GRÓ · GTP
Geothermal
Training
Programme

unesco
Centre
Under the auspices
of UNESCO

Predicting real-time geothermal well flow rate and enthalpy with machine learning techniques

Agata Rostrán Largaespada

Thesis of 60 ECTS credits
Master of Science (M.Sc.) in Sustainable Energy Engineering

August 2024

Predicting real-time geothermal well flow rate and enthalpy with machine learning techniques

Thesis of 60 ECTS credits submitted to the School of Science and Engineering at Reykjavík University in partial fulfilment of the requirements for the degree of
Master of Science (M.Sc.) in Sustainable Energy Engineering

August 2024

Supervisors:

María Sigríður Guðjónsdóttir
Supervisor Assistant Professor, Reykjavík University, Iceland

Egill Júlíusson, Co-Advisor
CTO Arctic Green Energy, Iceland

Examiner:

Pálmar Sigurðsson, Examiner
Project Manager at Reykjavik Energy, Iceland

Predicting real-time geothermal well flow rate and enthalpy with machine learning techniques

Agata Rostrán Largaespada

August 2024

Abstract

Geothermal energy is a sustainable energy source offering reliable and renewable energy solutions. However, accurately measuring geothermal well output like flow rate and enthalpy for wells that produce a two-phase fluid remains challenging due to the complexity and infrequency of traditional methods. This thesis addresses these issues by continuing the work of developing a real-time method to measure flow rate and enthalpy from geothermal wells without interrupting operations. The focus is on accurately estimating geothermal fluids' flow rate and enthalpy using advanced rule-based models and machine learning techniques.

This research integrates data-driven approaches for continuous monitoring and early detection of well performance changes by using measurements from Landsvirkjun's geothermal operations conducted in 2019, 2020, 2021, and 2023. The study employs a specialized differential pressure orifice plate meter setup at Theistareykir and Bjarnarflag Geothermal Power Plants, providing detailed measurements critical for the models.

The most effective model employed Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for noise reduction, Recursive Feature Elimination with Cross-Validation (RFECV) for precise feature selection, and Random Forest Regression (RFR) with five key features, achieving a Root Mean Square Error (RMSE) of 0.011. This approach can significantly enhance the efficiency and accuracy of geothermal power production measurements, offering insights into real-time monitoring and operational optimization.

Spá fyrir um rauntíma rennslishraða og entalpíu jarðvarmalinda með vélrænum námsaðferðum

Agata Rostrán Largaespada

ágúst 2024

Útdráttur

Jarðvarmaorka er sjálfbær orkuauðlind sem býður upp á áreiðanlegar og endurnýjanlegar orkulausnir. Hinsvegar getur verið erfitt að mæla nákvæm afköst borhola, eins og rennslishraða og vermi í borholum sem framleiða tvífasa vökva, vegna þess hve flóknar og lítt notaðar hefðbundnar mæliaðferðir eru. Þessi ritgerð er áframhald á þróun á rauntímamælinga á rennslishraða og vermi frá jarðhitaborholum án þess að trufla nýtingu þeirra. Megináherslan er á að áætla rennslishraða og entalpíu jarðvarmavökva með nákvæmni með notkun háþróaðra reglubundinna líkana og vélrænnar námsaðferða.

Rannsóknin samþættir gagnadrifnar nálganir til samfelldrar vöktunar og snemmbúinnar greiningar á breytingum á afköstum jarðvarmalinda, með því að nota mælingar frá Landsvirkjun sem gerðar voru á árunum 2019, 2020, 2021 og 2023. Sérstök mismunadrifninn þrýstingsþveropnumælir var notaður við Þeistareykja- og Bjarnarflagsvirkjun til að fá ítarlegar mælingar sem eru lykilatriði fyrir líkanagerðina.

Árangursríkasta líkanið sem notað var í þessari rannsókn, notaði þéttleikadrifna hópmyndun með hávaða (DBSCAN) til að minnka hávaða, endurtekið úrtaksaðdrátt með krossstaðfestingu (RFECV) fyrir nákvæma val á eiginleikum, og slembiskógarreiknivél (RFR) með fimm lykileiginleikum, sem náði rótum meðalvillutölu (RMSE) upp á 0,011. Þessi nálgun getur verulega aukið skilvirkni og nákvæmni mælinga á jarðvarmavinnslu og veitt innsýn í rauntímavöktun og rekstrarhagræðingu.

The undersigned hereby grants permission to the Reykjavík University Library to reproduce single copies of this Thesis entitled **Predicting real-time geothermal well flow rate and enthalpy with machine learning techniques** and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the Thesis, and except as herein before provided, neither the Thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatsoever without the author's prior written permission.

02/09/2024

date

Agata Rostrán
Largaespada
Master of Science

I dedicate this to my family, especially to my father.

Acknowledgements

I am deeply grateful to GRÓ GTP and ENEL for this incredible opportunity and for sponsoring my studies. I sincerely thank Friða Ómarsdóttir, Guðni Axelsson, and Ingimar Haraldsson for their constant support. I extend my sincere gratitude to my supervisor, María Sigríður Guðjónsdóttir, and my advisor, Egill Julíusson, for their invaluable guidance. Special thanks to Landsvirkjun for providing the essential data and to Hilmar Einarsson and Helgi Alfreðsson for their assistance. I am also thankful for the wonderful friends I made during this journey. Lastly, I am grateful to my family - my parents, sisters, and loving husband for their constant encouragement and love.

Preface

This dissertation is an original work by the author, Agata Rostran Largaespada.

Contents

List of Figures

List of Tables

List of Abbreviations

BJ	Bjarnaflag
DBSCAN	Density Base spatial cluster
DP	Differential pressure
EM	Experimental Measurement
MAPE	Mean Absolute Percentage Error
PC	Principal component
PCA	Principal component analysis
RF	Random forest
RFECV	Recursive Feature Elimination with Cross-Validation
RFR	Random forest regression
RM	Reference Measurements
RMSE	Root mean Squared Error
WCSS	Within-Cluster Sum of Square
WHP	Well head pressure
ÞG	Theistareykir

List of Symbols

Symbol	Description	Value/Units
A	Cross-sectional area	m^2
Cd	Discharge coefficient	
D	Pipe diameter	m
d	Orifice plate diameter	m
DR	Density ratio	
g	Gravitational constant	m/s^2
h	Enthalpy	kJ/kg
i	Sample	
L_1, L_2	Coefficient of tapping type	
$\dot{m}$	Mass flow rate	kg/s
n	Total number of samples	
P	Pressure	bar, Pa
P	Power	W
PLR	Pressio loss ratio	
PRR	Pressure recovery pressure	
RPR	Recovered DP to ppl Ratio	
$\dot{Q}$	Volumetric flow	m^3/s
Re	Reynolds number	
S	Slip ratio	
T	Temperature	°C
x	Steam quality	
β	Diameter ratio	
ϵ	Expansibility coefficient	
ρ	Density	kg/m^3
Δ	Differential	
μ	Dynamic viscosity	Pa S

Symbol	Description	Value/Units
Subscripts		
1,2,3	Pressure Taps	
s	steam phase	
w	water phase	
ppl	permanent pressure loss	
r	recovery	
t	traditional	

Chapter 1

Introduction

Geothermal energy plays an essential role as a reliable and renewable energy source. It harnesses the natural heat from the earth, providing a reliable and environmentally friendly energy supply. However, accurately measuring flow rate and enthalpy from geothermal wells, which produce a two-phase mixture of water and steam, presents a unique challenge. Traditional methods, such as the tracer dilution method, are complex, labour-intensive, and typically conducted a few times a year. Accurate and frequent measurements of flow rate and enthalpy from individual wells are crucial for balancing power demand with steam supply for geothermal power plants.

This project primarily focuses on developing a method that enables real-time measurement of flow rate and enthalpy from geothermal wells without requiring them to be taken offline. The project uses data from experiments conducted by Landsvirkjun in 2019, 2020 and 2021, along with new data collected in 2023.

The main research goal of this thesis is to model a real-time measurement method for estimating geothermal fluid flow rate and enthalpy in a wide range of geothermal conditions using rule-based and machine learning models. This approach seeks to enhance the accuracy and efficiency of well output measurements. This project is particularly relevant at an industrial level because it may enable continuous monitoring without shutting down wells for output measurement methods. Early detection of changes in well performance enables timely intervention, preventing potential issues and optimising energy production.

Experiments that were conducted by Landsvirkjun, Iceland's National Power Company, at Þeistareykir and Bjarnarflag Geothermal Power Plants resulted in a dataset organized into Reference Measurements (RM) and Experimental Measurements (EM). RM data includes field measurements of the total flow rate and enthalpy of the geothermal fluid using the separator and water tracer method, explained in Sections 2.1.1 and 2.1.2. EM data was gathered using a specialised differential pressure (DP) orifice plate meter setup, which uses three pressure taps instead of the traditional two, as described in Section 2.1.3. This setup may provide more detailed measurements and additional parameters, enhancing the accuracy of the data. Previous work in this field has primarily focused on traditional methods, which, although effective, have limitations regarding frequency and operational disruption. This project aims to provide a more efficient and accurate approach to real-time measurement by leveraging rule-based models and machine learning techniques. Rule-based models apply predefined rules to input data. In contrast, machine learning models use input and output data to learn and predict the relationships between variables, thereby improving the overall measurement process.

Chapter 2

Literature Review

2.1 Methods for Flow and Enthalpy Measurements for geothermal wells

In geothermal systems, the available power production is primarily determined by the flow rate and the enthalpy of the geothermal fluid. The geothermal industry employs various measurement techniques to assess these parameters, including total flow calorimetry, tracer dilution, separator methods, and lip pressure measurements. Alternative techniques such as differential pressure over an orifice plate, venturi meters, vortex meters, load cell sensors, and radio frequency methods have also been used. The choice of a specific measurement technique depends on the particular needs and circumstances of the project, as each method offers its unique set of advantages and limitations.

In this subchapter, the separator method, water tracer method, and orifice plate with differential pressure meter method will be explained in detail, as these were the techniques used in the experiments conducted in 2019, 2020, 2021, and 2023, which are used in the analysis in this work. The separator method and orifice plate with differential pressure meter were used in the 2019 - 2021 experiments, while the 2023 experiments used the water tracer and orifice plate methods. Other methods will also be briefly discussed, explaining why they were not selected for these experiments.

2.1.1 Separator Method

The separator method involves using a steam separator at the end of the flow line to measure the real-time output of geothermal systems, focusing on the single-phase flow rates of steam and water. Typically, the steam flow is measured using differential pressure (DP) across an orifice plate or other gas flow meters, while the volumetric mass flow of the separated water is assessed with a water weir. (Helbig & Zarrouk, 2012). The two-phase flow rate entering the separator is calculated by summing the steam and water flow rates exiting the separator. To determine the total enthalpy h, the dryness fraction (steam quality, x) is utilised in the following equation:

$$h = xh_s + (1 - x)h_w \tag{2.1}$$

Here, enthalpies of water (h_w) and steam (h_s) represent the saturation properties at the separator pressure. Steam quality, also known as the dryness fraction, indicates the proportion of vapour in the fluid within two-phase flow systems.

$$x = \frac{\dot{m}_s}{\dot{m}_s + \dot{m}_w} \tag{2.2}$$

Here, ṁs and ṁw represent the steam and water flow rates, respectively. This method yields dependable real-time results and does not require wells to be taken offline but involves a substantial initial investment. Typically, multiple wells are linked to centralised separators, posing challenges in monitoring the individual performance of each geothermal well.

2.1.2 Water Tracer Method

This thesis used the water tracer method in the experiments and will be discussed in more detail in this section. The total enthalpy and the flow rate of a mixture of steam and water from a well can be determined by introducing specific chemical tracers of known concentrations into the pipeline carrying the two-phase substance (Lovelock, 2001). The tracers utilised for the steam and water phases are different. This difference arises because water and steam have distinct physical properties, especially in terms of solubility. Tracers for the water phase are selected for their high solubility in water, allowing them to accurately track the flow of the liquid phase. In contrast, steam phase tracers are chosen for their ability to remain in the gaseous phase, ensuring they effectively measure steam flow without dissolving in water.

Hirtz et al., (2001) explained that measuring the water phase typically involves using common tracers like potassium fluoride (KF), sodium bromide (NaBr), fluorescein dye, sodium benzoate, rhodamine WT dye, 1,5-naphthalene disulfonate, and 2,7-naphthalene disulfonate. Meanwhile, the steam phase employs tracers such as propane, sulfur hexafluoride (SF6), freon-12, helium, and isopropanol. The choice of a tracer depends on considerations like cost, availability, and local expertise.

The equipment for this technique consists of two main components: an injection pump rig and a sampling setup. The process involves introducing chemical tracers from a tracer feed bottle into the upstream section of a two-phase pipeline through a positive-displacement dosing pump, which controls the injection rate. Subsequently, the tracers present in both the liquid and steam phases are collected at a sampling point situated downstream. Ideally, the steam sample is obtained at the upper part of the pipe, while the water sample is taken from a lower point (see Figure 3.4). It is crucial to ensure that the downstream sampling point is adequately distant from the injection point, allowing for thoroughly mixing the tracers with the geothermal fluids.

The mass flow rate of steam and water can be calculated if chemical sampling results are available using the methods from (Lovelock, 2001), (P. Bixley, N. Dench, and D.Wilson, 1998):

$$\dot{m}_w = \frac{\dot{m}_{wt}}{\dot{c}_{wt}}, \quad \dot{m}_s = \frac{\dot{m}_{st}}{\dot{c}_{st}} \tag{2.3}$$

Where $\dot{m}_{wt}$ is the flow rate of the water tracer injection, $\dot{C}_{wt}$ is the concentration of the tracer in water, $\dot{m}_{st}$ is the flow rate of the steam tracer injection, and $\dot{C}_{st}$ is the tracer concentration in steam. Lovelock, (2001) states that the steam mass flow rate ($\dot{m}_s$) in Equation (2.4) should be corrected for the steam tracer dissolved in the water:

$$\dot{m}_s = \frac{(\dot{m}_{st}T_s) - (\dot{m}_w T_w)}{T_s} \tag{2.4}$$

Where T_w and T_s are the concentrations of tracers in the water and steam phases, both steam flow ($\dot{m}_s$) and water flow ($\dot{m}_w$) are calculated at the pipeline pressure. Using the measured $\dot{m}_s$ and $\dot{m}_w$ values at pipeline pressure and the enthalpies h_s and h_w determined based on the saturated properties at the separator pressure, the total enthalpy (h) can then be computed:

$$h = \frac{(\dot{m}_s h_s) - (\dot{m}_w h_w)}{\dot{m}_s + \dot{m}_w} \tag{2.5}$$

Test results are usually available only after a few days, once the tracer sample analysis has been completed and returned from the laboratory. Helbig & Zarrouk, (2012) demonstrated that this method is the least precise and involves relatively high ongoing costs.

2.1.3 Orifice Plate Method

The orifice plate method will be discussed in detail since it was also utilised in the experiments to obtain results for this study. A typical orifice meter comprises a circular metal plate installed between pipe flanges featuring a square-edged round aperture within, shown inFigure 2.1. A depiction of the general sharp-edge orifice plate setup is illustrated in Figure 2.1.These orifice plate meters can vary in design, with alternatives like conical entrance, quadrant edge, segmental, eccentric, and multi-hole plates, each suited for different flow conditions and fluid types. There are three common types of differential pressure tapping shown in Figure 2.1, A shows a flange tapping, B represents a radius tapping (D-D/2), and C depicts corner tappings (Mubarok et al., 2019).

