Universitetet
i Stavanger

FACULTY OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS

Study program/ Specialization: Applied Data Science	Spring semester, 2022 Open

Writer: Akram Ourir

Faculty supervisor: Nejm Saadallah, University of Stavanger, NORCE

External supervisor: Cecilie Hiorth, Sval Energi A/S

Wim Lekens, GeoProvider A/S

Thesis title:

Applications of machine learning
in petroleum system characterization using mud gas data

Credits (ECTS): 30

Keywords:: Machine learning, mud gas, geosciences, permeability, well logs	Pages: 88 Stavanger 28. June 2022

Acknowledgement

First, I would like to express my gratitude to my family for their encouragement, motivation, gentle patience, and magnificent love that guides me joyfully every day.

I would like to thank my supervisors, Ms. Cecilie Hiorth, Dr. Wim Lekens and Dr. Nejm Saadallah, for their amazing mentorship, from insightful feedback and comments, knowledge sharing, engagement and guidance throughout this thesis. I would also want Geoprovider A/S and SVAL Energi A/S for providing the data that made this thesis possible.

Contents

Contents **i**

Abstract **xi**

1 Introduction **1**

1.1 Motivation 2

1.2 Objectives 2

2 Technical Background **4**

2.1 Regional Setting 4

2.2 Petroleum system 7

2.2.1 Source rock 7

2.2.2 Reservoir rock 9

2.2.3 Seal rock 12

2.2.4 Trap 12

2.2.5 Migration and timing 12

2.3 Dataset: Technical background 14

2.3.1 Mud gas data 14

2.3.2 Well logging 16

2.3.3 Coring 20

2.4 Causality analysis, literature and limitations 24

2.4.1 Causality analysis 24

2.4.2 Literature 26

2.4.3 Limitations 27

2.5 Summary and objectives 28

3 Theoretical Background: Machine Learning 29

3.1 Introduction 29

3.2 Data and data mining 29

3.2.1 Prediction vs Inference 30

3.2.2 Classification vs Regression 32

3.2.3 Data preprocessing 32

3.2.4 Data analysis 33

3.3 Machine Learning techniques 34

3.3.1 Supervised vs Unsupervised learning 34

3.3.2 Linear models 35

3.3.3 Tree-based models 38

3.3.4 SVM 40

3.3.5 Neural Network (NN) 40

3.4 Assessing model results 41

3.4.1 The Bias-Variance trade off 41

3.4.2 Training, validation and testing 42

3.4.3 Model selection 43

3.4.4 Model performance metrics . . . 44

4 Methodology 47

4.1 Dataset . . . 47

4.1.1 Data and data loading . . . 47

4.1.2 Data processing . . . 50

4.1.3 Data analysis . . . 53

4.1.4 Systematic error . . . 61

4.1.5 Data analysis summary . . . 62

4.2 Project workflow . . . 63

4.3 Machine learning methods and optimization . . . 65

4.3.1 Fluid classification tasks . . . 65

4.3.2 Permeability prediction task . . . 68

4.3.3 Hyperparameter tuning . . . 68

4.4 Results Assessment . . . 69

4.4.1 One well test set . . . 69

4.4.2 Leave One Well Out Cross Validation (LOWOCV) . . . 69

5 Results 71

5.1 First task: Water/Hydrocarbon classification . . . 71

5.1.1 Prediction . . . 71

5.1.2 Inference . . . 72

5.2 Second task: Oil/Gas classification . . . 73

5.2.1 Prediction . . . 73

5.2.2 Inference . . . 74

5.3 Third task: Permeability Prediction 76

5.3.1 Prediction 76

5.3.2 Inference 77

5.4 Petrel visualization 79

5.5 General comments 80

5.6 Limitations and potential improvements 82

6 Conclusion **83**

Bibliografy **88**

List of Figures

2.1 Regional map showing the Area Of Interest (AOI) 5

2.2 More focused map showing the wells used in this study 6

2.3 Stages of petroleum generation. 9

2.4 Porosity and permeability. 10

2.5 Regional lithostratigraphic chart in our AOI. 11

2.6 Structural traps. 13

2.7 Stratigraphic traps. 13

2.8 Simplified schematic of a working petroleum system. 14

2.9 Sketch of a possible gas monitoring system 15

2.10 Schematic representation of onshore logging operation and resulting composite log 17

2.11 Reservoir fluid zones in a normally pressurized petroleum reservoir. . . . 20

2.12 Diagram explaining parameters in the Darcy equation. 23

2.13 Example of poroperm relationship. 24

2.14 A simplified schematic of the processes involved while drilling. 26

3.1 Data mining as a confluence of many disciplines. 30

3.2 A representation of the tradeoff between flexibility and interpretability. . 32

3.3 Linear model classification region vs tree-based classification. 39

3.4 Bias variance trade off. 42

3.5 Example of ROC curve for the second classification task showing sensitivity as function of specificity and associated AUC value. 46

4.1 Example of a dataSet for a well used in this study visualised in Petrel. . 49

4.2 Example of dataSet for a well used in this study visualised in Jupyter-Notebook. 49

4.3 First raws of the data frame generated by merging all the data and resampling to 1 meter. 50

4.4 Missing data analysis. 52

4.5 Histogram of permeability and C1 before and after log transform. 53

4.6 Pairwise cross plots of mud gas data and HCInd response. 54

4.7 Pairwise cross plots of mud gas data and fluidType response. 55

4.8 Pairwise cross plots of mud gas data, well log data and log of permeability. 56

4.9 VIF when all the features are used in the model 57

4.10 Assessment of model residuals and leverage 57

4.11 Linear model summaries and ANOVA 58

4.12 Principal components biplots in reservoir 59

4.13 Principal components biplots in hydrocarbon bearing zones 60

4.14 Principal components biplots in seals 61

4.15 Box plots showing spread of density distribution and C1 distribution by well 62

4.16 Project workflow diagram 64

4.17 Lasso and Ridge optimisation 66

4.18 Neural network model architecture 67

4.19 Evolution of Loss and accuracy with epochs using the neural network model for both training and validation set 67

4.20 Subset selection optimization 68

5.1 Task-1 models inference. 73

5.2 Task-2 subset selection model summary. 75

5.3 Task-2 feature importance. 75

5.4 Cross plot of actual permeability function of the predicted permeability using linear models, colors representing different wells. 76

5.5 VIF of permeability prediction subset selection model features. 77

5.6 Peremability prediction subset selection model summary. 78

5.7 Permeability prediction ANOVA for the model with using CPOR and VShale, and the subset selection model. 78

5.8 Task-1 Petrel visualisation of the classification results 79

5.9 Task-2 Petrel visualisation of the classification results 79

5.10 Task-3 Petrel visualisation of the regression results 80

5.11 Evolution of mse with epochs using the neural network model for the permeability prediction for train set, validation set and LOWCV 81

List of Tables

2.1 Well Database 8

3.1 Confusion Matrix 45

4.1 Models optimized hyperparameters: * means manual tuning of the hyperparameter while the absence of * means grid search was implemented 69

5.1 Task-1 LOWOCV performance comparison for ML models. 72

5.2 Task-1 performance comparison: training set and test set balanced accuracy scores. 72

5.3 Task-2 LOWOCV performance comparison for ML models. 74

5.4 Task-2 performance comparison: training set and test set balanced accuracy scores. 74

5.5 Task-3 LOWOCV performance comparison for ML models. 77

Abstract

Mud gas data are continuous measurements of the different compounds of the gas released from the formation while drilling. While these data are cheap and extensive, as they are always recorded for safety reasons, they are mainly used qualitatively for fluid and reservoir characterization. It is only recently that mud gas data starts to be used quantitatively. The limited use of mud gas data is generally attributed to the difficulty in making a direct relation to reservoir and fluid properties.

The motivation of this study is to assess the potential use of mud gas data for quicker, better, and more cost-effective assessment of fluid and reservoir properties which would be beneficial for both exploration and development. Potentially, this could allow for better understanding of the reservoir, better development strategies, identifying missed net pay leading to new exploration and development opportunities, decrease data acquisition costs and more. In this master project, 40 wells in quad 35 of the Norwegian sector of Northern North Sea, were investigated using data analytics to explore the mud gas data and relate it to reservoir properties by evaluating model outputs and how they relate to real physical behavior. We focused on three predictive tasks (two classifications and one regression):

1. Identifying hydrocarbon bearing zones in the reservoir rock through supervised classification using features from mud gas data
2. Identifying fluid type (Gas or Oil) in hydrocarbon bearing zones through supervised classification using features from mud gas data
3. Estimating permeability in hydrocarbon bearing zones from combining mud gas data with logging data through supervised prediction

First, data were exported from different sources and loaded to R and Jupyter Notebook. The data were resampled in accordance with mud data resolution. Some of the

features were then log transformed and scaled to reduce skewness and improve prediction capability. Then, Multiple machine learning models (Linear with subset selection, Lasso, Ridge, Random Forest, Boosting, SVM and Neural Network) were generated and compared. For each model, hyperparameters were tuned either based on validation set or cross validation. Then, a more realistic approach was developed to test the different models by leave one well out cross validation (LOWOCV), which is less sensitive to systematic error from sampling and experience bias, to rank the different models. For the first and second classification tasks a separate well was also used as a test set.

The first classification results (distinguishing water from hydrocarbon) provided a moderately good prediction. The training set for this task consists of 2952 samples and 281 for the test set. Most of the models provided comparable results. The best model 'logistic regression with features selection' had a LOWOCV AUC=0.82, F1score=0.55, accuracy=0.77. The same model was the best for the test set with AUC=0.87, F1score=0.82, accuracy=0.96. The modest performance for distinguishing water from hydrocarbon using mud gas data is probably due to the presence of low saturation gas in some sand intervals.

The second classification showed much better results and allowed for a very good performance with regards to distinguishing gas and oil. The training set consists of 611 samples and 102 for the test set. The best model 'logistic regression with features selection' had a LOWOCV AUC=0.86, F1score=0.87 and accuracy=0.84, and a test set AUC=0.99, F1score=0.94 and accuracy=0.93. The main features important for the prediction are the different gas ratios which is in alignment with the finding of Pixler in 1969 [22].

The third prediction had a sample size of 282 observations and showed that adding mud gas data to porosity and shale volume improved the prediction of permeability compared to conventional approach of using only porosity or porosity with shale volume. The best model was linear regression with subset selection and R2 of the LOWOCV had improved from 0.65 to 0.85 by adding mud gas features as C1, and C1/(DEPTH*MW). Physically, C1/(DEPTH*MW) or (C1/MW) could be considered as a gas rate corrected by the mud weight or depth*mud weight. However, for this task the sample size is small making the generalization of this finding difficult.

The results of these different models for the three tasks, were loaded to Petrel software to allow better use and visualization by geoscientists.

While the data set is limited, we succeeded in obtaining valuable results from mud gas data which could be a base for further work and investigations.

Chapter 1

Introduction

Understanding reservoir properties is crucial in hydrocarbon exploration, development and production for reserve calculation. Thus, oil and gas companies invest a lot in data acquisition programs when drilling wells and try to develop new ways that provide better understanding of volumes and fluid flow inside the reservoir.

Mud gas data are continuous measurements of the different compounds of the gas released from the formation while drilling. These data are cheap and extensive as they are always recorded for safety reasons, but their use for reservoir and fluid characterization is usually limited and mainly used to qualitatively identify hydrocarbon bearing zones, migration pathways and sealing formations. Only recently [16] [29], mud gas data has started to be used quantitatively.

The limited quantitative use of mud gas data is probably due to the lack of a straightforward relationship between these data and reservoir properties contrary to logging data. In fact, logging data provide some direct measurements while mud gas data could be affected by multiple parameters making any inference difficult.

However, we think that mud gas data holds a wealth of information [1] and that advances in big data and data analytics would help us to unlock more value from the rich and abundant mud gas data. Thus, combining mud gas data with logging data could improve our prediction of multiple reservoir and fluid properties allowing better-informed decisions.

1.1 Motivation

The motivation for using mud gas data for reservoir and fluid characterization is that these data are extensive and low cost compared to petrophysical measurements in the borehole. Thus, if these data could improve our understanding of the petroleum system, it would be impressively beneficial in both exploration and development stages. In fact, this could allow for better development strategy, identification of missed net pay and thus new exploration and development opportunities, decreasing data acquisition costs…

1.2 Objectives

In this master project, we will try to explore mud gas data and use it to predict reservoir and fluid properties. We will use data analytics to help screen the main features for the prediction and try to explain the physics behind it. We will focus specially on three predictive tasks:

1. Identifying hydrocarbon bearing zones in the reservoir rock
2. Identifying fluid type (Gas or Oil) in hydrocarbon bearing zones
3. Estimating permeability in hydrocarbon bearing zones

In the first chapter, we will present the technical background. We will start by presenting the study area and the project dataset. Then, we will explain what a petroleum system is and how it relates to mud gas data. Next, we will explain the mud gas data, how it is acquired and why, its limitations and interpretation. Then, we will describe briefly the logging and coring data and how it is used for porosity, saturation estimation and permeability prediction. At the end, we will try to investigate the relationships between mud gas and logging data in relation to reservoir and fluid properties, we will show the latest literature updates on this field, and stipulate our research questions.

The second chapter will be about the theoretical background behind the different machine learning approaches and the data analytics techniques used during this study.

The third chapter will deal with the methodology that we adopted for the three prediction tasks. We will start by explaining the data extraction, cleaning and features generation method. Then, we will investigate the data through correlations, model result assessments, principal component analysis for both seal and reservoir. We will start to assess some feature importance and how it relates to real physical relations established in the previous chapter. Next, we will show the implementation of the different machine learning techniques, the features used, the optimization of the hyperparameters and the best models selection. Finally, we will show how the results were assessed and the implementation of the new LOWOCV cross validation method.

The fourth chapter will focus on the results from the predictive tasks. For each task, we will show the prediction results and the models ranking. Then, we will explain the inference obtained from the main important features. For permeability prediction, we will also compare the results with the conventional approach. Finally, we will provide some explanation of the results, the limitation of our procedure and the way forward.

Chapter 2

Technical Background

2.1 Regional Setting

The study area of interest (AOI) is part of the Norwegian sector of the northern North sea (NNS)2.1, specially the prolific area of quadrant 35, north of the Troll field, which includes major fields and discoveries as Gjøa , Duva, Nova, Vega, Byrding and Fram (Figure 2.2).40 released wells in this area containing mud gas data were used in this study (Table 2.1)

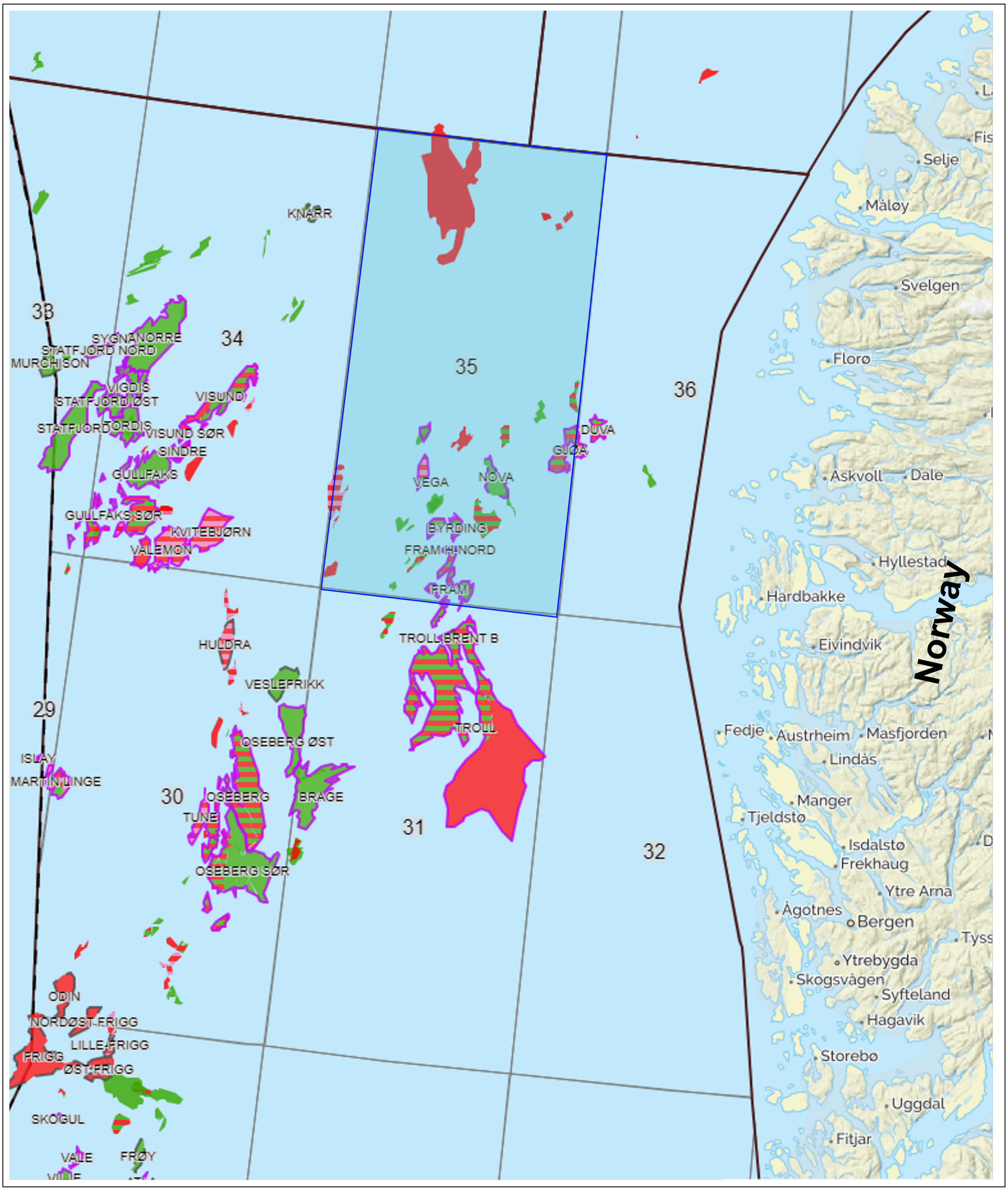


Figure 2.1: Regional map showing the Area Of Interest (AOI)

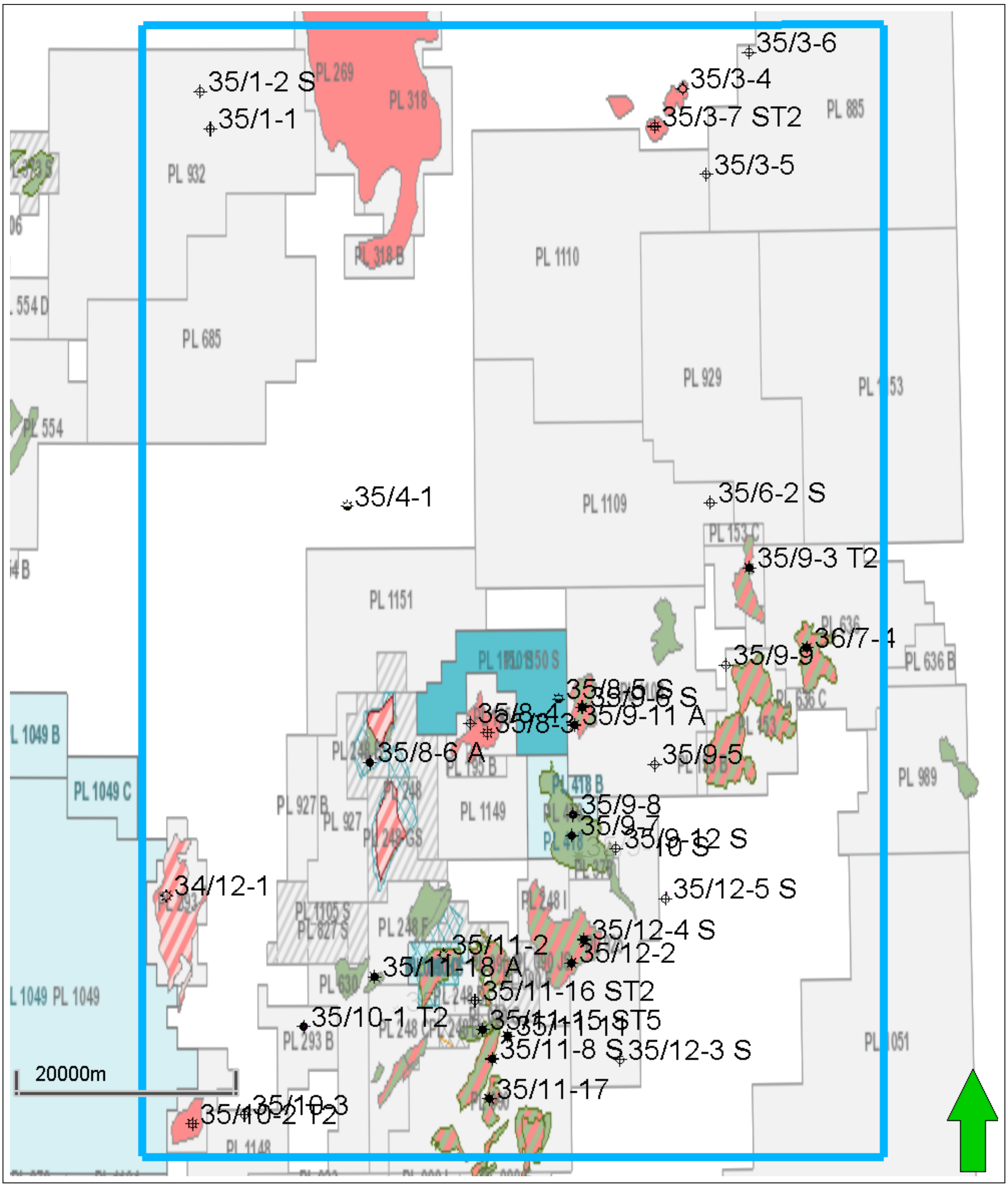


Figure 2.2: More focused map showing the wells used in this study

2.2 Petroleum system

For an oil and gas accumulation to exist, five factors need to be present.

- Presence of a source rock
- Presence of reservoir rock
- Presence of a sealing rock
- Hydrocarbon migration and timing
- Presence of a trap

In this section, we will define the five elements of the petroleum system.

2.2.1 Source rock

A source rock is a rock rich in organic matter that has been buried to a depth where pressure and temperature allows for generating movable quantities of hydrocarbons. The quantity of organic matter is commonly assessed by a measure of the total organic carbon (TOC) contained in a rock. Quality is measured by determining the types of kerogen (micro molecular material) contained in the organic matter. The change in temperature of the earth's interior per unit depth change is called geothermal gradient. With increase of temperature the kerogen cracks generating oil (Catagenesis) at a first stage, then gas (Methagenesis) at a second stage (Fig 2.3). The thermal maturity is most often estimated by using vitrinite reflectance measurements and data from pyrolysis analyses. The source rock is generally characterized by high GR2.3.2.1 values. But, laboratory tests on the rock allows an assessment the quality of the source rock. In our AOI, the main source rocks are Draupne and Heather shales which are part of the Viking group.

