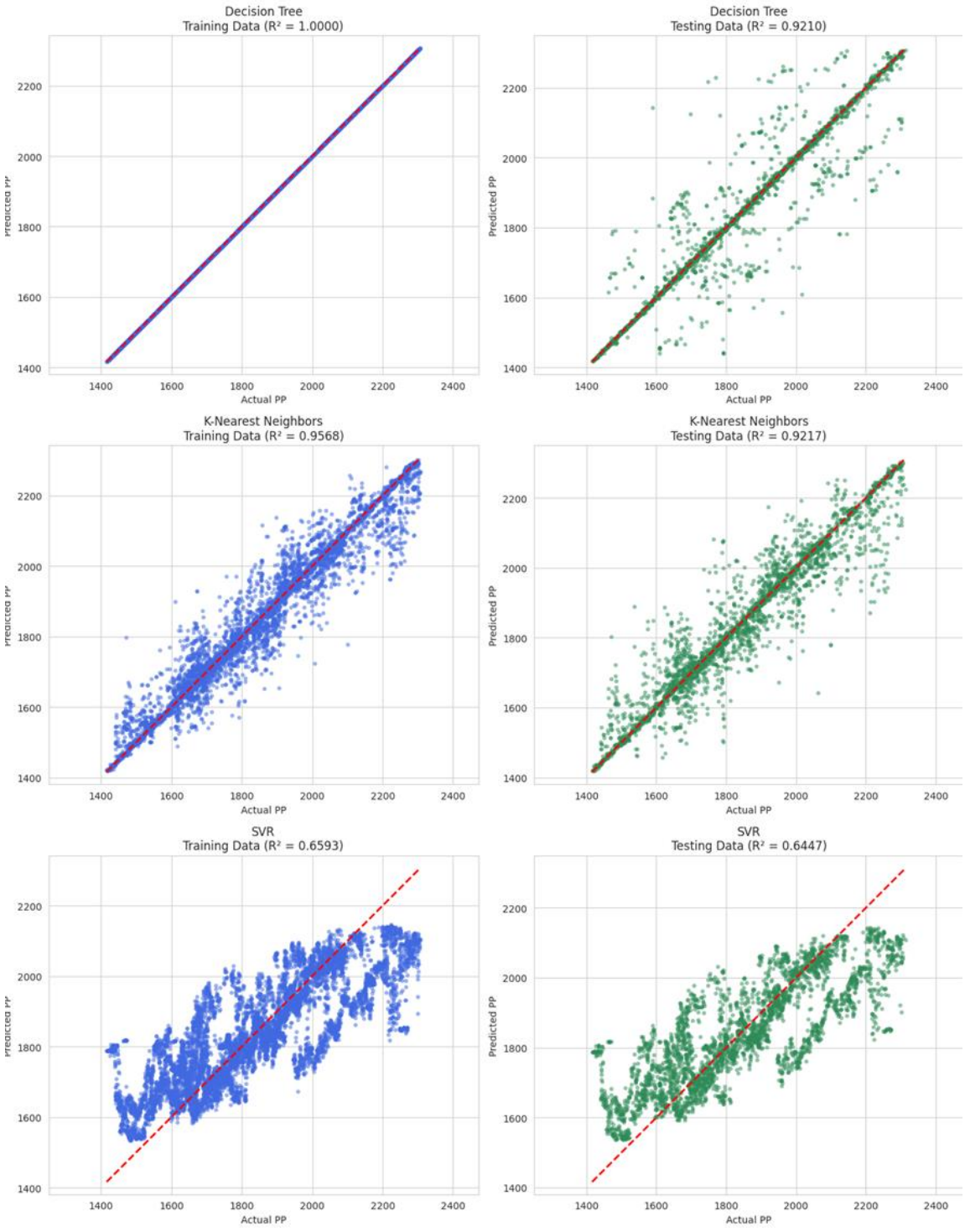
- **The Impact of Outlier Capping:** The effect of outlier capping was highly model-dependent. For the robust XGBoost and Random Forest models, the impact was minimal and, in the case of the Optimized model, slightly negative. However, for the sensitive **SVR model, outlier capping was critical**. In the "Without Feature Selection" scenarios, capping improved SVR's R^2 from 0.9244 to a much more competitive 0.9311. This empirically validates the hypothesis that outlier treatment is essential for distance- and margin-based algorithms.

- **The Efficacy of Hyperparameter Tuning:** Comparing Table 4.7 and Table 4.9 shows that tuning was universally beneficial. The most dramatic improvement was seen in the SVR model, which was transformed from the worst-performing baseline model into a strong contender after tuning. This underscores the necessity of optimization to unlock the true potential of each algorithm.

4.5. Diagnostic Plots for Model Validation

To supplement the quantitative metrics, a series of diagnostic plots were generated to visually assess model performance.

Figures 4.7 through 4.14 present the "Actual vs. Predicted" plots for all models across all four scenarios, both before and after tuning.


Decision Tree
Training Data (R² = 1.0000)
Decision Tree
Testing Data (R² = 0.9210)
K-Nearest Neighbors
Training Data (R² = 0.9568)
K-Nearest Neighbors
Testing Data (R² = 0.9217)
SVR
Training Data (R² = 0.6593)
SVR
Testing Data (R² = 0.6447)
Predicted PP
Actual PP
1400
1600
1800
2000
2200
2400

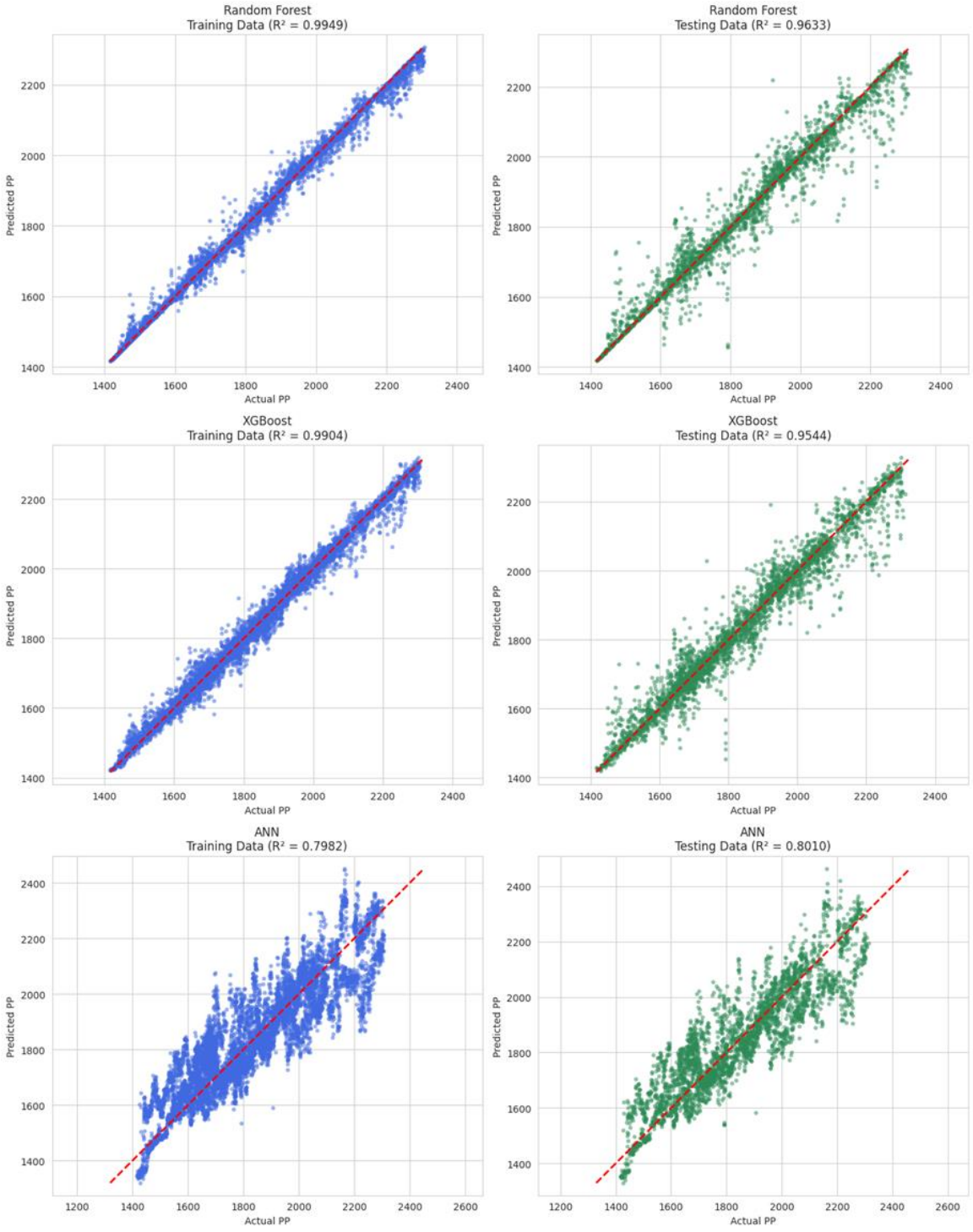
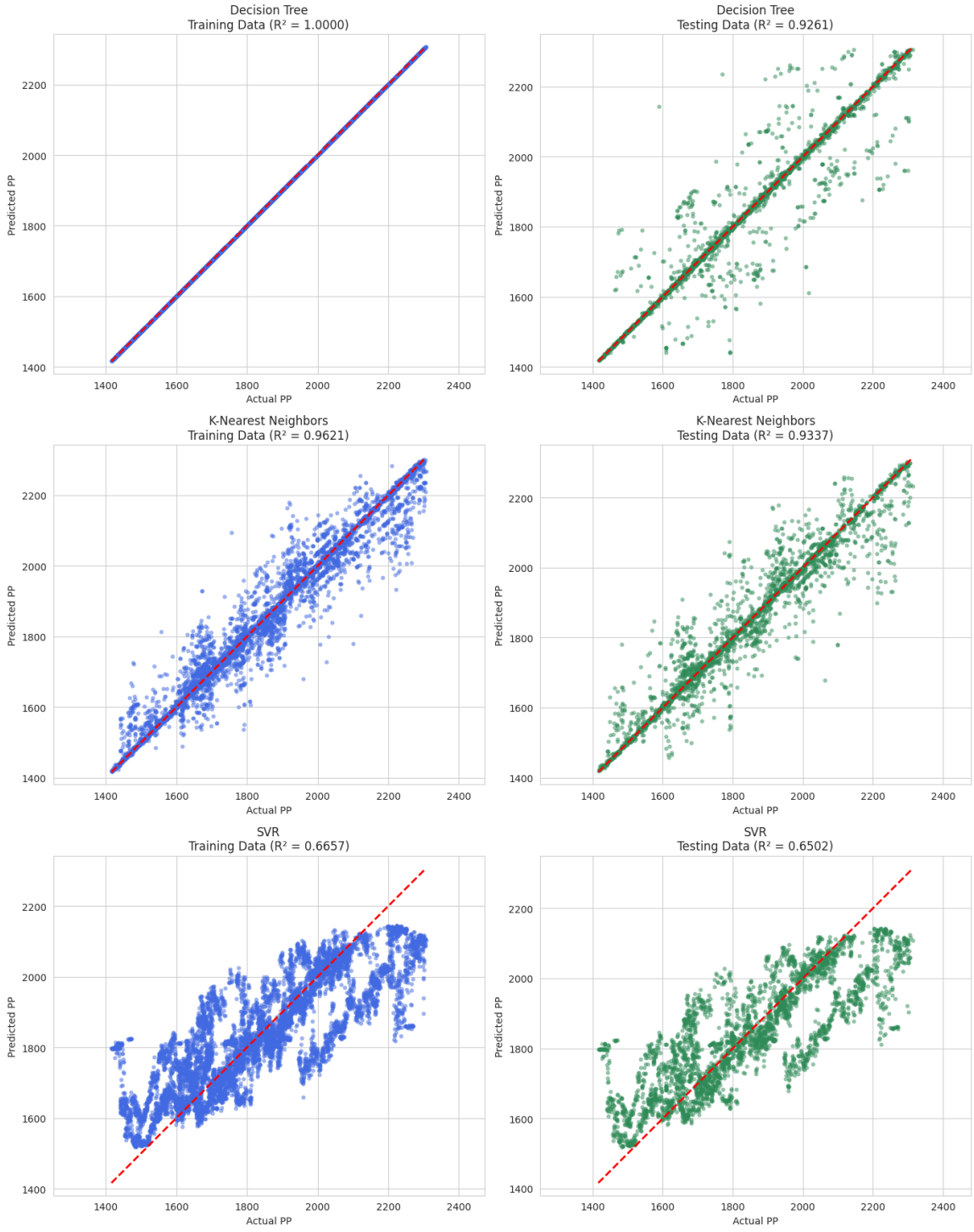