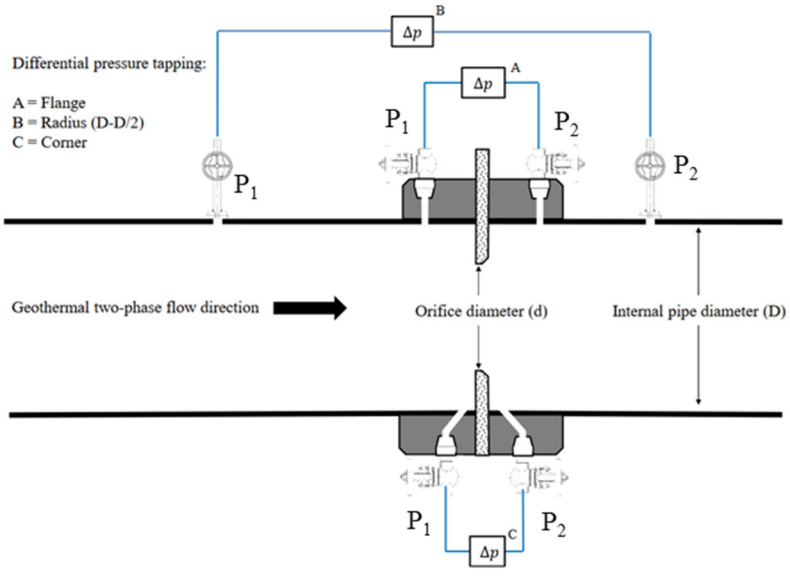


Figure 2.1 Cross-section of a sharp-edge concentric orifice plate and pressure tapping locations (Helbig & Zarrouk, 2012)

Orifice plate meters are widely used for measuring gas flow, operating on the principle of differential pressure. Research has been conducted to predict frictional pressure losses in pipes for two-phase flow using orifice meters. These meters have been extensively studied over the past decades, leading to the development of several ISO standards (J. Kinney and R. Steven, 2014). The total steam flow rate from orifice plate measurements is calculated

using the following equation:

$$\dot{m}_s = \frac{C_d}{\sqrt{1-\beta^4}} \epsilon \frac{\pi}{4} d^2 \sqrt{2\Delta P_t \rho_s} \tag{2.6}$$

Where ρ_s is the steam density, and ΔP_t denotes the pressure drop between the upstream pressure tap P_1 and the downstream pressure tap P_2 ($\Delta Pt = P_1 - P_2$, as shown in Figure 2.1). β, the diameter ratio is calculated by dividing the orifice diameter d, by the internal pipe diameter, D. Figure 2.1 illustrates these diameters. The formula for the β ratio is:

$$\beta = \frac{d}{D} \tag{2.7}$$

An iterative process is used to determine the discharge coefficient, C_d:

$$\begin{aligned} C_d = {} & 0.5961 + 0.0216\beta^2 + 0.216\beta^8 + 0.00521\left(\frac{10^6\beta}{Re_D}\right)^{0.7} \\ & + (0.0188 + 0.0063A)\beta^{3.5}\left(\frac{10^6}{Re_D}\right)^{0.3} \\ & + (0.043 + 0.08e^{-10L_1} - 0.123e^{-7L_1})(1 \\ & - 0.11A)\left(\frac{\beta^4}{1-\beta^4}\right) - 0.031(M_2 - 0.8M_2^{1.1})\beta^{1.3} \end{aligned} \tag{2.8}$$

Re_D, the pipe Reynolds number, is defined by the equation:

$$Re_D = \frac{4\dot{m}_s}{\pi\mu D} \tag{2.9}$$

Where μ is the dynamic viscosity of steam, D is the internal pipe diameter, and $\dot{m}_s$ is the steam flow rate, and the equations for substitution factors are as follows:

$$A = \left(\frac{1900\beta}{Re_D}\right)^{0.8} \quad , \quad M_2 = \frac{2L_2}{1-\beta} \tag{2.10}$$

The values of the L parameters depend on the pressure tap configuration; flange taps have L1 = L2 = 0.0254/D, radius taps have L1 = 1 and L2 = 0.47, and for corner taps, L=L2=0. Typically, the discharge coefficient, C_d, is about 0.6, so beginning the iterative process with $C_d = 0.6$ is recommended. For the water phase, the expansibility coefficient, ϵ used in Equation (2.6), is 1, but for steam, it is calculated using the following formula:

$$\epsilon = 1 - (0.351 + 0.256\beta^4 + 0.93\beta^8)\left[1 - \left(\frac{P_2}{P_1}\right)^{\frac{1}{1.3}}\right] \tag{2.11}$$

When the ratio (P_2/P_1) >0.75.

2.1.4 Other Methods

The following subsections will briefly discuss the total flow calorimeter, lip pressure, and load cell sensor methods. These methods, while effective in certain applications, were not selected for use in this study's experiments because some require the wells to be taken offline during measurements, making them unsuitable for the continuous monitoring needed in this research. Additionally, some challenges and impracticalities associated with these methods are still being investigated.

2.1.4.1 Total Flow Calorimeter

The total flow calorimeter is a commonly used method to measure the mass flow rate and enthalpy of geothermal fluids. In this process, geothermal fluid is discharged into an open-top tank, where it is mixed with cold water. By measuring the initial and final volumes and temperatures of the mixture, the mass flow rate and enthalpy of the fluid can be calculated. However, this method is generally only practical for wells with low flow rates (up to 25 kg/s) due to the limited tank capacity for low-enthalpy geothermal wells but is limited by tank capacity (Helbig & Zarrouk, 2012). It is commonly used for testing small exploration wells or low-enthalpy wells. Challenges include potential steam loss from the tank, heat loss through the tank walls, and flow restrictions in the connecting pipeline, which can affect the accuracy of the measurements. Additionally, the well needs to be taken out of operation during testing.

2.1.4.2 Lip Pressure Method

The lip pressure method described by James (1965), estimates the mass flow rate and enthalpy of geothermal fluids by measuring the pressure at the pipe-lip as the fluid is discharged to the atmosphere. There are two variations of this method: vertical and horizontal. In the vertical setup, the lip pressure pipe is connected directly to the top control valve, while in the horizontal setup, the pipe is connected horizontally and leads into a silencer. The vertical method is suited for wells with liquid-only feeds, using empirical correlations to calculate flow rate and enthalpy based on the pressure and temperature at the lip. The horizontal method, however, can be used for wells with two-phase feeds and provides more accurate measurements but requires additional equipment and setup, making it more expensive. Limitations include high noise levels during vertical discharge and environmental contamination from discharged fluids, and the well needs to be taken out of production during testing.

2.1.4.3 Load Cells Sensor

Based on the piezoelectric principle, the Load cell sensor is commonly used for real-time stress measurement by converting mechanical force into an electrostatic signal. There are two types: compression, where the object is placed on a platform connected to the sensor, and tension, where the sensor acts as a suspending hanger. This method has been tested in geothermal fields to measure the dryness fraction of two-phase flow. Compression load cells installed under the pipeline measure changes in liquid weight, which can indicate changes in steam dryness. However, the method requires an initial dryness value, and challenges include thermal expansion affecting pipe weight and the impracticality of field tests, which

need to be repeated at each location. The data from load cells can be used to infer parameters like velocity, dryness, and enthalpy of geothermal fluids (Mubarok et al., 2021).

2.1.5 Correlation Models for Two-Phase Flow

To address the complexities of two-phase flows, various correlation models have been developed using a differential pressure (DP) orifice plate meter as a key measurement tool. These models, described in studies like from Murdock (1962), James (1965), Lin (1982), Zhang et al. (1992), Helbig & Zarrouk (2012), and Campos et al. (2014) and Mubarok and Zarrouk (2018), are summarized in Table 2.1.

These correlation models are developed based on either separated or homogeneous flow assumptions. Separated flow models treat the phases distinctly, while homogeneous flow models consider them uniformly mixed. The models also vary in derivation approaches: phenomenological models rely on physical laws to describe flow behaviours, whereas empirical models are derived from experimental data. A key component of these models is the inclusion of correction factors, which are used in the two-phase flow correlation models to adjust key parameters based on empirical data or theoretical assumptions to enhance the model's accuracy for specific flow conditions. For instance, the James (1965) model uses a corrected dryness fraction (x^m), while the (Mubarok et al., 2019) model employs a corrected enthalpy coefficient (C_h).

In this study, the focus will be on the (James, 1965) and (Mubarok et al., 2019) models, as these have been determined to be most applicable to the geothermal data available. The details of these and other relevant models are summarised in Table 2.1 below, outlining the flow models and derivation approaches employed.

Table 2.1 Summary of two-phase correlations

Correlation	Flow Model	Derivation Approach
Murdock (1962)	Separated	Phenomenological model
James (1965)	Homogeneous	Empirical model
Lin (1982)	Separated	Phenomenological model
Zhang et al. (1992)	Homogeneous	Phenomenological model
Helbig and Zarrouk (2012)	Separated	Empirical and phenomenological model
Campos et al. (2014)	Homogeneous	Phenomenological model
Mubarok et al. (2019)	Separated	Empirical and phenomenological model

Models that rely on estimates of steam quality can also calculate it using the enthalpy (h) with the following formula:

$$x = \frac{h - h_w}{h_s - h_w} \tag{2.12}$$

Where h_s and h_w are the enthalpies for the steam and water, and h is the specific enthalpy of the mixture, respectively. In this thesis, steam quality was calculated from the enthalpy at the upstream pressure point (P_1) using PropsSI from the Coolprop library in Python (Bell et al., 2014), through the following function fit.

$$x = f(h, P_1) \tag{2.13}$$

Mubarok et al., (2021) utilised a large dataset from geothermal field tests to evaluate a correlation model introduced by Helbig & Zarrouk, (2012). For the entire range of geothermal reservoir enthalpies (600 - 2800 kJ/kg), they developed a new simplified correlation that provided increased accuracy and proposed an analytical model for predicting pressure drops across a two-phase orifice. The modified correlation by Helbig and Zarrouk is presented below:

$$\dot{m} = \frac{\left(\frac{p_1}{p_2}\right)^{D\sqrt{\frac{10^{-5}\Delta p}{D}}}\left(\frac{\pi d^2}{4}\right)C_h\sqrt{2\Delta p}}{\left(\sqrt{1-\beta^4}\right)} \quad , \quad C_h = (9.7 \times 10^5)(h)^{-1,72} \tag{2.14}$$

The following equation defines the James correlation model:

$$\dot{m} = \frac{\dot{m}_{s_{Apparent}}}{x^{1.5}\left(1 - \frac{\rho_s}{\rho_w} + \frac{\rho_w}{\rho_s}\right)} \tag{2.15}$$

The 'apparent' steam flow ($\dot{m}_{s_{Apparent}}$), is determined as described in the Equation (2.6). Steam quality (x) is calculated using Equation (2.13) while the steam and water densities (ρ_s and ρ_w) are obtained using PropsSI from the Coolprop library in Python (Bell et al., 2014). These densities are derived based on the upstream pressure (P_1) for both the steam and water states:

$$\rho_{s/w} = f(P_1, x_{s/w}) \tag{2.16}$$

2.2 Machine Learning Techniques

After collecting data from different geothermal flow measurement methods, applying correlation models, and identifying the important variables, machine learning techniques were used to develop prediction models. This Section 2.2 introduces and explains the various machine learning techniques and processes used in this study. A key distinction within these techniques is between supervised and unsupervised learning methods. Section 2.2.1 introduces supervised techniques, such as the Random Forest Algorithm, a robust tool for both regression and classification tasks. Section 2.2.2, then transitions to unsupervised learning techniques, including Principal Component Analysis (PCA) and K-means clustering, which are crucial for data reduction and pattern recognition.

2.2.1 Supervised Learning Techniques

Supervised learning relies on training data with targets, enabling models to learn the relationship between input features and the corresponding output. This approach is particularly effective for regression and classification tasks, where predictions are based on historical data. In this study, the Random Forest algorithm was employed for its robust regression capabilities, making it well-suited for tackling complex problems.

2.2.1.1 Random Forest

Random forest (RF) is a powerful algorithm used for regression and classification. It enhances the performance of individual decision trees by combining them to create a more robust and accurate model (Breiman, 2001). RF is particularly effective in handling outliers and noisy data, making it a preferred choice for complex datasets.

To comprehend random forest, it's essential to first understand decision trees. A decision tree classifies data by asking a series of questions at each decision point or node, where each node leads to subsequent branches based on the responses to these questions. The process starts at the root node and ends at the leaf nodes, representing the final outcomes. Internal nodes guide the branching decisions. As the dataset grows, the decision tree expands, making predictions by averaging the target values at the leaf nodes for regression tasks or by majority voting for classification tasks (Breiman, 2001).

To further clarify how a decision tree works, consider the following example, where the objective is to predict whether someone will play golf based on weather conditions. The key features are Outlook, Humidity, and Wind, while the target variable is whether the person will play golf ("Yes" or "No") (Milaan, 2021). The dataset used for this example is shown in Table 2.2.

Table 2.2 Small data set for a decision tree example

Day	Outlook	Humidity	Wind	Play Golf?
1	Sunny	High	Weak	No
2	Sunny	High	Strong	No
3	Overcast	High	Weak	Yes
4	Rain	Normal	Weak	Yes
5	Rain	Normal	Strong	No
6	Overcast	Normal	Strong	Yes
7	Sunny	Normal	Weak	Yes

The decision tree is constructed as follows:

- First Split: Outlook
  - If Outlook = Sunny, further splitting is done based on Humidity.
  - If Outlook = Overcast, the prediction is always "Yes" (the person will play golf).
  - If Outlook = Rain, further splitting is done based on Wind.
- Second Split (Sunny Branch): Humidity
  - If Humidity = High, the prediction is "No."
  - If Humidity = Normal, the prediction is "Yes."

- Second Split (Rain Branch): Wind
    - If Wind = Weak, the prediction is "Yes."
    - If Wind = Strong, the prediction is "No."

Based on the structure of the decision tree, the final predictions are as follows:

- If the Outlook is Sunny and the Humidity is High, the prediction is that the person will not play golf. However, if the Humidity is Normal, the prediction is that the person will play golf.
- If the Outlook is Overcast, the prediction is always that the person will play golf.
- If thee Outlook is Rain, the decision depends on the wind conditions: if the Wind is Weak, the person will play golf, but if the Wind is Strong, the prediction is that the person will not play golf.

These predictions show how the decision tree navigates through each condition combination to make a final decision. However, decision trees often face the issue of overfitting, where the model becomes overly complex and captures noise in the data. Random Forest mitigates this issue by constructing multiple decorrelated trees and averaging their predictions.

In Figure 2.2 the Random Forest algorithm is shown as it Figure 2.2 combines multiple decision trees to generate a final prediction. The algorithm depicted is showing for 600 trees:

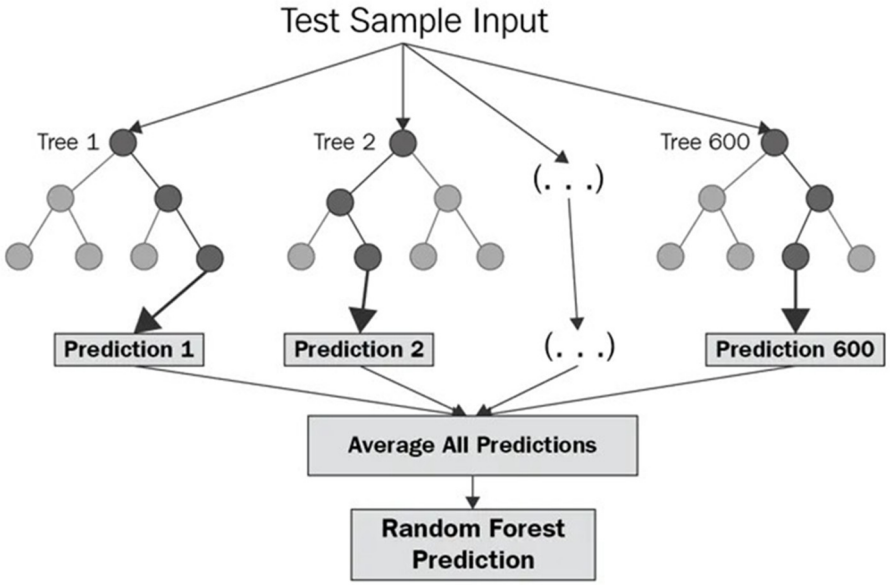


Figure 2.2 Random Forest Diagram (Rudd & Ray, 2020)

During training, the Random Forest algorithm works by generating several decision trees. A bootstrapped sample of the training data is used to create each tree, meaning the sample is drawn with replacement. During the construction of each tree, the algorithm selects a random subset of features at each split point, which helps to ensure the trees are decorrelated.

After all the trees are built, the Random Forest makes a prediction by aggregating the predictions of each individual tree. For regression tasks, it averages the predictions from all trees. For classification tasks, uses the majority vote of the trees to determine the final class (Breiman, 2001).

In practical terms, this process can be summarised as follows:

1. Bootstrap Sampling: Randomly sample the training data with replacement to create multiple subsets. This means some data points may be used multiple times while others may not be used at all. The number of trees is denoted by M, and each tree is built using a bootstrap sample from the training data.
2. Tree Construction: For each subset, build a decision tree by:
   - Randomly selecting a subset of features at each split.
   - Choosing the best split among the selected features.
   - Splitting the node and repeating the process until a minimum node size is achieved.
3. Prediction through Aggregation: For new data points, combine the predictions from all trees to determine the final prediction $\hat{y}(x)$:

$$\hat{y}(x) = \frac{1}{M}\sum_{m=1}^{M} T_m\,(x) \tag{2.17}$$

Where M is the total number of trees in the forest, $T_m(x)$ is the prediction made by the m-th tree for the input x. Additionally, the samples not selected for training a particular tree, known as out-of-bag (oob) samples, are used to evaluate the performance of that tree. This allows RF to provide an unbiased estimation of the generalisation error without needing a separate validation set. The error decreases with increasing the number of trees, which also prevents overfitting (Breiman, 2001). Random Forest also assesses the importance of different features by measuring how the accuracy decreases when a specific feature is switched while keeping the rest constant. This is particularly useful in high-dimensional datasets to identify the most influential features.

The RandomForestRegressor function from the Python library Scikit-learn (Pedregosa et al., 2011) was used in this thesis to implement the Random Forest algorithm.

2.2.2 Unsupervised Learning Techniques

Unsupervised learning relies on training data without targets, focusing on discovering hidden patterns or structures within the data. This method is used to group similar data points together or to reduce the dimensionality of data, making it easier to analyse. Techniques like clustering and dimensionality reduction are essential in scenarios where the goal is to explore data relationships without predefined outcomes. In this study, techniques like Principal Component Analysis (PCA) and K-means clustering are employed to simplify data complexity and reveal underlying relationships.

2.2.2.1 Principal Component Analysis

Principal Component Analysis (PCA) is a fundamental technique for reducing the dimensionality of large datasets while retaining most of the variation present in the data (Jolliffe, 2002);(Kherif & Latypova, 2020). This method is particularly useful for high-dimensional data, which refers to datasets with a large number of interrelated variables or features. Analysing and interpreting the relationships between variables can become complex in such datasets. PCA addresses this by converting the original variables into a new set of uncorrelated variables known as principal components (PCs), effectively simplifying

the data and making it more insightful and easier to analyse.

The main objectives of PCA are to extract the most important information from the data, reduce the number of variables to compress the dataset, simplify the data's description, and analyse the structure and relationships within the data (Jolliffe, 2002). PCA achieves dimensionality reduction by transforming the original variables into new variables that capture the maximum variance in the data. The first principal component accounts for the most variance, the second principal component accounts for the next largest variance and is orthogonal to the first, and this continues for the remaining components (Abdi & Williams, 2010).