Table 2.1: Well Database

	Well Name	**Well Results**	**Surface X**	**Surface Y**	**Latitude**	**Longitude**	**TD (TVDSS)**	**TD (MD)**	**Spud date**	**Operator**
1	34/12-1	Gas/Condensate	499935	6787698	61°13’19.8800"N	2°59’55.6198"E	4678	4704	Nov 03 2007	Eni Norge AS
2	35/1-1	Dry,	503863	6856062	61°50’9.1126"N	3°04’24.1065"E	4515	4540	May 28 2002	Phillips Petroleum Company Norway
3	35/1-2 S	Dry,	502943	6859374	61°51’56.1800"N	3°03’21.3702"E	4098	4202	Sep 12 2010	Statoil Petroleum AS
4	35/10-1 T2	Oil,	512198	6776028	61°07’2.0499"N	3°13’34.8701"E	3959	3986	Aug 01 1991	Den norske stats oljeselskap a.s
5	35/10-2 T2	Gas,	502260	6767363	61°02’22.6699"N	3°02’30.5997"E	4651	4677	Apr 16 1996	Den norske stats oljeselskap a.s
6	35/10-3	Dry,	506902	6768151	61°02’47.9483"N	3°07’40.0551"E	2224	2250	Jun 23 1999	Den norske stats oljeselskap a.s
7	35/11-11	Dry with shows	530382	6775183	61°06’31.1902"N	3°33’49.0798"E	3200	3225	Apr 16 1998	Norsk Hydro Produksjon AS
8	35/11-12	Dry with shows	520555	6777350	61°07’43.5299"N	3°22’53.6603"E	3353	3378	Apr 19 2000	Norsk Hydro Produksjon AS
9	35/11-15 S	Oil/Gas	528144	6775755	61°06’50.2801"N	3°31’19.9398"E	3186	3250	Apr 01 2007	Norsk Hydro Produksjon AS
10	35/11-15 ST5	Oil/Gas	528144	6775755	61°06’50.2801"N	3°31’19.9398"E	2961	3250	Apr 01 2007	Norsk Hydro Produksjon AS
11	35/11-16 ST2	Dry,	527464	6778351	61°08’14.3700"N	3°30’35.8702"E	3211	3554	Jan 10 2014	Statoil Petroleum AS
12	35/11-17	Oil/Gas	528745	6769599	61°03’31.1862"N	3°31’56.7464"E	2864	2889	Mar 24 2014	Statoil Petroleum AS
13	35/11-18 A	Oil/Gas	518510	6780460	61°09’24.4100"N	3°20’38.0898"E	3874	4020	Sep 27 2015	Wintershall Norge AS
14	35/11-2	Gas/Condensate	524675	6782417	61°10’26.4199"N	3°27’31.3598"E	4007	4031	Jul 20 1987	Mobil Exploration Norway INC
15	35/11-8 S	Oil/Gas	528987	6773139	61°05’25.5299"N	3°32’14.8498"E	3329	3624	Mar 03 1996	Norsk Hydro Produksjon AS
16	35/12-2	Oil/Gas	536069	6781689	61°09’59.7200"N	3°40’13.3502"E	2512	2541	May 30 2009	Wintershall Norge ASA
17	35/12-3 S	Dry,	540398	6773080	61°05’19.9996"N	3°44’56.3668"E	2729	2807	Dec 24 2010	Wintershall Norge ASA
18	35/12-4 S	Oil/Gas	537184	6783775	61°11’6.7601"N	3°41’29.3999"E	2737	3585	Apr 23 2011	Wintershall Norge ASA
19	35/12-5 S	Dry,	544455	6787415	61°13’1.6599"N	3°49’39.2202"E	3369	3570	May 10 2015	Wintershall Norge AS
20	35/3-4	Gas/Condensate	545990	6859632	61°51’54.5399"N	3°52’26.9899"E	4062	4089	Nov 30 1980	Saga Petroleum ASA
21	35/3-5	Dry,	548100	6851991	61°47’46.7099"N	3°54’44.0098"E	4095	4130	Dec 22 1981	Saga Petroleum ASA
22	35/3-6	Dry,	551916	6862883	61°53’36.8600"N	3°59’15.7897"E	3295	3321	Feb 06 2002	RWE Dea Norge AS
23	35/3-7 S	Gas,	543472	6856278	61°50’7.2400"N	3°49’31.8302"E	3946	4051	Jun 28 2009	VNG Norge (Operations) AS
24	35/3-7 ST2	Gas,	543472	6856278	61°50’7.2400"N	3°49’31.8302"E	3647	3777	Jun 28 2009	VNG Norge (Operations) AS
25	35/4-1	Dry with shows	516076	6822412	61°32’0.5499"N	3°18’8.2602"E	4898	4936	Dec 23 1996	Norsk Hydro Produksjon AS
26	35/6-2 S	Dry,	548461	6822727	61°32’0.9800"N	3°54’40.7202"E	3561	3700	Feb 11 2009	StatoilHydro Petroleum AS
27	35/8-3	Gas,	528567	6802217	61°21’5.3502"N	3°32’2.6303"E	3915	3944	Jul 06 1988	Norwegian Gulf Exploration Company AS
28	35/8-4	Dry,	527041	6803059	61°21’32.9499"N	3°30’20.4101"E	3700	3719	Jul 04 1999	BP Norway Limited U.A.
29	35/8-5 S	Dry with shows	534935	6805220	61°22’40.5000"N	3°39’13.2199"E	3803	4000	Jun 01 2003	Norsk Hydro Produksjon AS
30	35/8-6 A	Oil,	518097	6799557	61°19’41.6499"N	3°20’17.0598"E	3529	3800	Apr 22 2016	Wintershall Norge AS
31	35/9-10 S	Oil/Gas	536725	6791208	61°15’7.1001"N	3°41’3.8999"E	3108	3619	Oct 16 2013	Wintershall Norge AS
32	35/9-11 A	Oil/Gas	536315	6802911	61°21’25.4536"N	3°40’44.5727"E	3795	3860	Apr 15 2014	RWE Dea Norge AS
33	35/9-12 S	Dry,	540037	6791900	61°15’28.3128"N	3°44’46.5820"E	3423	3556	Nov 04 2014	RWE Dea Norge AS
34	35/9-3 T2	Oil/Gas	551937	6816925	61°28’51.9001"N	3°58’30.1003"E	2759	2783	Sep 23 1997	Norsk Hydro Produksjon AS
35	35/9-5	Dry,	543499	6799386	61°19’28.8722"N	3°48’45.1748"E	3511	3531	Jan 01 2010	Nexen Exploration Norge AS
36	35/9-6 S	Oil/Gas	537025	6804491	61°22’16.2708"N	3°41’33.5004"E	3664	3740	Sep 29 2010	RWE Dea Norge AS
37	35/9-7	Oil,	536101	6793049	61°16’6.8301"N	3°40’23.2802"E	2976	3006	Feb 28 2012	Wintershall Norge ASA
38	35/9-8	Oil,	536213	6794912	61°17’6.9726"N	3°40’32.1440"E	3231	3256	Feb 01 2013	Wintershall Norge AS
39	35/9-9	Dry,	549847	6808255	61°24’12.7199"N	3°56’0.4798"E	3298	3339	Oct 04 2013	GDF SUEZ E&P Norge AS
40	36/7-4	Oil/Gas	557054	6809820	61°24’59.7081"N	4°04’8.0051"E	2702	2726	Jul 18 2016	ENGIE E&P Norge AS

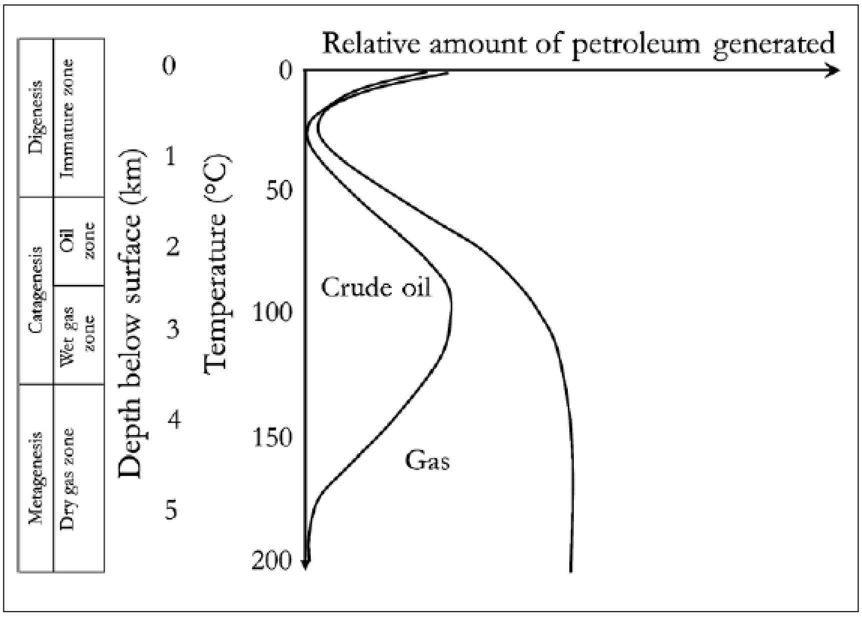


Figure 2.3: Stages of petroleum generation adapted from [Tissot and Welte, 1984; Selley, 1998]. [18]

2.2.2 Reservoir rock

A reservoir rock is a rock having sufficient porosity and permeability to store and transmit fluids [25]. In this study, we will focus on clastic sandstone reservoirs. The main reservoir properties are porosity and permeability.

The porosity of the reservoir is the amount of void space inside the rock. In fact, a rock is formed by smaller grains that get cemented together. The void between grains is the porosity. The bigger the grains that form the rock and the better sorting they have, the higher is the porosity. As the rock gets buried deeper, it gets more compacted and porosity is reduced. In addition, as temperature increases with burial, cementation (precipitation of minerals in the voids between the grains) occurs which also decreases the porosity (Fig 2.4). Porosity is determined through interpretation from log data2.3.2.2 like density, neutron porosity, or sonic logs and measured from cores.

Permeability is the measurement of fluid ability to flow inside the rock. The permeability relates to the porosity and the connectivity between pores. Presence of faults and fractures can also affect permeability. The main factors influencing permeability are the grain sorting, grain size and connection between pores surrounding the grains (Fig 2.4 C). Permeability generally is determined after lab test on core data or predicted from porosity.

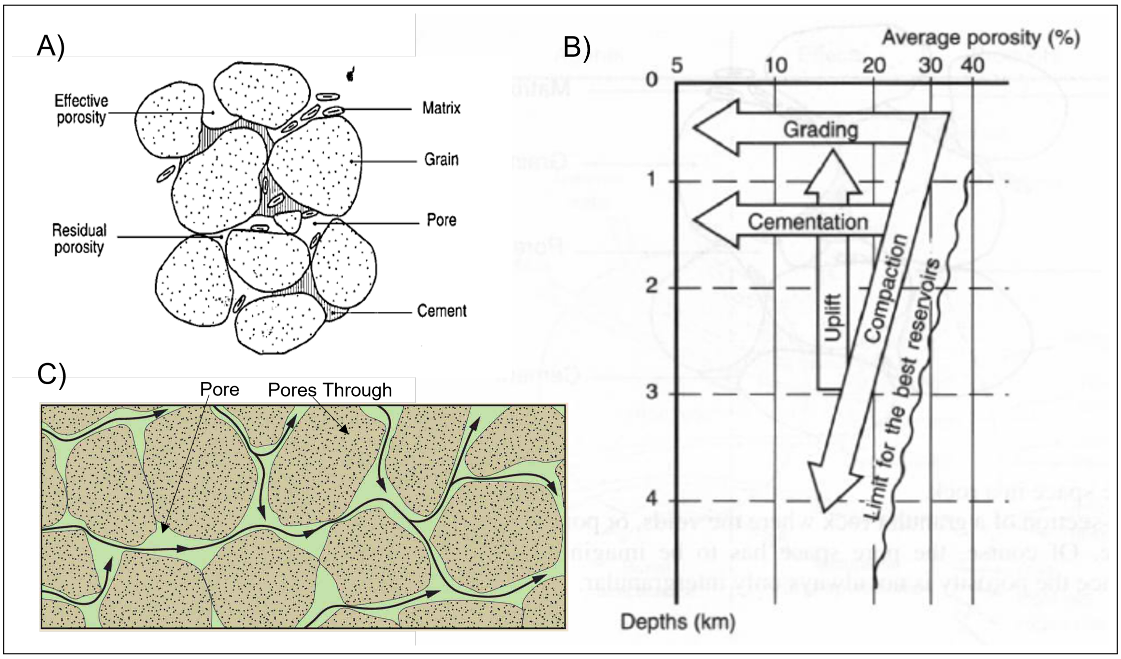


Figure 2.4: Porosity and permeability, (A) Schematic representation of granular rock showing the pore space, (B) Effect of burial on porosity and the main factors of porosity reduction [4], (C) The importance of connectivity. Connected pores (green) give rock its permeability, allowing fluid to flow (black arrows) [26]

In our AOI, the main discoveries are found in Cretaceous and Jurassic sandstone reservoirs2.5:

- Agat Formation: Agat formation is cretaceous sandstone deposited in marine environment influenced by gravity flows of sediment,
- Intra Heather Formation sandstone: The formation was deposited during upper Jurassic time in a coastal-shallow marine and deep marine environment,
- BRENT Group: The Brent group is a middle Jurassic group of formations. In our AOS, it is formed by Rannoch, Etive , Ness and Tarbert formations which could be reservoirs and are deposited in delta to shallow marine environments,
- Cook Formation: Consists mainly of shallow marine sandstone,

- Statfjord Group: A lower Jurassic group consists of formations deposited in continental environment

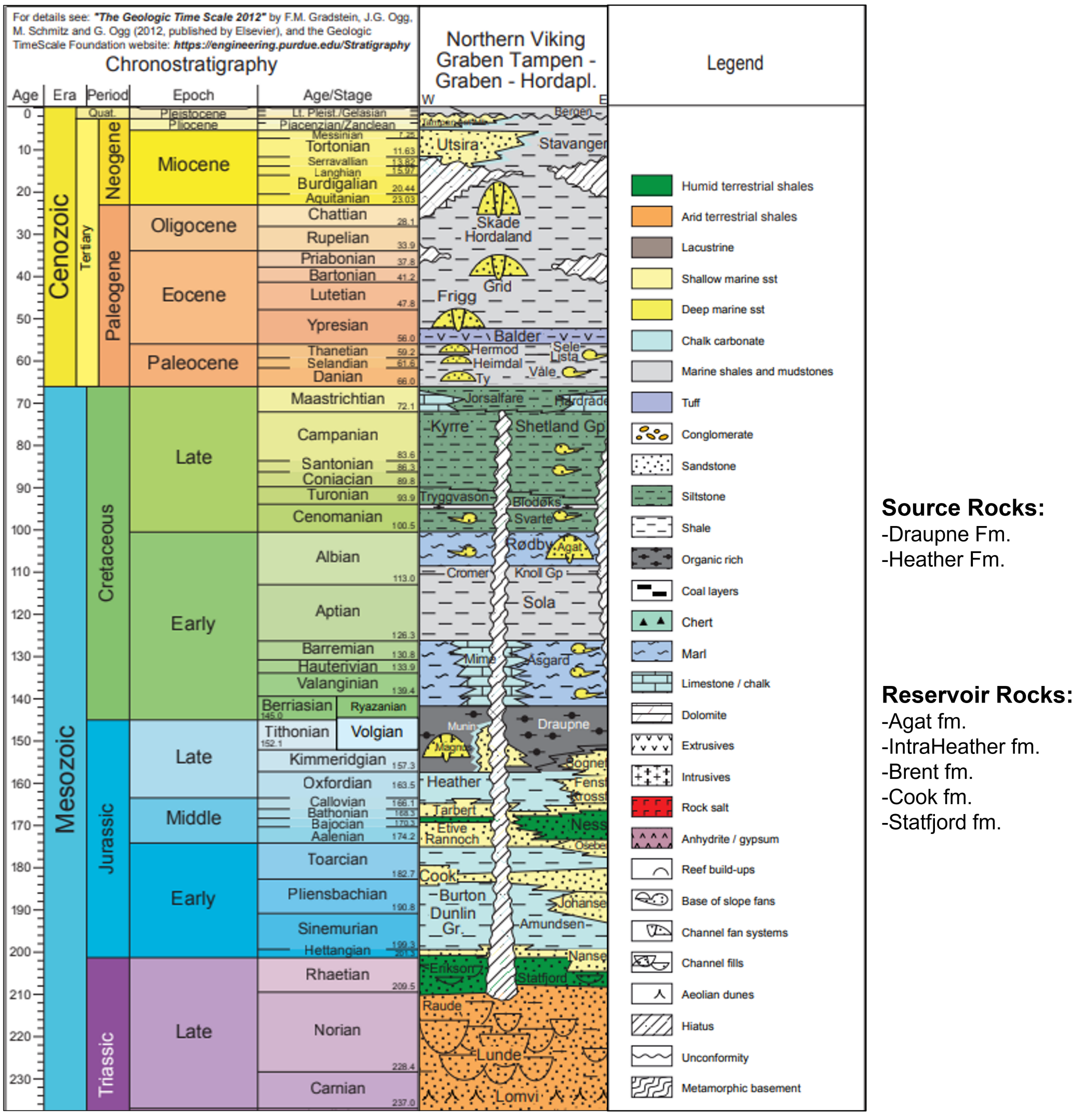


Figure 2.5: Regional lithostratigraphic chart in our AOI.[12]

2.2.3 Seal rock

The seal rock is a relatively impermeable rock. It should cap the reservoir rock to form a barrier and blocks the upward migration of hydrocarbon. The seal rock has a sealing capacity which determines how much hydrocarbon column it can hold. In fact, hydrocarbons accumulating below the seal start to generate pressure into the seal rock. While the seal is a relatively impermeable rock, after buildup of a certain pressure, which overcomes the capillary pressure, fluid could enter the seal rock and potentially start to flow very slowly (geological time) through the seal rock. Common seals include shale, clay, chalk, and evaporates.

2.2.4 Trap

The trap in a petroleum system is a geological structure formed by the combination of the seal and reservoir rock that allows accumulation and retainment of hydrocarbon in the reservoir. It could be structural (Fig 2.6) due to folding and faulting of the earth's strata, or stratigraphic (Fig 2.7) which occur due to lateral and/or vertical variations in reservoir properties.

2.2.5 Migration and timing

Migration is the movement of the hydrocarbons from the source rock to the reservoir rock. This process occurs over geological time. The primary migration is the expulsion from the source rock, it is mainly due to the overpressure generated in the pores of the source rock. Under compaction and high temperature, hydrocarbons start to be expelled to zones of lower pressure. The secondary migration is the movement of hydrocarbon through a carrier bed or through faults to a reservoir rock. Secondary migration is driven by the buoyancy forces of hydrocarbons which are generally less dense than water. Hydrocarbons will therefore displace the water and move upward to the surface unless stopped by an impermeable rock in a trap setting. It is important to notice that it is crucial that the timing of the trap formation precedes the time of migration. The figure 2.8 shows a simplified schematic of a working petroleum system.

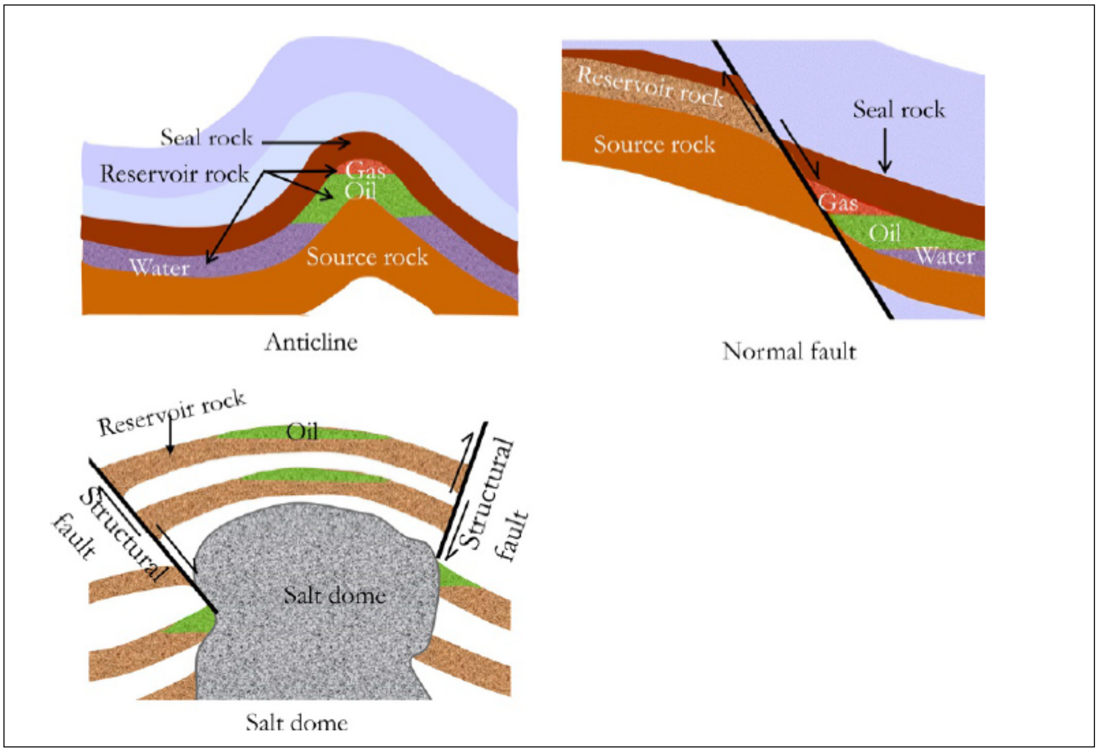


Figure 2.6: Structural traps [18].

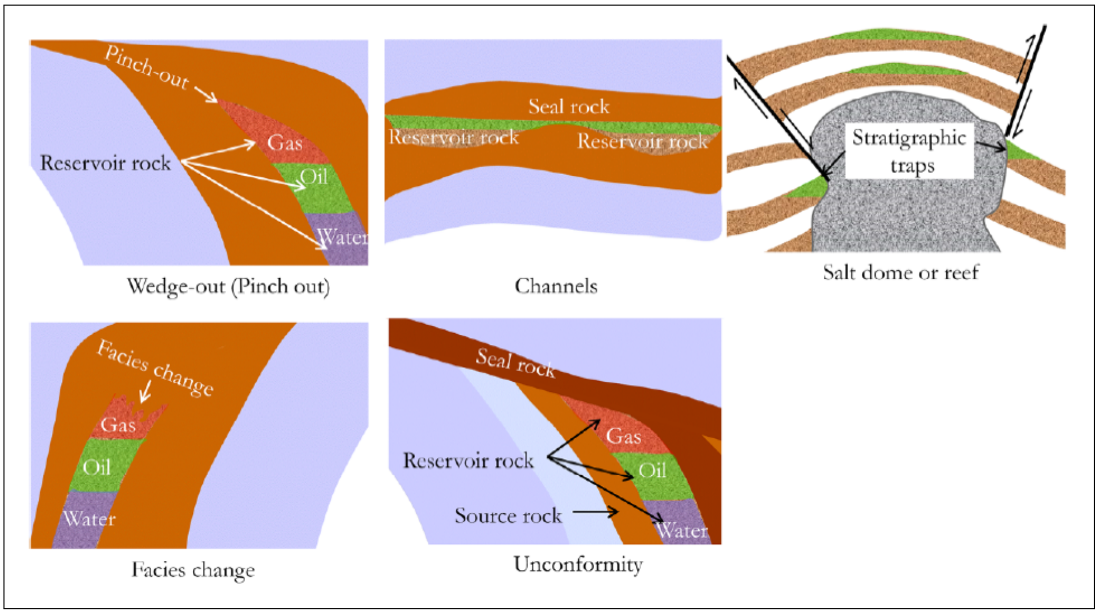


Figure 2.7: Stratigraphic traps [18]

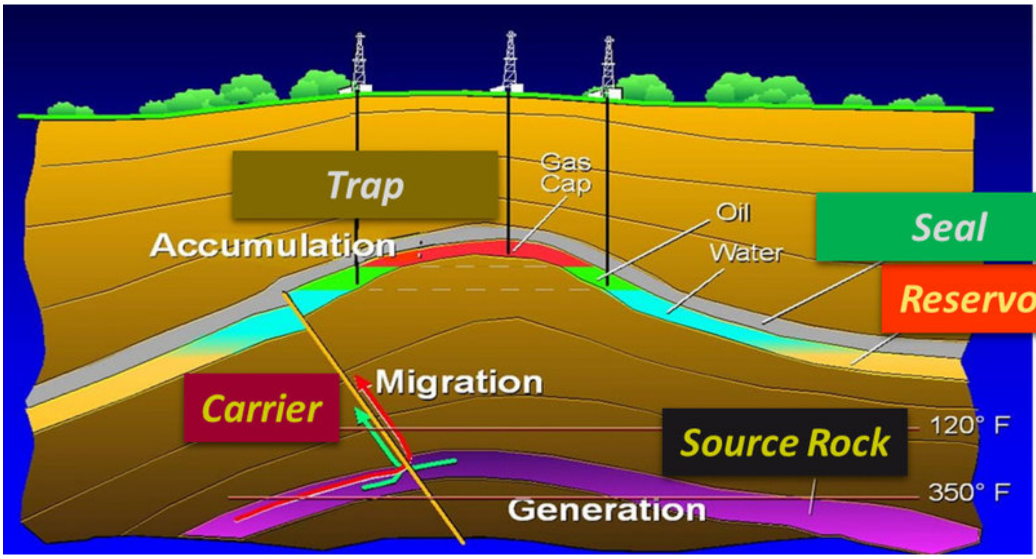


Figure 2.8: Simplified schematic of a working petroleum system [7].

2.3 Dataset: Technical background

The dataset used for this study consists of mud gas logs, well logs and core data from 40 wells in NNS. All of these data consist of continuous measurement of some properties along the well bore. These could be recorded either while drilling or after drilling by introducing some tools into the wellbore. In this section, we will provide the technical background regarding the datasets used in this study.

2.3.1 Mud gas data

Mud gas data are continuous measurements of the different compounds of gas released from the drilling mud while drilling. While drilling heavy mud is used to balance the pressure of the drilled formation. The mud used for drilling is called the drilling fluid and the primary role of drilling fluid is to ensure the safety of the drilling operations. Thus, the drilling fluid is used to avoid inflow from the formation, maintain borehole from collapsing, ensure formation stability, remove cuttings from the well, transmit hydraulic power to the drilling bit, and cool and lubricate the drill bit.

The drilling fluid is pumped from tanks at the surface (“mud pit”) by powerful pumps to the well. When the drill bit cuts through the different formations, the mud circulates from the surface to the bottom of the well then back to the surface again where it undergoes processing and treatments to maintain its properties before reaching the mud pit again where it can continue its circulation process. This circulation of the drilling fluid mobilizes the cuttings; fragments of rock that the drill bit cuts from the formation. Thus, the formation fluid in the cutting is released to the mud column as the mud travel up in the annulus. The gas trap extracts a sample of the gas in the mud. Depending on the gas trap system used for the well, the trap would have different efficiency. The sample is then accurately analyzed by the gas chromatography (Fig 2.9).

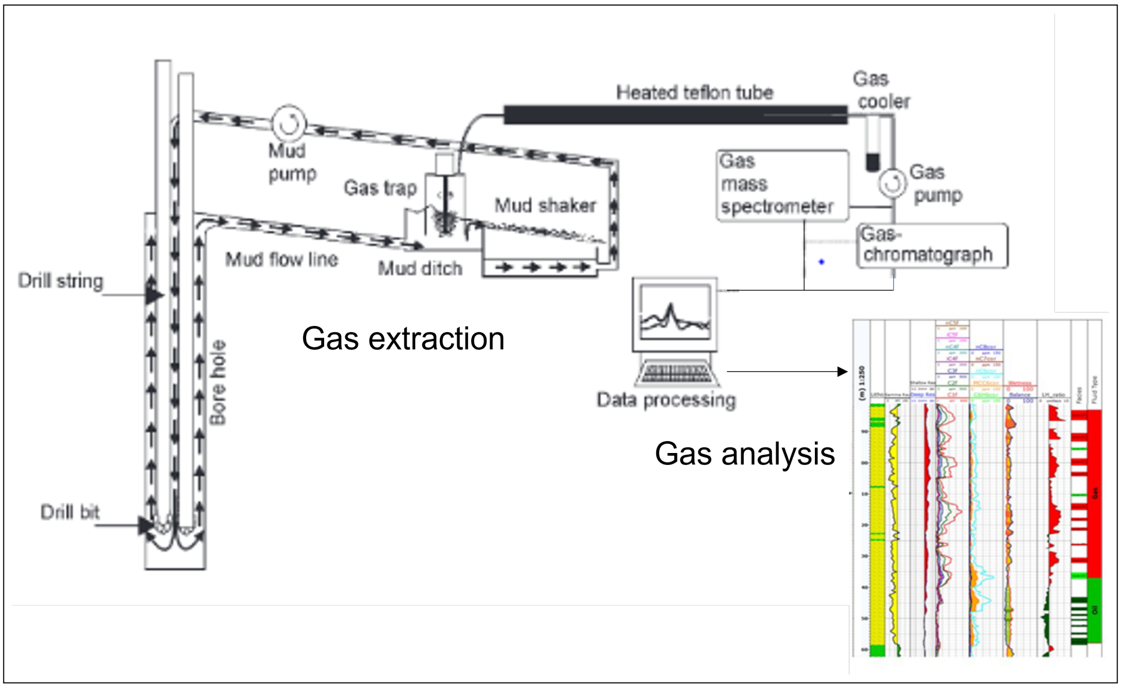


Figure 2.9: Sketch of a possible gas monitoring system [10]

The gas chromatography is a technique of separation of the gas components with the purpose of obtaining information about their molecular compositions and amounts [5]. Thus, in our case, (C1), (C2), (C3) and (C4) means the amounts of respectively methane, ethane, propane and butane in the mud gas sample. So, drilling into the hydrocarbon reservoir would generally be associated with an increase of these quantities. But an increase of these quantities does not always mean a hydrocarbon bearing reservoir. Residual gas in sand, high pressure reservoirs, gas in silty layers, source rock and fractured cemented zones can be associated with an increase of gas quantities as well. It is also important to notice that the mud gas quantities are sensitive to drilling parameters such as the rate of penetration (ROP) which is the drilling speed, weight on bit (WOB) which means the force exerted on the drill bit while drilling, drill bit diameter, etc.