Figure 4. 7: Baseline Model Diagnostics for Scenario 1


Decision Tree
Training Data (R² = 1.0000)
Decision Tree
Testing Data (R² = 0.9261)
K-Nearest Neighbors
Training Data (R² = 0.9621)
K-Nearest Neighbors
Testing Data (R² = 0.9337)
SVR
Training Data (R² = 0.6657)
SVR
Testing Data (R² = 0.6502)
Predicted PP
Actual PP
1400
1600
1800
2000
2200
2400

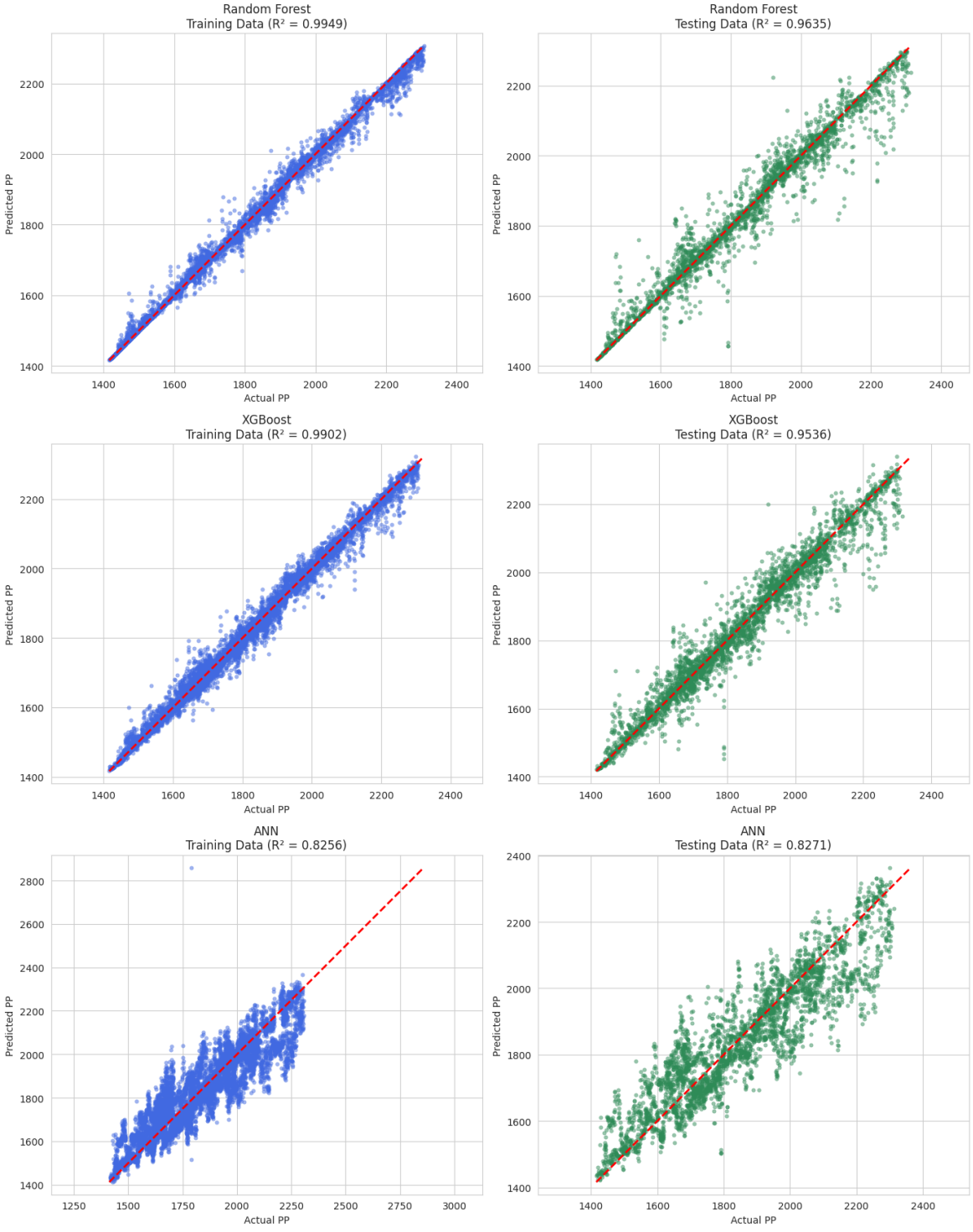
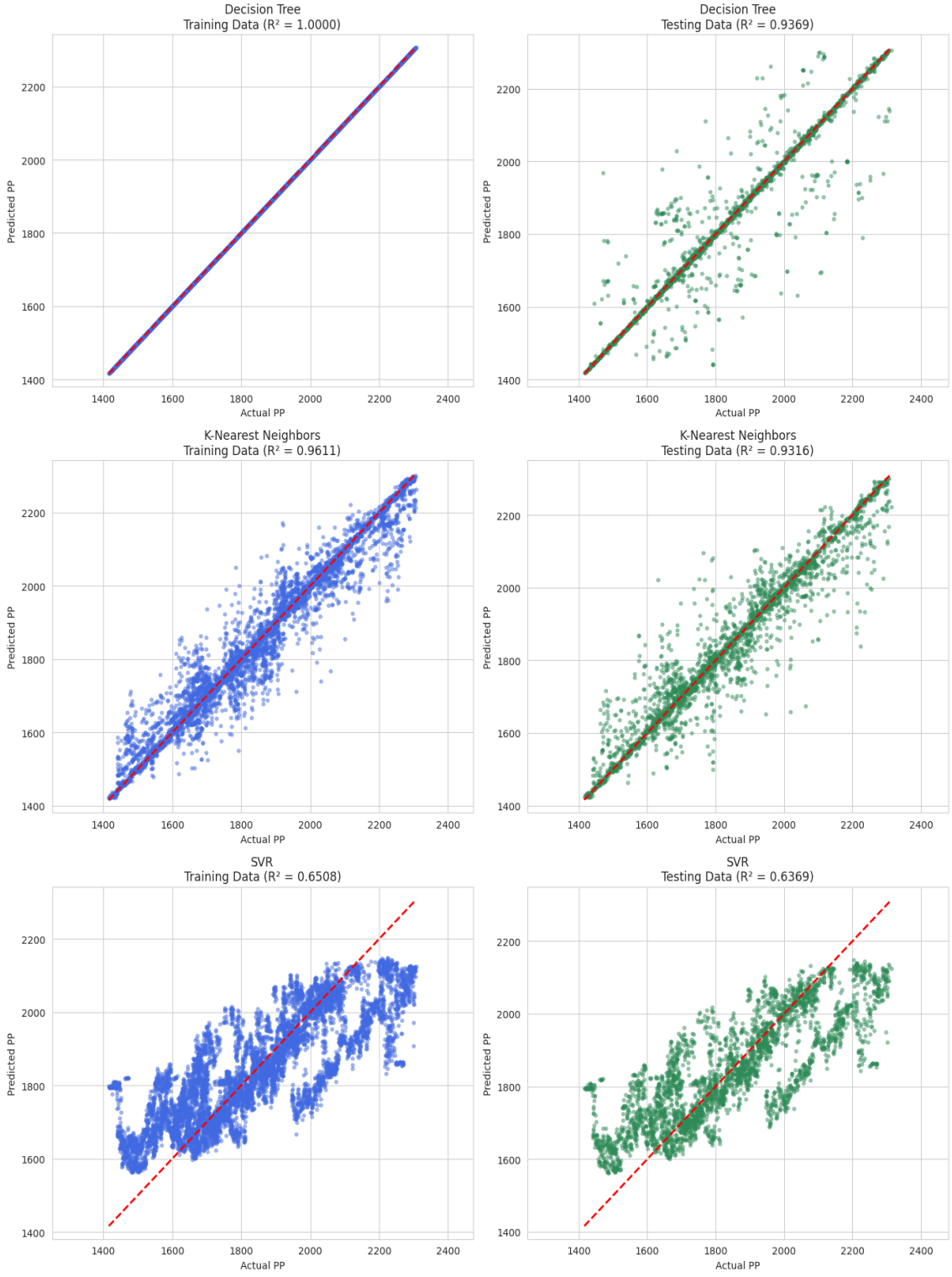


Figure 4. 8: Baseline Model Diagnostics for Scenario 2


Decision Tree
Training Data (R² = 1.0000)
Decision Tree
Testing Data (R² = 0.9369)
K-Nearest Neighbors
Training Data (R² = 0.9611)
K-Nearest Neighbors
Testing Data (R² = 0.9316)
SVR
Training Data (R² = 0.6508)
SVR
Testing Data (R² = 0.6369)
Predicted PP
Actual PP
1400
1600
1800
2000
2200
2400