Mathematically, PCA involves the eigen-decomposition of the covariance matrix of the data. The principal components are the eigenvectors of this matrix, with the eigenvalues indicating the variance captured by each component (Kherif & Latypova, 2020). The process includes standardising the data, computing the covariance matrix, performing eigen-decomposition, and forming principal components based on the top eigenvectors (Jolliffe, 2002).

Using PCA offers several benefits. It facilitates data visualisation by reducing the data to 2D or 3D, making complex relationships easier to interpret. It enhances storage efficiency by reducing the amount of data to store and eliminates multicollinearity among variables, improving the robustness of the analysis. Additionally, PCA helps reduce noise and enhance data clarity and quality. It also improves the performance of machine learning algorithms by minimising the risk of overfitting and increasing computational efficiency (Abdi & Williams, 2010).

In this thesis, PCA was implemented using the PCA function from the Python library Scikit-learn (Pedregosa et al., 2011). It was used to reduce the dimensionality of the dataset, making it more manageable and enhancing the subsequent analysis and modelling processes. This approach leverages PCA’s ability to capture the most significant patterns in the data, facilitating better insights and more efficient computations.

2.2.2.2 K-means

K-means is a widely used clustering algorithm in data mining and pattern recognition. Proposed by MacQueen, (1967), this unsupervised algorithm aims to partition a dataset into K distinct clusters, where each data point belongs to the cluster with the nearest mean, serving as the cluster centroid.

The K-means algorithm operates iteratively to minimise the variance within each cluster. The process begins by selecting K initial centroids, which can be chosen randomly or based on specific criteria. The algorithm then follows these steps: initialising by choosing K initial centroids randomly from the dataset, assigning each data point to the nearest centroid to form K clusters, recalculating the centroids as the mean of all data points assigned to each cluster, and repeating the assignment and update steps until the centroids no longer change significantly or a predefined number of iterations is reached. The primary objective of K-means is to minimise the sum of squared distances between data points and their respective cluster centroids, known as the within-cluster sum of squares (WCSS). This approach ensures that the clusters are as compact as possible (Na et al., 2010).

K-means clustering is known for its simplicity, efficiency, and speed, making it suitable for large datasets. However, the algorithm has some limitations. It is highly dependent on the choice of initial centroids, which can significantly affect the final clusters. Different initialisations can lead to different results, potentially causing the algorithm to converge to a local minimum rather than the global minimum. Additionally, the user must specify the number of clusters (K) in advance, which might not always be optimal for the given data. K-means can also be sensitive to outliers, as they can disproportionately affect the position of the centroids. Furthermore, the algorithm assumes that clusters are spherical and equally sized. However, real-world data can form clusters of various shapes and scales, which the algorithm might not accurately capture (Olukanmi et al., 2022). In practice, the K-means algorithm is commonly implemented using the K-means function from the Python library Scikit-learn (Pedregosa et al., 2011), which provides an efficient and user-friendly interface for clustering tasks.

2.2.2.3 Density-Based Spatial Clustering of Applications with Noise

DBSCAN is a robust clustering algorithm widely used for data mining and spatial data analysis. It excels in identifying clusters of arbitrary shape and size, even in the presence of noise and outliers, making it particularly effective for large and complex datasets such as GIS, satellite imagery, remote sensing, and environmental assessment. DBSCAN groups together closely packed points and marks outliers that lie in low-density regions. It relies on two key parameters: Epsilon (ε), which defines the radius within which points are considered neighbours, and MinPts, the minimum number of points required to form a dense cluster. A point is classified as a core point if it has at least MinPts neighbours within a radius of ε. Clusters are formed by core points, and all reachable points from core points (Ester et al., 1996). Figure 2.3 illustrates the clustering parameters used in DBSCAN, showing the classification of points as core, border, or noise based on their density.

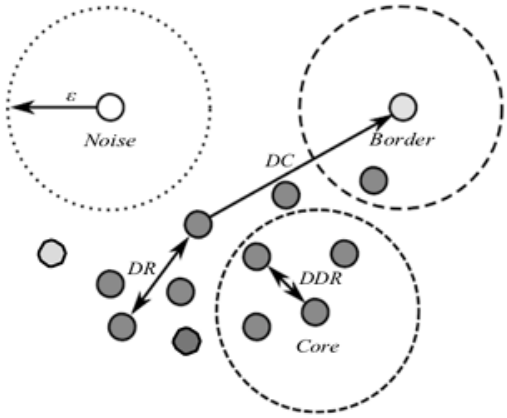


Figure 2.3 Description of DBSCAN Clustering Parameters (Götz et al., 2019)

The main advantages of DBSCAN include its ability to discover clusters of arbitrary shape, handle noise and outliers, and operate without needing to specify the number of clusters in advance. However, it requires careful selection of ε and MinPts, struggles with datasets containing clusters of varying densities, and can be computationally intensive. Unlike K-means, DBSCAN does not require a predefined number of clusters, instead identifying each point as a core, border, or noise point based on density. This flexibility allows DBSCAN to handle non-globular clusters and identify clusters within clusters,

providing more precise results (Ester et al., 1996). The DBSCAN algorithm can be implemented using the DBSCAN function from the Scikit-learn library.

2.2.3 Machine Learning Processes

2.2.3.1 Feature Selection

Feature selection is an important step in constructing effective machine-learning models. Its primary aim is to pinpoint the most relevant features that enhance the model's predictive accuracy while reducing computational costs. Using irrelevant or redundant features can degrade model performance and increase complexity. Different algorithms have built-in mechanisms for feature selection or dimensionality reduction (Theng & Bhoyar, 2024).

One method used in this study is Recursive Feature Elimination with Cross-Validation (RFECV). RFECV is a wrapper method that systematically removes less important features and uses cross-validation to evaluate model performance at each step. Starting with all features, RFECV ranks them based on their importance and iteratively eliminates the least important ones. This approach ensures the selection of the most relevant features while preventing overfitting (Awad & Fraihat, 2023).

Another approach involves using the SelectKBest method, which is a type of filter-based technique for feature selection. SelectKBest selects the top k features with the highest scores based on a statistical test. This method evaluates each feature individually against the target variable to determine its relevance (Pedregosa et al., 2011).

Additionally, Random Forests provide an inherent feature importance measure. During the training process, Random Forests evaluated the significance of each feature based on its contribution to improving model accuracy. This method is efficient and helps identify the most critical features for the model.

2.2.3.2 Standardization

Standardization is an essential preprocessing step in machine learning that ensures all features contribute equally by scaling them to have a mean of zero and a standard deviation of one. This step is particularly important for techniques like Principal Component Analysis (PCA) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN);(Shalev-Shwartz & Ben-David, 2014).

Standardization adjusts data with a mean of zero and a standard deviation of one. This study used Z-score standardization equation:

$$Z_x = \frac{X_i - \mu}{\sigma} \tag{2.18}$$

where Z_x is the standardized value, X_i is the original value, μ is the mean, and σ is the standard deviation. Standardization was implemented in this study using the StandardScaler function from the Python Scikit-learn library (Pedregosa et al., 2011). This function standardizes features by removing the mean and scaling to unit variance (Shalev-Shwartz & Ben-David, 2014).

2.2.3.3 Cross-Validation

Cross-validation is a statistical method to test how well a machine learning model performs. It involves partitioning the dataset into smaller subsets, using multiple subsets for training and one subset for testing. This process is repeated several times to make sure the model performs well on different parts of the data.

In this thesis, 5-Fold cross-validation was implemented. The dataset was divided into five groups, each used once as a testing set, while the remaining four groups are used for training. This process is repeated five times, and the final validation score is the average of the five testing scores. The choice of K=5 balances the trade-off between bias and variance, offering a reliable estimate of the model's predictive performance (Fushiki, 2009; Kohavi, 1995). The validation score from K-Fold cross-validation is used to evaluate model performance during feature selection and hyper-parameter tuning, ensuring the model's accuracy and generalization to unseen data.

2.2.3.4 Hyperparameter Tunning

Hyperparameter tuning is an essential step in optimizing machine learning models. It involves selecting the best set of hyperparameters for a learning algorithm to improve its performance on a specific dataset. In this study, hyperparameter tuning was performed for several algorithms including Random Forest, DBSCAN, K-means, and SelectKBest.

For the Random Forest algorithm, key hyperparameters such as the number of trees in the forest (n_estimators), the maximum depth of the trees (max_depth), the minimum number of samples required to split an internal node (min_samples_split), and the minimum number of samples required to be at a leaf node (min_samples_leaf) were tuned. Grid Search Cross-Validation (GridSearchCV) was employed to systematically explore different combinations of these hyperparameters and select the optimal set (Shalev-Shwartz & Ben-David, 2014).

The key hyperparameters for the DBSCAN algorithm are the maximum distance between two samples, known as eps, and the minimum number of samples needed for a point to be considered a core point, called min_samples. These parameters were tuned using GridSearchCV to ensure the algorithm correctly identifies clusters of varying densities.

For K-means clustering, the main parameter to set is the number of clusters (k). The best number of clusters was found using the Elbow Method and Silhouette Score. The Elbow Method looks at how much variance decreases as you add more clusters, and the best choice is where the decrease slows down. The Silhouette Score checks how well each point fits into its cluster compared to others, with higher scores meaning clearer clusters (Na et al., 2010).

The SelectKBest method, as described in Section 2.2.3.1, involves selecting the top k features based on statistical tests. The k parameter was tuned by evaluating the performance of the model with different numbers of top features, ensuring the best subset of features was chosen for model training.

Chapter 3

Methodology

This chapter presents an overview of the methodology used for data analysis in this study. Section 3.1 outlines the setup of the Landsvirkjun experiments for phases 1 and 2. Section 3.2 provides an overview of data gathering, preprocessing, and data cleaning. Section 3.3 explains how the models are evaluated based on performance criteria. Section 3.4 describes the python packages used for data handling throughout the study. Finally, Section 3.5 discusses the development and implementation of machine learning models described in Section 2.2 .

3.1 Landsvirkjun Real-Time Well Output Project

This research was carried out in partnership with Landsvirkjun, the leading electricity provider in Iceland. Playing a crucial role in producing 75% of the nation's electricity from hydro, wind, and geothermal resources, Landsvirkjun has been actively operating since 1965. It manages a network of 18 power stations located across five key regions within Iceland (Landsvirkjun, 2023a). This section will describe the background of the experiments gathered and goals of this study.

3.1.1 Overview

In 2019, Landsvirkjun began a project to determine a real-time measurement system for monitoring geothermal wells. The goal was to measure geothermal fluid's enthalpy and flow rate, allowing for continuous estimation of geothermal well output while the wells remain operational. This project aims to gather comprehensive data to improve reservoir management and optimize production processes.

In 2019 to 2021, experiments were conducted at Þeistareykir in northeastern Iceland, one of Landsvirkjun's three geothermal power stations with the highest electrical production capacity. The project explored several measurement approaches utilising robust sensors placed along the flow line, extending from the wellhead to the steam separator. According to Juliusson et al., (2023), the sensors considered for this testing phase included differential pressure (DP) sensors over an orifice plate, venturi meters, vortex meters, Coriolis meters, load cell sensors, and radio frequency sensors. To date, only the DP sensors with orifice plates and vortex meters have been tested in the Landsvirkjun project. While some of these techniques have been previously applied in geothermal environments, none have been confirmed for real-time measurement of flow and enthalpy.

The latest experiments in 2023 were conducted at the Bjarnaflag power plant (see Figure 3.1). The Bjarnarflag geothermal area is in Mývatnssveit and commenced operations on 5th March 1969. Bjarnarflag geothermal station is the smallest geothermal facility operated by

Landsvirkjun in Iceland and represents Iceland's first geothermal station of its type. Located near Námafjall Mountain, this station produces 5 MW of power. Besides producing 42 GWh of electricity annually, Bjarnarflag supplies steam for district heating and industrial use. It also provides geothermal water to the nature baths at Lake Mývatn (Landsvirkjun, 2023b).

Figure 3.1 Map showing the location of Bjarnarflag geothermal power plant (Image obtained from Google Earth, 2023)

3.1.2 Experiment Setup

The experiments aimed to create a practical range of experimental conditions in terms of flow rate, enthalpy, and pipe size. Based on the operating experience at Landsvirkjun's geothermal power plants, it was determined that each well would seldom produce more than 35 kg/s, and the enthalpy would generally range between 1000 and 2800 kJ/kg. These specific ranges are important because they represent the usual operating conditions of the geothermal wells at Landsvirkjun. By focusing on these conditions, the model developed in this study aims to accurately predict flow rates and enthalpy within these realistic and relevant limits.

For this thesis, data was collected from two sources: experiments conducted by Landsvirkjun between 2019 and 2021, referred to as phase 1, and a separate set of experiments carried out in 2023, referred to as phase 2, where the author collected data. The setups for these two phases of experiments are described in Sections 3.1.3 and 3.1.4.

There are two categories of measurements: reference measurements (RM) and experimental measurements (EM). RM data comes from the measured values of the fluid flow rate, and enthalpy gathered using the water tracer and separation methods described in Sections 2.1.1 and 2.1.2. The EM are recorded at the same time using differential pressure (DP) across orifice plates as described in 2.1.3.

3.1.3 Phase 1 - Experiments 2019 - 2021

Phase 1 of Landsvirkjun's enthalpy sampling experiments, conducted in 2019, 2020 and 2021 in the Þeisthareykir field, was divided into stages based on the sensor types used in the flow line. The initial stage involved experiments on wells ÞG-11, ÞG-15, and ÞG-18. These wells were connected to a steam separator via a 40-meter-long pipe. The first 20

meters had a diameter of 250 mm, and after the bend in the pipe, it narrowed to 200 mm, as shown in Figure 3.2. Four test runs were conducted on well ÞG-18, located on well pad F, as shown in Figure 3.2, a single line diagram, and an aerial photograph of this setup from (Juliusson et al., 2023) study. Additionally, six test runs were carried out on the combined output from wells ÞG-11 and ÞG-15, located on well pad B, as shown in Figure 3.3.

Fluid flow rate and enthalpy are measured in real-time at the end of the flow line using a steam separator as described in Section 2.1.1. These measurements, referred to as Reference Measurement (RM) data, include values obtained from separated steam and liquid flow measurements. The steam flow was measured by a vortex meter, while a water weir measured the liquid flow. As a backup, the steam flow was also measured using the lip pressure method as described in Section 2.1.4.2, and the liquid water flow is measured with a magnetic flow meter. The vortex meter uses fluid oscillation to measure the flow of gas, steam, or liquid flow and records the steam flow exiting the pipe at the separator's top. The separated liquid is directed into a weir box for mass flow measurements under atmospheric conditions. A v-notch water weir is then used to calculate the liquid phase flow rate (Einarsson, 2021).

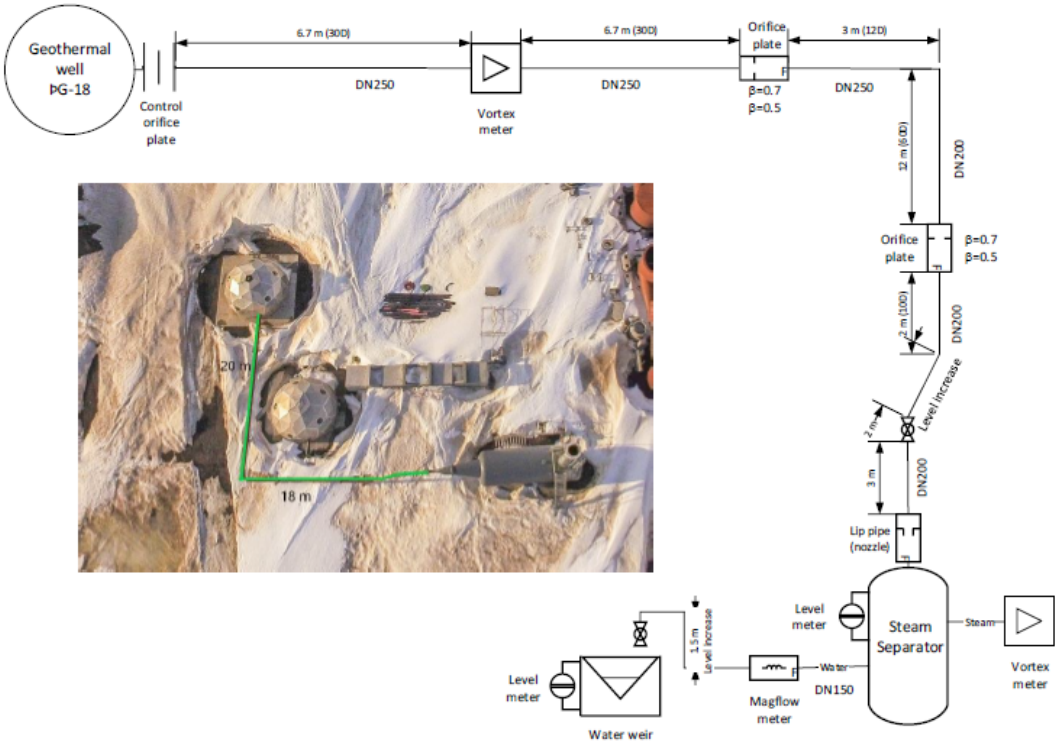


Figure 3.2 Setup of an experiment carried out on well ÞG-18, well pad F, at the Þeistareykir field (taken from Juliusson et al., 2023)

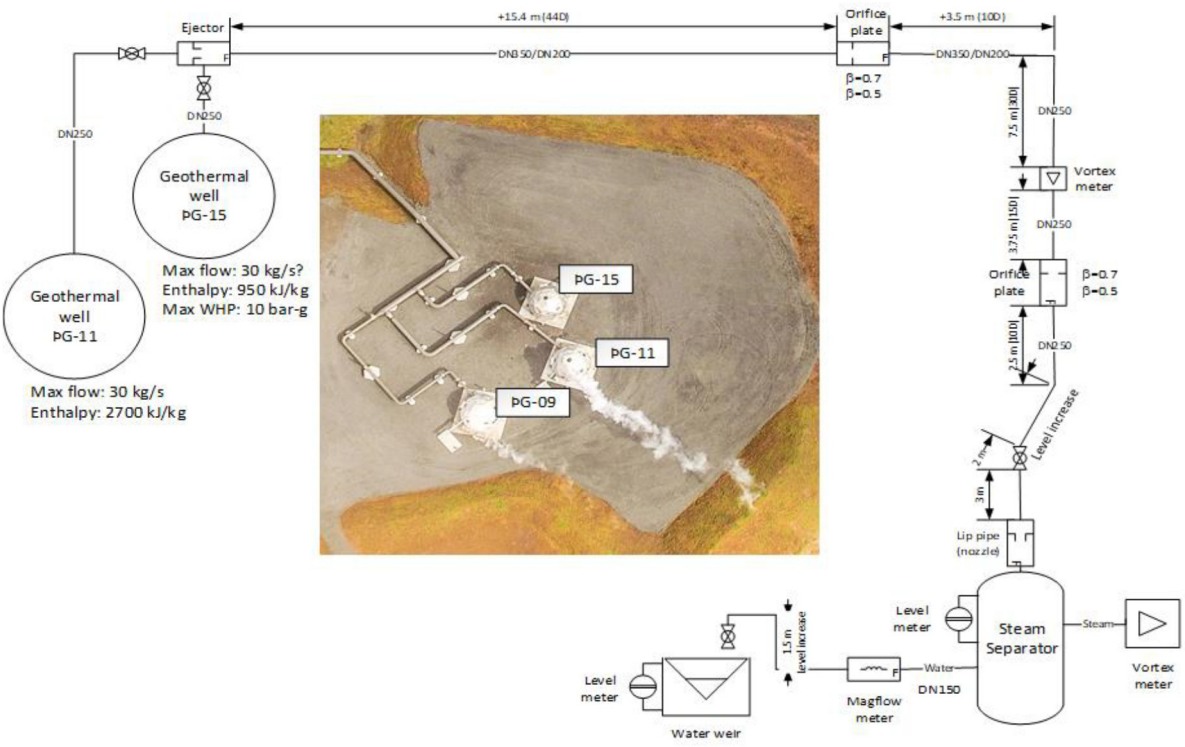


Figure 3.3 Setup of an experiment carried out on wells ÞG-11 and ÞG15, well pad B, at the Þeistareykir field (taken from Juliusson et al., 2023).