It is essential to monitor the mud gas quantities to keep control of the well. In fact, a sudden increase of gas into the well bore could mean entering a high-pressure zone and urgent measures needs to be undertaken to control the pressure with an ultimate risk of losing the well and gas blow out.

The well is mainly controlled by adjusting the drilling fluid properties. In fact, the drilling fluid would have multiple properties that differs from well to well and formation to formation. These properties are decided during planning of the well and adjusted while drilling in order to ensure the most safe and cost-effective operation. Some of these properties are the mud weight, viscosity, ph, fluid loss control, solids content and shale inhibition (Bloys et al., 1994). The control of these properties is done through multiple chemical additives. In addition, the drilling fluid could be water based or oil based. All of these parameters could affect the final mud gas data and compromise repeatability of the experiment.

Part of the mud logging procedure is also to analyze cuttings data. The mud log unit takes a sample of the cuttings at regular time interval corresponding to a regular change in formation depth (ex: 5m). The cutting sample is then analyzed with a microscope and a description is made and sent to the office. The description should contain information regarding grain properties, porosity, presence of hydrocarbon, ... In this study, information from cutting were not included due to the difficulties to combine these data with the logging and mud gas data.

2.3.2 Well logging

Well logging consists of recorded information from the borehole that would allow to get useful information regarding formation, fluid properties, fluid flow, etc. It could be recorded while drilling or after drilling. The well logging procedure consists of lowering tools into the borehole that measure several physical properties of the rocks (Fig2.10). The outputs of the logging procedure are these physical measurements function of depth which are called well logs. In this study, we are interested in well logs that relate to either reservoir or fluid properties. The spliced raw measurements in function of depth are generally called composite logs. The petrophysicist interpret these composite logs to generate interpreted logs.

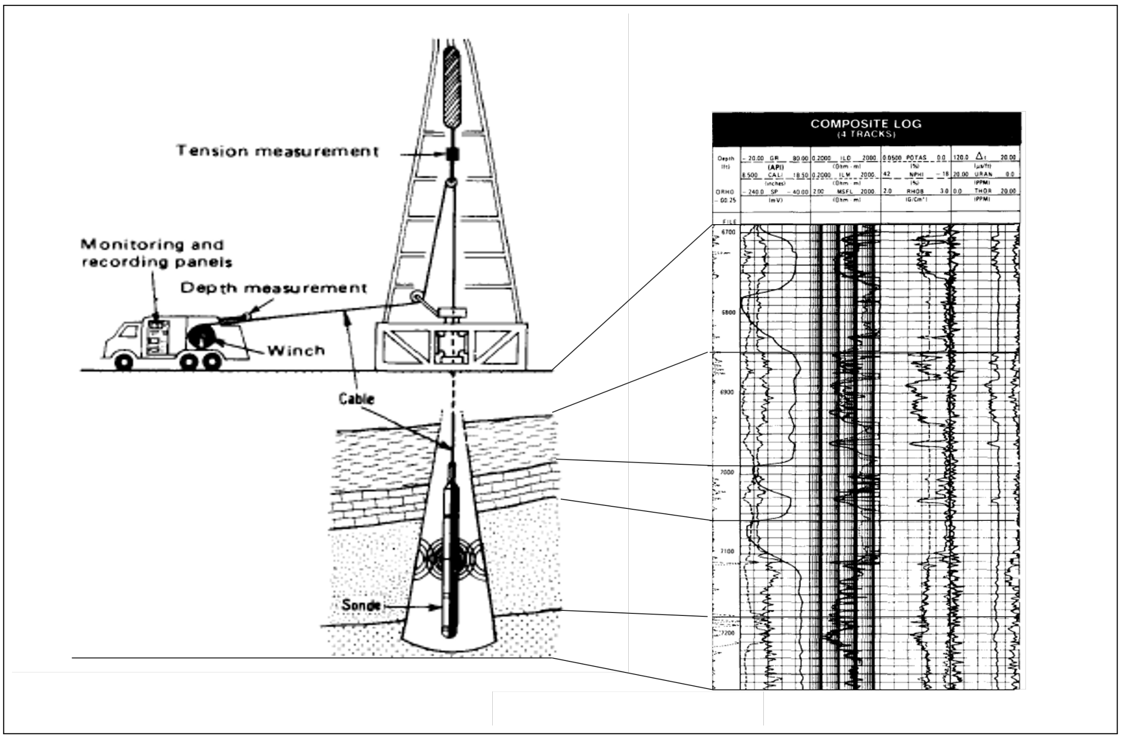


Figure 2.10: Schematic representation of onshore logging operation and resulting composite log [24]

2.3.2.1 Composite logs

Gamma Ray (GR): This tool measures the strength of the natural radioactivity present in the formation. As shales usually emit more gamma rays than other sedimentary rocks, GR tool is particularly useful in distinguishing sands from shales in siliciclastic environments and assessing the shale content in sandstones.

Caliper: This tool measures the geometry of the hole using either two or four arms[8]. It returns the diameter seen by the tool over either the major or both the major and minor axes. Caliper is used to detect borehole enlargements due to potential caving and drilling issues that could make the logging measurements unreliable or risky.

Density: Density tool emits Gamma Rays that interact with electron in the formation. The number of scattered gamma rays that reach the detector, placed at a set distance from the emitter, is related to the formation's electron density, which is related to formation density.

Neutron porosity: Neutron logs measure the hydrogen content in a formation. Neutron tool bombards the formation with neutrons from a chemical source. At collision with nuclei in the formation the neuron loses energy. After several collisions (specially against hydrogen atoms), the neutron is absorbed, and a gamma ray is emitted[19]. So, materials with large hydrogen content are the most effective on slowing down neutrons. As hydrogen and water are the fluids filling the pores, the absorbed energy can be related to porosity.

Sonic log: A sonic tool has a transmitter and receivers. The emitter emits a sonic wave that travel into formation and to the receivers. The time from the emitter to the receivers is the transit time and it is the inverse of the velocity. The transit time would depend on lithology and porosity. The denser the formation the lower is the transit time and the more porous the formation is. The sonic tool could provide measurements of the two types of sonic waves: compressional sonic (DTC) and shear sonic (DTS).

Resistivity logs: Resistivity logs are electrical well logs that record the resistivity (or inductivity) of a formation. Resistivity is usually recorded in ohm meters (Ωm). Three depths of resistivity can be logged (shallow, medium, and deep) that record the resistivity of the formation with increasing distance away from the borehole. Resistivity logs can be interpreted to infer information about the water saturation, and the presence of hydrocarbons

Formation pressure/sampling: Formation-testing tools are designed to measure the formation pressure and/or acquire formation samples at a discrete point in the formation [8]. It consists of isolating a part of the formation, either through probe or packers, then apply a pressure drawdown to allow fluid to move from the formation to the tool. By opening chambers in the tool and analyzing the fluids and pressures while the chambers are filled, it is possible to determine the true pressure of the formation (as distinct from the mud pressure). If needed, a sample will be retained for analysis at the surface. This will not only allow to assess the fluid type (oil , gas or water), but also to assess some fluid properties as gas oil ratio, viscosity, fluid density, etc.

2.3.2.2 Interpreted logs

After the logging data acquisition, the petrophysicist will interpret the composite logs to determine reservoir and fluid properties, by generating a set of interpreted log data and reservoir summation table. The main outputs of this interpretation are shale volume (Vshale), reservoir flag, porosity (PHIT, PHIE), water saturation (Sw) and fluid type.

Shale volume: First, the petrophysicist has to assess the presence of reservoir rock. We will assume that we are dealing with clastic reservoirs which are the reservoir rocks in our AOI. The most reliable indicator of reservoir rock will be from the behavior of the density/neutron logs, with the density moving to the left (lower density) and touching or crossing the neutron curve. In most cases, this will correspond also to a fall in the gamma ray (GR) log. Shales can be clearly identified as zones where the density lies to the right of the neutron. Generally, the greater the crossover between the density and neutron logs, the better the quality of the reservoir. But, it can also be a fluid effect as gas zones will exhibit a greater crossover for a given porosity than oil or water zones. The volume of shale (VShale) can be determined either by using density and neutron porosity or using GR. Using GR, Vshale could be determined using GR values in pure sand (GRsa) and GR values in pure shale (GRsh),

$$Vshale = \frac{GR - GRsa}{GRsh - GRsa}$$

A 60% Vshale could be used as cut off for sand, meaning higher values than 60% Vshale would be considered non reservoir. Also, porosity cut off could be used at a later stage.

Porosity (ϕ): Porosity could be calculated by multiple logs as sonic log and neutron porosity log. But, most often porosity is calculated from density log (Rho) log using this formula:

$$\phi_{den} = \frac{Rho_m - Rho}{Rho_m - Rho_f}$$

Rho_m is the matrix density and Rho_f is the fluid density. For sandstones, Rho_m typically lies between 2.65 and 2.67 g/cc. Where regional core data are available, Rho_m can be taken from the measures on conventional core plugs. Rho_f depends mainly on the density of the fluid filling the rock pores.

Porosity measurements from logs can be calibrated to core porosity if core material is available from the reservoir section.

Water saturation, fluid type and contacts: The main log data to identify hydrocarbons from water is the resistivity log. In fact, in reservoir section, the water is generally saline thus conductive, while the hydrocarbons are resistive. So, the presence of hydrocarbons in the reservoir would generally induce an increase of the resistivity. Generally speaking, the pores in the rock are either filled by water or hydrocarbon. Thus, the lower the water saturation, the higher is the hydrocarbon saturation. The resistivity tool allows the determination of water saturation through Archie equation[2]. Resistivity, however, does not allow distinguishing oil from gas. If the sandstone reservoir is clean and thick, the cross-over between density and neutron porosity (as discussed above) could allow to separate gas and oil. In case pressure points and fluid samples were acquired, they are extremely important to identify the fluid in the reservoir. Plotting pressure points as a function of depth, allows to identify pressure gradients which allow to identify the fluid and determine precisely the fluid contacts, which is the interface between the two-fluid fill (Fig 2.11).

In this master project, we will try to use mud gas data to assess fluid type while drilling. This would be extremely important to have an early assessment of the fluid in place and in the absence of fluid sample points.

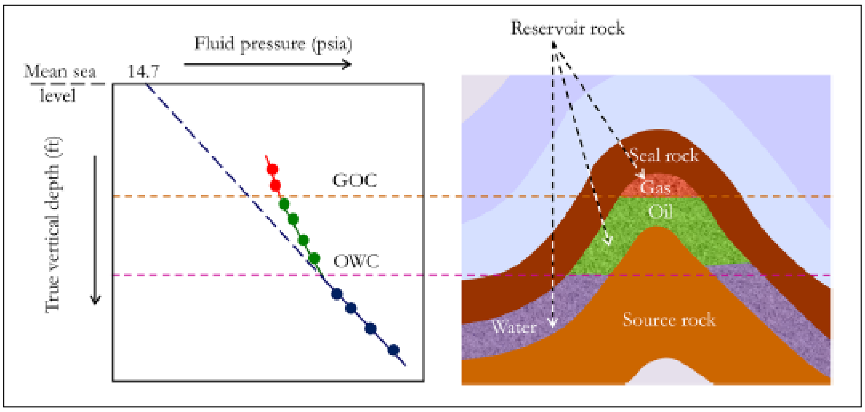


Figure 2.11: Reservoir fluid zones in a normally pressurized petroleum reservoir. [18]

2.3.3 Coring

2.3.3.1 Coring Procedures

Coring is the procedure of cutting a part of the formation with a special drill bit, allowing to move it to the surface to do more studies directly on the formation rock.

Particularly during the exploration phase of a field, coring presents an important means to calibrate the petrophysical model and gain additional information about the reservoir not obtainable by logs. Usually the decision of when and where to core will be made in conjunction with the geologist and operations department, considering the costs and data requirements. So, not all the wells have core data and generally it is acquired in a limited part of the reservoir zone.

It is considered essential to at least attempt to core a part of the main reservoir formation during the exploration and appraisal phases of drilling[8]. Coring is obligatory in Norway for new exploration wells when reservoir containing movable hydrocarbons are present.

2.3.3.2 Core data analysis

First, a core description is made by assessing visually its sedimentary features, grain sizes, colors, etc to infer information about depositional environment, fluid content, type of cementation, vertical heterogeneity, faults and fractures, etc.

Then, in routine core analysis (RCA) or special core analysis (SCAL), core plugs are cut to run multiple tests to assess different properties as porosity, permeability, formation salinity, saturation height, capillary pressure, etc. The main difference between RCA and SCAL, is that in RCA the plugs are at ambient conditions, while in SCAL, the plugs are at reservoir conditions[17].

Core Porosity: Porosity can be determined from both RCA and SCAL with different methods: destructive (the plug can't be used after) or non destructive methods. The most known method is the helium porosimeter test. Helium is used as the gas because it is an ideal gas at ambient conditions is inert and has a very small molecular size so that it can penetrate all the accessible pores in a rock. This test is a non destructive test and allows the determination of porosity and grain density. The empty container of the core plug would have a volume Vs. By introducing the plug, the volume become Vs-Vg, with Vg is the grain volume. A reference container filled with helium having initially a volume Vr and a pressure Pr, will be connected to the plug container. Thus, the pressure will expand. Using Boyles Law (P1*V1=P2*V2), the grain volume (Vg) can be determined by solving Pr*Vr=P*(Vr+Vs-Vg) [17].

The bulk volume of the plug (Vb) can be determined by mercury immersion. As the pores are too small for mercury to penetrate in. The volume increase by inducing the core plug in mercury would be equal to Vb. Then, porosity (Phi) could be determined by Phi*Vb=(Vb-Vg).

Core Permeability: There are three principle definitions of permeability in core analysis[17]:

- Absolute permeability: The permeability to a single phase where only that phase fills the pore space (e.g. air, oil or brine).
- Effective permeability: The permeability to one phase when two or three phases are present in the pore space. For example, oil permeability (Ko) and water permeability (Kw) at some specific water saturation.
- Relative permeability: The ratio of effective permeability to a base permeability which for SCAL laboratories is normally an endpoint effective permeability (Ko or Kg at Swir(irreducible water saturation)).

Permeability can also be split into horizontal and vertical permeability. Horizontal permeability is the ability to flow in the direction of the geological layer, the vertical permeability is the perpendicular direction to the geological layer.

The core permeability data used in this study is the absolute ambient horizontal permeability. So, we will not discuss the other type of permeability in this section.

The permeability test is based on the Darcy law which is a law relating pressure difference (P1-P2) and rate (Q) governing the fluid flow inside any medium. Q = k*A (P1-P2)/(mu*L). k, mu, L and A are respectively the permeability, the viscosity, length and area of the core plug (Fig 2.12).

The unit of permeability is the Darcy. But, generally due to the relative low permeability in the rocks, permeability is reported in mDarcy.

Steady state permeability test consists of injecting nitrogen gas into the core plugs. The core plugs should be isolated and a confining stress applied to avoid the nitrogen escaping. Then, the injection pressure, Pi, and flow rate, Qi, are recorded. The time to reach steady state in each measurement varies from only a few seconds to a few minutes, depending upon permeability. Once stable conditions are attained, the data are recorded either manually or automatically via a computer, and the probe is released and moved to the next location. Knowing the rate, pressures, length and area of the probe we can determine the permeability using Darcy's law (Fig 2.12).

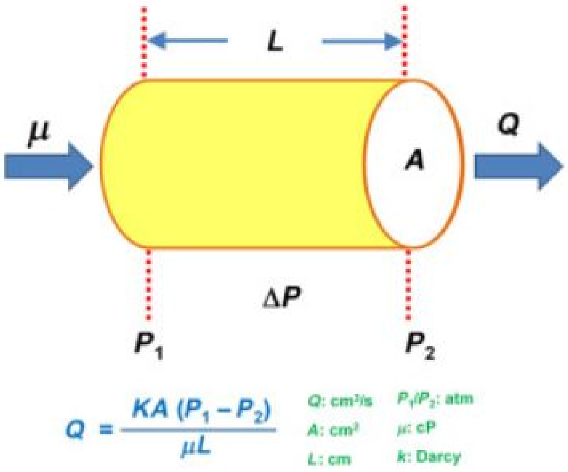


Figure 2.12: Diagram explaining parameters in the Darcy equation for incompressible liquid flow through a core plug allowing to determine the permeability. [17]

Porosity Permeability relation: Plotting core porosity vs. logarithm of permeability could show generally a linear relation. However, this linear relation generally changes with rock type. It means that it depends on cementation and pore structure of the rock. This linear relation is called poroperm relationship (Fig 2.13), and allows predicting the permeability in wells where we do not have core data, or predicting the permeability in new wells.

In addition, using this relation a cut off on porosity could be determined as net reservoir cut off. Example, a cut off of 10% porosity means that rocks having less than 10% porosity are assumed to not contribute to the hydrocarbon flow.

In this master project, we will try to use mud gas data in addition to the conventional method of using porosity and Vshale in order to assess if we can have better prediction of permeability.

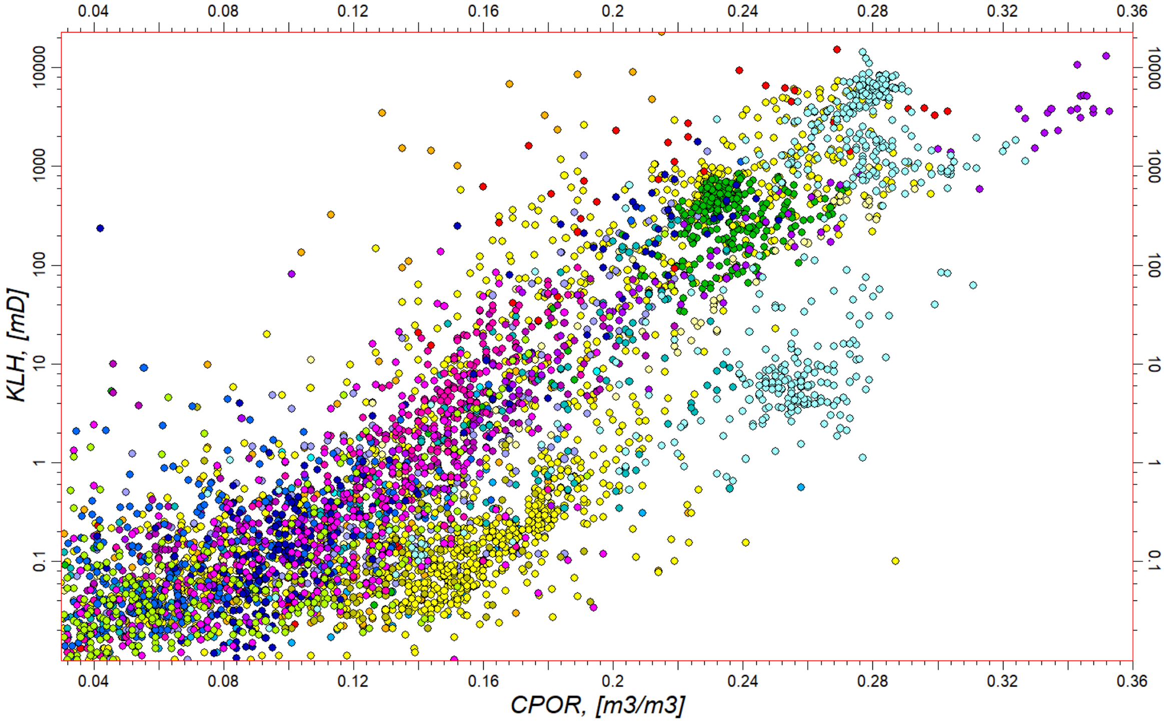


Figure 2.13: Example of poroperm relationship from AOI wells showing logarithmic horizontal permeability change with core porosity where each color correspond to a well

2.4 Causality analysis, literature and limitations

2.4.1 Causality analysis

Figure2.14 shows a simplified schematic of the processes involved while drilling.

First, the mud gas quantities, as they are a measure of gas quantities released from the cuttings of a particular formation in the mud, they should be related to the fluid composition in the same formation.

Thus, water filled reservoir rocks should not be associated to an increase of mud gas readings, compared to oil or gas reservoirs. So, mud gas quantities could be useful to separate water zones from hydrocarbon zones. However, very often, presence of residual gas due to a leaking trap, migration pathway, etc is associated with an increase of mud gas quantities.

Thus, our first prediction task in this study will be Water/ Hydrocarbon classification using mud gas data.

As oil contains heavier alkane compounds than gas, it is expected to have a higher ratio of heavier components in mud gas coming from oil zones than gas zones. To assess this relation different mud gas ratio will be used as features: (C1/C2), (C1/TG) with TG: total gas, (C1/(C2+C3)), (C3*100/(C1+C2)), wetness (Wh) and balance (Bh). Moreover, mud gas ratios would be less affected by drilling parameters and gas background trend, thus could be more reliable quantities.
Thus, our second prediction task in this study will be Oil/ Gas classification using mud gas data.

It could also be intuitive to say that the higher the porosity and permeability the higher the mud gas quantities. In fact, the higher the porosity the higher the amount of hydrocarbon per unit of volume and thus higher the quantity of gas liberated. Also, the higher the permeability the easier for the gas to leave the formation and the cuttings. In fact, due to low permeability the gas could be trapped inside the pores and do not escape to the gas trap.
Thus, our third prediction task in this study will be permeability prediction using mud gas and logging data.

Seal rocks are relatively impermeable and mud gas quantities would be small. However, the presence of migration route, overpressure, presence of hydrocarbon column below, could still fill the pores of a seal rock with gas. While drilling and breaking these rocks gas would be released to the well which increases the mud gas quantities.
Source rock would have high mud gas quantities due to the presence of hydrocarbon trapped in the pores.
Mud gas data effect seal and source rock will be assessed through clustering and PCA in the data analysis part.

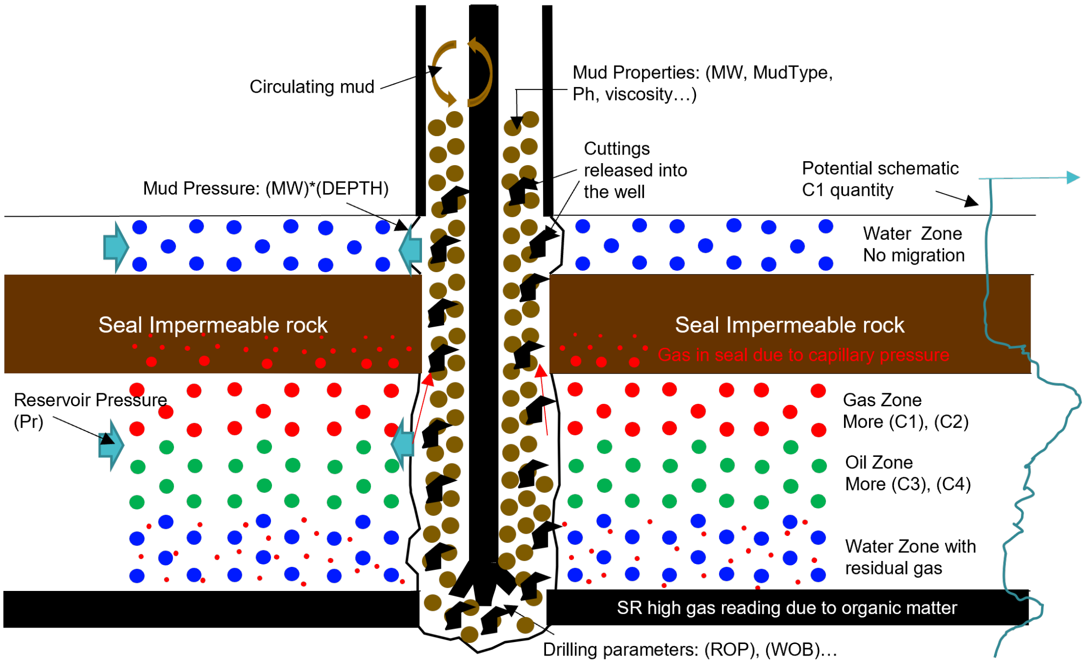

Circulating mud
Mud Properties: (MW, MudType, Ph, viscosity…)
Cuttings released into the well
Mud Pressure: (MW)*(DEPTH)
Potential schematic C1 quantity
Water Zone No migration
Seal Impermeable rock
Seal Impermeable rock
Gas in seal due to capillary pressure
Reservoir Pressure (Pr)
Gas Zone More (C1), (C2)
Oil Zone More (C3), (C4)
Water Zone with residual gas
SR high gas reading due to organic matter
Drilling parameters: (ROP), (WOB)…

Figure 2.14: A simplified schematic of the processes involved while drilling.

2.4.2 Literature

The presence of a relation between the ratio of C1 over heavier gas components and fluid type was established by Pixler on 1969[22]. These relations were used mainly qualitatively to assess hydrocarbon bearing zones, migration pathways and sealing formations Later on, using the wetness parameter, Haworth et al 1985[13] established the wetness (Wh), balance (Bh) and character (Ch) ratios.

$$
\begin{aligned}
Wh &= \frac{C2+C3+C4+C5}{C1+C2+C3+C4+C5} * 100\\
Bh &= \frac{C1+C2}{C3+C4+C5}\\
Ch &= \frac{C4+C5}{C3}
\end{aligned}
$$

Using the (Wh) (Haworth et al 1985) established the following fluid characters:

- Wh $<$ 0.5 very dry gas;
- 0.5 $<$ Wh $<$17.5 gas;
- 17.5 $<$ Wh $<$40 oil;
- Wh $>$ 40 residual oi

The more Wh increases the more dense the fluid. In fact, Wh represents the percentage of C2+ in the mud gas. These relations were mainly used qualitatively as the numbers changes from area to area.