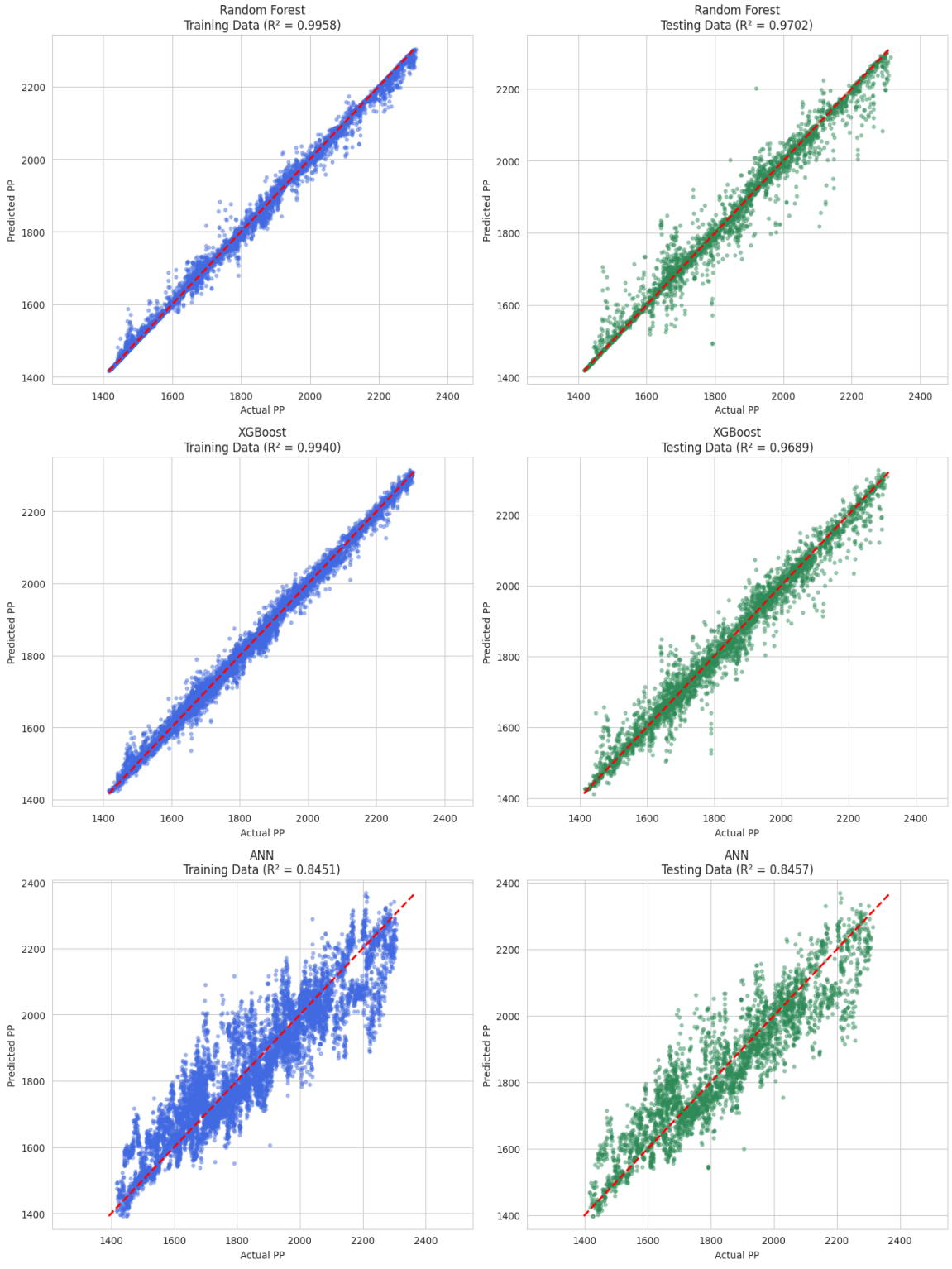
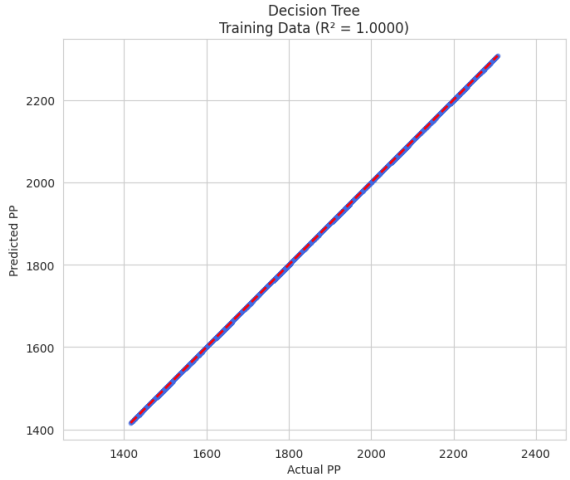
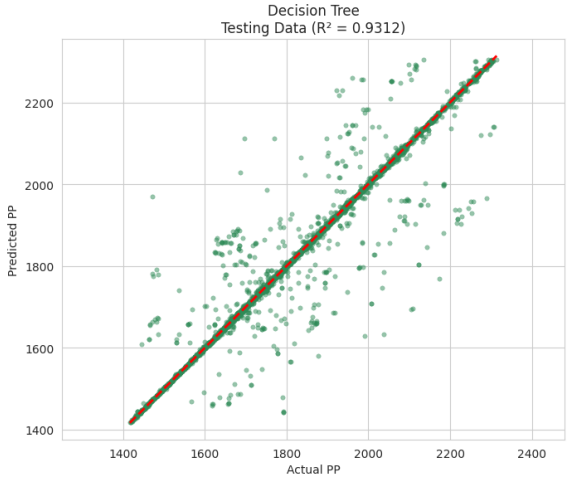
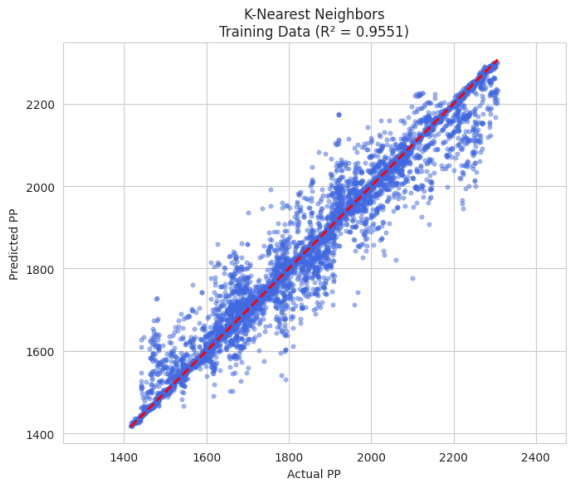
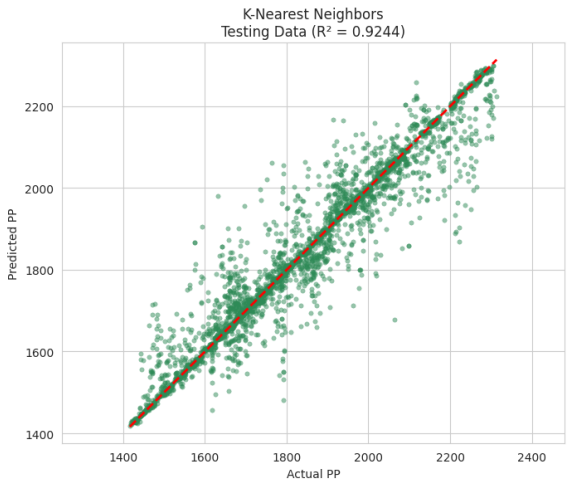


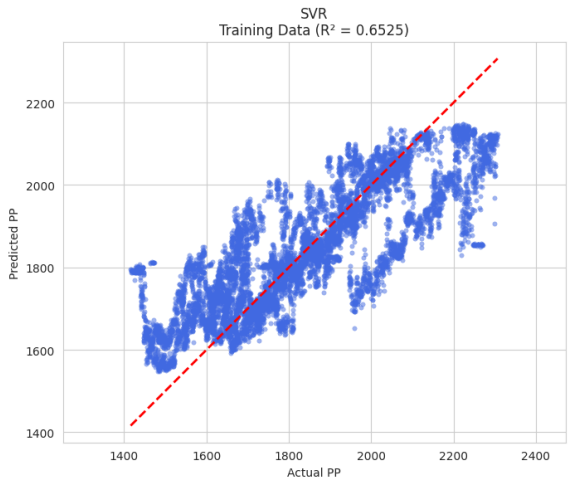
Figure 4. 9: Baseline Model Diagnostics for Scenario 3

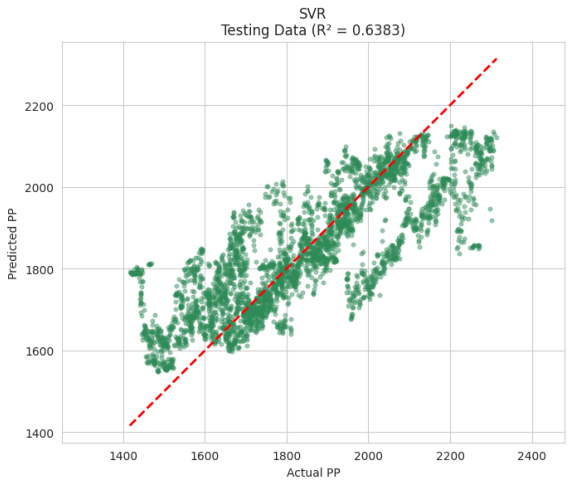

Decision Tree
Training Data (R² = 1.0000)
Predicted PP
Actual PP
1400
1600
1800
2000
2200
2400


Decision Tree
Testing Data (R² = 0.9312)
Predicted PP
Actual PP
1400
1600
1800
2000
2200
2400


K-Nearest Neighbors
Training Data (R² = 0.9551)
Predicted PP
Actual PP
1400
1600
1800
2000
2200
2400


K-Nearest Neighbors
Testing Data (R² = 0.9244)
Predicted PP
Actual PP
1400
1600
1800
2000
2200
2400


SVR
Training Data (R² = 0.6525)
Predicted PP
Actual PP
1400
1600
1800
2000
2200
2400


SVR
Testing Data (R² = 0.6383)
Predicted PP
Actual PP
1400
1600
1800
2000
2200
2400