3.1.4 Phase 2 - Experiments 2023

Phase 2 of Landsvirkjun experiments, where the author participated, conducted in 2023 in Bjarnaflag on well BJ-12, employed a modified water tracer method with an orifice plate, developed by Kemia and Landsvirkjun in 2011 (Hauksson, 2011). This approach integrates previous techniques with the water tracer method, described in Section 2.1.2, where a known concentration of sodium fluorescence dye is injected into the flow using a pump. The dilution of this dye downstream is then measured to calculate the water flow.

Pressure measurements were taken at various points along the flow line (P0, P1, and P2), as shown in Figure 3.4 and Figure 3.5. The wellhead pressure (P0) was controlled and varied, ranging from fully open to almost fully closed. During the first test run 20.03.2023, pressures ranged from 18.4 to 21.3 bar g. However, the second test run on 02.11.2023, which had pressures at 29.25 bar g, had to be discarded due to errors in the DP measurements.

A separator was connected after the orifice plate, with a metal cooling coil in a bucket before taking the liquid sample using a 300 mL glass flask (see Figure 3.4). A "blank" water sample was taken before the sodium fluorescence entered the stream. After ensuring proper flow, the pump was started; the pump connected to the wellhead is shown in Figure 3.5. Once the fluorescence was visible, samples were collected in glass bottles every 3 minutes. Each bottle was rinsed three times before filling, and the lid was screwed tightly once full. The bottles were then placed in a cardboard box and kept out of sunlight to prevent the dye from breaking down. This process was repeated until all samples were collected.

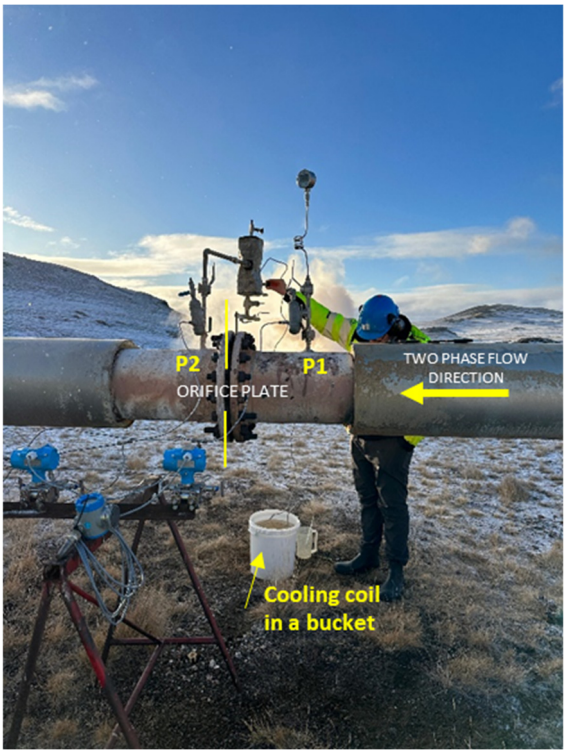


Figure 3.4 Setup of water tracer method

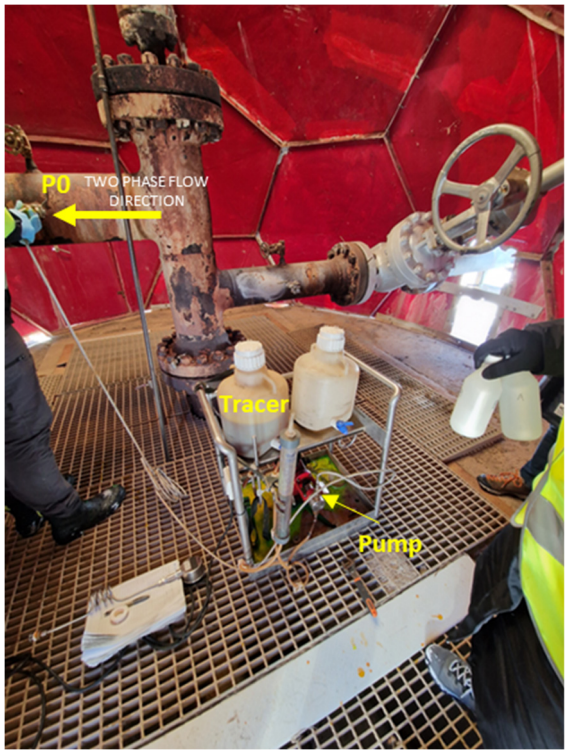


Figure 3.5 Pumping tracer into well BJ-12

In Figure 3.6, a sketch of the entire setup is shown, illustrating both the internal and external components. Inside the well dome, the pump responsible for injecting the tracer into the flow line is located, and the wellhead pressure (P0) is measured at this point. Outside the dome, pressures P1 and P2 are measured along the flow line, which is then used to calculate the pressure drop across the orifice plate, and the water sample is collected after it passes through the cooling coil.

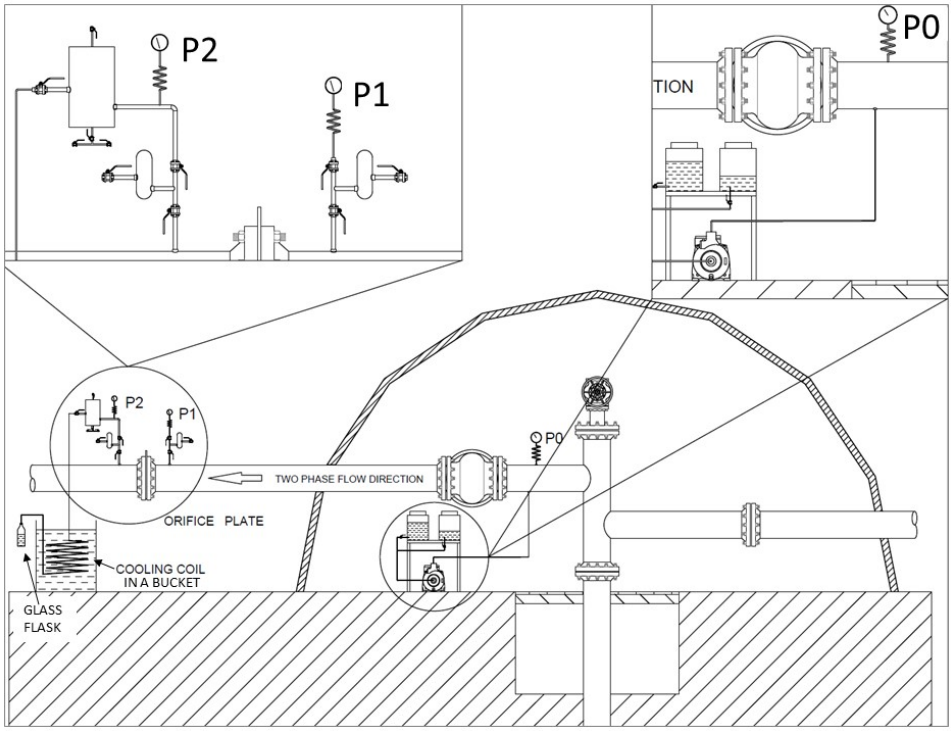


Figure 3.6 Water tracer method setup

The author conducted the sampling, and Landsvirkjun carried out the analysis using a fluorometer to determine the dye concentration. The results were processed using Landsvirkjun's classified calculation method, which is based on a combination of correlation models described in Section 2.1.5. Typically, such tests are performed once a year during the summer. However, for this research, they were conducted twice, in March and November, with different wellhead pressure configurations to cover a broader range. Landsvirkjun provided the data for use in this thesis.

3.1.5 Differential Pressure Meter Setup

Differential Pressure (DP) meters, such as orifice plate meters, are commonly used in the gas industry for measuring flow rates, particularly for natural gas and other hydrocarbon fluids (Upp & LaNasa, 2014). The experimental measurement (EM) data is obtained from a DP orifice plate meter setup, illustrated in Figure 3.7. This setup includes three pressure taps: one at a distance of 1D upstream (P1), another at 1/2D downstream (P2), and a third at 6D downstream (P3), with D representing the pipe diameter. This configuration measures the pressure drop between P1 and P2, as well as the pressure recovery between P2 and P3. By analysing these pressure changes, we can determine the flow characteristics and understand the fluid dynamics within the pipe.

Unlike traditional DP orifice plate meters that use two pressure taps, this setup utilises three pressure taps, offering more measurements and additional parameters. The aim is to investigate whether there exists a relationship between the measured physical properties (like pressure readings from the taps P1, DP12, DP13, and DP32) (see Figure 3.8) recorded during the experimental measurements (EM) and the calculated properties (such as enthalpy h and mass flow rate $\dot{m}$) derived from reference measurements (RM), described by the function:

$$f(h, \dot{m})_{RM} = g(P1, DP12, DP13, DP32)_{EM} \tag{3.1}$$

This approach aims to correlate the experimental data with the reference measurements, improving the accuracy and reliability of the output measurements.

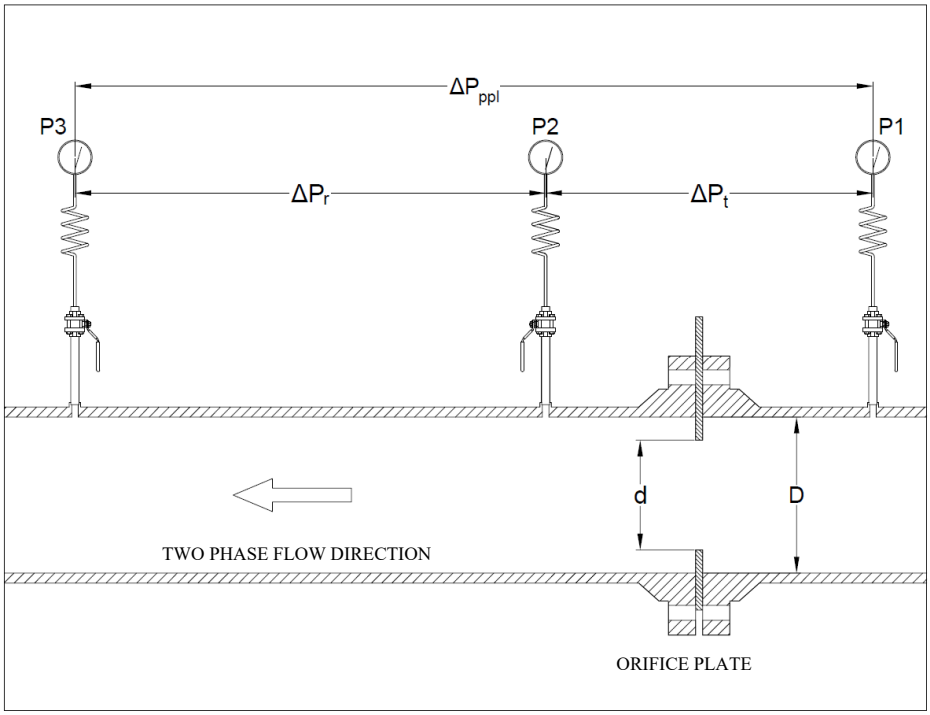


Figure 3.7 Configuration of the DP Orifice Plate Meter

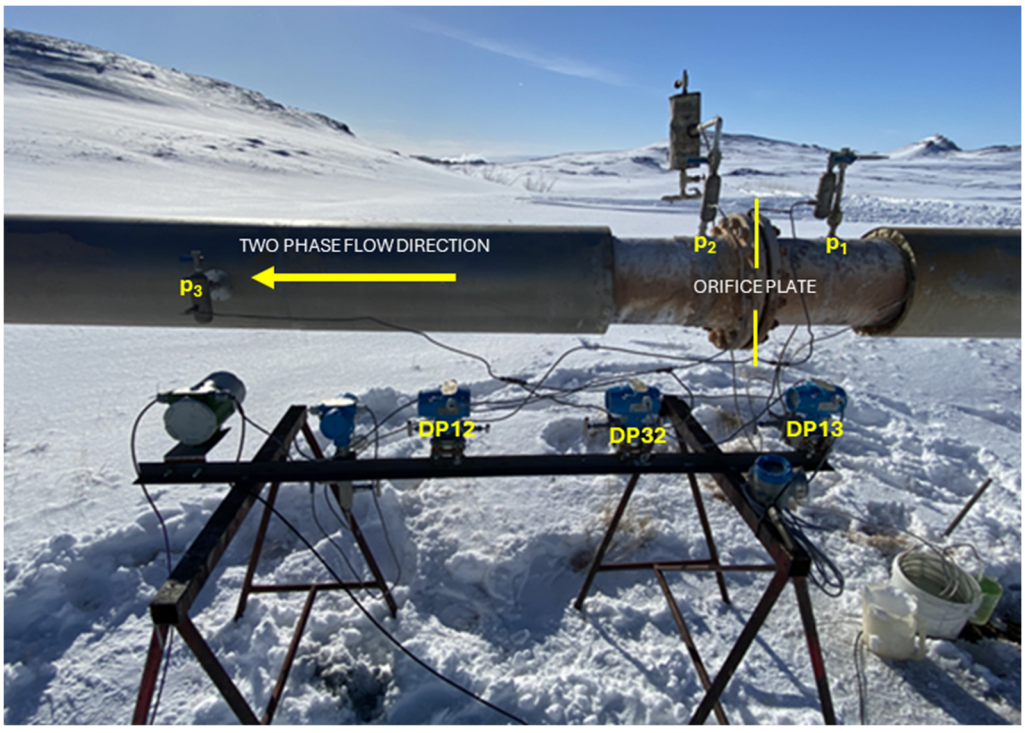


Figure 3.8 Setup on well BJ-12 DPs

3.2 Data Overview

3.2.1 Data Gathering

The data used in this thesis comes from experiments of phases 1 and 2, described in Sections 3.1.3 and 3.1.4, conducted by Landsvirkjun. These experiments were designed to find practical methods for measuring the flow rate and enthalpy of geothermal fluid in real-time. Table 3.1 provides a summary of the test runs for phases 1 and 2 used for this research.

Phase 1 dataset was provided, cleaned, and processed (Section 3.2.4) by Juliusson et al., (2023) and ready to be used in this study. For phase 2, the new dataset was collected by Landsvirkjun, with the author assisting during the measurements. The author was present with the chemical team during both test runs in phase 2. Landsvirkjun recorded the data, ran it through their system for calculations, and shared the results in Excel format. Additionally, data from the differential pressure meters (DP) was downloaded and shared. This new dataset needed to be validated, cleaned, and processed for further analysis.