Only recently, mud gas data was used quantitatively and it was to determine fluid properties. (Liqiang et Al, 2014)[16] generated a mathematical formula to calcualte GOR in Yingdong Oil/Gas Field in the Qaidam Basin. This was done using star chart of gas components[16], the coefficients on the formula were determined by mathematical fitting of the data. Next, (Tao Yang et Al, 2014)[30] showed that all PVT inputs can be derived reasonably from only surface gas composition in a shale gas reservoir. Finally, (Tao Yang et Al, 2020)[31], used machine learning methods to quantitatively predict GOR from advanced mud gas data in conventional reservoir.

For permeability, based on fields on US, Pixler[22] established qualitatively a potential relation between these gas ratios and the formation permeability assessing if it is tight or productive. It was also established that in permeable zone the amount of lighter gas increases compared to the heavier fraction [Giovanni N. Pinna 2008][11]. Thus a low K'=(C1+C2)/(C4+C5), would mean lower permeability.

In this study, we will investigate these relations by using data analytics to validate or discard these inferences in our area of study and we will investigate more features that could be important for the predictions. Thus, we will try to estimate quantitatively the reservoir permeability.

2.4.3 Limitations

The main uncertainty on using mud gas for fluid and reservoir characterization is that the amount of extracted gas from the mud would depend on multiple factors and not only the fluid properties on the rock:

- Gas trap and extraction equipment: While the gas trap technology had shown a lot of progress through the years. In our area of study traditional gas extractor with different efficiency were used which would be a bias in our data. Moreover, efficiency of gas traps can vary between 20% and 50% depending upon design, location and mud properties and maintenance. Efficiency will also depend on the composition of gas present, distribution of gas in the mud, viscosity, gel strength of the mud, and the mud flow rate.In addition, traditional gas trap would underestimate heavy alkane compounds (C3+) amounts.

- Drilling parameters: For repeatability and coherent results, it is needed to have a constant mud flow. However, change on drilling parameters while drilling could make difficult maintaining a consistent mud flow specially on old gas trap[23].
- Mud types and properties: Oil-based mud would have much greater interaction with the fluid in the reservoir than water-based mud.
- Temperature: Temperature affects solubility

2.5 Summary and objectives

In this chapter, first, we defined the petroleum system. Then, we provided the technical background of the mud gas data and the multiple logging data that we acquire to characterize reservoir and fluid properties and that will be used in this study. We have presented the current approaches to assess permeability which require to cut core through the formation and generate a poroperm relationship. We have explained that the main way to assess reservoir fluid type is to acquire discrete pressure or sample data. Next, we tried to identify how mud gas data could help to predict permeability and fluid through causality assessment and using most recent literature studies. Finally, we presented some potential limitations of mud gas data that could affect our machine learning models.

In this study, we will explore through data analytics and machine learning techniques the mud gas data. Despite its limitation, we will assess if we can extract from conventional mud gas data insights on reservoir and fluid properties through three predictive:

1. Identifying hydrocarbon bearing zones in the reservoir rock
2. Identifying fluid type (Gas or Oil) in hydrocarbon bearing zones
3. Estimating permeability in hydrocarbon bearing zones

Chapter 3

Theoretical Background: Machine Learning

3.1 Introduction

In the last chapter, we discussed the technical background behind our data set acquisition and the limits of mud gas data. In this chapter, we will discuss briefly the theory of data mining and the machine learning methods used in this project.

3.2 Data and data mining

A dataset consists of multiple objects (records) characterized by a collection of attributes. These attributes could be quantitative or qualitative. Data mining is the process of automatically discovering useful information and patterns from dataset. Data mining techniques also provide capabilities to predict the outcome of a future observation. Recent progresses on data mining allowed to develop algorithms to use data that were before unusable due to its size or complexity. In fact, data mining integrated disciplines from statistics and machine learning, to solve the challenges from big, complex and high dimensions datasets[27].

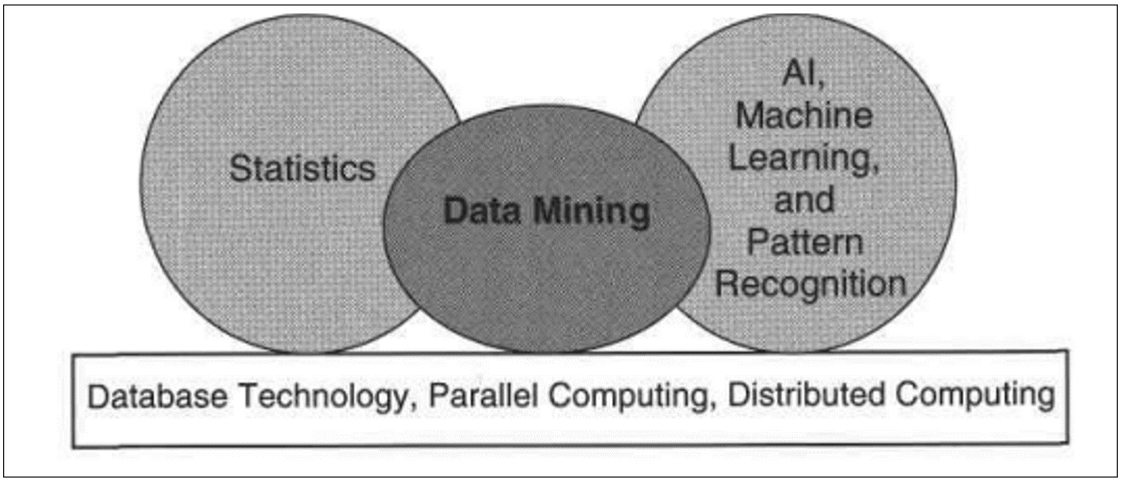

Statistics
Data Mining
AI,
Machine
Learning,
and
Pattern
Recognition
Database Technology, Parallel Computing, Distributed Computing

Figure 3.1: Data mining as a confluence of many disciplines[27].

3.2.1 Prediction vs Inference

Suppose that we observe a quantitative response Y and p different predictors, X1, X2,...,Xp. We assume that there are some relationships between Y and X = (X1, X2,...,Xp), which can be written in the very general form Y = f(X)+ ε . ε is a random error term. f is some fixed but unknown function. Our main goal in machine learning is to estimate the function f for two purposes prediction and inference. We will refer to this function f through the chapter.

3.2.1.1 Prediction

In order to predict f, we want to find a function $\hat{f}$ as $\hat{Y} = \hat{f}(X)$, that minimizes the error between predictions and real observations E($\hat{Y}$-Y). This estimate of error is formed by two terms a reducible error that we want to minimize and an irreducible error that is due to the features, even in an ideal condition, won't be able in to perfectly explain the response. Thus, the irreducible error will always provide an upper bound on the accuracy of our prediction for Y[14]. Machine learning algorithms are used to learn from X to estimate Y using $\hat{f}$.

3.2.1.2 Inference

If our purpose is only prediction, we are interested only on $\hat{Y}$ and knowing the $\hat{f}$ is not important. However, if our goal is also inference, it is needed to know $\hat{f}$. This would allow to understand which predictors are associated with the response, the type of relationship between the predictor and the response (positive or negative and at what extent) and is it linear or more complicated[14].

3.2.1.3 Trade-off between inference and prediction

Later we will explain the different machine learning methods used in this study. But, first it is important to explain the trade-off between inference and prediction. In fact, there are different machine learning methods to estimate the function f. These algorithms would have different degree of flexibility. A linear model for example would be inflexible approach because it constrains the prediction to a linear relation to the features. However, being so simple, it is very efficient for interpretability as we can assess quantitatively how a positive or a negative change of the feature value would change the response. In the other end, neural network with multiple layers are very flexible models as it allows for non-linearity and can fit any form of functions. But it makes any inference very difficult. Figure 2.7 provides an illustration of the trade-off between flexibility and interpretability for the methods that we used in this study.

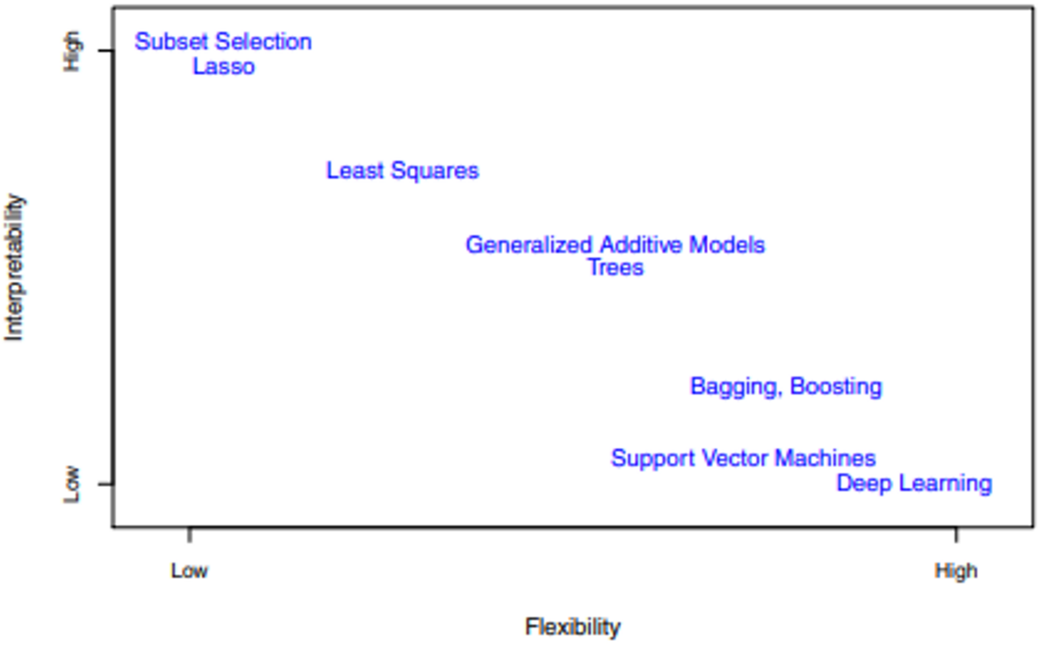


Figure 3.2: A representation of the tradeoff between flexibility and interpretability for the methods used in this study showing that generally if the method flexibility increases, its interpretability decreases[14]

3.2.2 Classification vs Regression

Data attributes could be measurable (quantitative) or categorical (qualitative). When the response variable is quantitative, we use regression methods, while if the response variable is qualitative, we use classification methods. In some cases, we can transform quantitative variables to categorical and use classification methods (for example using threshold and apply a binary classification). Similarly, in some instances, we can transform a categorical variable to quantitative and use regression methods, for instance, if the categorical variable describe order (ex: 1-low, 2-medium and 3-high).

3.2.3 Data preprocessing

Data preprocessing consists in transforming the raw input data into an appropriate format for further analysis. This includes merging data from multiple sources, cleaning data to remove noise and duplicate observations, selecting records and features that are relevant to the data mining task at hand, resampling if needed, assessing and handling of missing values, data transformation as applying logarithmic or exponential,etc.

3.2.4 Data analysis

3.2.4.1 Statistics and visualization

Part of the data analysis is to look to features statistics. Generally, we can assess statistics for the whole range of data or select subset of the data. First, we can start by assessing individual statistics for each variable as max, min, mean, median, standard deviation (which shows the spread of the data) and percentiles. Also, at this step data histograms can be plotted and skewness of the data assessed. For categorical variables we can assess frequencies and mode.
The variables generally are not independent. To assess the dependencies between variables we use multivariate statistics. We can assess covariance and correlation between the different variables. Good correlation between features and response could show that there is a relation between the response and the variable. However, high correlations between features, could show redundant information and could harm model stability and interpretability. For visualization we can cross-plot each pair of features, we can use scatter plots for three variables using colors or shape for the third variables, and for categorical variable we can use boxplots.

3.2.4.2 Model summary

At this step, linear models could be generated to assess features statistical importance by assessing F-statistic and to examine the associated p-value. If we conclude based on the p-value that at least one of the predictors is related to the response, we could look at the individual p-values to assess which ones (will be discussed further in linear model section).

In addition, assessment of the variance inflation factor (VIF) allows to detect multicollinearity problem. In fact, while correlation allows to identify collinearity between two features, (VIF) allow to assess collinearity between three and more features even if no pair of features has a particularly high correlation. The VIF is defined as the ratio of the variance of $\hat{\beta}_j$ when fitting the full model divided by the variance of $\hat{\beta}_j$ if fit on its own[14]. The smallest possible value for VIF is 1, which indicates the complete absence of collinearity. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity, specially for linear models.

3.2.4.3 PCA for data analysis

As discussed earlier, generally the variables would have some correlations which would correspond to a redundant information in the data. This is especially true in high dimensionality data. In fact, the more features we use, sparsity of the data increases and the model becomes more prone to over fitting as adding features that are not associated to the response would add more noise. This concept is called the curse of dimensionality. In addition, the more features there are the more difficult to visualize and explain the results. The goal of the Principal Component Analysis (PCA) is to reduce the dimensions by finding a new set of attributes that better captures the variability of the data removing redundant information. In PCA, this is done by projecting the features space into a new space. Thus, the first dimension is chosen to capture as much of the variability as possible. The second dimension is orthogonal to the first and captures as much of the remaining variability as possible, and so on. The least dimension, would have the least variability. Then, dimensionality reduction could be done by selecting a few principal components that explain most of the information of the data.

PCA could be used for data analysis. In fact, reducing the number of dimensions allows for better visualization of the main variability of the data. Thus, displaying pairwise both the principal component scores (projection of the data on the principal components) and the principal component loadings (projection of the original features on the principal components) of the first principal components would already provide an understanding of the information present in the data by identifying the strongest patterns. This figure is known as a biplot.

PCA could be also used for clustering and regression. The risk using PCA for prediction, is that it is not certain that the response variable would be associated to the chosen main principal components.

3.3 Machine Learning techniques

3.3.1 Supervised vs Unsupervised learning

Unsupervised learning or clustering consists of automatic classification of the data based on a measure of certain distance between the data points. The data points that have the smallest distance between them are clustered together. No response variable is used. The unsupervised learning could show some patterns in the data. In this study, we will focus on supervised learning. In supervised learning setting, for each data point ‘i’ with a set of features ‘Xi’ there is a response ‘Yi’ associated to it. The goal in supervised learning is to train a model to be able for a new data point ’j’ with feature vector ’Xj’ to predict the associated response ’Yj’. The machine learning techniques that will be presented below corresponds to techniques used in supervised leaning setting and that we had used during this study.

3.3.2 Linear models

3.3.2.1 Linear regression

In the linear regression model, our estimate function f̂ is restricted to be linear. Thus, $\hat{\beta}_j\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + + \hat{\beta}_p x_p$. The parameters are estimated using the least squares approach. Thus, we choose β_0, β_1,..., β_p to minimize the sum of squared residuals (RSS):

$$
\begin{aligned}
RSS &= \sum_{i=1}^{n} (y_i^2 - \hat{y_i}^2) \\
&= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2
\end{aligned}
$$

The values $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ $\cdots$ $\hat{\beta}_p$ that minimize the RSS are the multiple least squares regression coefficient estimates. Linear regression models allow for high degree of interpretability. In fact, we interpret these results as follow: if all the features values are fixed, a unit increase of feature i value leads to an increase of $\hat{\beta}_i$ in the response. This is extremely important as it allows to understand the relation between the response and the features. In fact, if all the $\hat{\beta}_i$ are equal to 0, It means that the response has no relation with the predictors. This hypothesis is called the null hypothesis and is assessed by computing the F-statistics. Similarly, for each individual predictor a t-statistic and a p-value are computed. These provide information about whether each individual predictor is related to the response, after adjusting for the other predictors. Interaction between two predictors could also be evaluated by multiplying the two predictors in the linear model. We say there is an interaction effect, when the increase of the response value, by an increase of feature i, depends on feature j.

These relations between features and response could be highly more complicated in case of high number of features, confounding relation between variables and multicollinearity issues which affects prediction and interpretability.

3.3.2.2 Subset selection and regularization

Subset selection, regularization and dimensionality reduction techniques allow to improve linear model prediction and interpretability in case of high features number compared to the sample size. The improvement on prediction is due to reducing overfitting effect by removing irrelevant features or reducing their contribution. The improvement in interpretability is also due to reducing the number of features and thus the multicollinearity.

Subset selection The subset selection consists of identifying the p predictors that we believe to be related to the response. Then we fit a linear model on the reduced set of variables.

In subset selection we try out different subsets of the p predictors and pick the subset which gives the best model. The best model selection could be based on cross-validation error, or directly on the training set by assessing Cp, AIC, BIC, or Adjusted R2. In fact, as the training set error would keep decreasing by increasing complexity as adding features3.4.1, assessing MSE, R2 or error rates directly on the training set underestimates the true error as it won't detect overfitting situations. To simplify, Cp, AIC, BIC, and Adjusted R2 are methods to estimate the test set performance by adding a penalty to the training error. The penalty term increases with the number of predictors in the model.

Different algorithms exist to determine the best features combination by either going through each possible combinations which is very expensive or selecting more smartly by adding (or removing) the features that improve the best (or the least) the results.

Regularization We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as regularization) has the effect of reducing variance and can also perform variable selection reducing overfitting effect. This is done by modifying the least square optimization by adding a shrinkage factor.

Ridge: For Ridge, the regression coefficient estimates $\hat{\beta}_j$ are the ones that minimize

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

Lasso: For Lasso, the regression coefficient estimates $\hat{\beta}_j$ are the ones that minimize

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

Lasso tend to remove some features by setting their $\hat{\beta}_j$ equal to 0, while Ridge only decreases $\hat{\beta}_j$ values without nullifying them.
The shrinkage factor λ is a positive number which could be optimized through cross validation error.

3.3.2.3 Logistic regression

When it comes to classification, we will consider binary classification formed by two classes class (0) and class (1). Linear regression is not a good technique for classification. Instead, logistic regression is used. Logistic regression uses the logistic function

$$p(X) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}$$

In logistic regression, the β_i are determined by fitting the training data by maximizing likelihood function. It can be shown that P(X) is then an estimate of the probability of the sample being in class (1).

Logistic regression is considered a linear model because logit of the probability called also log-odds is linear function of the features.

$$log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Similarly to linear models, regularization and features selection are used in logistic regression.

3.3.3 Tree-based models

In this project, we used boosting and random forest which are tree-based models.

Trees are set of non-linear methods for regression and classification which split the predictor space into regions, and use the mean (or mode, median, . . .) of the training observations in each region for prediction. The split sequence (feature and associated value) is based on optimizing a criterion (Gini index, Entropy, RSS). Based on this optimization, it is possible to assess the most important features for the splits.

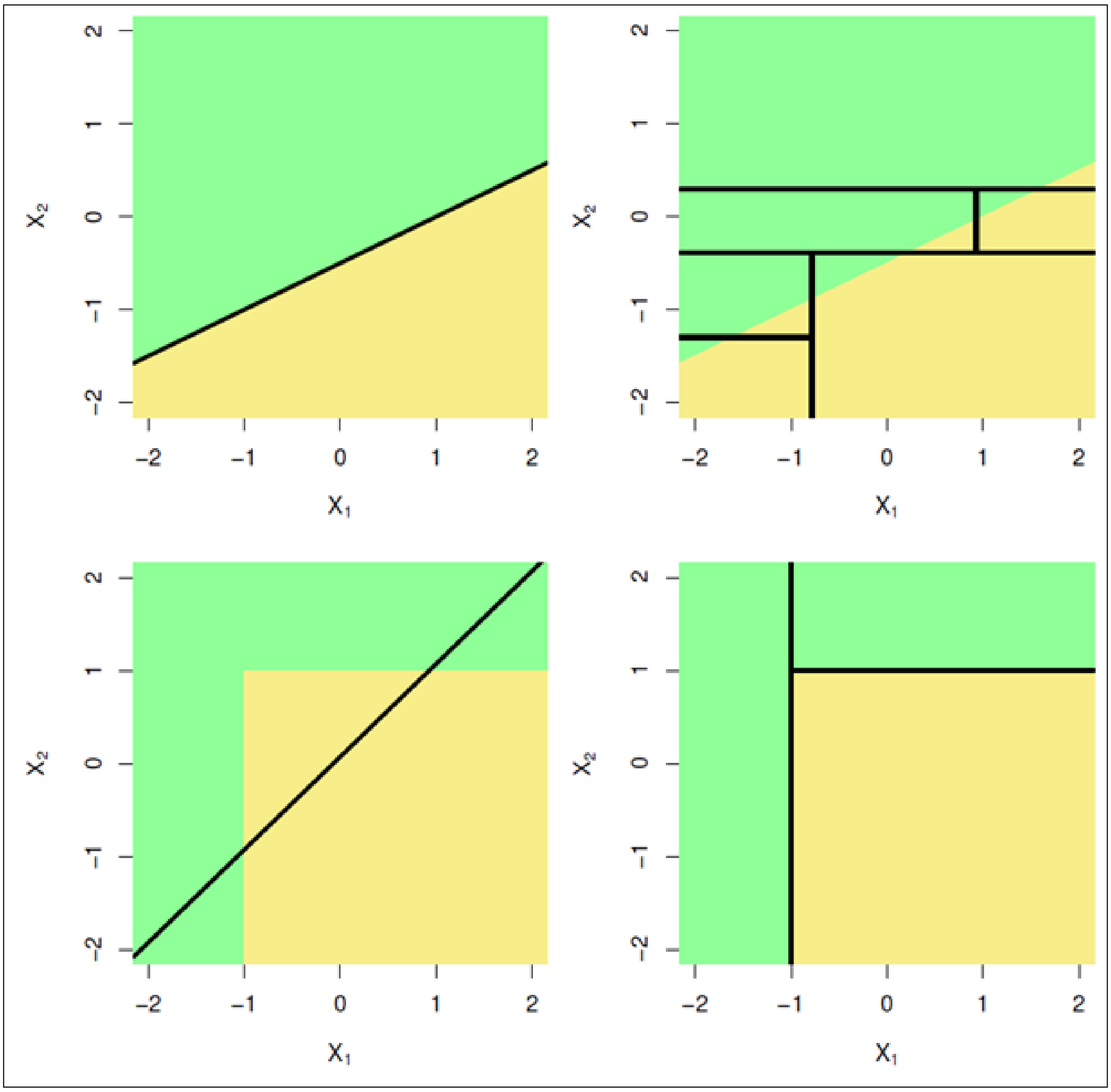


Figure 3.3: Linear model classification region (left) vs tree-based classification region (right) with the decision boundaries showing (top) better classification using linear model if the true boundary between classes is linear, (bottom) better classification using tree based model if the true boundary between classes is a square

3.3.3.1 Random forest

The ensemble method consists of aggregating multiple models and averaging the results to improve the prediction. It is based on the fact that averaging uncorrelated predictors improves the prediction uncertainty by reducing the variance[28]. In fact, decision trees often suffer from high variance as the trees are sensitive to small changes in the predictors. In fact, changing the observation set may lead to a very different tree. The main idea of random forest is to reduce the variance by averaging the results from multiple prediction trees. This is done by bootstrapping (sampling with replacement) the training data and fitting a decision tree for each bootstrap, then averaging the results. If all features are used this is called Bagging. The random forest method would reduce further the variance by restricting the number of features to use at each split as this would reducing the correlation between the trees.

3.3.3.2 Boosting

Boosting is another tree-based approach which uses trees sequentially. It consists of fitting a small tree to the data, then, fitting a small tree to the residuals of the model. The first tree is then updated based on the residual tree with a weight. Then, the procedure is repeated till a certain stop criterion.

3.3.4 SVM

SVM method is used for classification, it started by finding the best hyperplane separating two classes based on maximizing a margin between the two classes data points. This method is called maximal margin classifier. SVM was then extended to nonlinear boundary by using different types of Kernels as polynomial or radial. This method is not a probabilistic method; thus, class prediction probability cannot be directly obtained. In addition, features need to be scaled.

3.3.5 Neural Network (NN)

The human brain consists primarily of neurons linked together with other neurons. Analogous to human brain structure, a Neural Network (NN) is composed of an interconnected assembly of nodes and directed links[32].

There are multiple types of neural network models. In this study, we used the multilayers perceptron which we think is the most adapted to our data and was used in similar context.

3.3.5.1 Multi-Layers Perceptron model

A multi-layer perceptron model is composed by an input layer, intermediate layers, and a final output layer. Each layer l is formed by nodes i. The node (i) from layer (l) is connected to the node (j) in layer (l+1) through a weight θ^l_{ij}. Example, the activation node a^2_1 is written as follow:

$$a^2_1 = g(\theta^1_{10}x_0 + \theta^1_{11}x_1 + \theta^1_{12}x_2 + \ + \theta^1_{1p}x_p)$$

a^l_i is called the activation of node (i) in layer (l) and g is the activation function. The problem comes up to finding all the θ^l_{ij} through optimization. First, we use an arbitrary θ^l_{ij} and calculate all the a^l_i. This is called the forward propagation. Then, we backpropagate the error term, which is a(L)-y(i), a(L) is the last layer activation and y(i) is the response for the training sample (i). Then gradient descent algorithm is used to update the θ^l_{ij}.

3.4 Assessing model results

3.4.1 The Bias-Variance trade off

An important concept in model fitting is the Bias-Variance trade off. It is possible to fit the data almost perfectly. But, generally, the prediction capabilities of such models are week, this is called overfitting. In this case, the estimate predictive function would have high complexity that it starts fitting patterns that are not in reality associated with the response. Thus, minimizing the error on the training set data does not guarantee the best performance on new data point. In fact, it could be proven that[14] for a new observation (X_0, Y_0) the mean squared error (MSE) is equal:

$$MSE = E[(Y_0 - \hat{f}(x_0))^2] = [Bias(\hat{f}(x_0))]^2 + Var(\hat{f}(x_0)) + Var(\varepsilon)$$

The (MSE) is formed by two competing terms. The bias term is minimized by better fitting the training data. The variance term measures the variability of the estimate of f by changing the training data. However, to reduce the bias term, the variability of the estimate function f is generally increased. Thus, the variance term increases. In other term, as the predictive function becomes more complex the bias term will decrease and the variance term will increase. This is bias-variance dilemma (Fig3.4). High bias leads to underfitting, and high variance leads to overfitting. The best model is generally the one which has the correct trade off between bias and variance.