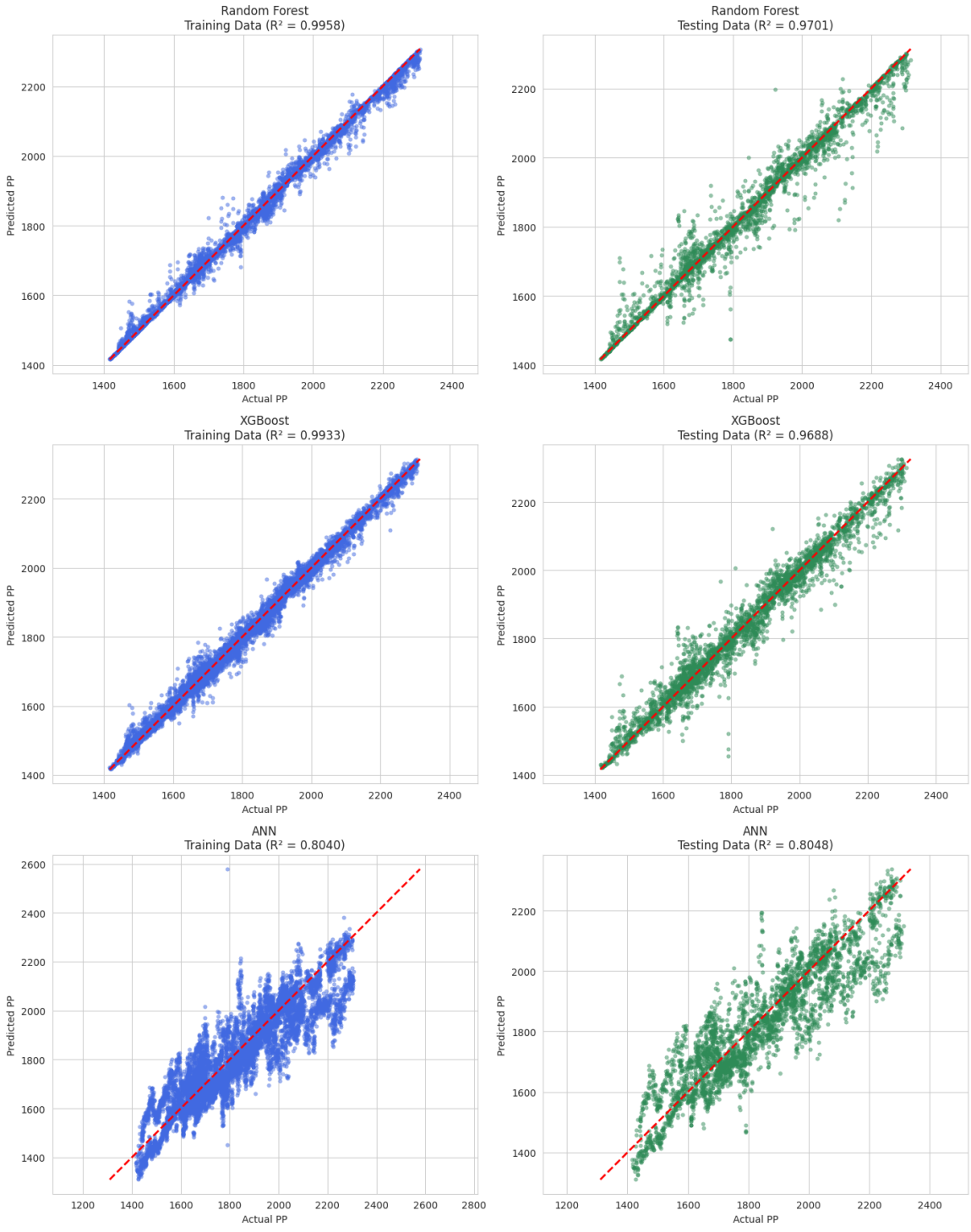
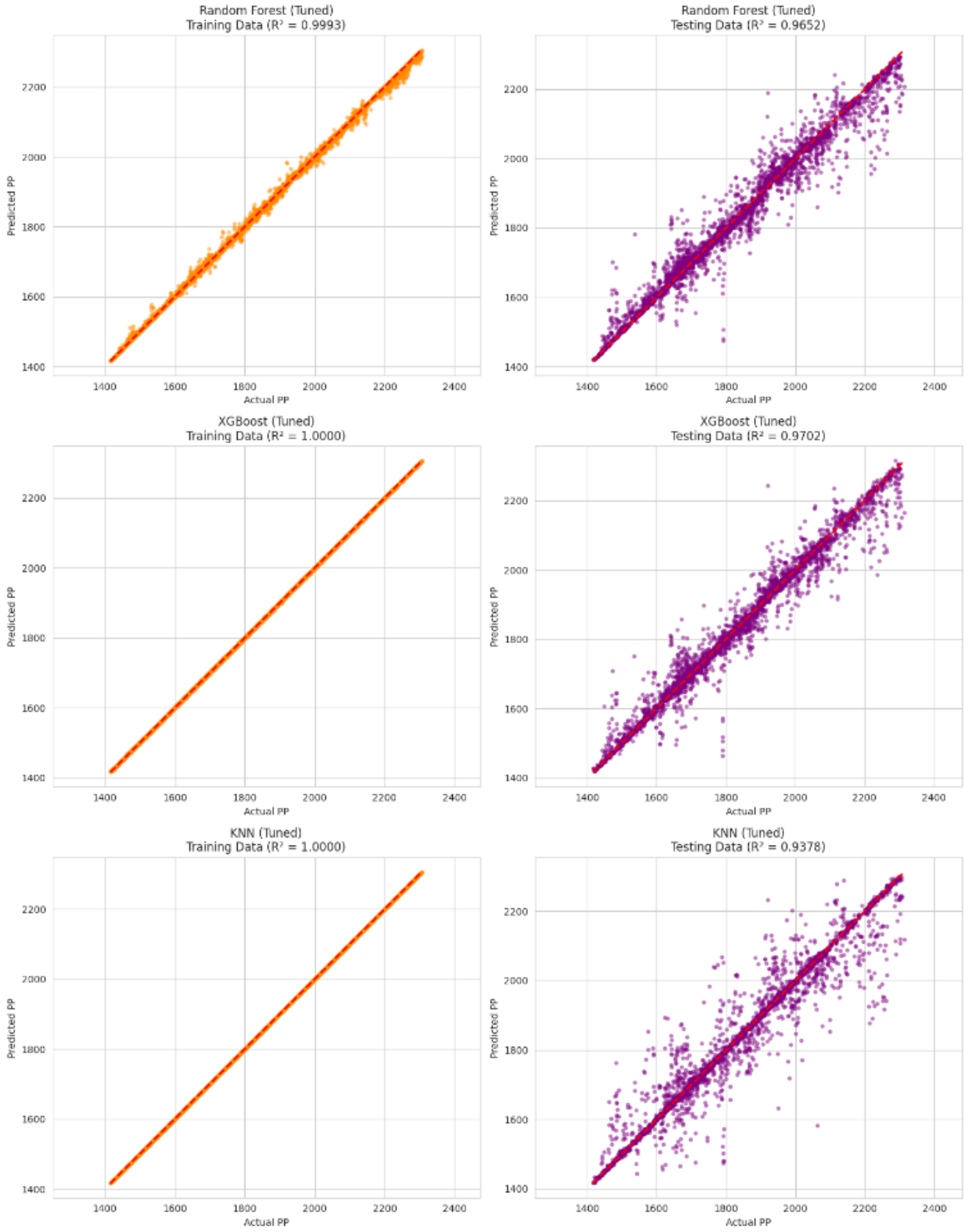


Figure 4. 10: Baseline Model Diagnostics for Scenario 4


Random Forest (Tuned)
Training Data (R² = 0.9993)
Random Forest (Tuned)
Testing Data (R² = 0.9652)
XGBoost (Tuned)
Training Data (R² = 1.0000)
XGBoost (Tuned)
Testing Data (R² = 0.9702)
KNN (Tuned)
Training Data (R² = 1.0000)
KNN (Tuned)
Testing Data (R² = 0.9378)
Predicted PP
Actual PP
1400
1600
1800
2000
2200
2400