Table 3.1 Summary of Phase 1 and 2 test runs

	Experiments	Time Period		Pipe Section 1 (See Figures 3.2,3.3)		Pipe Section 2 (See Figures 3.2,3.3)	
		Start	End	Pipe diameter	Orifice diameter	Pipe diameter	orifice Diameter
Phase 1 - Separator method	ÞG - 18 Test Run 1	18.09.2019, 17:02	16.11.2020, 08:02	254 mm (10")	178 mm (7")	203 mm (8")	142 mm (5.5")
	ÞG - 18 Test Run 2	22.06.2020, 17:01	10.07.2020, 08:01	254 mm (10")	178 mm (7")	203 mm (8")	142 mm (5.5")
	ÞG - 18 Test Run 3	30.07.2020, 17:00	04.08.2020, 08:00	254 mm (10")	178 mm (7")	203 mm (8")	142 mm (5.5")
	ÞG - 18 Test Run 4	04.08.2020, 15:30	07.08.2020, 12:00	254 mm (10")	127 mm (5")	203 mm (8")	102 mm (4")
	ÞG - 11 & ÞG - 15 Test Run 1	12.09.2020, 10:51	16.09.2020, 15:20	340 mm (13.4")	238 mm (9.38")	254 mm (10")	178 mm (7")
	ÞG - 11 & ÞG - 15 Test Run 2	16.09.2020, 23:00	18.09.2020, 15:20	340 mm (13.4")	238 mm (9.38")	254 mm (10")	178 mm (7")
	ÞG - 11 & ÞG - 15 Test Run 3	28.09.2021, 16:45	07.10.2020, 12:00	340 mm (13.4")	238 mm (9.38")	254 mm (10")	178 mm (7")
	ÞG - 11 & ÞG - 15 Test Run 4	21.06.2021, 17:00	25.06.2021, 14:00	340 mm (13.4")	238 mm (9.38")	254 mm (10")	178 mm (7")
	ÞG - 11 & ÞG - 15 Test Run 5	5.07.2021, 21:00	09.07.2021, 15:00	340 mm (13.4")	238 mm (9.38")	254 mm (10")	178 mm (7")
	ÞG - 11 & ÞG - 15 Test Run 6	01.09.2021, 10:00	08.09.2021, 17:00	340 mm (13.4")	165 mm (6.5")	254 mm (10")	178 mm (7")
Phase 2 - Water tracer method	BJ -12 Test Run 1	20.03.2023, 8:30	21.03.2023, 17:30	345 mm (13.6")	139 mm (5.5")		
	BJ -12 Test Run 2	02.11.2023, 10:30	02.11.2023, 23:59	345 mm (13.6")	239 mm (9.4")		

One significant difference between phase 1 and phase 2 test runs is the number of data points collected in each run. The separator method as described in Section 2.1.1 allow for the collection of as many data points as the differential pressure sensors (DP) can register. However, the water tracer method described in Section 2.1.2 used in phase 2 tests resulted in fewer data points. In the first test run, five points were collected, and in the second, six points were collected. Unfortunately, not all data points were usable due to the DP not registering reliable data at the time. Specifically, the BJ-12 test run 1 had to be discarded. Nevertheless, phases 1 and 2 data points are presented in the results Section 4.1.1.

The diameter ratio, β (see in Equation (2.7)), for phase 1 test runs, it was 0.7, except for the fourth test run on ÞG-18, where it was 0.5. For phase 2 experiments, the first test run had a diameter ratio of 0.4, while the second test run had a ratio of 0.7.

The flow rate and enthalpy ranges also differed between phase 1 and 2 test runs. In the phase 1 tests, the flow rate ranged up to 35 kg/s, and the enthalpy varied between 1000 and 2800 kJ/kg. In contrast, phase 2 tests recorded flow rates between 20-30 kg/s and enthalpy values between 1600 and 2000 kJ/kg, which fall within the ranges observed in phase 1 but contribute additional data within these specific intervals.

To illustrate the difference in the number of data points collected, Table 3.2 compares the quantity of phase 1 and 2 data points:

Table 3.2 RM Data points cleaned

Experiments methods	Number of Data Points
Separator Method	19454
Water Tracer Method	5

3.2.2 Data Preprocessing

The datasets for each test run are organised by aligning the reference measurements (RM) and experimental measurements (EM) based on their timestamps. The next section will explain the methods used to estimate the enthalpy and total flow rate.

3.2.2.1 Reference Measurements

The RM for phase 1 experiments were provided pre-processed by Juliusson et al., (2023). In these experiments, the flow rate and enthalpy are derived from the separator method Equations (2.1) and (2.2) in Section 2.1.1. where the geothermal fluid flow enters a steam separator, which separates it into vapour and liquid phases. A vortex meter at the top of the separator measures the steam flow rate, while the separated liquid is directed into a V-notch water weir box to find the water flow rate. Additionally, the water level in the separator is monitored, with fluctuations converted into volume changes to correct the total water flow entering the separator (Einarsson, 2021).

For phase 2 experiments, the enthalpy and flow rate values were provided directly by Landsvirkjun water tracer analysis described in Section 2.1.2, requiring no additional calculations.

All modelling of fluid properties in this study assumes that the geothermal fluid consists of pure water and steam. The steam quality is used to calculate the total enthalpy of the geothermal fluid. The enthalpies for the steam and water phases are determined using the PropsSI function in Python (see Equation ((**2.1**)).

3.2.2.2 Experimental Measurements

In this thesis, three different pressure modes are utilised: absolute, gauge, and differential. These modes are essential for various calibration methods, and all pressure values are expressed in either pascals (Pa) or bars (bar). The orifice plate meters recorded pressure readings using various units based on the type of differential pressure sensors. If the pressure was measured in pounds per square inch (psi) or inches of water (inH_2O), it was converted to bar units to ensure consistency. Differential pressure values were kept as bars, while gauge pressures (bar-g) were converted to absolute pressure (bar-a) by adding atmospheric pressure (P_{atm}). This atmospheric pressure was obtained from hourly logs recorded at a nearby weather station. This conversion ensures all pressure readings are in a consistent and comparable format for analysis.

3.2.3 Dataset and Parameters

The dataset, organised according to the layout shown in Table 3.3, which is presented without specific values to illustrate key reference measurements such as enthalpy (h) and mass flow rate (ṁ), alongside experimental measurements like upstream pressure (P_1) and differential pressures, ΔP_t (see Equation (3.2)), ΔP_{ppl} (see Equation (3.3)) and ΔP_r (see Equation (3.4)) and also shown in Figure 3.7. These parameters are essential for the analysis, enabling the calculation of PRL (see Equation (3.5)), PRR (3.6)) and RPR (3.7)).

Table 3.3 Dataset Structure, including RM and EM and Setup Configuration

Data samples	Reference Measurements		Experimental Measurements				Configuration		
	h	$\dot{m}$	P_1	ΔP_t	ΔP_{ppl}	ΔP_r	D	d	β
ÞG well Test Run 1									
ÞG well Test Run 2									
ÞG well Test Run 3									
ÞG well Test Run 4									
ÞG well Test Run 5									
BJ well Test Run 1									
BJ well Test Run 2									

The difference in pressure between taps P_1 and P_2 is used to determine the "traditional" pressure loss (ΔP_t), and it is calculated as follows:

$$\Delta P_t = P_1 - P_2 \tag{3.2}$$

"Permanent" pressure loss (ΔP_{ppl}) is calculated as the pressure difference between P_1 and P_3:

$$\Delta P_{ppl} = P_1 - P_3 \tag{3.3}$$

"Recovery" pressure (ΔP_r) is measured by calculating the pressure difference between P_3 and P_2.

$$\Delta P_r = P_3 - P_2 \tag{3.4}$$

Before proceeding with the analysis, it is necessary to define additional parameters. The differential pressure loss ratios PLR, PRR, and RPR are expressed by the following equations:

$$PLR = \frac{\Delta P_{ppl}}{\Delta P_t} \tag{3.5}$$

$$PRR = \frac{\Delta P_r}{\Delta P_t} \tag{3.6}$$

$$RPR = \frac{\Delta P_r}{\Delta P_{ppl}} \tag{3.7}$$

Where PLR represents the pressure loss ratio, PRR is the pressure recovery ratio, and PPR is

the ratio between the recovery and permanent loss. The densities as described in Equation (2.16) are used to determine the ratio of steam to water density, denoted as DR, defined as:

$$DR = \frac{\rho_s}{\rho_w} \tag{3.8}$$

3.2.4 Data Cleaning and Quality

The phase 1 data preprocessing of EM, as described by Einarsson, (2021); Juliusson et al., (2023) involved detecting and correcting inaccuracies based on several quality criteria. Completeness was ensured by removing samples with missing values, often due to non-operational equipment. Measurements that fell outside specified range constraints were discarded. Consistency checks led to the removal of unresolvable inconsistencies. Spikes in data were addressed using moving averages, and excessively noisy data was excluded. DP data was validated against set criteria, and corrections were applied when discrepancies were minimal. Backup measurements were used for additional corrections, and the water level in the separator was monitored to adjust for fluctuations. Following the cleaning process, the dataset was reduced to 19454 datapoints, as shown in Table 3.2

For phase 2 experiment, the data cleansing process was faster due to the fewer data points, with only 11 points collected from the water tracer method. However, ensuring that the DP sensors provided accurate measurements simultaneously with the water tracer sampling was critical to correctly classify these points as Experimental Measurements (EM). During the experiments, adjustments to the wellhead pressure (WHP) required time for the well to stabilise, which was reflected in the DP meters as spikes. Moreover, the 6 points from the second test run had to be discarded because the DP measurements indicated inaccurate readings, with trends that were physically unrealistic and inconsistent with the expected operational behaviour, reducing the dataset to 5 points, as shown in Table 3.2.

3.3 Performance Criterion

This study used the Random Forest Regression model to predict steam quality (x) based on features derived from DP orifice plate measurements, resulting in a numerical output. The performance of the model was evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). RMSE measures the average magnitude of errors, penalising larger discrepancies more heavily, while MAPE provides a scale-independent percentage error, making it easy to compare across different datasets. These metrics are essential for understanding the model's accuracy in predicting steam quality.

This thesis evaluated the effectiveness of various predictive methods and models for estimating steam quality, which is essential for determining the geothermal fluid's flow rate and enthalpy. The performance of these models was assessed by comparing their numerical outputs to the reference data (RM) using two key metrics:

1) The Mean Absolute Percentage Error (MAPE) is widely utilised as a metric for its scale-independent characteristics, making it advantageous for comparing errors across different scales. MAPE provides insight into the accuracy of predictions by

expressing errors as a percentage. It is calculated as follows:

$$MAPE = \frac{\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|}{n} \times 100\% \tag{3.9}$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations (Tayman & Swanson, 1999)

2) The root mean Squared Error (RMSE) is another essential metric frequently used during the development phase. RMSE is beneficial for comparing the same method under various internal settings, as it penalises larger errors more than smaller ones (Chai & Draxler, 2014). It is calculated as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{3.10}$$

3.4 Python Packages used for Analysis

This study utilised a range of Python libraries to facilitate data handling, statistical analysis, machine learning, and thermodynamic calculations. The following libraries and tools were instrumental in the execution of this research:

Jupyter Notebooks: All analysis and model development were conducted using Jupyter Notebooks. This tool was chosen for its flexibility and ability to combine code execution, data visualisation, and narrative text in a single document. Jupyter Notebooks allowed for an iterative and interactive approach to the research, making it easier to document and reproduce the analysis (Kluyver et al., 2016).

Pandas: The Pandas library was extensively used for data manipulation, including data cleaning, transformation, and aggregation. It provided powerful data structures like DataFrames, which were crucial for managing the complex datasets used in this study (McKinney, 2010).

NumPy: NumPy was employed for numerical operations, particularly in handling large arrays and performing mathematical computations. Its integration with Pandas and other libraries made it a fundamental tool for efficient data processing (Harris et al., 2020).

Scikit-learn: This library was the basis for all machine learning tasks, including model training, validation, and evaluation. Scikit-learn was used for:

- **Random Forest Regression**: Employed for predicting steam quality based on features derived from DP orifice plate measurements.

- **DBSCAN**: Used for clustering data points and identifying noise, particularly useful in handling large and complex datasets.

- **K-means Clustering**: Applied to group data into clusters, facilitating the analysis of similar data points.

- **Principal Component Analysis (PCA)**: Used for dimensionality reduction, helping to simplify the dataset while retaining most of its variance.

- **SelectKBest**: Utilized for feature selection, allowing the model to focus on the most relevant features.

- **Recursive Feature Elimination with Cross-Validation (RFECV):** Applied to iteratively remove less important features, ensuring that the model retains only the most relevant variables for improved accuracy and to prevent overfitting.

- **GridSearchCV:** Employed for hyperparameter tuning, GridSearchCV systematically worked through combinations of model parameters, using cross-validation to find the best set of parameters that optimise model performance (Pedregosa et al., 2011).

Matplotlib and Seaborn: Matplotlib and Seaborn were utilised to create detailed plots and graphs for data visualisation. These visualisations were essential for exploratory data analysis and for interpreting the results of machine learning models (Hunter, 2007; Waskom, 2021).

CoolProp: Thermodynamic calculations were performed using the CoolProp library, specifically the PropsSI function, which provided accurate water properties required for the rule-based models (Bell et al., 2014).

All parameters for rule-based models and thermodynamic calculations were applied in either base units or derived units of the SI system, ensuring consistency and accuracy across all computations.

3.5 Model Development

Data preprocessing plays an important role in machine learning because the quality of the data and the insights derived from it greatly influence the model's learning effectiveness. Before training the model, the dataset must undergo preprocessing, which has been described in Section 3.2.2. In this study, the following steps were implemented to build the machine learning model:

1. Hold-out set creation: A portion of the dataset was reserved as a "hold-out" set for final evaluation.
2. Data labelling: The target variable, steam quality (x), was labelled based on features derived from DP orifice plate measurements.
3. Normalization: Data was scaled to ensure all features contribute equally, typically by adjusting values between 0 and 1 or by having a mean of 0 and standard deviation of 1 (Section 2.2.3.2).

4. Cross-Validation (Section 2.2.3.3): Used to assess model performance, with specific steps including:
    a. Feature Selection: RFECV and Random Forest feature importance methods were employed to identify the most relevant features (Section 2.2.3.1).
    b. Hyperparameter Tuning: Optimization of model parameters through Grid Search Cross-Validation (GridSearchCV) (Section2.2.3.4).
5. Final evaluation on hold-out set: The model's performance was tested on the "hold-out" set to ensure its generalizability and prevent overfitting.

The goal is to create a regression model using Random Forest to predict steam quality (x) based on features derived from DP orifice plate measurements, resulting in a numerical output. The primary aim of constructing an optimal model is to test its performance on new data, preventing overfitting. Overfitting happens when a model achieves high accuracy with training data but poorly on new, unseen data due to excessive adjustment to the training data's specific patterns and noise. This is a significant challenge in machine learning.

A cross-validation method was applied to mitigate overfitting and ensure reliable evaluation. This involved training the models on part of the dataset and evaluating their performance on unseen data. A "hold-out" set, representing 20% of the full dataset, was separated and not used until the final performance evaluation. This approach ensures that the model's performance is assessed on truly unseen data, providing a more accurate measure of its predictive capabilities.

The features used to train the models are derived from the DP orifice plate meter, as defined in Section 3.2.3. These features are crucial for the model to learn and make accurate predictions. For the regression problem, the target variable is the estimated steam quality (x), as seen in Figure 3.9. Proper labelling of these target variables is essential for effective model training and accurate predictions. In this context, the differential pressures were renamed as follows: ΔP_t to DP12, ΔP_{ppl} to DP12 and ΔP_r to DP32.

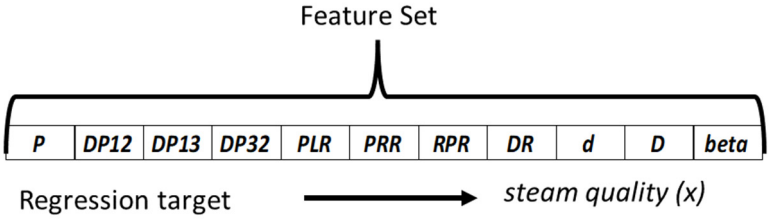


Figure 3.9 Feature set and target variable

For feature selection in this study, recursive feature elimination with cross-validation (RFECV) and random forest feature importance were used to optimise the model's performance, as described in Section 2.2.3.1. By selecting the most relevant features and reducing computational complexity, these methods help to enhance the model's efficiency and accuracy. RFECV works by systematically eliminating the least important features and using cross-validation not only to evaluate model performance but also to reduce prediction error. Similarly, Random Forest determines feature importance by assessing how much each feature decreases prediction error across all trees during training. Both methods, implemented using Scikit-learn, ensured that the optimal number of features was selected, thus improving the model's overall performance. In addition, hyperparameter tuning was performed for several algorithms, including Random Forest, DBSCAN, K-means, and

SelectKBest.

For the Random Forest algorithm, described in Section 2.2.1.1, key hyperparameters were tuned to optimise model performance. The number of trees (n_estimators), which represents how many decision trees are built in the forest, was initially set to 100 and incrementally increased. The performance metrics improved up to 600 trees, after which no significant gains were observed, indicating that 600 trees provided the optimal balance between accuracy and computational efficiency. Other parameters, such as the maximum depth of the trees (max_depth), which controls how deep each tree can grow, were left at the default value of None, allowing the trees to expand until all leaves are pure or contain fewer than the minimum samples required to split a node (min_samples_split). This parameter, min_samples_split, determines the minimum number of data points needed to split an internal node, while the minimum number of samples at a leaf node (min_samples_leaf) defines the smallest number of samples that a leaf node can contain. These parameters were also kept at their default values. To identify the optimal combination of these hyperparameters, Grid Search Cross-Validation (GridSearchCV) was employed, which systematically explores various parameter combinations to find the most effective ones. Additionally, the random_state parameter was set to 42 to ensure reproducibility. The random_state acts as a seed value, which is a starting point for the random number generator that controls the randomness of data splitting. The choice of 42 was arbitrary, but it is commonly used in the machine learning community as a standard example. Using the same seed value (42), the model produces the same random splits and results each time it is run, ensuring consistency and allowing others to replicate the analysis exactly.