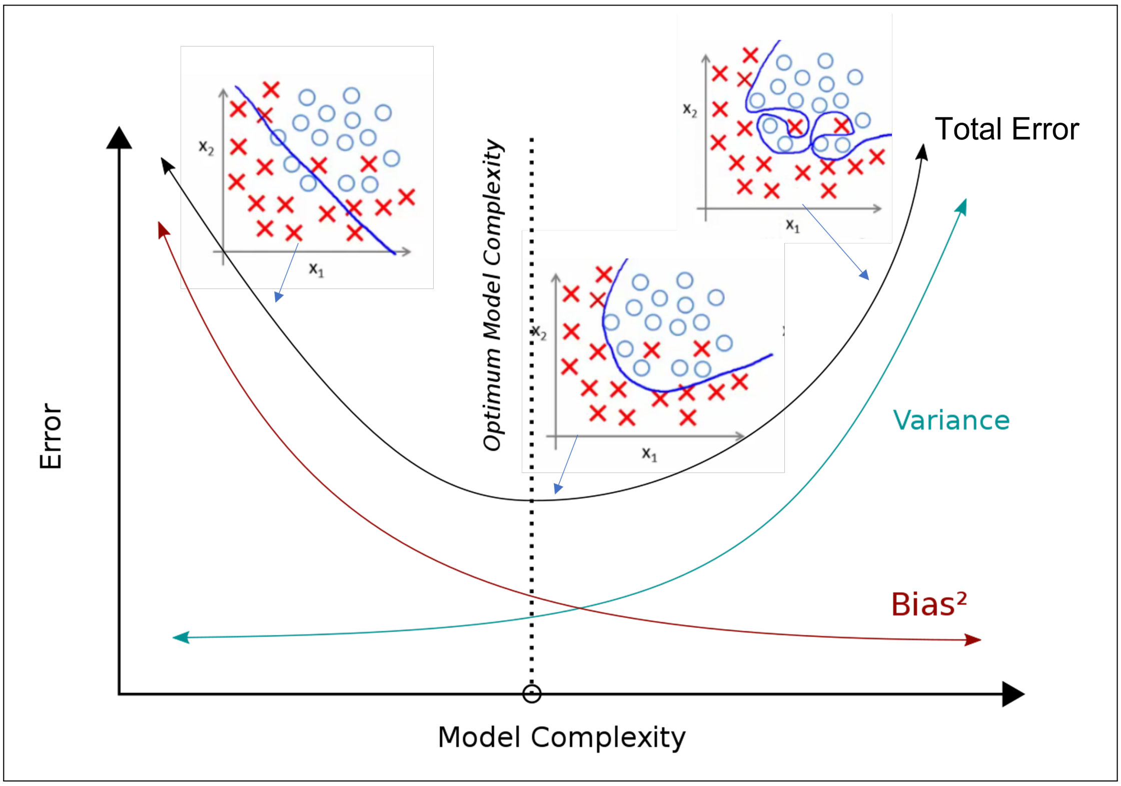


Figure 3.4: Typical bias variance trade off: squared bias (red), variance(blue) and total error(black) curves for a data set. The vertical dotted line indicates the optimal complexity level minimising the total error.

3.4.2 Training, validation and testing

The best model is the model which has the best predictive power on unseen data. So, to find the best model, we split the data randomly in different data sets.

- The training set is used to train the model. In fact, the model uses these data to learn and extract the relevant patterns and association between the features and the response. Thus, it decreases the model bias.

- The validation set is used to optimize model parameters to obtain the best trade off between bias and variance. In fact, we will explain later that most of the machine learning models would have some parameters to tune allowing to adjust the complexity and the variance of the models. The best model parameters are generally the ones that have the best predictions on the validation set as it provides an estimate of the test set error. However, using a validation set could be limited specially in case of small data set. In fact, first, the estimate of the validation set error can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set. So that a different split could change noticeably the validation error. Second, using a validation set means that less data are used for training which gives less robust models. Cross validation (CV) techniques allows to overcome these limitations by using the whole training data for both training and validation. The k-fold CV is done by dividing the training data into k groups. One of the groups is treated as a validation set, the remaining k-1 groups as training set. The error is then computed in the held-out group. This procedure is repeated k times; each time, a different group of observations is treated as a validation set. This process results in k estimates of the test error, The k-fold CV estimate is computed by averaging these values. A special case of K-fold CV is leaving one out cross validation (LOOCV) where k equal 1 which consists of leaving one observation out and training on all remaining observations. Then assess the error on that observation. Then, we repeat the previous operation for the whole training data. The average error on all the left-out observations is the LOOCV error.

- Test set is used for a final comparison of the different models having the optimal parameters. As the validation set was used to optimize the models parameters it could be biased. The test set in a completely independent data set never seen by the models. Thus, it is a good estimate for unseen new data.

3.4.3 Model selection

Machine learning models allow to control the balance between bias and variance through adjustment of the models hyperparameters. Practically, this could be done exhaustively by varying one parameter at a time and assess the validation set or cross validation performance, or manually. The best model's parameters are the ones that provide the best results in the cross validation, or the validation set.

Examples of hyperparameters are the shrinkage factor for Lasso and Ridge, the number of decision trees for random forest, the number of trees, the learning rate and tree max depth for boosting, cost, gamma, degree and kernel for SVM and number of iterations, batch size, number of layers and regularization parameters for the neural network.

3.4.4 Model performance metrics

The goal of the machine learning models is to be able to predict the value of new data. As the test set is an estimate of new unseen data. Our model performance should be assessed on test set. Thus, to compare the performance of different models, we need to quantify the extent to which the predicted response value for a given observation in the test set is close to the true response value for that observation. The selection of the metrics would depend on the task (prediction or classification) and the data. Generally, these metrics measure a certain distance between the predicted and the true response. For prediction the most used metrics are mean squared error (MSE) and R2.

- MSE stands for mean squared error, it is the mean square of the Euclidian distance between the prediction and true response. Thus, the lower the MSE the better the fit.
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$
- R2 statistics takes values between 0 and 1.
$$R^2 = \frac{TSS - RSS}{TSS}$$
where $TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2$ and $RSS = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$.
Thus, R2 measures the proportion of variability in Y that can be explained using X[ref stat]. The higher the R^2 value the better is the fit.

For binary classification (for simplicity, we assume the classes are labeled “-” and “+”), there are multiple metrics depending on the data and the objective. Example of classification metrics used in this project are:

- The Confusion matrix: The confusion matrix is a practical tool to visualize the data the columns are the predicted classes while the rows are the actual classes and each cell in the matrix would be the count of predictions vs actuals (Table3.1);

Table 3.1: Confusion Matrix

	True -	**True +**	**Total**
Predicted-	True Negative TN	False Negative FN	N*
Predicted +	False Positive FP	True Positive TP	P*
Total	N	P	

- Misclassification rate: the fraction of the predictions that were wrong, without distinguishing between positive and negative predictions. $\frac{FN+FP}{N+P}$
- Accuracy: the fraction of the predictions that were correct, without distinguishing between positive and negative predictions $\frac{TN+TP}{N+P}$
- Recall or Sensitivity or True Positive Rate: is the proportion of correctly classified positive observations $\frac{TP}{P}$
- Specificity or True Negative Rate is the proportion of correctly classified negative observations $\frac{TN}{N}$
- Precision is the proportion of correctly predicted positive observations from the predicted positive observations $\frac{TP}{P*}$
- Balanced accuracy[20] is equal to the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate), which avoids inflated performance estimates on imbalanced datasets

$$balanced-accuracy = \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$$

- F1score: Is the harmonic mean of precision and recall

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

- ROC curve: The receiver operating characteristics (ROC) curve gives a graphical display of the sensitivity against specificity, as the threshold value (cut-off on probability of success or disease) is moved over the range of all possible values (Fig3.5) . An ideal classifier will give a ROC curve which hugs the top left corner, while a straight line represents a classifier with a random guess of the outcome.

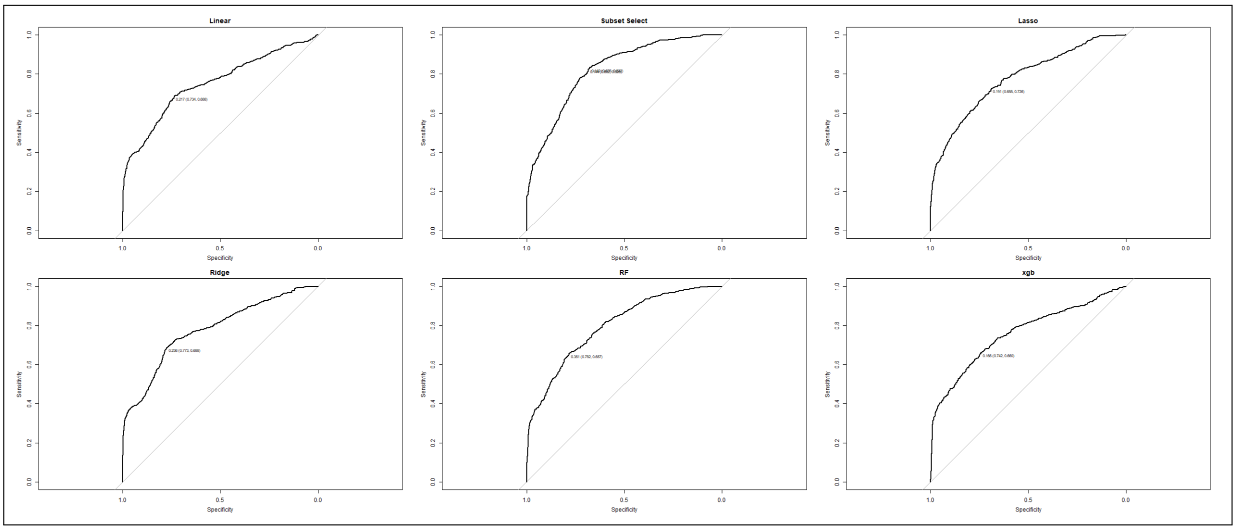


Figure 3.5: Example of ROC curve for the second classification task showing sensitivity as function of specificity and associated AUC value.

- AUC score is the area under the ROC curve(Fig3.5). It ranges between the values 0 and 1, where a higher value indicates a better classifier. The AUC score is useful for comparing the performance of different classifiers, as all possible threshold values are considered.

Chapter 4

Methodology

4.1 Dataset

4.1.1 Data and data loading

The dataset used for this study consists of mud gas logs, well logs and core data from 40 wells. The normal display of these data is function of depth. A data point would be a record of certain properties for a given well at a certain depth.

The well logs consist of composite logs (measured) and interpreted logs (interpretation of the different composite log)2.3.2:

- The composite logs are direct measurements of some physical properties that allows later to identify lithology, fluid and porosity. These logs include gamma ray (GR), caliper (CAL), resistivity (shallow, medium and deep), density (ROHB), neutron (NPHI), shear sonic (DTS) and compressional (DTC) and were provided by Spirit energy through Petrel Studio.
- The interpreted logs are interpretation of the composite log by a petrophyisicist to relate to reservoir properties or fluid content. This includes porosity (PHIE), shale volume (VSHALE) and water saturation (SW). These were interpreted by Spirit-Energy petrophysicist and provided by Spirit-Energy through Petrel Studio.

The mud gas and drilling data were provided by Geo-Provider. It consists of cleaned mud gas quantities and ratios that are generally used for qualitative reservoir and fluid assessments (C1), (C2), (C3), (C4), (C5), (C1/C2), (C1/TG) with TG: total gas, (C1/(C2+C3)), (C3*100/(C1+C2)), wetness (WH): (C2+C3+C4+C5)/(TG)*100 and balance (BH): (C1+C2)/(C3+C4+C5), and drilling parameters as mud weight (MW), rate of penetration (ROP) and weight on bit (WOB). These data were provided in form of a petrel project. For completion some of the missing drilling data as (MW) was imported from NPD website, and (WOB) and (ROP) imported from some final well drilling reports in order to decrease number of missing data.

The core data consists mainly of core porosity (CPOR) and core horizontal permeability (KLH). These data were provided as .las files by Spirit-Energy petrophysicist.

Petrel is a software developed by Schlumberger. It is mainly used in Spirit Energy by geoscientists to integrate works from petrophysicsit, geoscientist and reservoir engineers. Petrel Studio allows to load all sort of different data from a common database to Petrel.

A new Petrel Project was created in which all the data needed for the current study were collected and checked.Then, all the data were exported as multiple .las files. Figure 4.1 shows an example of the data used in this study for a well 'A' visualized with Petrel software.

A .las file corresponds to all the data from one well with a header specifying some well attributes as well location, number of logs, their units, maximum depth... Next, using 'lasio' library[15] that handles .las files, these .las files were loaded both into R and Jupyter notebook (Fig4.2). Then, a dataframe was generated merging all these log data and adding a column for the well name (Fig4.3).

In this study, Python, R and Petrel were used to visualize the data.

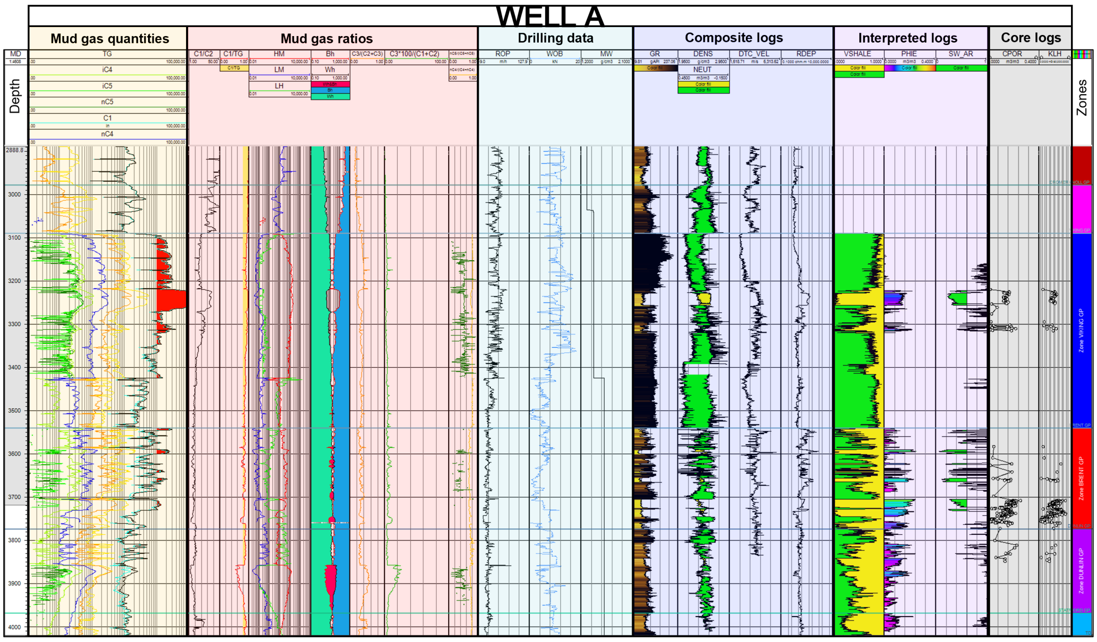
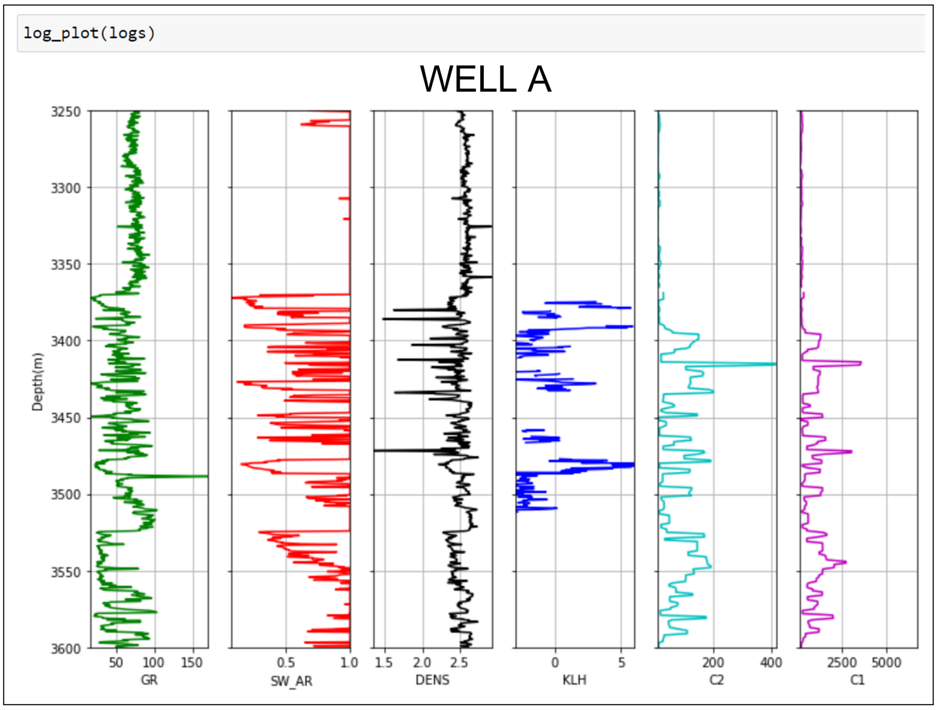
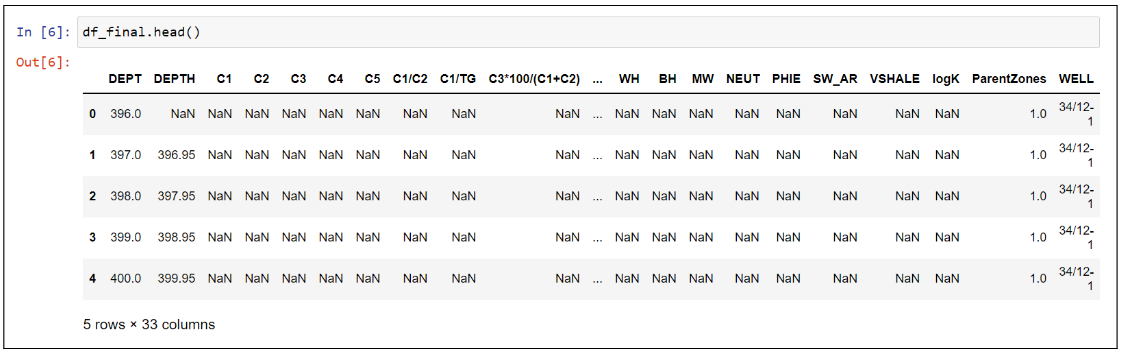


Figure 4.1: Example of data-set for a well used in this study visualised in Petrel.



```
log_plot(logs)
```

Figure 4.2: Example of data-set for a well used in this study visualised in JupyterNotebook.



In [6]:
```
df_final.head()
```

Out[6]:

	DEPT	DEPTH	C1	C2	C3	C4	C5	C1/C2	C1/TG	C3*100/(C1+C2)	...	WH	BH	MW	NEUT	PHIE	SW_AR	VSHALE	logK	ParentZones	WELL
0	396.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	34/12-1
1	397.0	396.95	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	34/12-1
2	398.0	397.95	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	34/12-1
3	399.0	398.95	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	34/12-1
4	400.0	399.95	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	34/12-1

5 rows × 33 columns

Figure 4.3: First raws of the data frame generated by merging all the data and re-sampling to 1 meter.

4.1.2 Data processing

Logging data has a resolution of 0.15m. Permeability data are punctual and discrete measures from core plugins thus distance between the measures depends on quality of core and interest. Mud gas data have a resolution of around 1m. To merge all these data, a rolling median filter was generated to have a sampling rate of 1m.

Two binary categorical response variables were generated for each well:

- (HCind) to distinguish water (0) from hydrocarbon (1) samples: this response variable was generated by cut off of 0.65 on water saturation (Sw). So that if water saturation is more than 0.65, the point is classified as water and classified hydrocarbon if not.
- (FluidType) to distinguish between oil (0) and gas (1) in hydrocarbon bearing area: this response variable was generated using well reports and from NPD web site. This data generally comes from well test reports or pressure and fluid sampling data.

A cut off of 10% on porosity were applied for the data for the Hydrocarbon/Water classification. In fact, this cut off is generally used in petrophysical evaluation to consider the rock as reservoir rock. Using this cut off, we have 2952 samples for the training set and 281 samples for the test set corresponding to well: "35/12-2”. The data is imbalanced: class-0 (water) has 2323 samples and class-1 (hydrocarbon) has 676 samples.

A cut off of 5% on porosity and 60% on water saturation was used for fluid type classification. In fact, 60% is generally used by petrophysicist as hydrocarbon flag (considers the rock to be hydrocarbon bearing). Using this cut off the training set consists of 611 observations from which 382 oil (0) and 229 water(1), and 102 observations (52 for oil and 50 for gas) for the test set corresponding to the well “36/7-4”.

The same cut off as in task-2 (5% on porosity and 60% water saturation) was used for permeability prediction. In addition, not all the wells had core data and the core data does not cover generally the whole reservoir sections. So, for this task, we had 282 observations.

An analysis of the generated data regarding missing values were done (Fig4.4) (number of NaN values by features, number of NaN values by features and by wells). Some missing data were then checked and recollected as the mud weight from NPD web site. The features that have high percentage of missing values were dropped as some drilling data ROP and WOB.

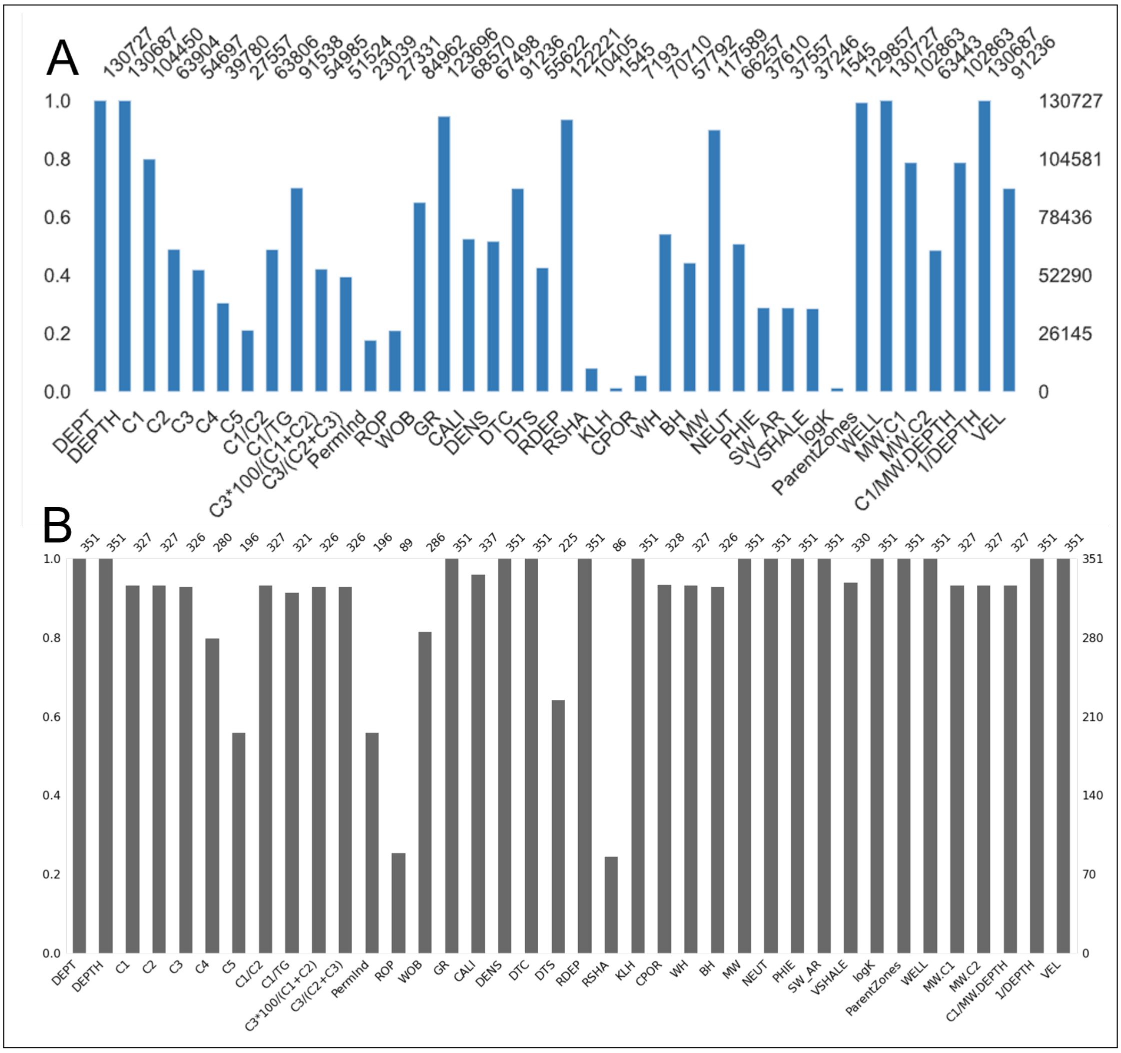


Figure 4.4: Missing data analysis: A) using the whole data-set B) using only the data where we have permeability measures.

Considering that the dataset is small, other features with lower percentage of missing values were also dropped when it was noticed that they are not important for the prediction.