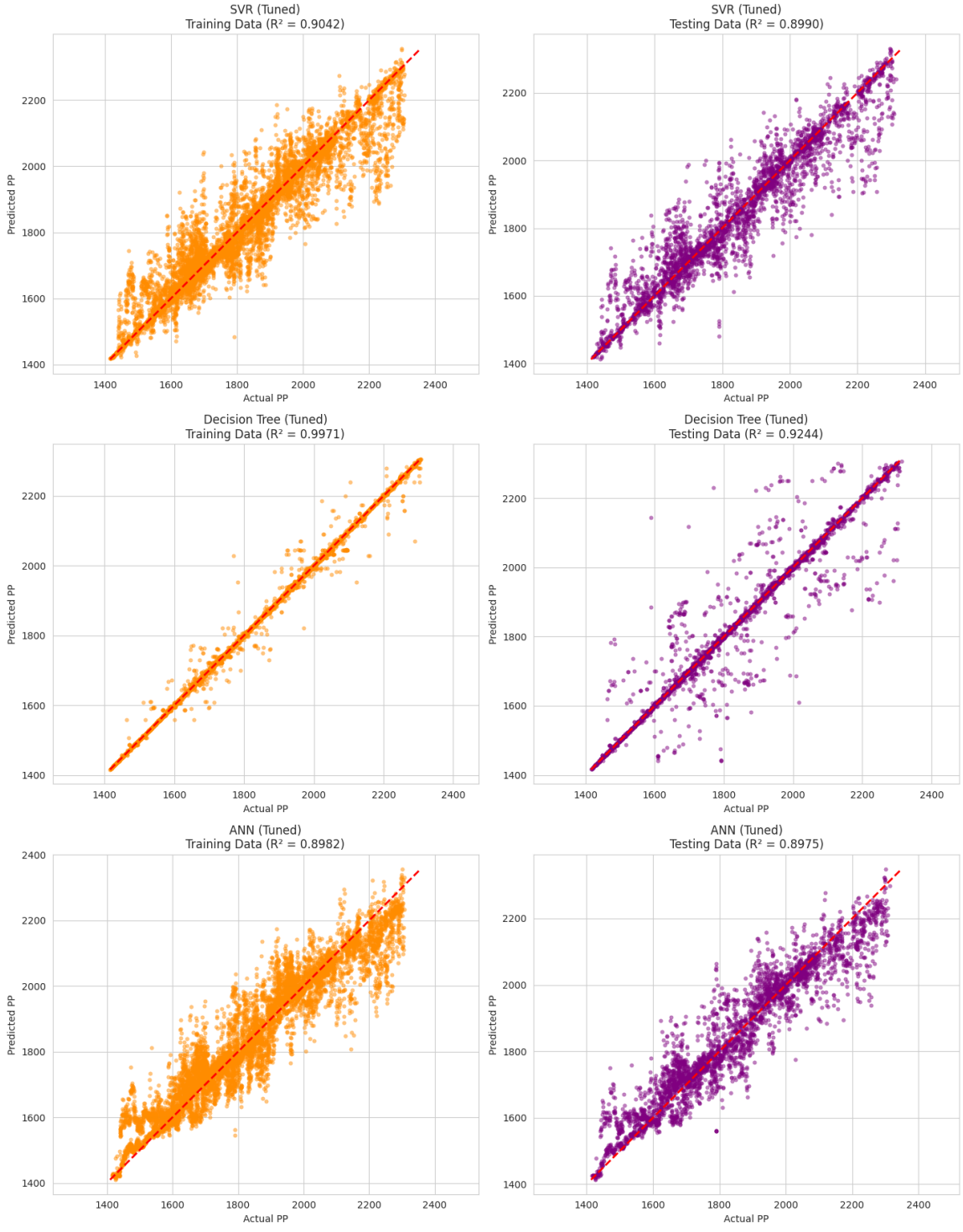
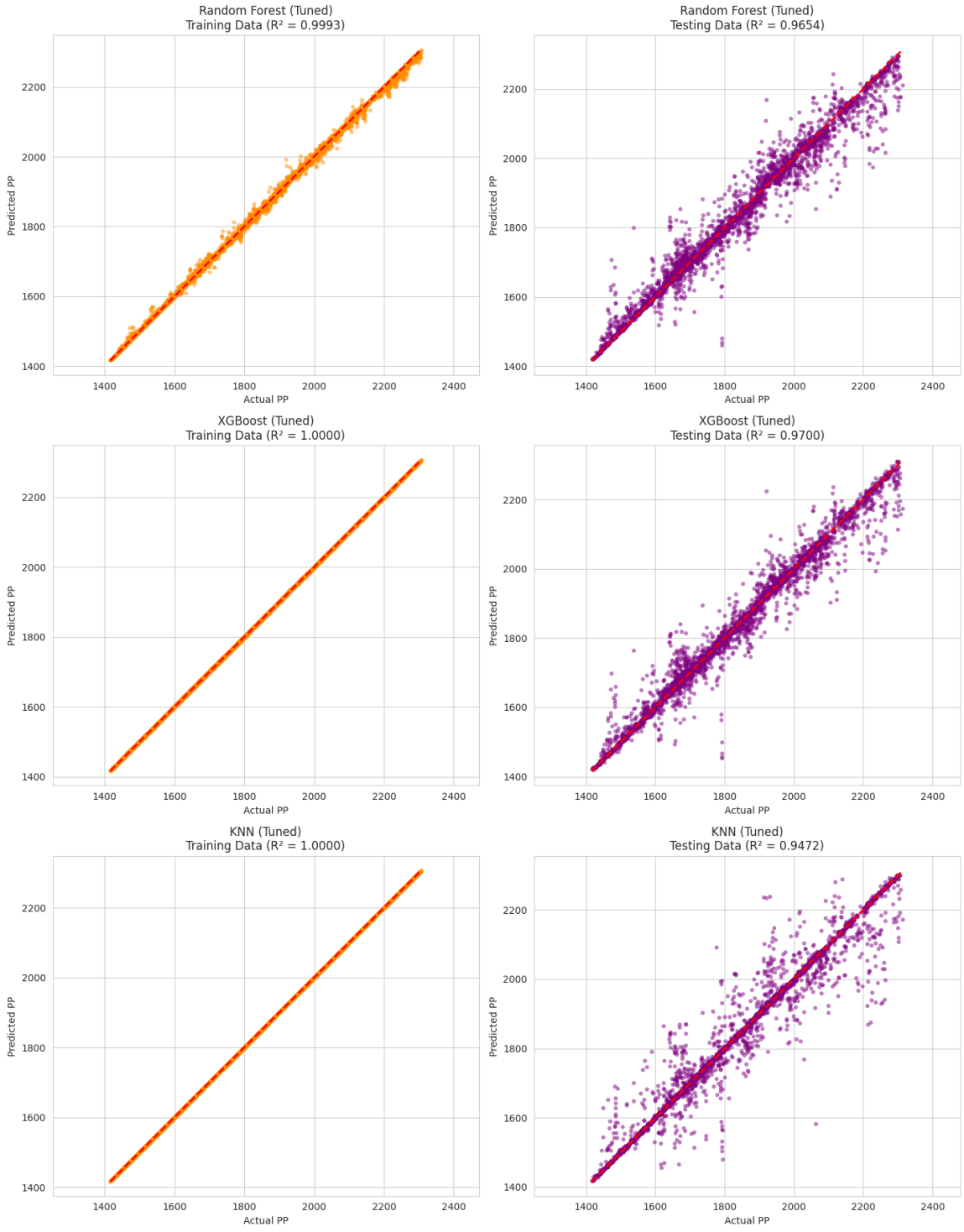


Figure 4. 11: Tuned Model Diagnostics for Scenario 1


Random Forest (Tuned)
Training Data (R² = 0.9993)
Random Forest (Tuned)
Testing Data (R² = 0.9654)
XGBoost (Tuned)
Training Data (R² = 1.0000)
XGBoost (Tuned)
Testing Data (R² = 0.9700)
KNN (Tuned)
Training Data (R² = 1.0000)
KNN (Tuned)
Testing Data (R² = 0.9472)
Predicted PP
Actual PP
1400
1600
1800
2000
2200
2400

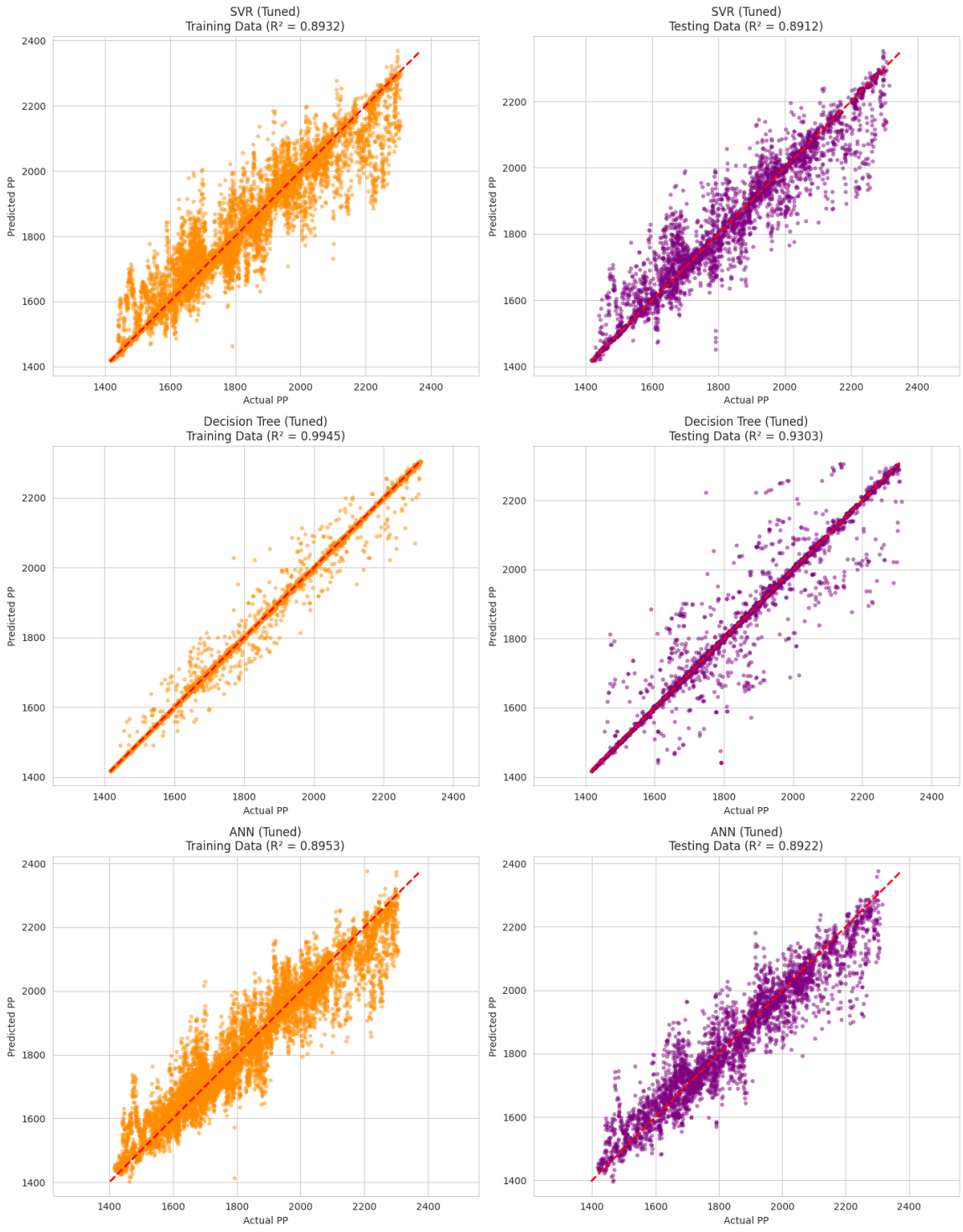
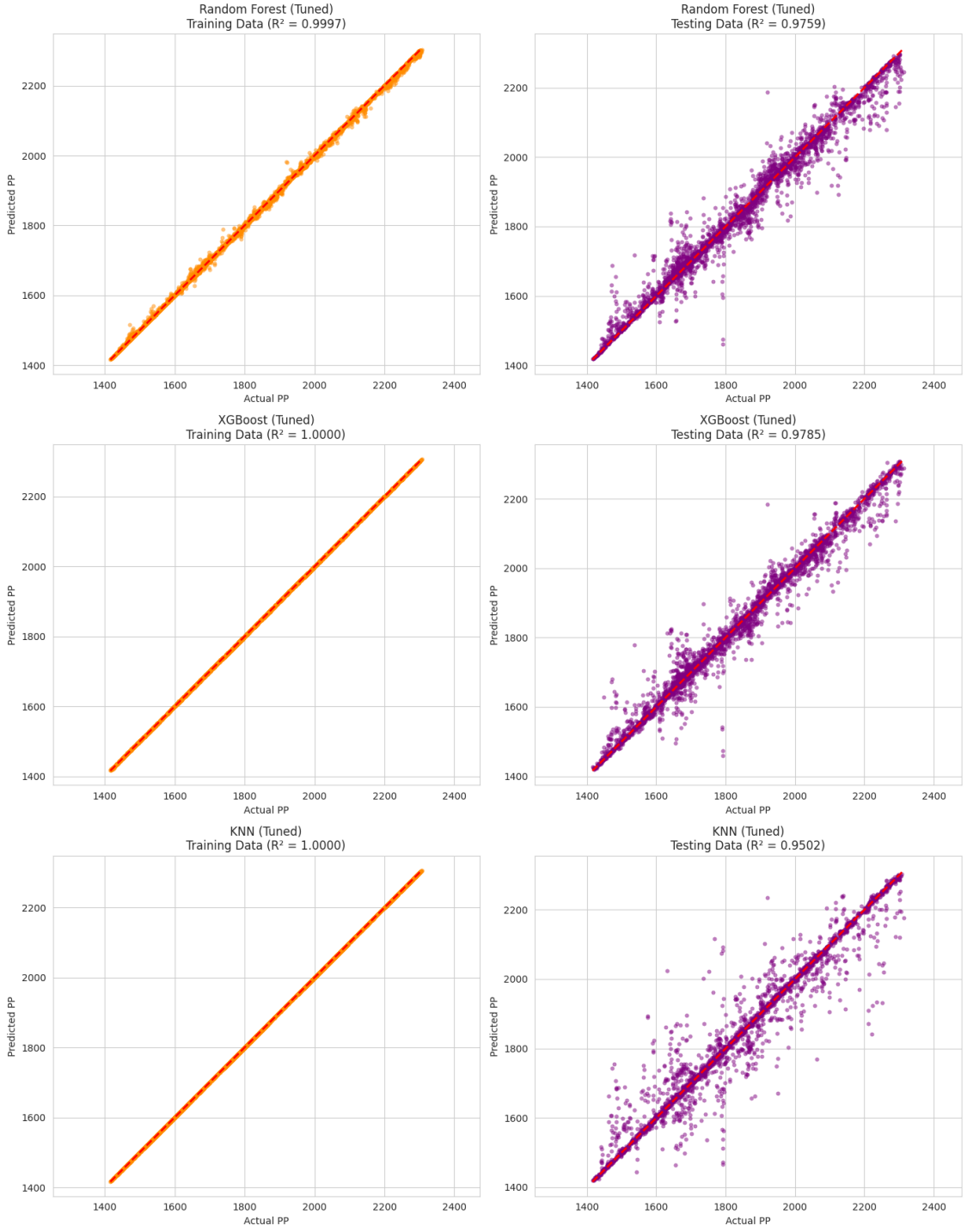