For the DBSCAN algorithm, as described in Section 2.2.2.3 and Figure 2.3, the key hyperparameters are 'eps', epsilon (ε), which specifies the radius within which points are considered neighbours, and 'min_samples,' which determines the minimum number of points required to form a dense cluster (core point). In this study, the optimal values for these parameters were determined through GridSearchCV, a systematic search process. After evaluating different combinations, eps was set to 0.10 and min_samples to 10, ensuring the algorithm accurately identified clusters of varying densities. This tuning process is essential for DBSCAN to effectively group closely packed points and distinguish outliers in large, complex datasets.

For K-means clustering, as described in Section 2.2.2.2, the primary hyperparameter is the number of clusters (k). In this study, the value of k was kept at 300, consistent with the methodology by Juliusson et al., (2023) to facilitate direct comparison of results. The decision was made to ensure consistency and comparability with earlier studies.

For the SelectKBest method, as described in Section 2.2.3.1, the key parameter tuned was k, which specifies the number of top features to select based on their statistical significance. This tuning process involved evaluating the model's performance with different values of k to identify the optimal subset of features. The final selected features included:

- the pressure upstream of the orifice (P1),
- the traditional differential pressure over the orifice (DP12),
- the pressure loss ratio (PLR = DP13/DP12),
- the pipe diameter (D) and the orifice to pipe diameter ratio (β)

These features were identified as the most relevant in Juliusson et al., (2023) research, showing the best performance results.

Chapter 4

Results and Discussion

This chapter outlines the findings from all the activities carried out in this study. Section 4.1 organizes and presents the data after preprocessing and cleaning. Section 4.2 discusses predictions for the two-phase flow mass rate using predefined correlation models. Lastly, Section 4.3 provides a detailed discussion on the results of the machine learning prediction models.

4.1 Data

Table 4.1 summarizes the key details of the experiments conducted during the two phases of this study, including the wells used, the measurement methods applied for Reference Measurements (RM) and Experimental Measurements (EM), and the β ratios employed. This overview provides context for understanding the experimental setups and the corresponding results that will be discussed in the following sections.

Table 4.1 Summary of the Experiments

Experiments	Year	Wells	RM	EM	β
Phase 1	2019-2021	ÞG-11, ÞG-15, ÞG-18	Separator method	DP meters	0.5 , 0.7
Phase 2	2023	BJ -12	Water tracer method	DP meters	0.4 , 0.7

4.1.1 Total Flow Rate and Enthalpy

The results from the phase 1 experiments, conducted between 2019 and 2021, are presented in Figure 4.1, which includes all data test runs data points including varying the β values. The phase 2 experiments, carried out in 2023, are shown in Figure 4.2. As noted in the previous chapter, fewer data points were collected during the phase 2 experiments.

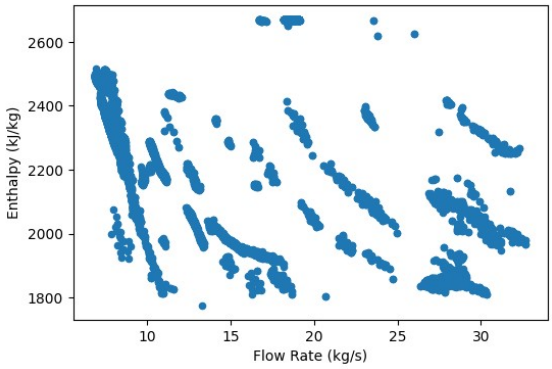


Figure 4.1 Data points from phase 1 experiments in 2019-2020

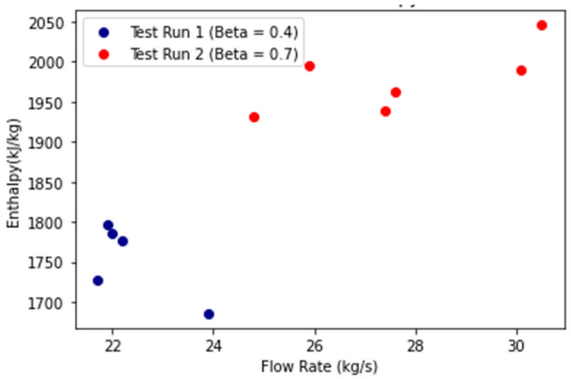


Figure 4.2 Data points from phase 2 experiments in 2023

4.1.2 Orifice Plate Measurements

Figure 4.3 presents the Experimental Measurement (EM) data from the 2019-2021 experiments, focusing on the differential pressure parameters DP12, DP13, and DP32, as defined in Equations (3.2), (3.3) and (3.4). Figure 4.4 shows the DP measurements taken during the 2023 experiments. The highlighted blocks in this figure indicate the periods during which the water tracer method was performed for test run 2 on well BJ-12. This data reflects the orifice meter readings for differential pressures DP12, DP13, and DP32.

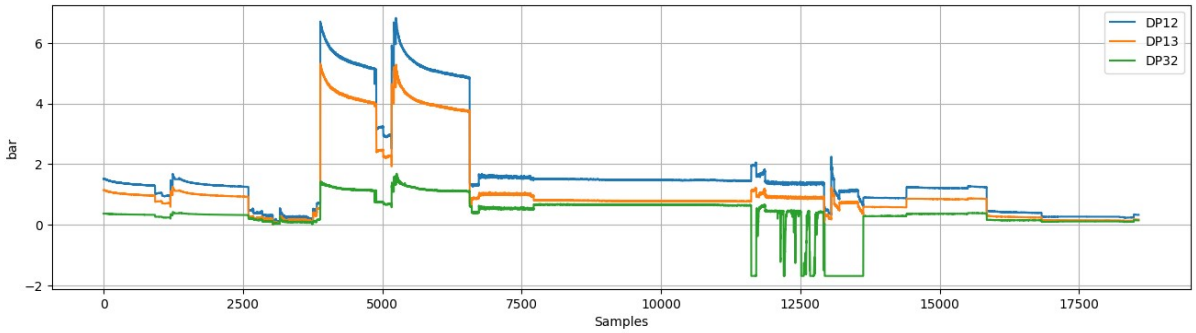


Figure 4.3 Phase 1 DP measurements in Þeistareykir - all test runs

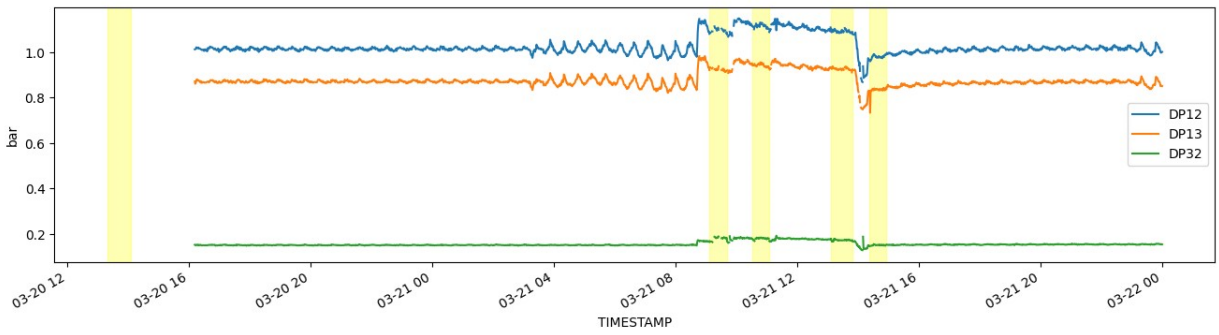


Figure 4.4 Phase 2 DP measurements BJ12- test run 1 with highlighted zones when the water trace method was performed

4.2 Rule Base Model

4.2.1 Two-phase Flow Correlation Models

In this section, the performance of two widely used two-phase flow correlation models Mubarok et al., (2021) and James, (1965)**,** described in Section 2.1.5, was evaluated. These models are essential for calculating flow rates in geothermal wells.

The evaluation was based on comparing the predicted flow rates from each model with the measured field data. This comparison was quantified using two error metrics: Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) using Equations (3.9) and (3.10), which are standard measures to assess the accuracy of model predictions against observed data. Table 4.2 presents the MAPE and RMSE for the two models.

The Mubarok et al., (2019) model requires accurate enthalpy values, which were used to calculate the flow rates as described in Equation (2.14).

For the James, (1965) model, the flow rates were calculated using Equations (2.15) and (2.16). These equations rely on the pressure P_1 from RM and enthalpy being an integral part of the calculation.

The James, (1965) model consistently provides more accurate predictions, as indicated by its significantly lower MAPE and RMSE values compared to the model from Mubarok et al., (2019). In both phases and both models, the reference measurements (RM) parameters were used to calculate the flow rates.

Table 4.2 Two phase flow correlation models performance

Correlation Model		Full Data Set	BJ-12
Mubarok (2019)	MAPE %	189,6	18,7
	RMSE	38,8	6,1
James (1965)	MAPE %	12,1	23,5
	RMSE	2,7	7,4

The results indicate that the James, (1965) model provides more accurate predictions with a significantly lower MAPE and RMSE compared to the Mubarok et al., (2019) model for the full data set. The James model is especially effective when the steam quality (x) is close to 1, as the flow behaviour closely resembles single-phase steam. For the phase 2 dataset, the results are similar, with less pronounced differences between the two models. The other models mentioned in Section 2.1.5, could not be applied due to the lack of collected parameters for this dataset.

Figure 4.5 shows the comparison of measured and calculated flow rates for the Mubarok (2019) and James (1965) models, respectively.

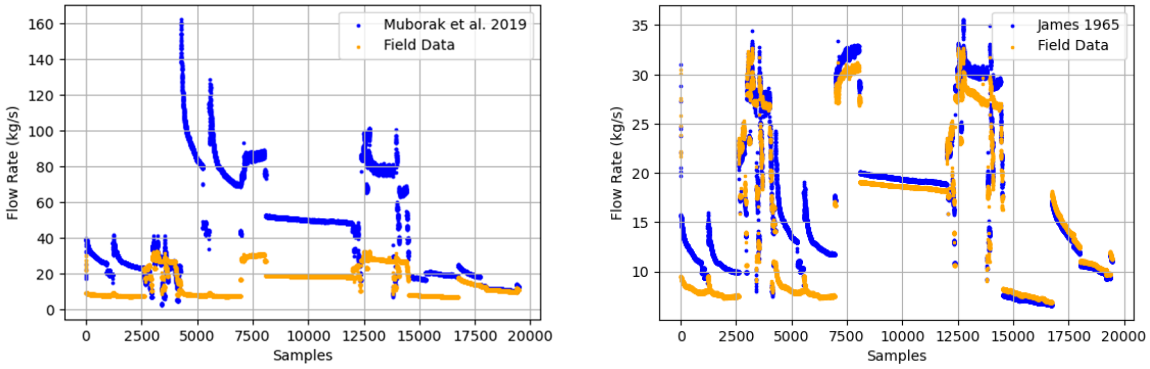


Figure 4.5 Measured vs calculated flow rates using two phase flow correlations

From these figures, it can be observed that the James, (1965) model aligns more closely with the measured field data, demonstrating better overall accuracy compared to the Mubarok et al., (2019) model. The Mubarok model shows larger deviations, especially at higher flow rate regions. This conclusion is further supported by the significantly lower MAPE and RMSE values for the James model, indicating better performance and accuracy.

4.3 Machine Learning Models

This section provides the results and discussion of all the procedures implemented for machine learning models in this research. The focus is on the development and evaluation of various scenarios using different algorithms, with a primary emphasis on the Random Forest regression model.

The study involved updating previous work, referred to as the baseline, done by Juliusson et al., (2023), with phase 2 data points. The initial step was to run the Juliusson et al., (2023) code with the new data points to assess performance.

To enhance the model performance, several cases were created in this study by altering different components of the machine learning processes. These cases included variations in clustering methods, data reduction techniques, and feature selection methods while consistently using Random Forest as the regression model. The aim was to determine the optimal combination of these techniques to improve the accuracy and reliability of the predictions.

The following cases were modelled:

Case 1: Targeted Feature Selection Using K-means and Grid Methods in Random Forest Regression (Section 4.3.1) - This case focused on evaluating whether incorporating new data points would enhance the prediction performance of the model by using specific features identified in previous studies by (Juliusson et al., 2023) .

Case 2: SelectKBest for All Features Using K-means and Grid Methods in Random Forest Regression (Section 4.3.2) - In this case, the selection of different features was explored to improve the model's prediction accuracy, using SelectKBest to identify the most relevant features.

Case 3: Dimensionality Reduction and Clustering with Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Principal Component Analysis (PCA) in Random Forest Regression (Section 4.3.3) - This case investigated the combination of DBSCAN for clustering and PCA for dimensionality reduction to refine the model's performance.

Case 4: Fractional Data Reduction, Clustering, and Feature Elimination with Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Recursive Feature Elimination with Cross-Validation (RFECV) in Random Forest Regression (Section 4.3.4) - This final case examined the impact of reducing data points and using RFECV for feature selection to enhance the model's predictive capabilities.

Each of these cases provided insights into the effectiveness of different data processing and feature selection techniques, ultimately contributing to the development of a robust Random Forest regression model for predicting steam quality.

4.3.1 Case 1: Targeted Feature Selection Using K-means and Grid Methods in Random Forest Regression

For this case, the objective was to evaluate if incorporating new data points would enhance the prediction performance of the model. The original data set from phases 1 and 2 consisted of 19459 data points. Figure 4.6 shows the flow diagram of the model, which follows a series of steps beginning with data collection, preprocessing, and scaling.

The initial step involved data collection, where data from Phase 1 and Phase 2 were collected. Data preprocessing was the next step, combining the datasets from both phases and performing necessary cleaning and preparation tasks. This step also involved calculating necessary parameters, such as differential pressures and ratios, to prepare the data for model training. Once preprocessed, the data was standardized by scaling, as described in Section 2.2.3.2, ensuring all features had a mean of zero and a standard deviation of one.

The subsequent steps focused on data reduction, which was done using two methods: K-means (Section 2.2.2.2) clustering and the Grid method similar to the approach used in the previous work by Juliusson et al., (2023).

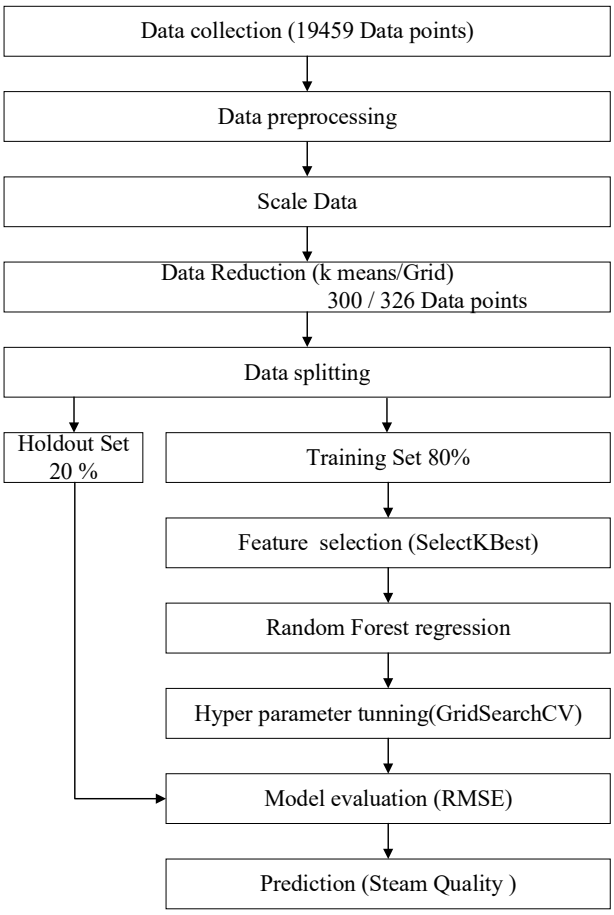


Figure 4.6 Flow diagram Case 1: Targeted Feature Selection Using K-means and Grid Methods in Random Forest Regression

K-means clustering assigns each observation to a cluster based on its proximity to the nearest cluster center. In this study, 300 clusters were created using features such as pipe diameter, beta ratio, upstream pressure (P1), differential pressure over the orifice (DP12), and the pressure loss ratio (PLR=DP13/DP12). The Grid method selected one data point closest to the center of each block defined by flow rate, enthalpy, pipe diameter, and β ratio. This reduced the data from 19 459 to 326 points, ensuring a representative subset.

Feature selection was performed using the same features identified in Juliusson et al., (2023) research: P1, DP12, PLR, D, β. These features were selected based on a combination of insights from the data and results from a linear regression algorithm (Einarsson, 2021).

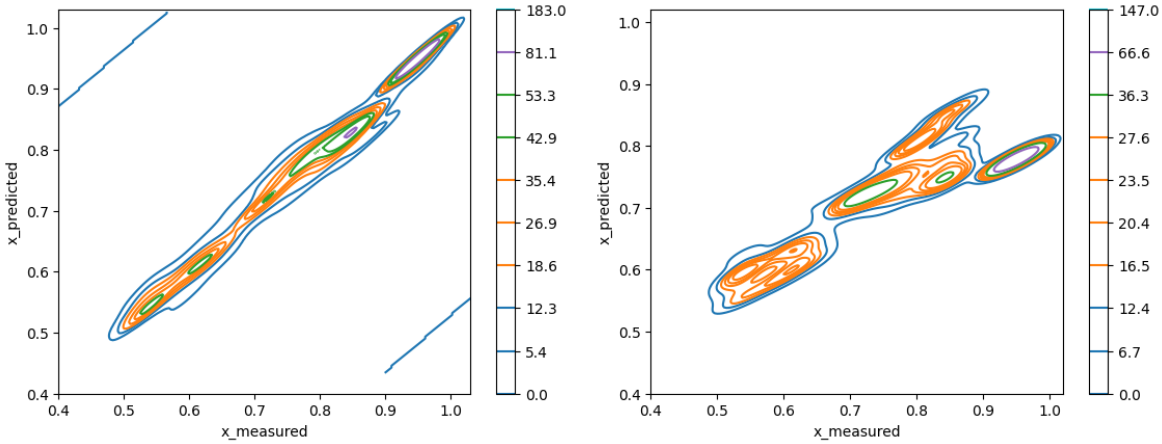
After data reduction and feature selection, the dataset was split into a training set (80%) and a holdout set (20%). Random Forest regression (Section 2.2.1.1) was used as the predictive model, followed by hyperparameter tuning (Section 2.2.3.4) and model evaluation using RMSE (Section 3.3). The results including all data (phase 1 and 2) for K-means clustering and the Grid method were compared against the results by Juliusson et al., (2023) in Table 4.3.