After plotting the correlations between the different features and the response variables and generating features statistics, some features were log transformed to have more adequate distribution and better prediction (Fig 4.5)

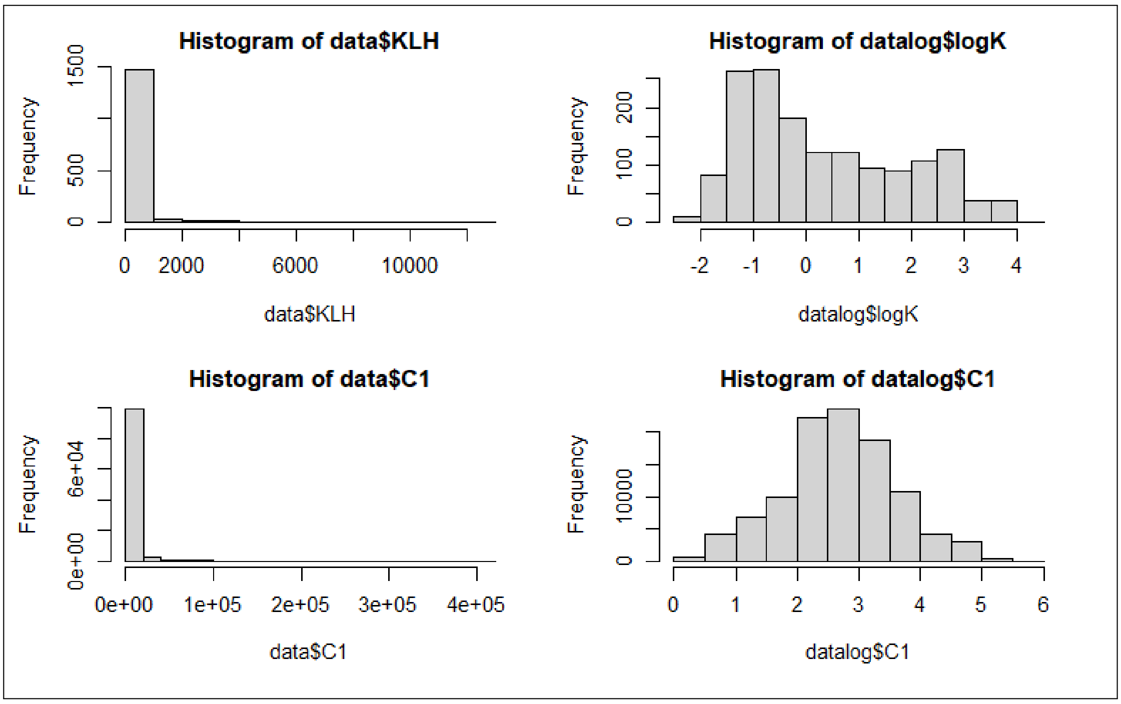


Figure 4.5: Histogram of permeability (top) and C1 (bottom) before (left) and after log transform (right) showing better spread of the distributions.

4.1.3 Data analysis

4.1.3.1 Corrleations

A pairwise comparison of the features and response variables and their associated Pearson correlation coefficient were generated using GGally library ([9]). We notice high correlation between some features, specially between the gas quantities (C1), (C2), (C3) and (C4), shallow resistivity and deep resistivity, (GR) and (VShale), (DTS) and (DTC), . . .

We notice from the boxplot in figure 4.6, that the hydrocarbon and water classes could be well separated by the mud gas quantities (C1), (C2) and (C3), while the gas and oil classes (Fig 4.7) could be well separated using both mud gas quantities (C1), (C2) and (c3) and/or ratios (C1/C2), (WH) and (BH).

We notice that log permeability (response variable), porosity, density, core porosity are highly correlated. No strong correlation were observed between individual mud gas data and permeability (Fig 4.8)

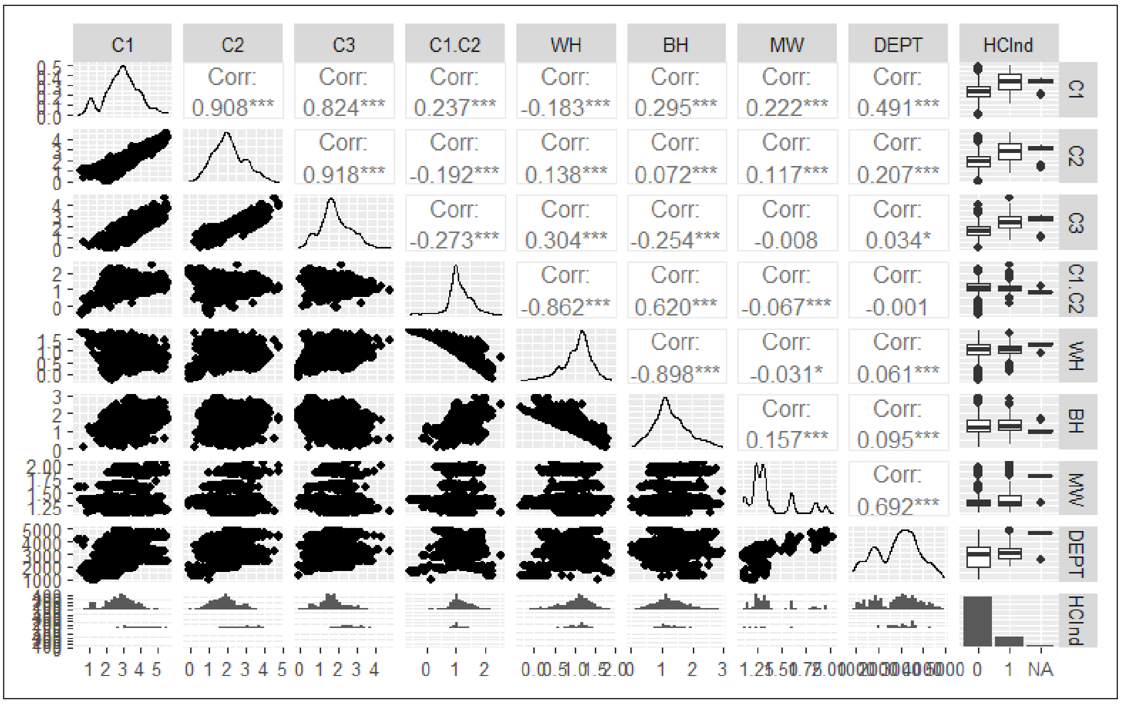


Figure 4.6: Pairwise cross plots of the mud gas data and HCInd response showing high correlation between mud gas quantities and that the mud gas quantities could allow for a good separation between water and hydrocarbon.

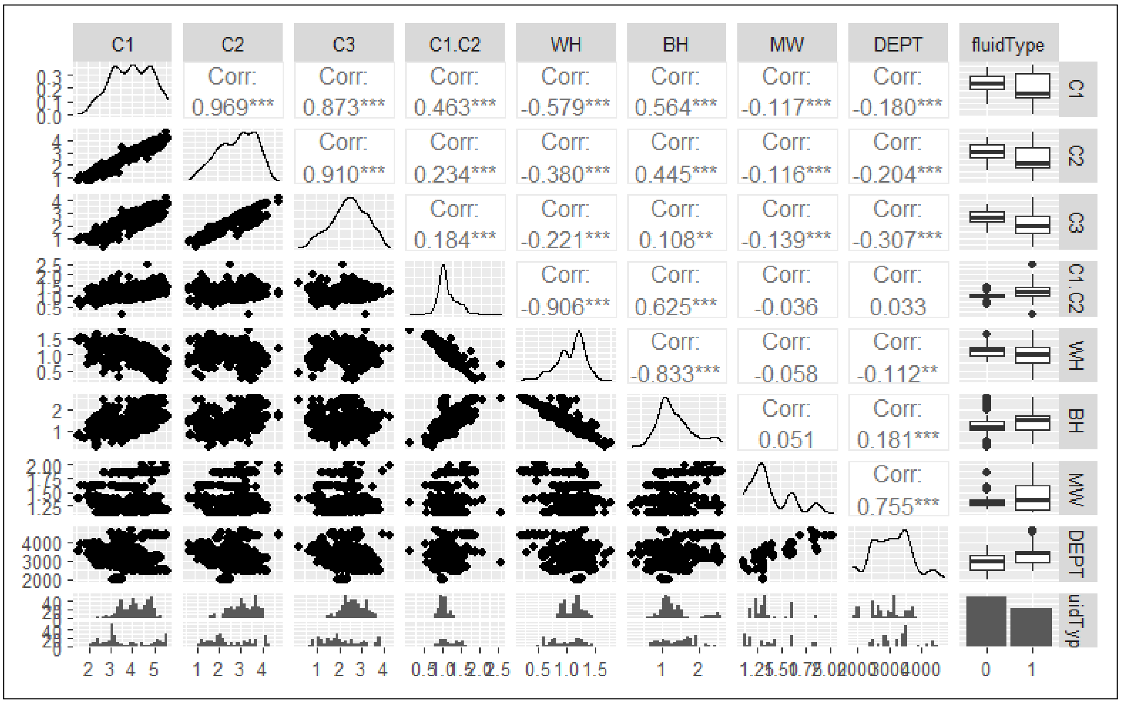


Figure 4.7: Pairwise cross plots of the mud gas data and fluidType response showing that the mud gas quantities and/or ratios could allow for a good separation between oil and gas.

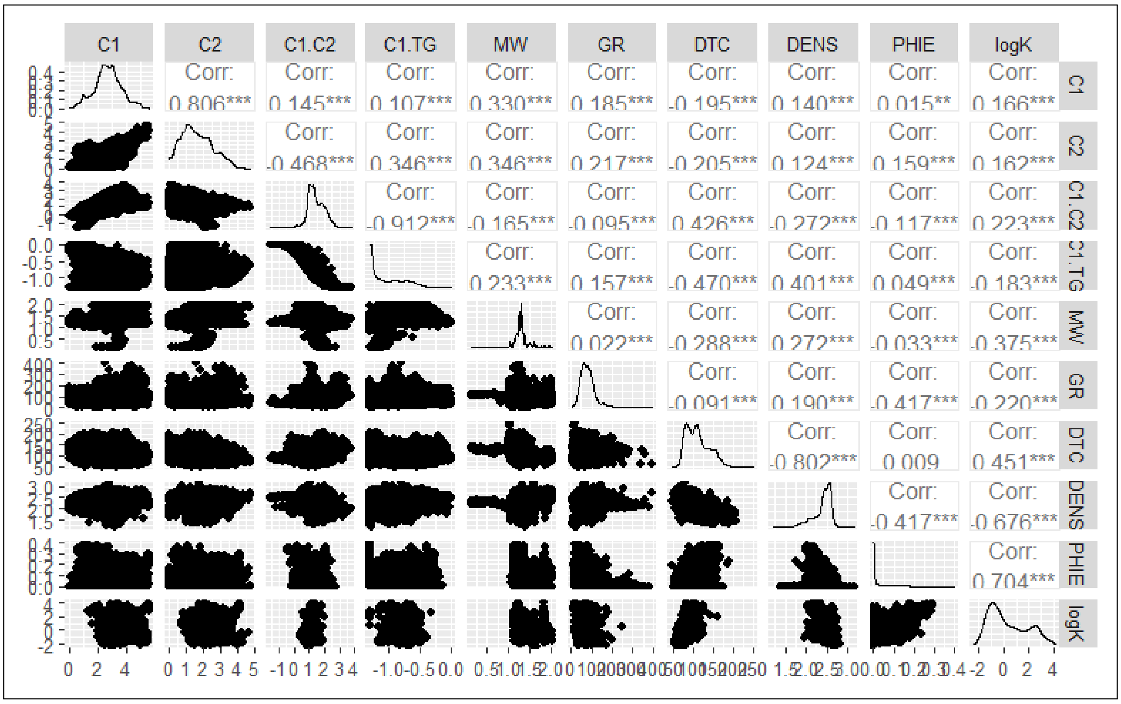
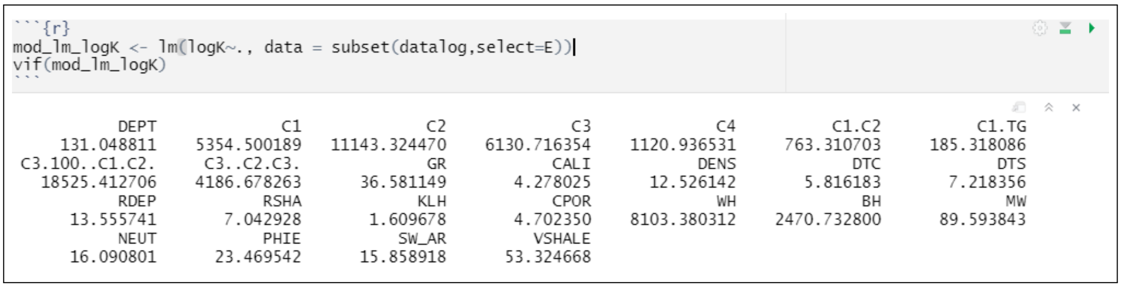


Figure 4.8: Pairwise cross plots of mud gas data, well log data and log of permeability showing that log permeability, porosity, density, core porosity are highly correlated. No strong correlation were observed between mud gas data and permeability.

4.1.3.2 Model summary

Some linear models were generated to assess features statistical importance and multicollinearity through model summary and variance inflation factor (VIF) (Fig 4.9). Analysis of the variance (ANOVA) on adding some mud gas data to porosity and shale for predicting permeability (which is the standard approach for permeability assessment) showed that mud gas data are important and improves the fit (Fig5.7). No special high leverage and high residuals data points were observed that we think need to be removed (Fig4.10).



~~~
```{r}
mod_lm_logK <- lm(logK~., data = subset(datalog,select=E))
vif(mod_lm_logK)
```

         DEPT           C1           C2           C3           C4        C1.C2        C1.TG
   131.048811  5354.500189 11143.324470  6130.716354  1120.936531   763.310703   185.318086
C3.100..C1.C2.  C3..C2.C3.           GR         CALI         DENS          DTC          DTS
 18525.412706  4186.678263    36.581149     4.278025    12.526142     5.816183     7.218356
         RDEP         RSHA          KLH         CPOR           WH           BH           MW
    13.555741     7.042928     1.609678     4.702350  8103.380312  2470.732800    89.593843
         NEUT         PHIE        SW_AR       VSHALE
    16.090801    23.469542    15.858918    53.324668
~~~

Figure 4.9: VIF when all the features are used in the model showing high multicolinearity problem between mud gas data thus, features selection is needed to reduce this VIF value.

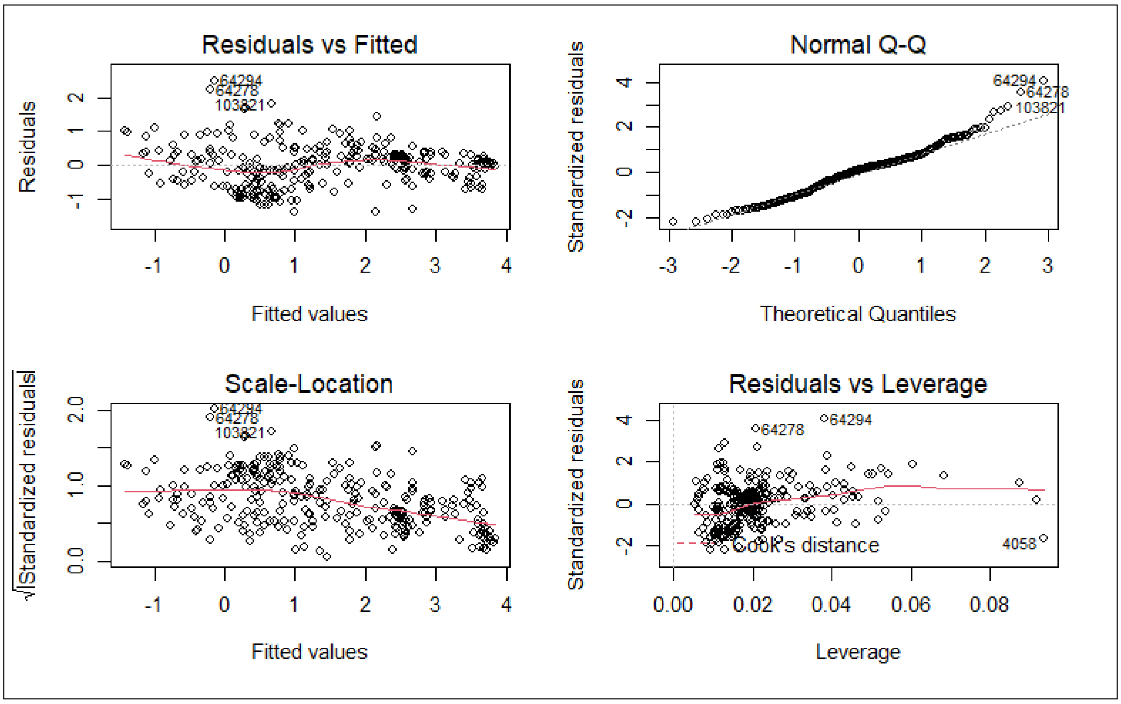
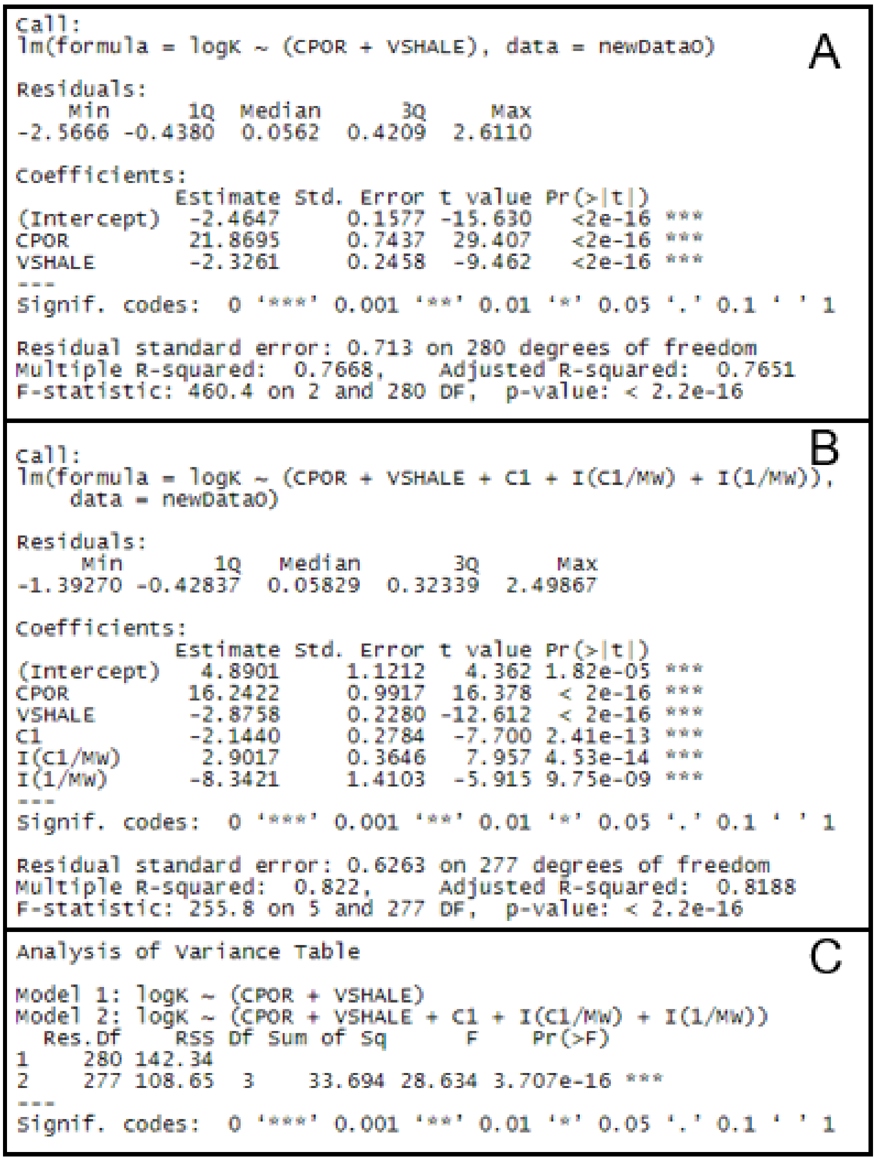


Figure 4.10: Assessment of model residuals and leverage through multiple cross plots.

Similarly, model output was used to assess statistical importance of mud gas data to classify FluidType.

Assessment of collinearity using VIF, shows high correlation between features and specially mud gas quantities and ratios which could be a problem for inference and model stability(Fig 4.9). A lot of work will be done to reduce this multicollinearity effect by reducing the number of features.