Figure 4. 12: Tuned Model Diagnostics for Scenario 2


Random Forest (Tuned)
Training Data (R² = 0.9997)
Random Forest (Tuned)
Testing Data (R² = 0.9759)
XGBoost (Tuned)
Training Data (R² = 1.0000)
XGBoost (Tuned)
Testing Data (R² = 0.9785)
KNN (Tuned)
Training Data (R² = 1.0000)
KNN (Tuned)
Testing Data (R² = 0.9502)
Predicted PP
Actual PP
1400
1600
1800
2000
2200
2400

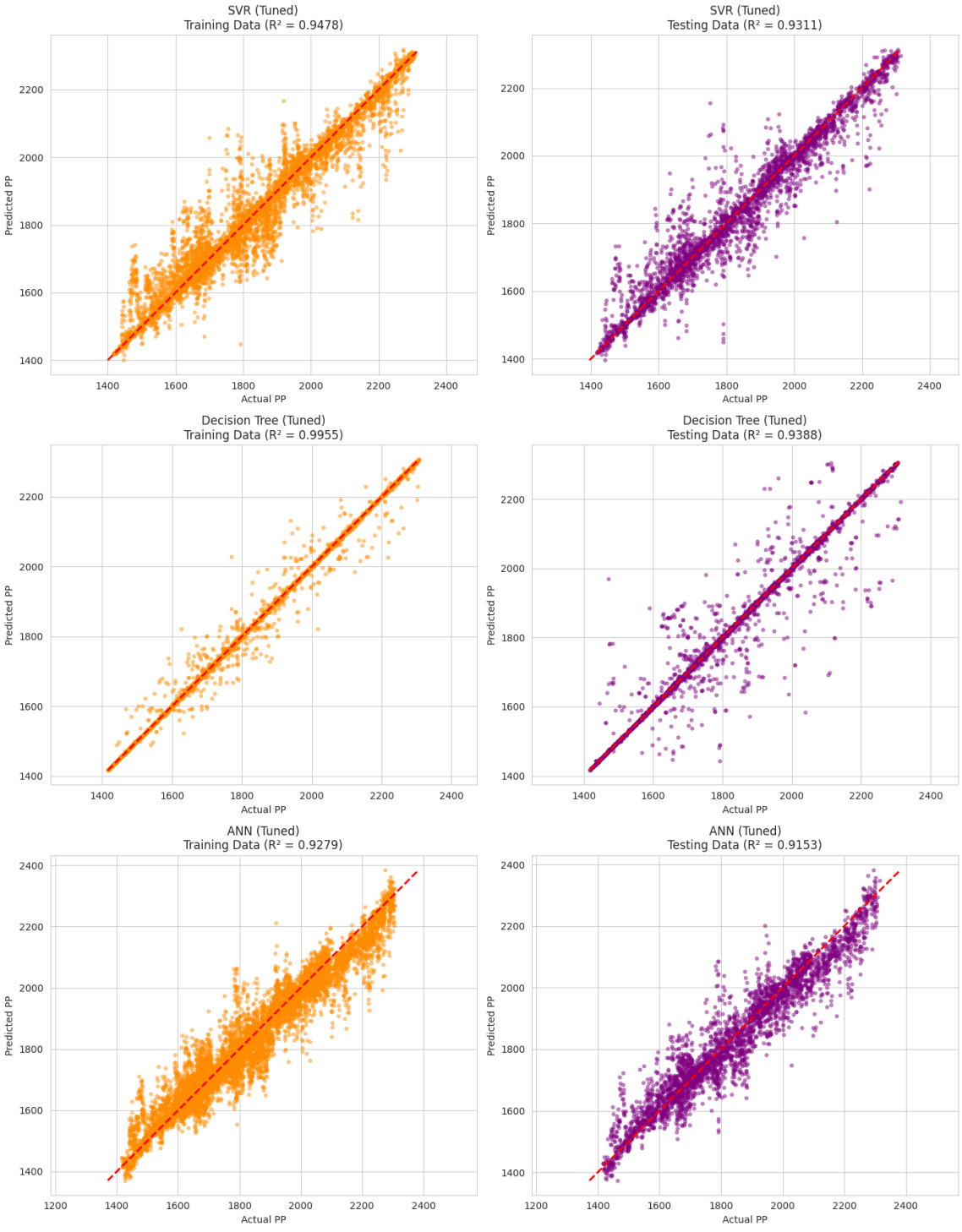
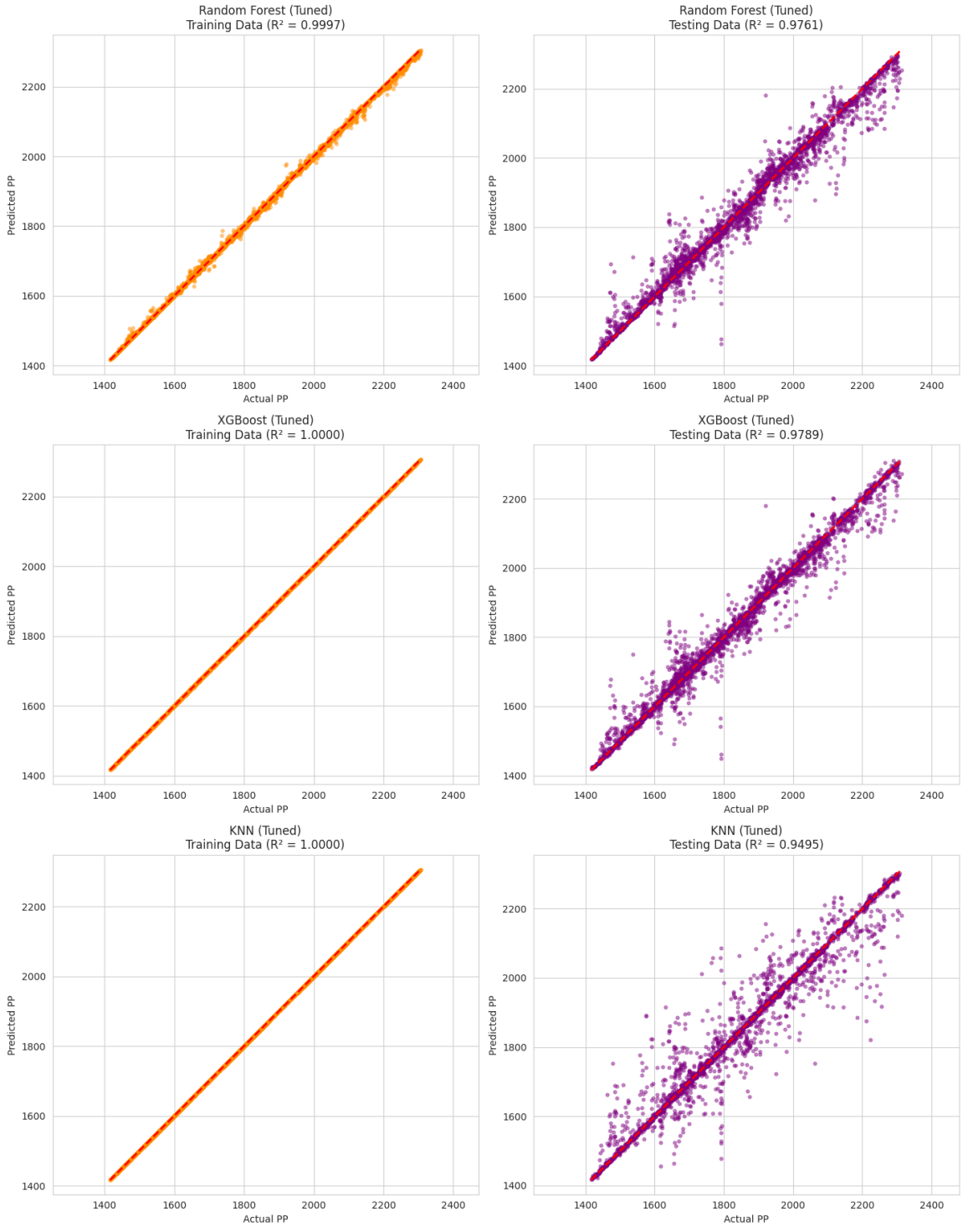


Figure 4. 13: Tuned Model Diagnostics for Scenario 3


Random Forest (Tuned)
Training Data (R² = 0.9997)
Random Forest (Tuned)
Testing Data (R² = 0.9761)
XGBoost (Tuned)
Training Data (R² = 1.0000)
XGBoost (Tuned)
Testing Data (R² = 0.9789)
KNN (Tuned)
Training Data (R² = 1.0000)
KNN (Tuned)
Testing Data (R² = 0.9495)
Predicted PP
Actual PP
1400
1600
1800
2000
2200
2400