Table 4.3 Comparison of RMSE results

Method	RMSE Phase 1 + 2	RMSE Phase 1 (Juliusson et al., 2023)
K-means Clustering	0.036	0.030
Grid Method	0.054	0.075

The updated results which included new data of Phase 2 indicate a slight improvement in the RMSE for the Grid method, decreasing from 0.075 to 0.054. However, the K-means clustering method showed a minor increase in RMSE, from 0.030 to 0.036. This suggests that while both methods provide relatively accurate predictions, the Grid method benefited more from the additional data points.

Additionally, a useful approach to visualize the quality of the predictions is plotting the measured and predicted steam quality using contour lines, representing areas of similar data point density, showing where the predicted and measured steam quality values frequently occur. In the plot, the x-axis represents the measured steam quality (x_measured) obtained from field data, while the y-axis shows the predicted steam quality (x_predicted) generated by the model. The closer these lines are to the diagonal line x_predicted = x_measured, the better the model's predictions align with the actual measurements. The color intensity of the contour lines shows the density of these points, with darker regions indicating a higher concentration of predictions and measurements in close proximity, suggesting more frequent occurrence of those values. Figure 4.7 shows these contour plots for the Grid method and K-means clustered data, respectively. These visualizations highlight how effectively K-means clustering creates a representative subset of data for making accurate predictions.



(a) Using K-means method (b) Using Grid method

Figure 4.7 Measured vs predicted steam quality values

The K-means clustering method proved to be highly effective, maintaining strong performance even with large datasets and various changes in data. These findings emphasize how crucial data reduction techniques and careful feature selection are for boosting the performance of models.

Overall, the combination of scaling, clustering, feature selection, and hyperparameter tuning has proven effective in developing a reliable Random Forest regression model for predicting steam quality. The improvements observed in the Grid method's RMSE demonstrate the value of incorporating new data points and refining data processing techniques.

Given these results, it is crucial to explore the selection of other features or consider all features to see how this could further improve the model's performance. The next case will focus on this modified approach, aiming to identify the optimal set of features that maximize predictive accuracy.

4.3.2 Case 2: SelectkBest for all features using K-means and Grid Methods in Random Forest Regression

In this case, the objective was to explore the selection of different features to enhance the prediction performance of the model. The process used the same flow diagram as in Figure 4.6 , and followed the same initial steps as described in Case 1 (Section 4.3.1.), including data preprocessing, scaling, and splitting the dataset into a training set (80%) and a holdout set (20%).

The SelectKBest method (Section 2.2.3.1) was used to select different combinations of features based on their relevance. Random Forest regression (Section 2.2.1.1) was used as the predictive model, followed by hyperparameter was applied to optimize the model. The performance of each feature combination was evaluated using the RMSE metric, and both RMSE values and the importance of selected features were recorded as shown in Table 4.4.

In Table 4.4 the Selected Features column presents the features identified by the SelectKBest method for each value of k, where k represents the number of top features selected. The RMSE column displays the model's performance for each set of features, with lower RMSE (see Equation (3.10)) values indicating better predictive accuracy. The Feature Importance column provides the importance scores of the selected features, ranked by their contribution to minimizing the RMSE. A higher importance score signifies that the feature has a more significant impact on reducing the model's predictive error, thereby enhancing the accuracy of the predictions.

Table 4.4 RMSE and selected features and feature importance

Number of Features (k)	Selected Features	RMSE	Feature Importance
1	['DP32']	0.0687	DP32: 1.0
2	['DR', 'DP32']	0.0234	DP32: 0.6281, DR: 0.3719
3	['DR', 'DP32', 'P1']	0.0233	DP32: 0.6271, DR: 0.1948, P1: 0.1780
4	['DR', 'DP32', 'P1', 'RPR']	0.0169	RPR: 0.6889, DP32: 0.2347, P1: 0.0388, DR: 0.0376
5	['DR', 'DP32', 'P1', 'RPR', 'd']	0.0164	RPR: 0.5809, d: 0.3194, P1: 0.0355, DR: 0.0331, DP32: 0.0311
6	['D', 'DR', 'DP32', 'P1', 'RPR', 'd']	0.0164	RPR: 0.5808, d: 0.3184, P1: 0.0348, DR: 0.0337, DP32: 0.0310, D: 0.0013
7	['D', 'DR', 'DP32', 'P1', 'RPR', 'PRR', 'd']	0.0154	RPR: 0.5630, d: 0.3183, P1: 0.0344, DR: 0.0324, DP32: 0.0294, PRR: 0.0212, D: 0.0013
8	['D', 'DR', 'DP12', 'DP32', 'P1', 'RPR', 'PRR', 'd']	0.0123	RPR: 0.5621, d: 0.3158, DP12: 0.0404, P1: 0.0211, DP32: 0.0206, DR: 0.0198, PRR: 0.0197, D: 0.0006
9	['beta', 'D', 'DR', 'DP12', 'DP32', 'P1', 'RPR', 'PRR', 'd']	0.0123	RPR: 0.5614, d: 0.3144, DP12: 0.0401, P1: 0.0213, DP32: 0.0206, DR: 0.0200, PRR: 0.0201, beta: 0.0015, D: 0.0006
10	['beta', 'D', 'DR', 'DP12', 'DP13', 'DP32', 'P1', 'RPR', 'PRR', 'd']	0.0123	RPR: 0.5634, d: 0.3143, DP12: 0.0355, P1: 0.0209, DR: 0.0195, DP32: 0.0180, PRR: 0.0177, DP13: 0.0085, beta: 0.0016, D: 0.0006
11	['beta', 'D', 'DR', 'DP12', 'DP13', 'DP32', 'P1', 'PLR', 'RPR', 'PRR', 'd']	0.0125	RPR: 0.4888, d: 0.3173, PLR: 0.0880, DP12: 0.0228, P1: 0.0210, DR: 0.0202, PRR: 0.0170, DP32: 0.0122, DP13: 0.0067, D: 0.0055, beta: 0.0005

The exploration of feature selection highlighted several key findings. As the number of selected features increased, the RMSE generally decreased, indicating improved model performance. The best RMSE was achieved using 9 and 10 features, with a value of 0.0123. This demonstrates the importance of selecting an optimal set of features for accurate predictions.

The selected features varied across different values of k, but certain features such as RPR, d, DP12, and DP32 consistently appeared among the top features, indicating their significance in predicting steam quality. Including these features contributed to a more robust and accurate Random Forest regression model.

Conversely, features like β and D were among the least important, as indicated by their lower importance scores in the Random Forest model. This suggests that while these features are part of the dataset, they do not significantly contribute to the predictive accuracy of steam quality.

While the SelectKBest method is useful for identifying the most relevant features, it has some limitations. One of the main disadvantages is that it evaluates each feature independently of the others. This means it might miss combinations of features that are only significant when considered together. For example, a pair of features might not be particularly informative on their own but could be highly predictive when combined. SelectKBest does not account for such interactions, potentially overlooking optimal feature combinations.

Overall, the results suggest that a careful selection of features, combined with effective data reduction techniques and hyperparameter tuning, can significantly enhance the performance of the model. The limitations of the SelectKBest method highlight the need for more sophisticated feature selection techniques that consider feature interactions. In the next cases, methods such as Recursive Feature Elimination with Cross-Validation (RFECV) will be explored.

4.3.3 Case 3: Dimensionality Reduction and Clustering with Based Spatial Clustering of Applications with Noise (DBSCAN) and Principal Component Analysis (PCA) in Random Forest

In this case, involves using DBSCAN (Section 2.2.2.3) as a clustering method and PCA (Section 2.2.2.1) for dimensionality reduction to address some limitations observed with previous methods and enhance the model's predictive accuracy. The combination aims to improve the identification of data structures and enhance the overall predictive accuracy of the model.

The overall process, including the steps taken and the final evaluation, is depicted in Figure 4.8.

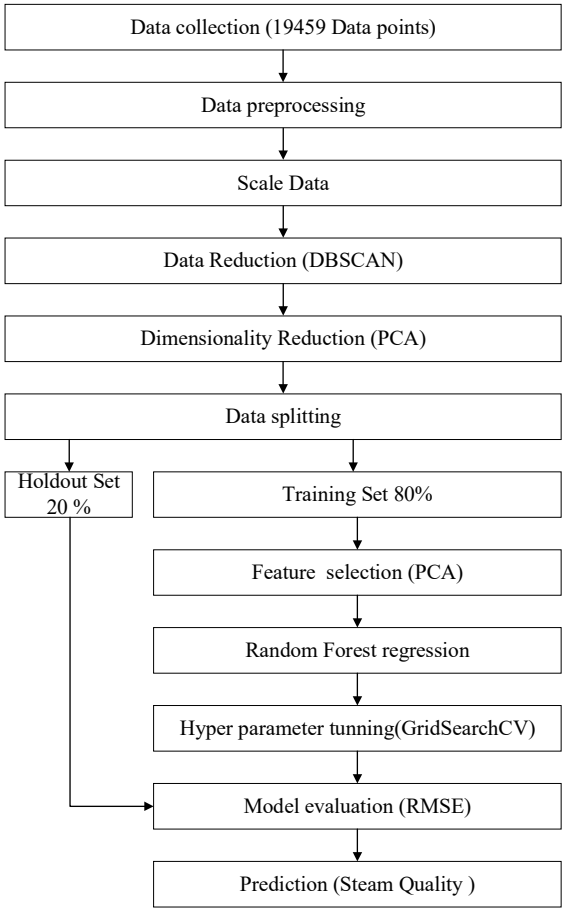


Figure 4.8 Flow diagram Case 3: Dimensionality Reduction and Clustering with Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Principal Component Analysis (PCA) in Random Forest Regression

The dataset was prepared by combining phase 1 data points with phase 2 data points. The data preprocessing was performed as described in Section 4.3.1, involving scaling the data to ensure consistency across features.

DBSCAN (Section 2.2.2.3) was then applied to identify clusters, filtering out noise and outliers. After applying DBSCAN, PCA (Section 2.2.2.1) was used to transform the filtered data into principal components., as described in Section 2.2.2.1. The PCA was designed to retain 95% of the total variance in the data, which means that the selected components capture the most significant patterns and variations. This reduced the

dimensionality from 11 features to 5 principal components (PC1, PC2, PC3, PC4, PC5).

Principal components are defined as linear combinations of the original features that capture the maximum variance in the data. The first principal component (PC1) captures the most variance, the second principal component (PC2) captures the second most variance, and so on. In this study, the first five principal components were selected because they collectively explained 95% of the total variance in the dataset, effectively summarising the information contained in the original 11 features.

The transformed dataset was split into training (80%) and hold-out (20%) sets. A Random Forest regression model was trained using GridSearchCV (Section 2.2.3.4) to optimise hyperparameters. The best model was selected based on the lowest RMSE of 0.0134.

The results of the feature importance analysis, as shown in Figure 4.9 , indicated that PC2 and PC1 were the most significant principal components for predicting steam quality. This highlights the effectiveness of PCA in capturing the most relevant information from the dataset.

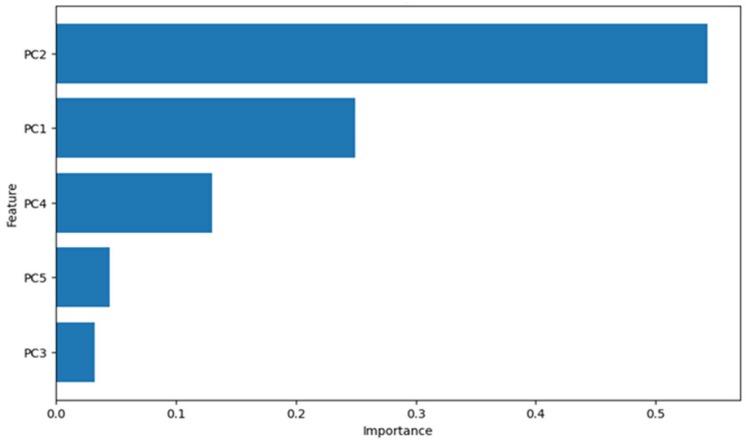


Figure 4.9 PCs feature importance results

The detailed loading scores for the principal components are presented in Table 4.5. For example, PC2 had high loadings for DP32 (0.372) and PRR (0.547), indicating their significant impact on steam quality predictions. PC1 showed high loadings for DP12 (0.437) and DP13 (0.443). PC4's high loading for PLR (0.857). While PC5 and PC3 were less important, PC5's notable loading for the orifice to pipe diameter ratio (beta) was at 0.711.

Table 4.5 Principal components loading scores

Feature	PC1	PC2	PC3	PC4	PC5
P1	0,028	-0,328	0,596	-0,058	-0,077
DP12	0,437	-0,101	0,041	-0,202	0,272
DP32	0,323	0,372	0,197	-0,200	0,064
DP13	0,443	-0,111	0,049	-0,127	0,234
beta	-0,370	0,064	0,069	-0,262	0,711
PRR	0,081	0,547	0,281	0,165	0,025
RPR	0,046	0,558	0,247	-0,166	-0,166
DR	0,028	-0,328	0,596	-0,056	-0,081
d	-0,431	0,073	0,186	-0,077	0,324
D	-0,382	0,067	0,227	0,182	-0,279
PLR	0,170	0,017	0,125	0,857	0,372

In this case, the application of DBSCAN and PCA demonstrated a significant improvement in model performance, as evidenced by the reduced RMSE. The results emphasise the importance of advanced clustering and dimensionality reduction techniques in refining predictive models.

Figure 4.10 shows the contour plot for the Case 3 (DBSCAN - PCA – RFR) model, demonstrating a highly linear relationship between predicted and measured steam quality values. The levels in the contour graph represent different density thresholds, which each level indicating the concentration of data points at that specific density, as shown by the colour bar on the side. The close alignment of these lines with the diagonal (x_predicted = x_measured) line indicates that the model's predictions closely match the actual measurements.

This linearity is a direct result of the PCA's effect on the data, where the dimensionality reduction process compresses the original feature set into a few principal components that encapsulate the main variance in the dataset. PCA effectively captures the most critical linear relationships among the features, leading to the observed alignment in the predictions. This transformation allows the Random Forest Regression model to focus on the most impactful variations, contributing to its predictive accuracy.

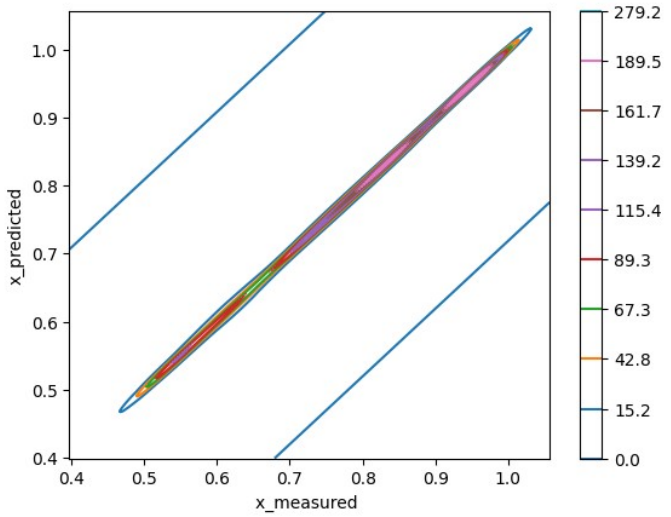


Figure 4.10 Measured vs predicted steam quality with DBSCAN -PCA -RFR model

However, it is important to acknowledge the limitations of using PCA for feature reduction. Although PCA is effective at capturing the majority of variance in the data, it may overlook complex interactions between the original features that could be significant for the model. To address this, future work will focus on using Recursive Feature Elimination with Cross-Validation (RFECV) to systematically evaluate different combinations of features, aiming to enhance model performance. Additionally, investigate how varying the fractions of the dataset used for the training can further improve the model's accuracy.

4.3.4 Case 4: Fractional Data Reduction, Clustering and Feature Elimination with Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Recursive Feature Elimination with Cross Validation (RFECV) in Random Forest Regression

This case evaluates the effect of reducing data points and applying Recursive Feature Elimination with Cross-Validation (RFECV) (Section 2.2.3.1)for feature selection on the model's predictive performance. The aim is to explore the relationship between data quantity, feature selection, and model accuracy, and enhance predictive capabilities.

The process begins by applying the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Section 2.2.2.3) algorithm to clean the dataset, removing noise and outliers, and identifying clusters. This leads to a refined set of data points, which is then used to create different fractions of the dataset (2%, 10%, 25%, 50%, 75%, and 100%). These fractions are sampled from each identified cluster to assess the impact of dataset size on model performance. The Random Forest Regressor is then trained on each subset, initially using all 11 features (see Figure 3.9) described in the study. RFECV (Section 2.2.3.1) is employed to systematically identify and retain the most influential features, thereby optimising model performance by minimising the Root Mean Squared Error (RMSE) (Equation (3.10). The flow diagram depicting this process is shown in Figure 4.11.