Dimensionality reductions were done through PCA to explore the data. This method was used separately in seal (PHIE<0.1) and reservoir rocks (PHIE>0.1) through porosity cut offs, to assess any trends in the data that could characterize seals and reservoirs.



```
Call:
lm(formula = logK ~ (CPOR + VSHALE), data = newDataO)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5666 -0.4380  0.0562  0.4209  2.6110

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.4647     0.1577 -15.630   <2e-16 ***
CPOR         21.8695     0.7437  29.407   <2e-16 ***
VSHALE       -2.3261     0.2458  -9.462   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.713 on 280 degrees of freedom
Multiple R-squared:  0.7668,    Adjusted R-squared:  0.7651
F-statistic: 460.4 on 2 and 280 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = logK ~ (CPOR + VSHALE + C1 + I(C1/MW) + I(1/MW)),
    data = newDataO)

Residuals:
     Min       1Q   Median       3Q      Max
-1.39270 -0.42837  0.05829  0.32339  2.49867

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.8901     1.1212   4.362 1.82e-05 ***
CPOR         16.2422     0.9917  16.378  < 2e-16 ***
VSHALE       -2.8758     0.2280 -12.612  < 2e-16 ***
C1           -2.1440     0.2784  -7.700 2.41e-13 ***
I(C1/MW)      2.9017     0.3646   7.957 4.53e-14 ***
I(1/MW)      -8.3421     1.4103  -5.915 9.75e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6263 on 277 degrees of freedom
Multiple R-squared:  0.822,     Adjusted R-squared:  0.8188
F-statistic: 255.8 on 5 and 277 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Model 1: logK ~ (CPOR + VSHALE)
Model 2: logK ~ (CPOR + VSHALE + C1 + I(C1/MW) + I(1/MW))
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    280 142.34
2    277 108.65  3    33.694 28.634 3.707e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4.11: Linear model summaries fitting permeabiltiy with A) porosity and Vshale; B) prosity, Vshale, C1, 1/MW and C1/MW, and C) ANOVA showing that some mud gas features could be statistically important to predict permeability.

4.1.3.3 PCA analysis

In Reservoir Rock: The figure 4.12 represents both the principal component scores and the loading vectors for the reservoir zones in a single biplot display. We notice that:

- PC1 gives most weight to Depth, MW, GR, Neutron, Vshale. An increase of depth is associated with a decrease of Vshale. This could be explained by the general trend that the deeper reservoirs are cleaner than shallower (upper Jurassic) reservoirs as we know from the geology and depositional environments.
- PC2 gives clearly more weight to gas quantities (C1, C2, C3), resistivity and water saturation. An increase of gas quantities is associated with an increase of resistivity and a decrease on water saturation. PC2 clearly relates more to fluid content. In fact, it captures the negative trend between the gas quantities and (Sw), which shows that these mud gas quantities could allow to distinguish hydrocarbon zones from water zones which is our first classification task.
- PC3 gives more weight to gas ratios, density and sonic. We see as expected positive correlation between amount of heavy component in the mud gas with velocity and density.

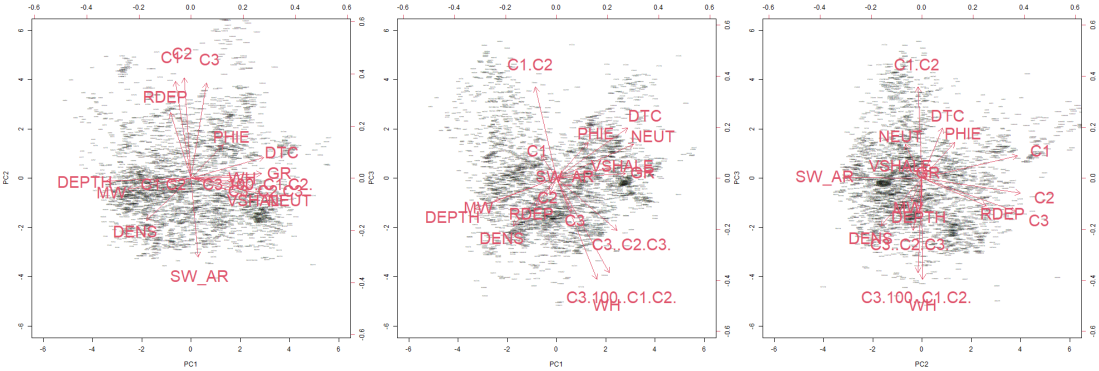


Figure 4.12: Principal components biplots in reservoir. Left: biplots of PC1 and PC2, middle:biplots of PC2 and PC3 and right: biplots of PC1 and PC3 showing the major variability directions in the data. Example: PC2 shows the negative relation between mud gas quantities and water satruation

In Hydrocarbon bearing reservoir rock The figure 4.13 represents both the principal component scores and the loading vectors for the reservoir zones in a single biplot display. We notice that:

- PC1 gives most weight to C1, C2 and C3, followed by density, depth and MW. It relates to compaction. Higher the depth, higher the density and the mud weight.
- PC2 gives most weight to gas ratios, followed by neutron porosity. It could relate to fluid in the reservoir as Neutron porosity is quite affected by fluid. If the reservoir is gas bearing, we have lower neutron density. While the proportion of heavy component (like 'Wh') will decrease as we see from PC2. This shows that these mud gas quantities could allow to distinguish between oil and gas zones which is our second classification task.
- PC3 gives most weight saturation followed by shale volume, porosity and resistivity. It relates to how both shale volumes and porosity would affect the hydrocarbon saturation. The lower shale volume, higher the porosity, higher the hydrocarbon saturation.

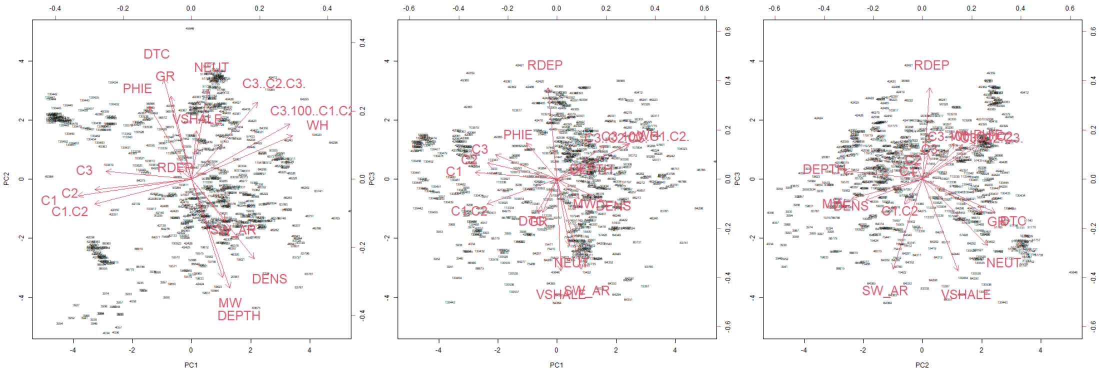


Figure 4.13: Principal components biplots in hydrocarbon bearing zones. Left: biplots of PC1 and PC2, middle:biplots of PC2 and PC3 and right: biplots of PC1 and PC3 showing that PC1 could be more related to compaction while PC2 could be more related to fluid content

In Seal Rock The figure 4.14 represents both the principal component scores and the loading vectors for the reservoir zones in a single biplot display. We notice that:

- PC1 gives most weight to mud gas ratios and MW.
- PC2 gives most weight to mud gas quantities and GR
- PC3 gives most wight to NEUT, RDEP, DTC and GR

No clear relation in the seal was found, probably due to the geological complexity and lack of repeatability of the acquisition. Generally, while drilling the seal rock, the goal is to drill safely as fast as possible. So, the focus on data quality is minimum.

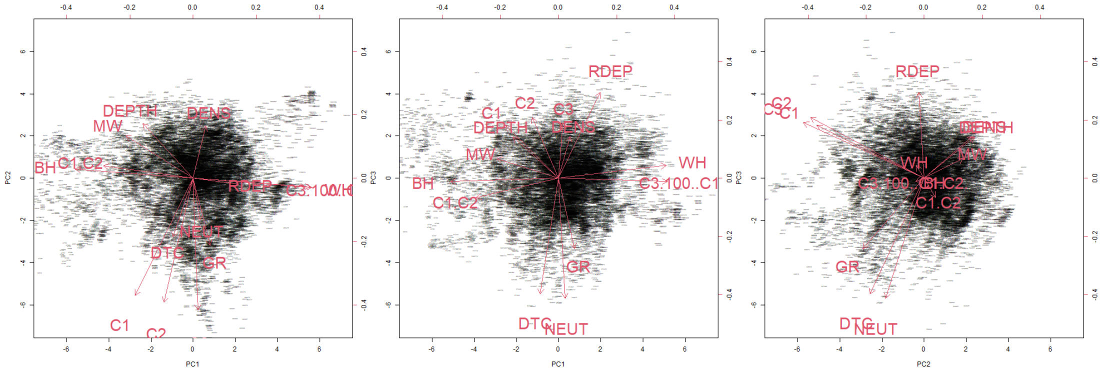


Figure 4.14: Principal components biplots in seals: Left: biplots of PC1 and PC2, middle:biplots of PC2 and PC3 and right: biplots of PC1 and PC3 showing the major variability directions in the data. No clear relation in seals was found.

4.1.4 Systematic error

Figure 4.15 shows the distribution of C1 from well to well compared to the density log. We can notice that while density has similar distribution in most wells, the distribution from mud gas are more variables. In fact, the distribution of C1 is more sensitive to the wells. This could be an indication of systematic error[3] on the mud gas quantities. In fact, in section2.4.3 we discussed that mud gas data are influenced by multiple parameters as drilling parameters and the gas trap system used, reservoir pressure, the mud sample extracted, etc. This is inducing a systematic error that would change from well to well, or from formation to formation. A tentative to reduce this error was done by generating corrected mud gas quantities, by dividing by mud wight and depth. In addition, mud gas ratios could be less affected by the systematic error.

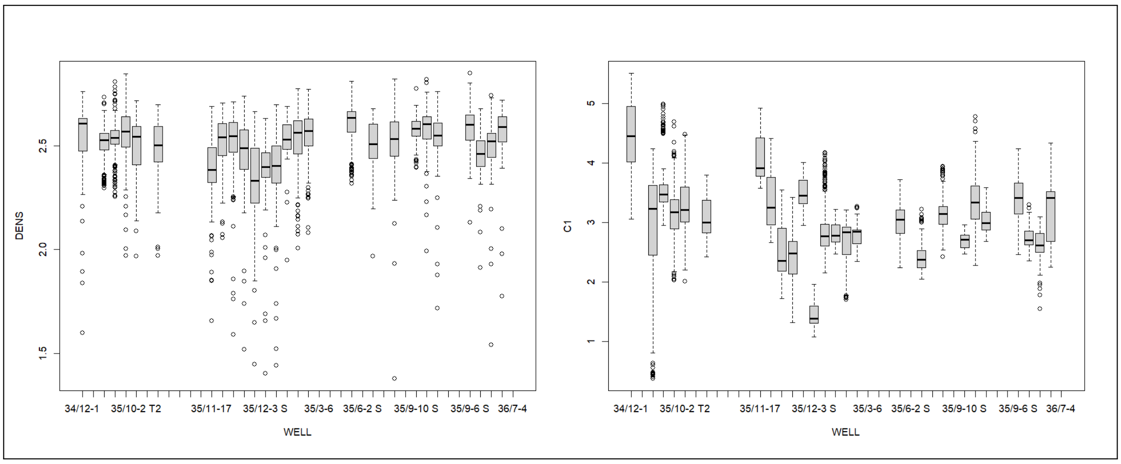


Figure 4.15: Box plots showing spread of density distribution (left) and C1 distribution (right) by well. While density has similar distribution in most wells. The distribution of C1 are more sensitive to the wells which could show a systematic error in the data changing from well to well

4.1.5 Data analysis summary

To summarize, from the data processing and analysis, we can already notice some interesting features that could be important for the prediction tasks. We noticed that:

- Mud gas quantities could be important to separate water from hydrocarbon bearing observations,
- Mud gas quantities and ratios could be important to separate water from oil and gas,
- Corrected mud gas quantities by mud weight could be important for permeability prediction,
- The data (specially mud gas and drilling data) are affected by systematic error that could be due to drilling conditions and parameters, mud gas equipment's, etc.

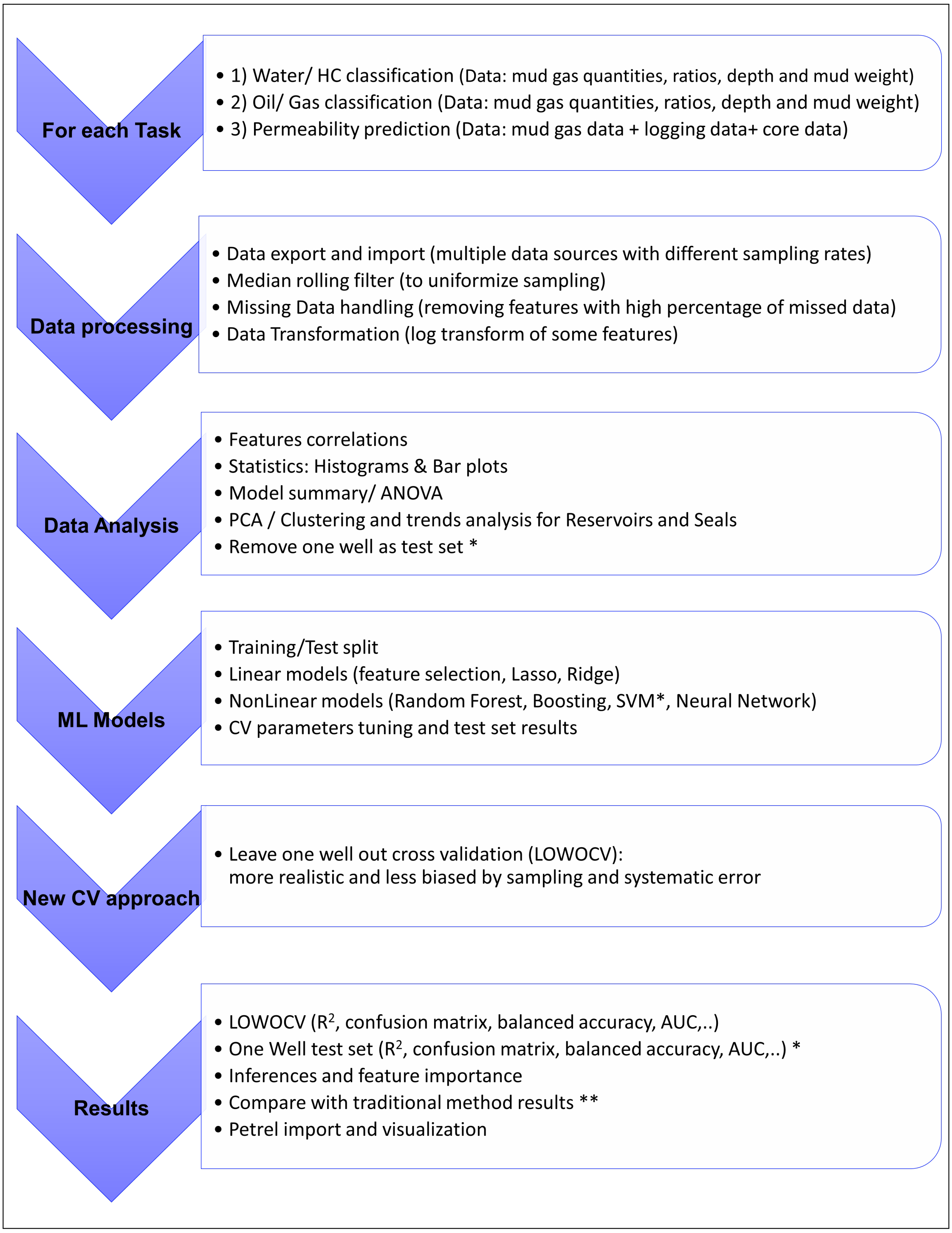
4.2 Project workflow

After data cleaning, processing and analysis, one well was removed from the data as a test set for the two classification tasks. This was not done for the permeability prediction due to the small number of data points.

Then for the classification tasks seven models were assessed (logistic regression with feature selection, Lasso, Ridge, Random Forest, Boosting, SVM and Neural Network) and for the prediction task six models were generated (linear regression with feature selection, Lasso, Ridge, Random Forest, Boosting and Neural Network)

Next, for each model, model hyperparameters were tuned through measuring performance on validation set or cross validation.

Final models were then run with these optimal parameters to assess cross validation error using a new technique which consists of removing one well out and training on the other wells, then iterate through wells. This is similar to Leave one out cross validation (LOOCV), but instead of leaving one observation we leave one well with all the observations associated to it (LOWOCV). This is more realistic than a random sample split to avoid bias from data from the same well. In addition, for the classification tasks 1 and 2, these models were also run on the test set well. Both metrics on the LOWOCV set and test set were used to judge on the best model 4.16.



For each Task

- 1) Water/ HC classification (Data: mud gas quantities, ratios, depth and mud weight)
- 2) Oil/ Gas classification (Data: mud gas quantities, ratios, depth and mud weight)
- 3) Permeability prediction (Data: mud gas data + logging data+ core data)

Data processing

- Data export and import (multiple data sources with different sampling rates)
- Median rolling filter (to uniformize sampling)
- Missing Data handling (removing features with high percentage of missed data)
- Data Transformation (log transform of some features)

Data Analysis

- Features correlations
- Statistics: Histograms & Bar plots
- Model summary/ ANOVA
- PCA / Clustering and trends analysis for Reservoirs and Seals
- Remove one well as test set *

ML Models

- Training/Test split
- Linear models (feature selection, Lasso, Ridge)
- NonLinear models (Random Forest, Boosting, SVM*, Neural Network)
- CV parameters tuning and test set results

New CV approach

- Leave one well out cross validation (LOWOCV):
  more realistic and less biased by sampling and systematic error

Results

- LOWOCV (R^2, confusion matrix, balanced accuracy, AUC,..)
- One Well test set (R^2, confusion matrix, balanced accuracy, AUC,..) *
- Inferences and feature importance
- Compare with traditional method results **
- Petrel import and visualization

Figure 4.16: Project workflow diagram showing the steps and tasks followed during the project

4.3 Machine learning methods and optimization

4.3.1 Fluid classification tasks

4.3.1.1 Logistic regression

For the first and second task, logistic regressions were done in R using generalized linear models (glm). Multiple combinations of mud gas data were fed to the model. (Fluidtype) and (HcInd) were the response variables. Subset selection, Ridge regularization and lasso regularization were then tried, in order to have simpler model for better interpretability and prediction.

- Subset selection: Step wise selection of the features were done based on AIC using 'stepAIC' function,
- Lasso Regularization: Lasso regularization with lamda optimization using cross validation 4.17,
- Ridge Regularization: Ridge regularization with lamda optimization using cross validation 4.17.

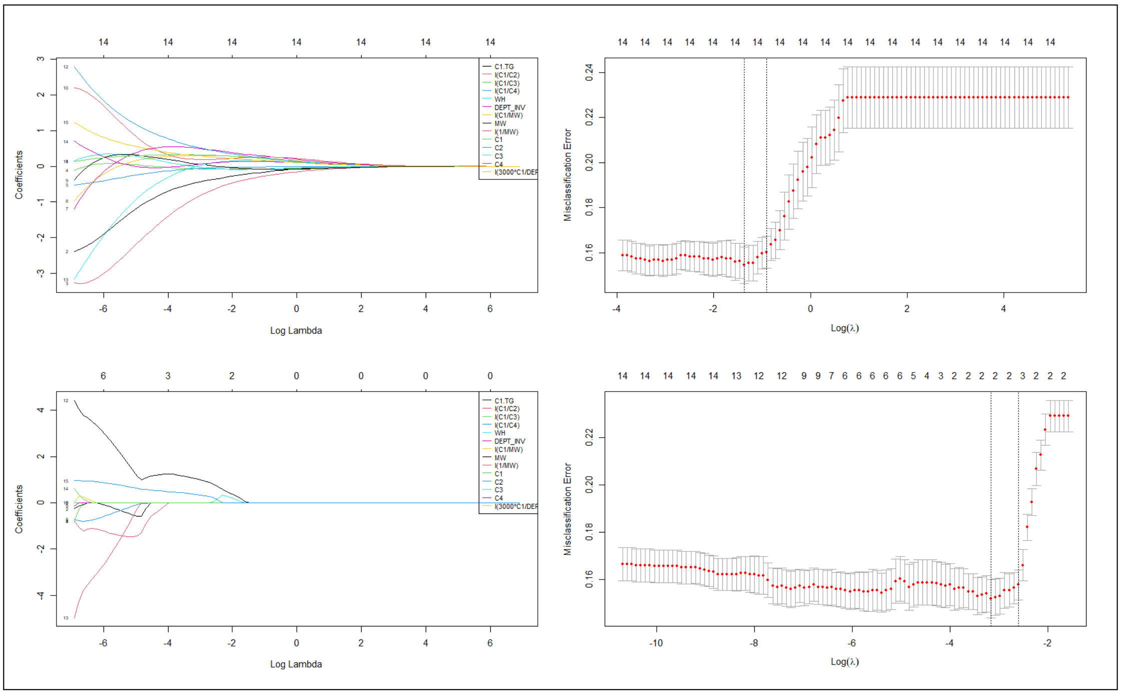


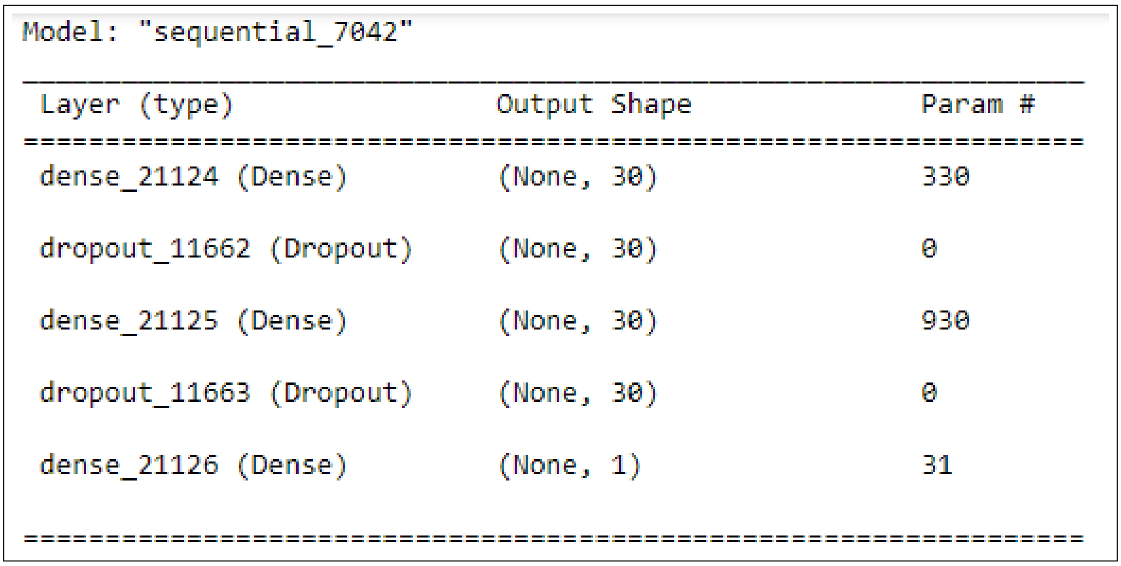
Figure 4.17: Lasso and Ridge optimisation

4.3.1.2 Tree based methods and SVM

Tree based methods (Random Forest (RandomForest) and Boosting (xgboost)), and SVM (svm) were implemented using R. (Xgboost) iteration number was optimized through cross validation. SVM linear, polynomial and radial were implemented and optimized using grid search and validation set misclassification rate. SVM with radial kernel showed the lowest misclassification rate of the validation set.

4.3.1.3 Neural Network

Neural Network were implemented in Python (Jupyter notebook) using Keras library[6], two hidden layers perceptron model with 30 neurons and regularization was used (Fig 4.18). First, all featrues were normalized. Three combination of the normalized features were used on the models. Optimization of number of epochs were done through train/validation loss and accuracy evaluation (Fig 4.19). However, we notice that the loss and accuracy of the validation keep improving at high iterations number this is believed to be due to the systematic error in the data and that our model start early to overfit, learning some patterns that won't be present in a new well. This bias will be discussed more on chapter5.5.



```
Model: "sequential_7042"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_21124 (Dense)         (None, 30)                330

 dropout_11662 (Dropout)     (None, 30)                0

 dense_21125 (Dense)         (None, 30)                930

 dropout_11663 (Dropout)     (None, 30)                0

 dense_21126 (Dense)         (None, 1)                 31

=================================================================
```

Figure 4.18: Neural network model architecture

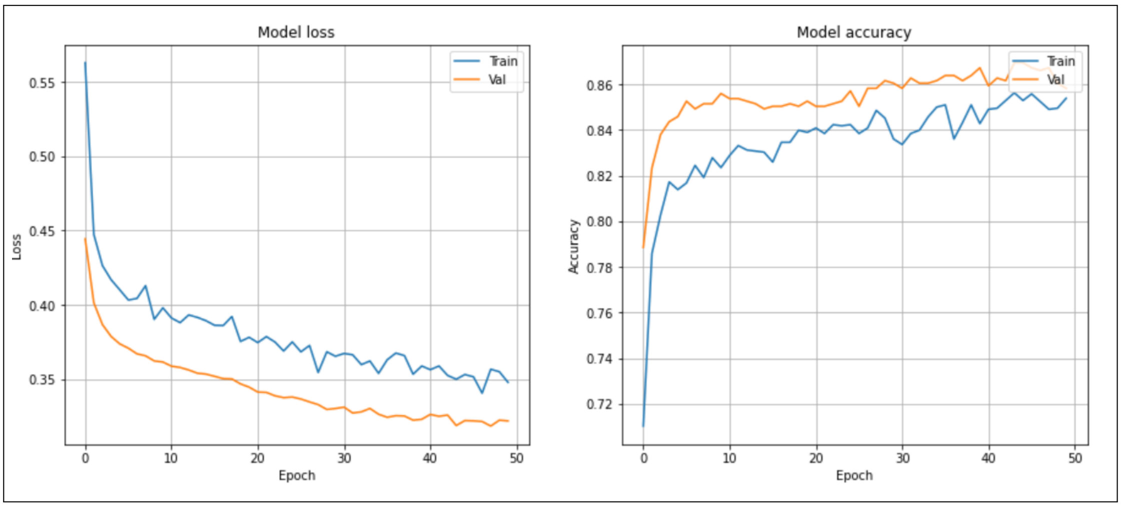
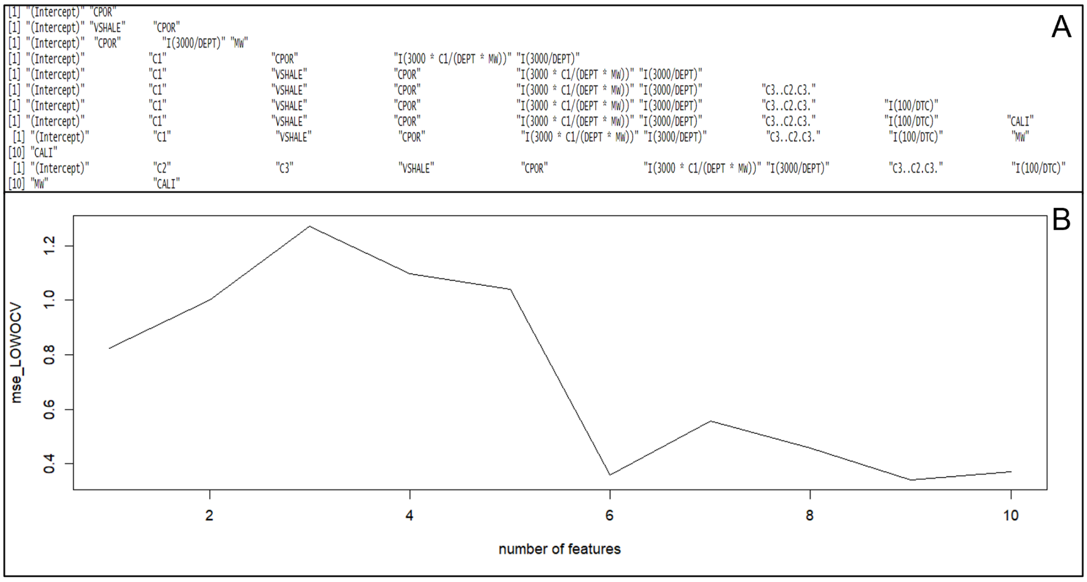


Figure 4.19: Evolution of Loss and accuracy with epochs using the neural network model for both training and validation set showing questionable continuous improvement in the validation set

4.3.2 Permeability prediction task

For permeability prediction, similar techniques with parameter optimization were used as linear prediction using subset selection optimized through cross validation and through LOWOCV mean squared error (Fig 4.20); Lasso and Ridge, Random Forest and boosting, and Neural Network.