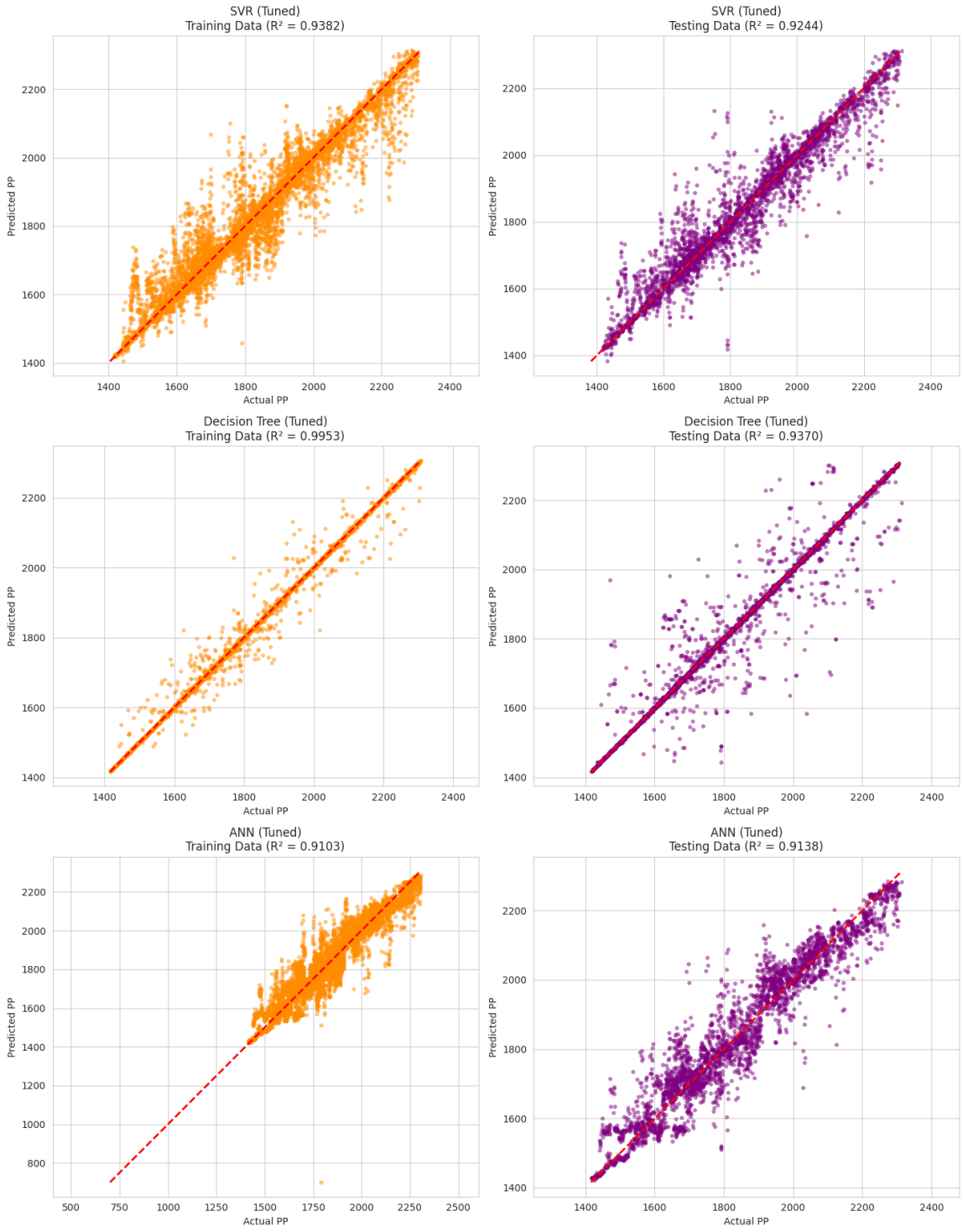


Figure 4. 14: Tuned Model Diagnostics for Scenario 4

Table 4. 9: Summary of Diagnostic Plot Analysis for Baseline and Tuned Models Across All Scenarios

Scenario	Model Type	Key Observations
Scenario 1 – Full Preprocessing	**Baseline (Fig. 4.7)**	Decision Tree shows severe overfitting (Train R^2=1.0, Test R^2=0.9197); ensembles (RF, XGBoost) generalize well; SVR poor fit (Test R^2=0.7131) with wide scatter despite cleaned data.
	Tuned (Fig. 4.11)	Universal improvement: SVR shows major boost (Test R^2=0.9325) with tight clustering; tuned ensembles achieve exceptionally strong fits; near-perfect train R^2 values but excellent generalization maintained.
Scenario 2 – Feature Selection Only	**Baseline (Fig. 4.8)**	Similar trends to Scenario 1; Decision Tree overfits; ensembles perform well; SVR performance drops further (Test R^2=0.6502) with more scatter, highlighting higher sensitivity to outliers than redundant features.
	Tuned (Fig. 4.12)	Ensembles remain strong; SVR and ANN show visibly wider scatter than in outlier-capping scenarios, confirming negative impact of outliers on tuned performance.
Scenario 3 – Outlier Capping Only	**Baseline (Fig. 4.9)**	Highest baseline performance for ensembles (RF Test R^2=0.9706, XGBoost Test R^2=0.9680) with tight clustering; SVR slightly better than in raw data cases, reinforcing its outlier sensitivity.
	Tuned (Fig. 4.13)	Peak performance for several models; Tuned SVR achieves best R^2=0.9568 with tight clustering, nearly matching ensembles; shows competitiveness with outliers controlled.

Scenario 4 – Raw Data	**Baseline (Fig. 4.10)**	Ensembles remain robust to both outliers and multicollinearity; SVR and KNN plots highly scattered, confirming vulnerability to unprocessed data.
	Tuned (Fig. 4.14)	Tuned XGBoost emerges as champion model (Test R^2=0.9789); exceptionally tight and well-distributed clustering confirms strong generalization even on unprocessed data.

4.6.Interpretation of Key Findings

The four-scenario experimental design yielded a rich dataset on model performance, leading to several key insights that address the core research questions of this thesis.

Across all four experimental scenarios, both before and after tuning, the ensemble models Random Forest and XGBoost consistently demonstrated superior performance over the other algorithms. The final results conclusively identified the **Tuned XGBoost model, trained on the raw, unprocessed data (Scenario 4), as the optimized model**, achieving the highest R^2 (0.9789) and lowest RMSE (31.98).

The success of these ensemble methods, particularly XGBoost, can be attributed to their inherent design. XGBoost's sequential, gradient-boosted approach, where each new tree is trained to correct the errors of its predecessors, allows it to build a highly accurate and nuanced predictive function (Hastie et al., 2009). The results strongly suggest that this methodology is exceptionally well-suited to capturing the complex, non-linear relationships between petrophysical logs and pore pressure.

One of the most significant findings of this study is the impact of feature selection. The analysis consistently showed that for the top-tier models (XGBoost and Random Forest), performance was **better when no features were removed**. For instance, after tuning, the XGBoost model's R^2 on capped data improved from 0.9694 (with feature selection) to 0.9792 (without).

This result challenges the conventional wisdom that removing multicollinear features is always beneficial. Our feature selection was based on linear correlation, identifying depth as redundant with Stress. However, the higher performance of the models using the full feature set suggests that Depth must contain unique, **non-linear information** that Stress does not capture. While the two features are linearly coupled, Depth may encode subtle, localized variations in compaction trends or lithological boundaries that a powerful non-linear model like XGBoost can exploit. By removing it, we were inadvertently "blinding" the model to this valuable, albeit non-linear, signal. This highlights a critical principle: for complex, non-linear algorithms, features that are redundant in a linear sense may not be redundant in a non-linear one.

The experimental design provided a clear verdict on the utility of outlier capping. The impact was highly dependent on the chosen algorithm. For robust, tree-based models like XGBoost and Random Forest, the effect of capping was minimal and, in the case of the optimized model, slightly negative.

In stark contrast, for the margin-based SVR model, outlier capping was absolutely critical. In the "Without Feature Selection" scenarios, applying the capping procedure improved the Tuned SVR's R^2 score from 0.9244 to 0.9568. This empirically validates the theoretical understanding that algorithms whose loss functions are sensitive to the magnitude of errors (like SVR) are disproportionately affected by extreme values. The outliers distort the optimal placement of the hyperplane and support vectors, leading to a globally suboptimal model. This finding demonstrates that outlier treatment is not a universally required step but a **targeted intervention necessary to enable sensitive models to perform competitively**.

4.7. Addressing Methodological Considerations

This section directly addresses two important critiques regarding the interpretation of the data and the justification for the preprocessing methods employed.

A critical perspective on data preprocessing posits that statistical outliers in geological data are often not measurement errors but rather manifestations of genuine, albeit infrequent, formation properties (e.g., thin, highly porous sand streaks within a thick shale sequence).

From this viewpoint, altering these "outliers" could be seen as removing valid geological information.

This study acknowledges the validity of this perspective. The methodological choice to apply outlier capping in two of the four scenarios was therefore **not an attempt to "clean" the data of valid information or to correct perceived errors**. Instead, it was a **practical and necessary step to facilitate a fair and comprehensive comparison across all selected algorithms**. As the results conclusively demonstrated, without outlier treatment, the SVR model's performance was severely compromised. To have evaluated SVR only on raw data would have been to compare a fundamentally disadvantaged algorithm against others that are inherently robust to such data.

Therefore, the capping procedure should be understood as a form of **data normalization for methodological purposes**, enabling a level playing field to compare the optimal potential of every algorithm. The fact that the ultimate Optimized model (XGBoost) performed best on the un-capped data reinforces the initial premise: the outliers are indeed valuable information, but only for algorithms robust enough to handle them.

4.8. Implications of Research

The findings of this thesis have significant practical and academic implications.

- **For Practitioners:** The results suggest that for pore pressure prediction tasks, using a state-of-the-art, robust model like XGBoost on raw, complete well log data may be a more effective and efficient workflow than engaging in complex preprocessing. It prioritizes the algorithm's power over manual data manipulation. However, if simpler or more interpretable models are required, the study highlights which preprocessing steps (like outlier capping for SVR) are mandatory.
- **For Academia:** This research contributes a clear, empirical case study demonstrating that the "best practices" of data preprocessing are not universal. It highlights the need to move beyond a one-size-fits-all approach and instead consider the interaction between data characteristics, preprocessing techniques, and algorithm choice as a core part of the experimental design.

CHAPTER 05: CONCLUSION AND FUTURE WORK

This thesis embarked on a systematic investigation into the application of machine learning for pore pressure prediction, with a unique focus on the interplay between data preprocessing strategies and algorithm performance. This concluding chapter synthesizes the key findings derived from the comprehensive four-scenario analysis. It outlines the primary contributions of this research to the field of applied data science in geomechanics, candidly discusses the limitations of the study, and provides concrete recommendations for future research that can build upon the insights gained.

5.1. Summary of Findings

The research set out to answer four key questions, and the experimental results have provided clear and, in some cases, counter-intuitive answers.