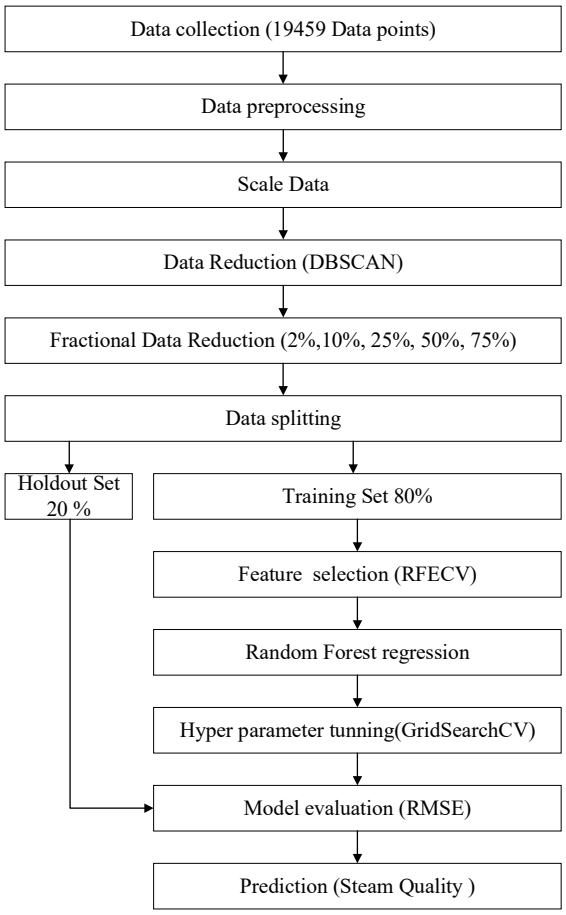


Figure 4.11 Flow diagram Case 4: Fractional Data Reduction, Clustering and Feature Elimination with Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Recursive Feature Elimination with Cross Validation (RFECV) in Random Forest Regression

The initial step involved combining datasets from phase 1 and phase 2 experiments. This combined dataset was used for further analysis. The selected features for clustering included PRR, RPR, PLR, P1, DP12, DP13, DP32, beta, D, d, and DR. DBSCAN was applied to remove noise and outliers from the dataset, ensuring a cleaner and more reliable set of data points for analysis. DBSCAN was applied to the scaled features, identifying 18 clusters and 43 noise points. The distribution of data points across these clusters is shown in the following Table 4.6

Table 4.6 Fraction and points in each cluster

DBSCAN method 18 Clusters	
2%	387 Points
10%	1942 Points
25%	4857 Points
50%	9712 Points
75%	14564 Points
100%	19421 Points
Noise	43 Points
Total Data Set	19464 Points

Different fractions (2%, 10%, 25%, 50%, 75%, and 100%) of data points from each cluster were sampled to evaluate their impact on model performance. RFECV (Section 2.2.3.1) was implemented to evaluate different combinations of features systematically and identify the optimal subset that contributes to the highest model accuracy. A Random Forest regression (Section 2.2.1.1) model was trained using the selected features, with hyperparameter tuning applied to optimise performance.

Figure 4.12 and Figure 4.13 illustrate the comparison of all the data points (phase 1 and 2 points) with the 2% and 10% sampled data. The decision to sample 2% of the data was based on the approach taken by (Juliusson et al., 2023), where the data was similarly reduced to approximately 2%. These visualisations show that even with a small percentage of sampled data, the general trends and patterns are preserved, allowing for effective model training.

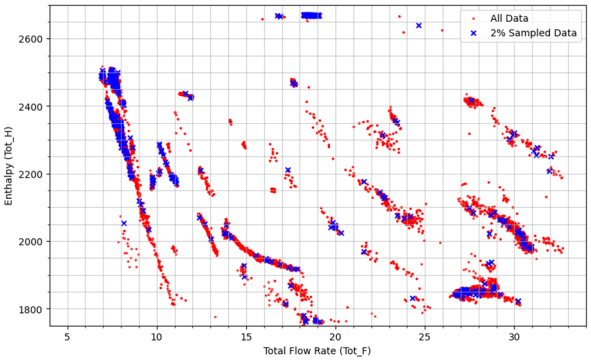


Figure 4.12 Enthalpy vs flow rate - 2% sampled Data

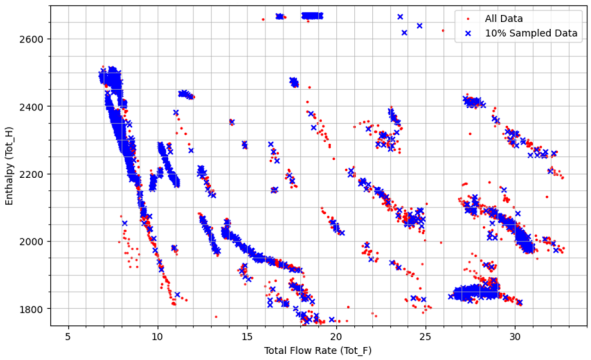


Figure 4.13 Enthalpy vs flow rate - 10% sampled Data

The results indicate that reducing the data to as little as 25% still maintains a high level of predictive accuracy, as evidenced by the RMSE result. This finding highlights the efficiency of DBSCAN in noise reduction and the robustness of the Random Forest model when combined with RFECV for feature selection. The ability to achieve comparable performance with reduced data points suggests potential for faster training times and lower computational costs.

The RFECV method, combined with Random Forest Regression, demonstrates the potential to optimise feature selection and improve predictive accuracy significantly. The results for different data fractions are summarised in Table 4.7.

Table 4.7 Results for Case 4

Fraction	Optimal Features Test	RMSE
2%	[PRR, RPR, PLR, DP32, d, DR]	0,0280
10%	[PRR, RPR, PLR, P1, DP12, DP32, d, DF	0,0316
25%	[RPR, PLR, P1, DP12, d]	0,0187
50%	[RPR, PLR, P1, DP12, DP32, d, DR]	0,0137
75%	[RPR, PLR, P1, DP13, d]	0,0117
100%	[PRR, RPR, PLR, P1, DP32, d, DR]	0,0122

The study demonstrates that feature importance varies across different data fractions, indicating that feature relevance can change with dataset size. However, features such as RPR, PLR, P1, DP12, and d consistently appeared among the top selected, underscoring

their significance in predicting steam quality. As the dataset size increased, the RMSE decreased, highlighting the importance of larger datasets for achieving better predictive accuracy. The lowest RMSE of 0.0117 was achieved with a 75% dataset fraction using the features [RPR, PLR, P1, DP13, d]. The effectiveness of RFECV was evident in its ability to identify the most relevant features, improving model accuracy by minimising noise and irrelevant variables. Nonetheless, RFECV has limitations as it does not account for interactions between features, potentially missing combinations that are collectively significant. Overall, this study illustrates the value of systematic feature selection and varying dataset sizes in developing robust machine-learning models, underscoring the potential of these techniques in predicting geothermal fluid properties and providing a basis for future research and industrial applications.

4.3.5 Summary

Table 4.8 presents a summary of all the machine-learning models evaluated in this study, focusing on predicting steam quality. The study applied various methodologies, including clustering, feature selection, and regression techniques, to identify the most effective model configurations.

The best-performing model was Case 4: Fractional Data Reduction, Clustering, and Feature Elimination with DBSCAN and RFECV in Random Forest Regression. Using a 75% dataset fraction with the features [RPR, PLR, P1, DP13, d], this model achieved the lowest RMSE of 0.0117. This result highlights the success of advanced feature selection and data reduction strategies.

The most significant features across different models were RPR, PLR, P1, DP12, and d, which consistently proved crucial in enhancing predictive accuracy. Looking at the data Reduction Impact, both K-means and Grid-based clustering methods highlighted the importance of representative data sampling, with K-means generally showing better adaptability. In dimensionality reduction, PCA, when paired with DBSCAN, effectively captured essential variance, although it may overlook complex interactions between features.

Table 4.8 Summary of results from machine learning models

Case description	RMSE	Brief Summary
Baseline: K-means - SelectKBest - RFR by Juliusson et al. (2023)	0.030	Phase 1 data, K-means for clustering for data reduction and SelectKBest for feature selection, with RFR.
Baseline: Grid - SelectKBest - RFR by Juliusson et al. (2023)	0.075	Phase 1 data, a grid-based method for clustering for data reduction, followed by feature selection with SelectKBest and RFR.
Case 1: Targeted Feature Selection Using K-means in Random Forest Regression	0.036	All data (phase 1 and 2) K-means for clustering for data reduction, specific features and RFR.
Case 1: Targeted Feature Selection Using Grid Methods in Random Forest Regression	0.054	All data (phase 1 and 2) Grid method for clustering for data reduction, specific features and RFR.
Case 2: SelectKBest for All Features Using K-means in Random Forest Regression	0.0123	All data (phase 1 and 2) ,K-means for clustering , modifications in feature selection led to a significant reduction in RMSE, and RFR.
Case 3: Dimensionality Reduction and Clustering with DBSCAN and PCA in Random Forest Regression	0.0134	All data (phase 1 and 2) , used DBSCAN for clustering and removing noise and PCA for dimensionality reduction, with RFR.
Case 4: Fractional Data Reduction, Clustering, and Feature Elimination with DBSCAN and RFECV in RFR	0.0117	All data (phase 1 and 2) , combined DBSCAN with RFECV for feature selection and varied dataset sizes and RFR.

To implement this model in real future applications, the process starts by placing a Differential Pressure (DP) orifice plate meter in the flow line where the two-phase geothermal fluid is present. The algorithm then reads the data from these meters and estimates the steam quality using the model developed. With this estimated steam quality, enthalpy and the total flow rate of the geothermal fluid can be computed. Finally, the total available power is calculated using the formula:

$$P_{th} = h\dot{m} \tag{4.1}$$

Where enthalpy h and the total flow rate ṁ are determined from the predicted steam quality, this method can give an estimation of the geothermal well output.

Chapter 5

Conclusion and Recommendations

The objective of this thesis was to develop and evaluate machine learning models for predicting steam quality in flow from geothermal wells. The study applied various data preprocessing, feature selection, and regression techniques to achieve accurate predictions. The findings and insights obtained from this research provide valuable contributions to the field of geothermal energy. The following were the key findings:

- The model using DBSCAN for noise reduction, RFECV for feature selection, and Random Forest Regression (RFR) for prediction achieved an RMSE of 0.011 using five key features: [RPR, PLR, DP13, P1, d].
- Applying RFECV and RFR to the 75% sampled dataset resulted in an RMSE of 0.012 using all available features, demonstrating high predictive accuracy even with reduced data.
- Using Principal Component Analysis (PCA) for dimensionality reduction and RFR for prediction achieved an RMSE of 0.013 using the top 5 principal components, indicating PCA's effectiveness in compressing feature information while maintaining predictive performance.
- DBSCAN effectively reduced noise in the dataset, improving model accuracy and robustness.
- Principal Component Analysis (PCA) was effective in compressing feature information, allowing for efficient and accurate predictions with fewer features.
- Larger datasets generally led to better model performance, as evidenced by lower RMSE values. This underscores the importance of utilising as much relevant data as possible for training machine learning models.
- Using more than five features generally improves prediction accuracy, highlighting the importance of feature selection in building robust models.

To enhance the robustness and generalizability of the models, future research should focus on testing them across diverse geothermal datasets from different fields. This approach will ensure that the models are not overfitted to the specific dataset used in this study and can be reliably applied to other contexts.

Integrating traditional methods, which are typically conducted a couple of times a year, into the model for continuous calibration could greatly improve accuracy over time. By updating the model with fresh data each year, its predictive performance can be refined and enhanced, providing more precise and reliable measurements.

The demand for real-time measurements is not exclusive to Iceland's geothermal fields. This model has the potential to be adapted and implemented in other countries with high-temperature geothermal resources. Conducting similar studies in different geothermal fields worldwide would provide valuable insights into the factors influencing the model's performance, leading to broader applications and advancements in geothermal energy management.

Bibliography

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/wics.101

Awad, M., & Fraihat, S. (2023). Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems. *Journal of Sensor and Actuator Networks*, *12*(5), 67. https://doi.org/10.3390/jsan12050067

Bell, I. H., Wronski, J., Quoilin, S., & Lemort, V. (2014). Pure and Pseudo-pure Fluid Thermophysical Property Evaluation and the Open-Source Thermophysical Property Library CoolProp. *Industrial & Engineering Chemistry Research*, *53*(6), 2498–2508. https://doi.org/10.1021/ie4033999

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chai, T., & Draxler, R. R. (2014). *Root mean square error (RMSE) or mean absolute error (MAE)?* https://doi.org/10.5194/gmdd-7-1525-2014

Einarsson, H. (2021). *Real-time measurement of geothermal well output* [Thesis, Reykjavik University]. https://skemman.is/handle/1946/39453

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231.

Google Earth. (2023). *Location of Bjarnarflag geothermal power plant* [Map]. https://earth.google.com/

Götz, M., Kononets, M., Bodenstein, C., Riedel, M., Book, M., & Palsson, O. P. (2019). Automatic water mixing event identification in the Koljö fjord observatory data. *International Journal of Data Science and Analytics*, *7*(1), 67–79. https://doi.org/10.1007/s41060-018-0132-z

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hauksson, T. (2011). *Landsvirkjun, Kröflustöð. Afkastamælingar borhola með þynningaraðferð og tvífasa mæliblendu. Innleiðing aðferða* (LV-2011/018).

Helbig, S., & Zarrouk, S. J. (2012). Measuring two-phase flow in geothermal pipelines using sharp edge orifice plates. *Geothermics*, *44*, 52–64. https://doi.org/10.1016/j.geothermics.2012.07.003

Hirtz, P. N., Kunzman, R. J., Broaddus, M. L., & Barbitta, J. A. (2001). Developments in tracer flow testing for geothermal production engineering. *Geothermics*, *30*(6), 727–745. https://doi.org/10.1016/S0375-6505(01)00023-2

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

J. Kinney and R. Steven. (2014). *Effects of Wet Gas Flow on Gas Orifice Plate Meters* (p. 7). Colorado Engineering Experiment Station, Inc.

James, R. (1965). Metering of Steam-Water Two-Phase Flow by Sharp-Edged Orifices. *Proceedings of the Institution of Mechanical Engineers*, *180*(1), 549–572. https://doi.org/10.1243/PIME_PROC_1965_180_038_02

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer - Verlag.

Juliusson, E., Sveinsson, K. E., Þórhallsson, R., & Steven, R. (2023). Real-Time Monitoring of Geothermal Wells using the Dual-Differential Pressure Method. *Proceedings World Geothermal Congress*, 16.

Kherif, F., & Latypova, A. (2020). Principal component analysis. In *Machine Learning* (pp. 209–225). Elsevier. https://doi.org/10.1016/B978-0-12-815739-8.00012-2

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., & team, J. development. (2016). Jupyter Notebooks? A publishing format for reproducible computational workflows. In F. Loizides & B. Scmidt (Eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (pp. 87–90). IOS Press. https://eprints.soton.ac.uk/403913/

Landsvirkjun. (2023a). *Act on Landsvirkjun*. The National Power Company of Iceland. https://www.landsvirkjun.com/act

Landsvirkjun. (2023b). *Bjarnarflag Geothermal Station*. https://www.landsvirkjun.com/powerstations/bjarnarflag

Lovelock, B. (2001). Steam flow measurement using alcohol tracers. *Geothermics*, *30*, 641–654. https://doi.org/10.1016/S0375-6505(01)00020-7

MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (pp. 281–297). University of California Press.

McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Milaan, P. (2021). *Decision Tree PlayGolf CART*. GitHub. https://github.com/milaan9/Python_Decision_Tree_and_Random_Forest/blob/main/002_Decision_Tree_PlayGolf_CART.ipynb

Mubarok, M. H., Zarrouk, S. J., & Cater, J. E. (2019). Two-phase flow measurement of geothermal fluid using orifice plate: Field testing and CFD validation. *Renewable Energy*, *134*, 927–946. https://doi.org/10.1016/j.renene.2018.11.081

Mubarok, M. H., Zarrouk, S. J., Cater, J. E., Mundakir, A., Bramantyo, E. A., & Lim, Y. W. (2021). Real-time enthalpy measurement of two-phase geothermal fluid flow using load cell sensors: Field testing results. *Geothermics*, *89*, 16. https://doi.org/10.1016/j.geothermics.2020.101930

Na, S., Xumin, L., & Yong, G. (2010). Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm. *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, 63–67. https://doi.org/10.1109/IITSI.2010.74

Olukanmi, P., Nelwamondo, F., Marwala, T., & Twala, B. (2022). Automatic detection of outliers and the number of clusters in k-means clustering via Chebyshev-type inequalities. *Neural Computing and Applications*, *34*(8), 5939–5958. https://doi.org/10.1007/s00521-021-06689-x

P. Bixley, N. Dench, and D.Wilson. (1998). Development of well testing methods at wairakei 1950-1980. *Proceedings 20th Geothermal Workshop*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,

Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rudd, J. M., & Ray, H. "Gene." (2020). An Empirical Study of Downstream Analysis Effects of Model Pre-Processing Choices. *Open Journal of Statistics*, *10*(05), 735–809. https://doi.org/10.4236/ojs.2020.105046

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781107298019

Tayman, J., & Swanson, D. A. (1999). *On the validity of MAPE as a measure of population forecast accuracy*.

Theng, D., & Bhoyar, K. K. (2024). Feature selection techniques for machine learning: A survey of more than two decades of research. *Knowledge and Information Systems*, *66*(3), 1575–1637. https://doi.org/10.1007/s10115-023-02010-5

Upp, E. L., & LaNasa, P. J. (2014). *Fluid flow measurement: A practical guide to accurate flow measurement* (3d edition). Butterworth-Heinemann, an imprint of Elsevier.

Waskom, M. L. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021