- **The Best Algorithm:** The primary objective was to identify the most accurate and robust algorithm for pore pressure prediction. Across all scenarios, the ensemble models consistently outperformed other methods. The **Tuned XGBoost model emerged as the best model**, particularly when trained on raw, unprocessed data, where it achieved the highest predictive accuracy with a testing R^2 of 0.9789 and the lowest RMSE of 31.98. The Tuned Random Forest and Tuned K-Nearest Neighbors models also proved to be top-tier performers.
- **The Impact of Preprocessing:** The study revealed that the utility of preprocessing is highly context dependent.
    - **Feature Selection:** The removal of linearly correlated features (Depth, GR, Porosity) was found to be **counter-productive for the top-performing models**. Both XGBoost and Random Forest achieved higher accuracy when trained on the full feature set, suggesting they were able to extract unique, non-linear information from features that appeared redundant under linear analysis.
    - **Outlier Capping:** The impact of outlier treatment was highly model-specific. For robust tree-based ensembles like XGBoost, capping had a

negligible effect. However, for the sensitive Support Vector Regression (SVR) model, outlier capping was **a critical and necessary step**, dramatically improving its R^2 score from 0.71 to 0.93 in the baseline case and transforming it into a competitive model after tuning.

- **The Optimal End-to-End Pipeline:** Contrary to common assumptions, the optimal pipeline for achieving maximum accuracy was **not the one with the most extensive preprocessing**. The best overall result was obtained using the **Tuned XGBoost model trained directly on the raw data**, with no outlier capping or feature selection applied. This highlights that for a sufficiently powerful algorithm, extensive data manipulation may be unnecessary.
- **Influential Petrophysical Parameters:** The feature importance analysis, derived from the robust Random Forest and XGBoost models, consistently identified Stress and Resistivity as the two most influential predictors of pore pressure, followed by Vp. This aligns with petrophysical principles, where overburden stress is a primary driver of pressure, and resistivity and velocity logs are key indicators of formation properties and fluid content.

5.2. Contribution to Knowledge

This thesis makes several key contributions to the field of applied machine learning in the geosciences:

- **Methodological Framework for Preprocessing Evaluation:** The primary contribution is the four-scenario experimental design itself. This study provides a clear and replicable framework for moving beyond a single, assumed "best practice" pipeline. It empirically demonstrates a methodology for testing the interaction between data preparation techniques and algorithm choice, which can be adopted by future researchers to produce more nuanced and reliable findings.
- **Empirical Evidence on the Nuances of Preprocessing:** This research provides strong, data-driven evidence that challenges one-size-fits-all approaches to data cleaning. It proves that:

- Removing features based on linear correlation can be detrimental to complex, non-linear models.
- The necessity of outlier treatment is not a universal rule but is instead a function of the chosen algorithm's sensitivity.

- **A Robust Comparative Analysis for Pore Pressure Prediction:** By systematically training, tuning, and evaluating six different machine learning models across four distinct data scenarios, this thesis offers one of the most comprehensive comparative analyses on this specific problem to date, providing a clear ranking and justification for the use of XGBoost in practical applications.

5.3. Limitations of the Study

While this research was conducted with methodological rigor, it is important to acknowledge its limitations:

- **Data Provenance and Generalizability:** The study was conducted on an open-source dataset with no specific geographical or geological provenance. While this does not invalidate methodological comparisons, it does limit the direct applicability of the final trained model to a specific, known oil field. The model's ability to generalize to different geological basins remains untested.
- **Limited Feature Space:** The analysis was constrained to the well log data provided in the dataset. Other potentially valuable data sources, such as seismic attributes (e.g., interval velocity), detailed mud logging data, or offset well information, were not available. The inclusion of these features could further enhance model accuracy.
- **Scope of Hyperparameter Tuning:** While extensive, the RandomizedSearchCV process samples from a predefined hyperparameter space. It is possible that a more computationally intensive search, such as a Bayesian Optimization or a random search with more iterations, could discover a slightly better configuration, particularly for the complex ANN model.

5.4. Recommendations for Future Work

Based on the findings and limitations of this study, several promising avenues for future research are recommended:

- **Model Validation on Characterized Datasets:** The next logical step is to apply the Optimized model (Tuned XGBoost on raw data) and the methodological framework from this thesis to proprietary datasets from well-characterized geological basins. This would test the generalizability of the findings and move towards creating a field-specific predictive tool.
- **Integration of Multi-Modal Data:** Future research should focus on incorporating a wider range of data types. Integrating seismic attributes could provide valuable information on rock properties between wells, potentially improving the model's spatial prediction capabilities.
- **Advanced Feature Engineering:** Based on the observation of a bimodal Vp distribution, future work could explore unsupervised clustering techniques (e.g., K-Means) to automatically generate a "lithology" or "rock type" feature. Explicitly providing this information to the models could lead to further performance gains.
- **Exploration of Model Interpretability:** While the XGBoost model was the most accurate, it is often considered a "black box." Future research should apply model interpretability techniques, such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations), to better understand *how* the Optimized model makes its predictions. This would increase confidence in the model and could reveal novel insights into the complex relationships between petrophysical logs and pore pressure.

REFERENCES

Abdelaal, A., Elkatatny, S., & Abdulraheem, A. (2021). Data-driven modeling approach for pore pressure gradient prediction while drilling from drilling parameters. *ACS Omega*, *6*(21), 13807–13816. https://doi.org/10.1021/acsomega.1c01340

Ahmed, A., Elkatatny, S., Ali, A. Z., Abdulraheem, A., & Mahmoud, M. (2019, March). *Artificial neural network ANN approach to predict fracture pressure* [Paper presentation]. SPE Middle East Oil and Gas Show and Conference, Manama, Bahrain. https://doi.org/10.2118/194852-MS

Azadpour, M., Shad Manaman, N., Kadkhodaie-Ilkhchi, A., & Sedghipour, M.-R. (2015). Pore pressure prediction and modeling using well-logging data in one of the gas fields in south of Iran. *Journal of Petroleum Science and Engineering*, *128*, 15–23. https://doi.org/10.1016/j.petrol.2015.02.022

Blanco, Y., & Turner, M. (2011, October). *Application & evolution of formation pressure while drilling technology (FPWD) applied to the Gulf of Mexico* [Paper presentation]. SPE Annual Technical Conference and Exhibition, Denver, Colorado, USA. https://doi.org/10.2118/147556-MS

Eaton, B. A. (1975). The theory of abnormal pore pressure prediction. *Journal of Petroleum Technology*, *27*(1), 21–26. https://doi.org/10.2118/5173-PA

Edwards, N. (n.d.). *Box and whisker plot: Definition, how to draw a box and whisker plot? | Example*. BYJU'S. Retrieved July 12, 2025, from https://byjus.com/maths/box-and-whisker-plot/

Feng, J., Wang, Q., Li, M., Li, X., Zhou, K., Tian, X., & Sun, M. (2024). Pore pressure prediction for high-pressure tight sandstone in the Huizhou Sag, Pearl River Mouth Basin, China: A machine learning-based approach. *Journal of Marine Science and Engineering*, *12*(5), 703. https://doi.org/10.3390/jmse12050703

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann Publishers.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Kaiser, S. (2017). Incorporating rock brittleness for fracture pressure prediction in sandstones. *Journal of Geomechanics*, *8*(2), 125-136. https://doi.org/10.1016/j.jog.2016.10.002

Li, H., Tan, Q., Deng, J., Dong, B., Li, B., Guo, J., & Bai, W. (2023a). A comprehensive prediction method for pore pressure in abnormally high-pressure blocks based on machine learning. *Processes*, *11*(9), 2603. https://doi.org/10.3390/pr11092603

Li, H., Tan, Q., Li, B., Feng, Y., Dong, B., Yan, K., & Chen, J. (2023b). Physically-data driven approach for predicting formation leakage pressure: A dual-drive method. *Applied Sciences*, *13*(18), 10147. https://doi.org/10.3390/app131810147

Ogbu, A. D., Iwe, K. A., Ozowe, W., & Ikevuje, A. H. (2024). Advances in machine learning-driven pore pressure prediction in complex geological settings. *Computer Science & IT Research Journal*, *5*(7), 1648–1665. https://doi.org/10.51594/csitrj.v5i7.1350

Ramatullayev, S., Makhmotov, A., Zhabagenov, M., Cesari, M., Torrisi, S., Capone, G., & Ferrari, F. (2019, November). *Formation pressure while drilling technology: Game changer in drilling overpressured reservoirs* [Paper presentation]. Abu Dhabi International Petroleum Exhibition & Conference, Abu Dhabi, UAE. https://doi.org/10.2118/198367-MS

Sanei, M., Ramezanzadeh, A., & Asgari, A. (2024). Applied machine learning-based models for determining the magnitude of pore pressure and minimum horizontal stress. *Arabian Journal of Geosciences*, *17*(210). https://doi.org/10.1007/s12517-024-11997-2

Sun, Y., Pang, S., Zhang, Y., & Zhang, J. (2024). Application of the dynamic transformer model with well logging data for formation porosity prediction. Physics of Fluids, 36(3), 036620. https://doi.org/10.1063/5.0193903

Tammy Reservoir. (2022). *Pore-Pressure-Prediction-for-Oil-and-Gas* [Source code]. GitHub. https://github.com/tammyreservoir/Pore-Pressure-Prediction-for-Oil-and-Gas/tree/main/Dataset/raw

Tan, Q., Hao, X., Luo, C., Zhang, J., & Weng, H. (2020). Application of Bowers Model in Abnormal Pore Pressure Prediction in Deepwater Drilling. *IOP Conference Series: Earth and Environmental Science*, *513*(1), 012062. https://doi.org/10.1088/1755-1315/513/1/012062

Zhang, J., & Yin, S.-X. (2017). Fracture gradient prediction: An overview and an improved method. *Petroleum Science*, *14*(4), 720–730. https://doi.org/10.1007/s12182-017-0182-1

APPENDICES

- ✓ Appendix A: Python Code for Scenario 1 (Full Preprocessing)
  https://colab.research.google.com/drive/1p2uCK6kGlNK4aK1El7iR3376z-t1357r?usp=sharing

- ✓ Appendix B: Python Code for Scenario 2 (Feature Selection Only)
  https://colab.research.google.com/drive/142Rlbp5rz0D41C7EfP5-lXXRAwQbcE1j?usp=sharing

- ✓ Appendix C: Python Code for Scenario 3 (Outlier Capping Only)
  https://colab.research.google.com/drive/14Y9tFYMvQ94jHv_rnlv2O1eJGqKAU69J?usp=sharing

- ✓ Appendix D: Python Code for Scenario 4 (Raw Data)
  https://colab.research.google.com/drive/1dmNll4onTSF0uVR3T5kB10nelWgKbLaf?usp=sharing