```
[1] "(Intercept)" "CPOR"
[1] "(Intercept)" "VSHALE"      "CPOR"
[1] "(Intercept)"  "CPOR"         "I(3000/DEPT)" "MW"
[1] "(Intercept)"              "C1"                       "CPOR"                     "I(3000 * C1/(DEPT * MW))" "I(3000/DEPT)"
[1] "(Intercept)"              "C1"                       "VSHALE"                   "CPOR"                     "I(3000 * C1/(DEPT * MW))" "I(3000/DEPT)"
[1] "(Intercept)"              "C1"                       "VSHALE"                   "CPOR"                     "I(3000 * C1/(DEPT * MW))" "I(3000/DEPT)"             "C3..C2.C3."
[1] "(Intercept)"              "C1"                       "VSHALE"                   "CPOR"                     "I(3000 * C1/(DEPT * MW))" "I(3000/DEPT)"             "C3..C2.C3."               "I(100/DTC)"
[1] "(Intercept)"              "C1"                       "VSHALE"                   "CPOR"                     "I(3000 * C1/(DEPT * MW))" "I(3000/DEPT)"             "C3..C2.C3."               "I(100/DTC)"               "CALI"
 [1] "(Intercept)"              "C1"                       "VSHALE"                   "CPOR"                     "I(3000 * C1/(DEPT * MW))" "I(3000/DEPT)"             "C3..C2.C3."               "I(100/DTC)"               "MW"
[10] "CALI"
 [1] "(Intercept)"              "C2"                       "C3"                       "VSHALE"                   "CPOR"                     "I(3000 * C1/(DEPT * MW))" "I(3000/DEPT)"             "C3..C2.C3."               "I(100/DTC)"
[10] "MW"                       "CALI"
```

Figure 4.20: Subset selection optimization: A) best features for each number of features used, B) mse for LOWOCV evolution with number of features showing that six features provide the best results

4.3.3 Hyperparameter tuning

Some Hyperparameters for each model were hyper tuned. Table4.1 summarizes the tuned parameters for each prediction task.

Table 4.1: Models optimized hyperparameters: * means manual tuning of the hyperparameter while the absence of * means grid search was implemented

	Task-1&2	Task-3
Subset Selection		
Lasso	lambda (shrinkage factor)	lambda (shrinkage factor)
Ridge	lambda (shrinkage factor)	lambda (shrinkage factor)
Random Forest	ntree* (number of trees), criterion: (gini or entropy)	ntree* (number of trees), criterion: (gini or entropy)
XGBoost	nrounds (number of trees), eta* (learning_rate), max_depth*	nrounds (number of trees), eta* (learning_rate), max_depth*
SVM	cost, gamma, degree, kernel	
Neural Network	epochs, number of layers*, batch_size*, regularization parameter*	epochs, number of layers*, batch_size*, regularization parameter*

4.4 Results Assessment

Two methods were used to assess the results of the models: one well test set and leave one well out cross validation. The main metrics were F1 score and balanced accuracy for the classification tasks and R2 for the prediction task. These metrics were implemented using scikit-learn[21].

4.4.1 One well test set

One well was removed from the data prior to training for the classification tasks and used as a test set. This well was chosen to have both positive and negative classes in a way representative of the training distribution.

4.4.2 Leave One Well Out Cross Validation (LOWOCV)

LOWOCV was implemented to have a more realistic assessment of the prediction. In reality, we want to predict the fluid and reservoir properties on a complete new well. A classical train/test split would mix the data from all wells. So, data from the same well would be used for both training and testing. Thus, the testing error would be misleading (underestimating the real error). In fact, the models would be learning patterns that are not real and that would be present also in the test set and this idea was validated by our results (figure5.11).

The LOWOCV is similar to Leave one out cross validation (LOOCV), but instead of leaving one observation we leave one well with all the observations associated to it.

This is more realistic than a random sample split to avoid bias from data from the same well.

Chapter 5

Results

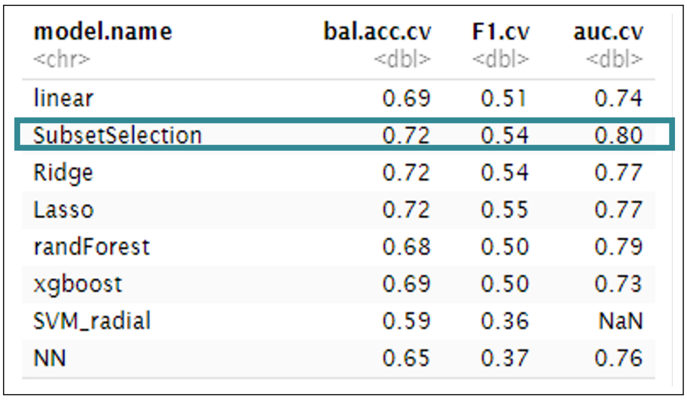
5.1 First task: Water/Hydrocarbon classification

For the first classification (Water/Hydrocarbon classification in reservoir rock), the training set consists of 2952 samples and 281 for the test set corresponding to the well data of the "35/12-2" well. As the data is imbalanced: class-0 (water) has 2323 and class-1 (hydrocarbon) has 676 samples, main metrics used to assess results are F1 score, balanced accuracy and AUC.

5.1.1 Prediction

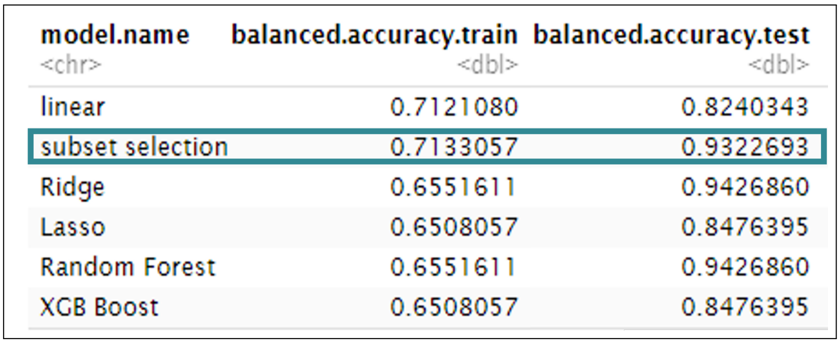
Most of the models provided comparable results. The best model 'logistic regression with subset selection' had a LOWOCV AUC=0.80, F1score=0.54, balanced accuracy=0.72(Table5.1). The same model was the best for the test set with AUC=0.87, F1score=0.82, accuracy=0.96 (Table5.2). The prediction performance were moderate. We could not get a good prediction probably due to probably the presence of low saturation gas in some sand interval which responses are similar to hydrocarbon bearing reservoirs and at this stage we failed to separate these responses with mud gas data.

Table 5.1: Task-1: LOWOCV performance comparison of the ML models showing that subset selection provided the best performance.



model.name	bal.acc.cv	F1.cv	auc.cv
<chr>	<dbl>	<dbl>	<dbl>
linear	0.69	0.51	0.74
SubsetSelection	0.72	0.54	0.80
Ridge	0.72	0.54	0.77
Lasso	0.72	0.55	0.77
randForest	0.68	0.50	0.79
xgboost	0.69	0.50	0.73
SVM_radial	0.59	0.36	NaN
NN	0.65	0.37	0.76

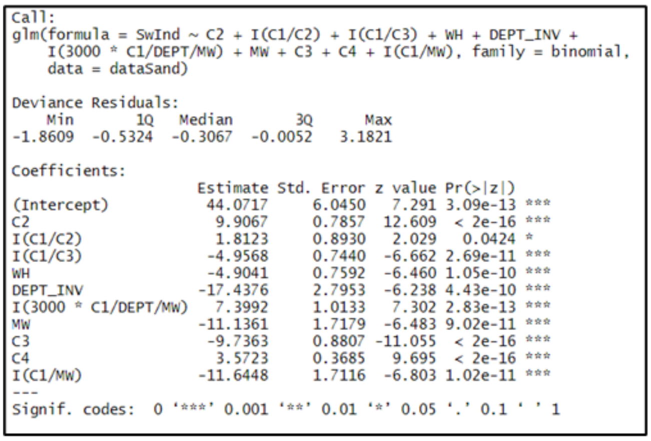
Table 5.2: Task-1 performance comparison: training set and test set balanced accuracy scores showing that subset selection provided the best test set performance.



model.name	balanced.accuracy.train	balanced.accuracy.test
<chr>	<dbl>	<dbl>
linear	0.7121080	0.8240343
subset selection	0.7133057	0.9322693
Ridge	0.6551611	0.9426860
Lasso	0.6508057	0.8476395
Random Forest	0.6551611	0.9426860
XGB Boost	0.6508057	0.8476395

5.1.2 Inference

Subset selection model had removed the following features C1/C4, C1/TG, 1/MW and C1 judged not important for the prediction. As showed in figure 5.1, it looks that most of the remaining features have statistical importance. So, both gas quantities and ratios were important. However, it is still difficult to make some statistical inference due to collinearity between these features. No clear separation using cut offs on ratios2.4.2 were established as done by Pixler 1969[22] and Haworth et al.1985[13].



```
Call:
glm(formula = SwInd ~ C2 + I(C1/C2) + I(C1/C3) + WH + DEPT_INV +
    I(3000 * C1/DEPT/MW) + MW + C3 + C4 + I(C1/MW), family = binomial,
    data = dataSand)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8609  -0.5324  -0.3067  -0.0052   3.1821

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           44.0717     6.0450   7.291 3.09e-13 ***
C2                     9.9067     0.7857  12.609  < 2e-16 ***
I(C1/C2)               1.8123     0.8930   2.029   0.0424 *
I(C1/C3)              -4.9568     0.7440  -6.662 2.69e-11 ***
WH                    -4.9041     0.7592  -6.460 1.05e-10 ***
DEPT_INV             -17.4376     2.7953  -6.238 4.43e-10 ***
I(3000 * C1/DEPT/MW)   7.3992     1.0133   7.302 2.83e-13 ***
MW                   -11.1361     1.7179  -6.483 9.02e-11 ***
C3                    -9.7363     0.8807 -11.055  < 2e-16 ***
C4                     3.5723     0.3685   9.695  < 2e-16 ***
I(C1/MW)             -11.6448     1.7116  -6.803 1.02e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

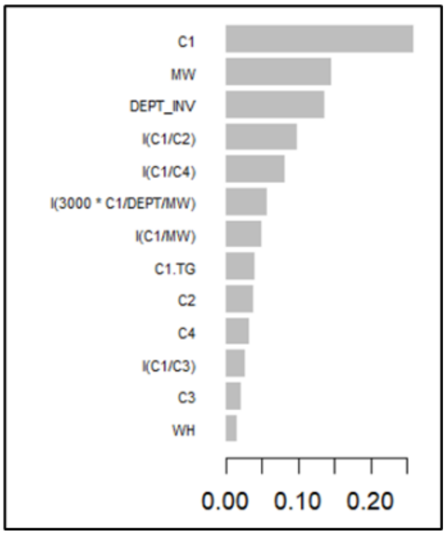


Figure 5.1: Task-1 model inference: subset selection model summary (left) shows remaining important features however inference still difficult due to multicolinearity problem; XGboost features importance (right) shows most important features

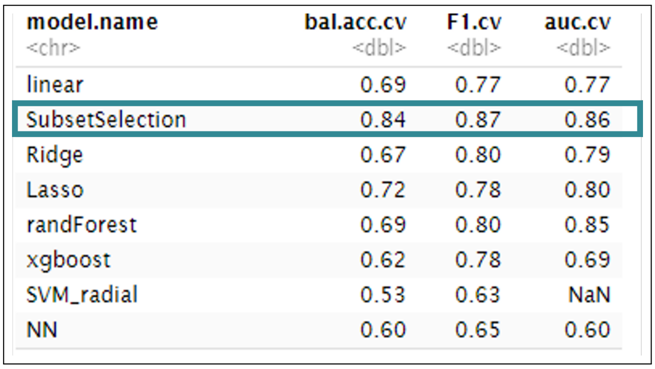
5.2 Second task: Oil/Gas classification

For the second task (Identifying fluid type (Gas or Oil) in hydrocarbon bearing zones), the training set consists of 611 samples and 102 for the test set. Main metrics was F1 score, balanced accuracy and AUC.

5.2.1 Prediction

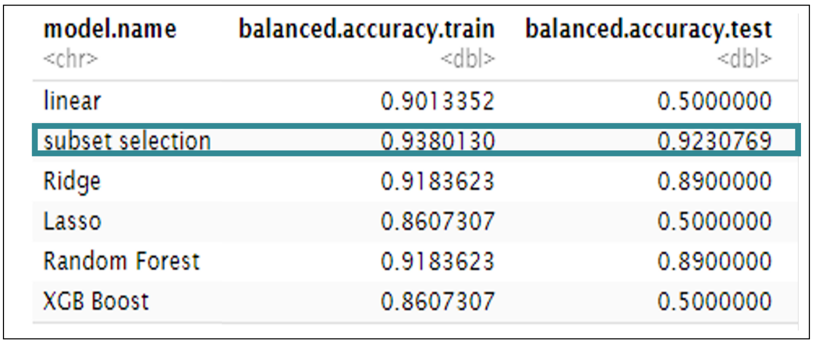
The Oil/Gas classification classification showed much better results than Water/Hydrocarbon classification and allowed for an excellent performance on distinguishing gas and oil. The best model ‘logistic regression with features selection’ had a LOWOCV AUC=0.86, F1score=0.87, balanced accuracy=0.84(Table5.4). Best model for test set provided AUC=0.99, F1score=0.94, accuracy=0.93 (Table5.3).

Table 5.3: Task-2: LOWOCV performance comparison of the ML models showing that subset selection provided the best performance.



model.name	bal.acc.cv	F1.cv	auc.cv
<chr>	<dbl>	<dbl>	<dbl>
linear	0.69	0.77	0.77
SubsetSelection	0.84	0.87	0.86
Ridge	0.67	0.80	0.79
Lasso	0.72	0.78	0.80
randForest	0.69	0.80	0.85
xgboost	0.62	0.78	0.69
SVM_radial	0.53	0.63	NaN
NN	0.60	0.65	0.60

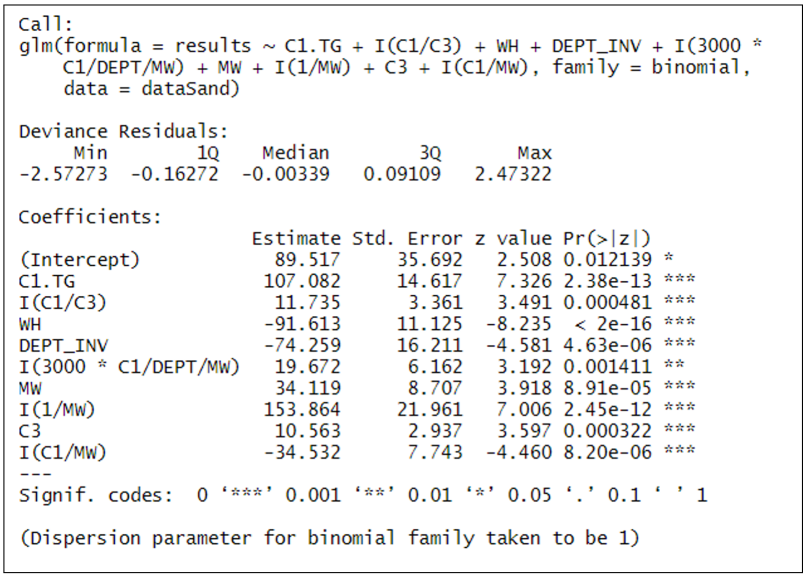
Table 5.4: Task-2 performance comparison: training set and test set balanced accuracy scores showing that subset selection provided the best test set performance.



model.name	balanced.accuracy.train	balanced.accuracy.test
<chr>	<dbl>	<dbl>
linear	0.9013352	0.5000000
subset selection	0.9380130	0.9230769
Ridge	0.9183623	0.8900000
Lasso	0.8607307	0.5000000
Random Forest	0.9183623	0.8900000
XGB Boost	0.8607307	0.5000000

5.2.2 Inference

Subset selection and further manual tests allowed to remove multiple features (C1, C2, C4, DEPT, I(C1/C2), I(C1/C4))5.2. The main features important for the prediction are the multiple gas ratios5.3 which is in alignment with the finding of Pixler 1969[22] and Haworth et al 1985[13]. Thus, an increase of the proportion of the heavy components increases the probability of oil rather than gas. However, using directly Pixler or Haworth cut offs for fluid classification2.4.2 did not allow for good results which shows that these cut offs need calibration to the study area. It was also noticed that gas rates (C3) and mud weigh (MW) are important but collinearity between features make further inference difficult.



```
Call:
glm(formula = results ~ C1.TG + I(C1/C3) + WH + DEPT_INV + I(3000 *
    C1/DEPT/MW) + MW + I(1/MW) + C3 + I(C1/MW), family = binomial,
    data = dataSand)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.57273  -0.16272  -0.00339   0.09109   2.47322

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)            89.517     35.692   2.508 0.012139 *
C1.TG                 107.082     14.617   7.326 2.38e-13 ***
I(C1/C3)               11.735      3.361   3.491 0.000481 ***
WH                    -91.613     11.125  -8.235  < 2e-16 ***
DEPT_INV              -74.259     16.211  -4.581 4.63e-06 ***
I(3000 * C1/DEPT/MW)   19.672      6.162   3.192 0.001411 **
MW                     34.119      8.707   3.918 8.91e-05 ***
I(1/MW)               153.864     21.961   7.006 2.45e-12 ***
C3                     10.563      2.937   3.597 0.000322 ***
I(C1/MW)              -34.532      7.743  -4.460 8.20e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

Figure 5.2: Task-2 subset selection model summary showing that mud gas ratio are important for the fit, from the model coefficients we can infere that an increase of the proportion of the heavy components increases the probability of oil rather than gas..

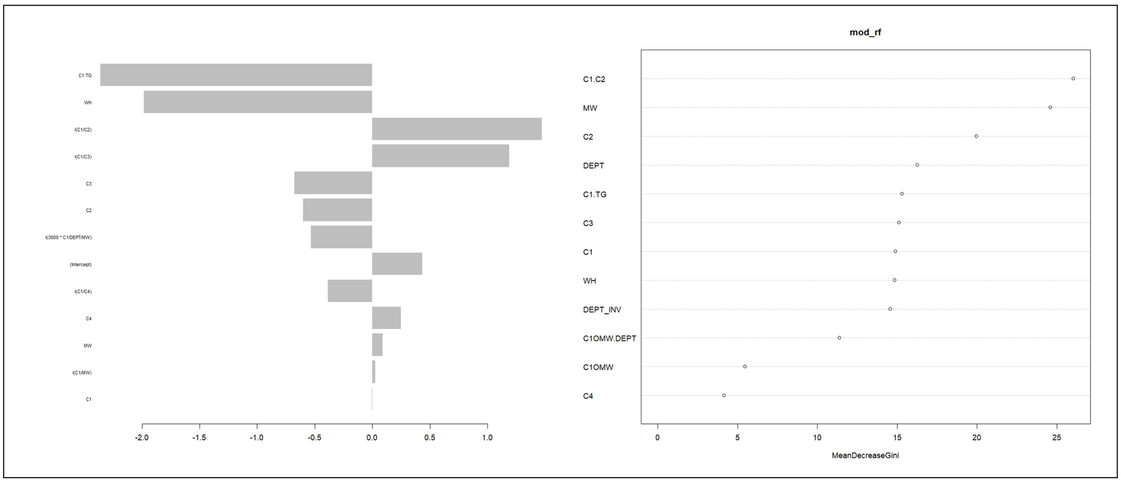


Figure 5.3: Task-2 feature importance from XGBoost(left) and random forest (right), showing, specially XGBoost, that mud gas ratios then mud quantities are the most important features.

5.3 Third task: Permeability Prediction

The third prediction task (predicting permeability in hydrocarbon bearing zones) consists of a sample size of 282 observations. Main metrics used to assess the results is R^2.

5.3.1 Prediction

Combining mud gas data with porosity and shale volume improved the prediction of permeability compared to conventional approach of using only porosity or porosity with shale volume. Linear models performed well on the prediction as subset selection, Ridge and Lasso5.5. The best model was linear regression with subset selection and R^2 of the LOWOCV had improved from 0.65 to 0.85 by adding mud gas features: C1, C1/(DEPTH*MW) and 1/DEPTH5.4.

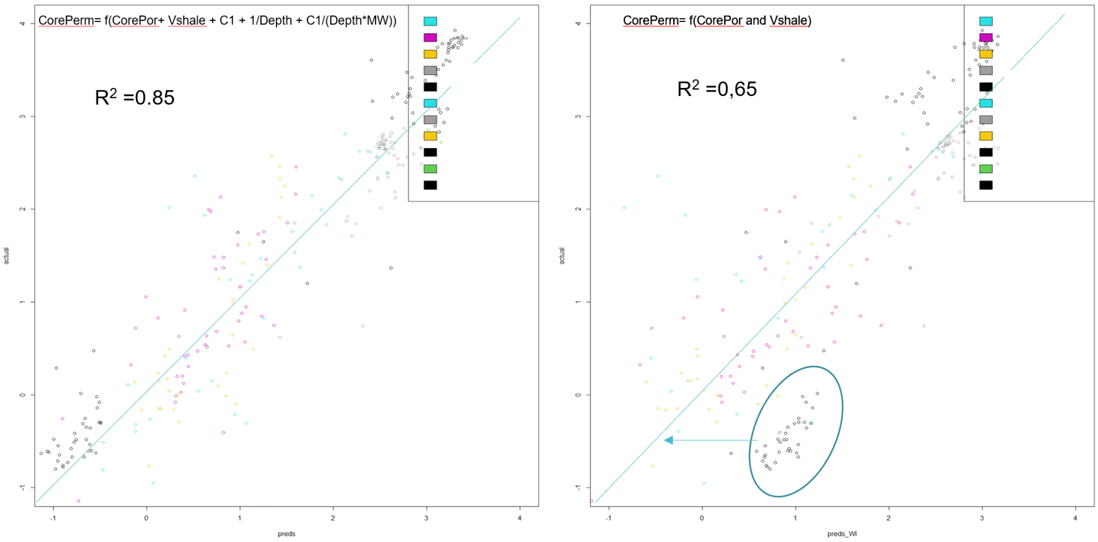
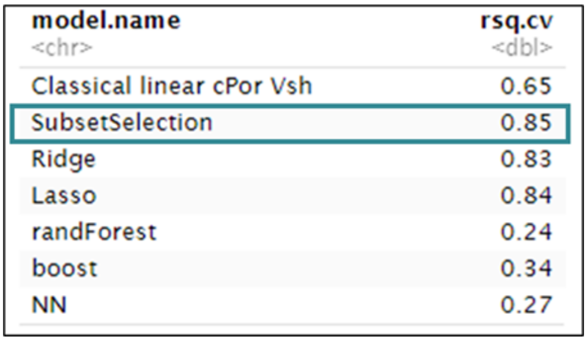


Figure 5.4: Cross plot of actual permeability function of the predicted permeability using linear model: using only porosity and Vshale in the right, using porosity, Vshale, C1, 1/Depth and C1/(MW*Depth) in the left showing an improvement of the LOWOCV R_2 from 0.65 to 0.85, colors represent different wells

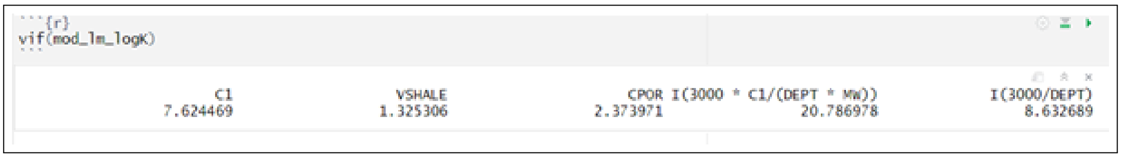
Table 5.5: Task-3: LOWOCV performance comparison of the ML models showing that subset selection provided the best performance.



model.name <chr>	rsq.cv <dbl>
Classical linear cPor Vsh	0.65
SubsetSelection	0.85
Ridge	0.83
Lasso	0.84
randForest	0.24
boost	0.34
NN	0.27

5.3.2 Inference

Some of the generated features as C1/(DEPT*MW (or C1/MW) led to an interesting improvement of the prediction. These features could be considered as corrected gas rates to the mud weight and depth and needs further investigation. As seen in the subset selection results (Fig5.6), porosity, as expected, is the most important feature and it is the one that should be used if we want to fit a linear regression with one feature, followed by Vshale. Then, the model starts to see the importance of mud gas data, especially (C1) and (C1) corrected by MW and depth. ANOVA(Fig5.7) and the improvement of the prediction on the LOWOCV5.4 show the importance of these features. In addition, we succeeded to reduce the number of features that we can have a robust model and some inference by reducing the collinearity as max VIF is around 20 (Fig5.5).

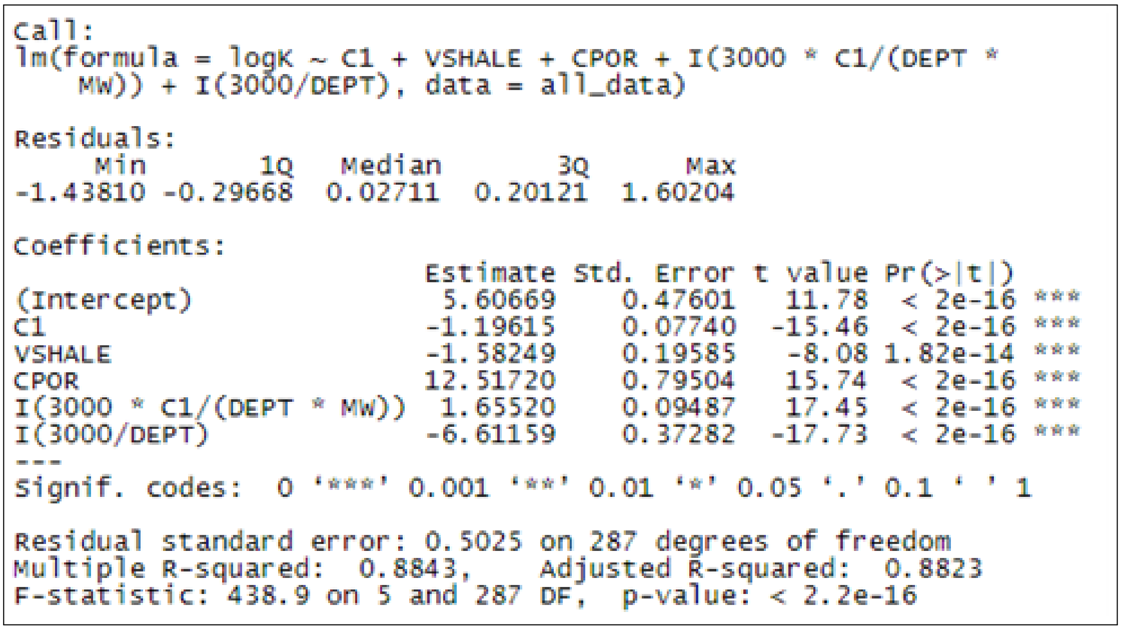


~~~
```{r}
vif(mod_lm_logK)
```

                C1             VSHALE              CPOR I(3000 * C1/(DEPT * MW))          I(3000/DEPT)
          7.624469           1.325306          2.373971                20.786978              8.632689
~~~

Figure 5.5: VIF of permeability prediction subset selection model features showing a very big improvement on the multicolinearity problem by reducing the VIF allwing for more stability and interpretabaility.

As expected, an increase of core porosity and a decrease of Vshale would tend to increase the permeability. Figure5.6 shows that if depth*MW $<$ 4125, we have an increase of C1 associated with an increase of permeability and the opposite effect if depth*MW$>$4125 (high depth and high MW). This could be due to the overpressure that we have at high depths in this area. So, that even with low permeability we have higher gas quantities. A way forward for better prediction is to try to include pressure data (Figure 2.14) as features.



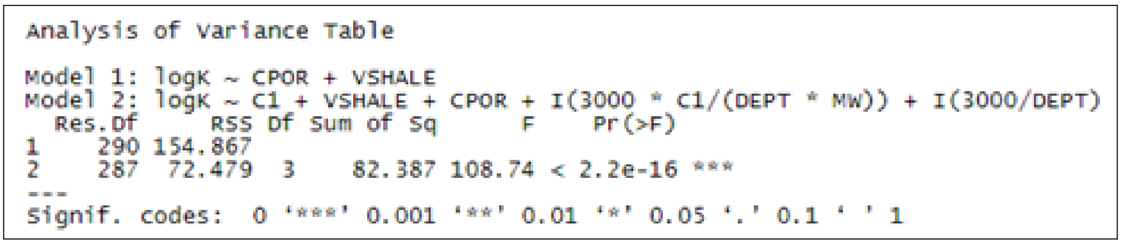
```
Call:
lm(formula = logK ~ C1 + VSHALE + CPOR + I(3000 * C1/(DEPT *
    MW)) + I(3000/DEPT), data = all_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43810 -0.29668  0.02711  0.20121  1.60204

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                5.60669    0.47601   11.78  < 2e-16 ***
C1                        -1.19615    0.07740  -15.46  < 2e-16 ***
VSHALE                    -1.58249    0.19585   -8.08 1.82e-14 ***
CPOR                      12.51720    0.79504   15.74  < 2e-16 ***
I(3000 * C1/(DEPT * MW))  1.65520    0.09487   17.45  < 2e-16 ***
I(3000/DEPT)              -6.61159    0.37282  -17.73  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5025 on 287 degrees of freedom
Multiple R-squared:  0.8843,    Adjusted R-squared:  0.8823
F-statistic: 438.9 on 5 and 287 DF,  p-value: < 2.2e-16
```

Figure 5.6: Peremability prediction subset selection model summary showing that an increase of core porosity and a decrease of Vshale would tend to increase the permeability. In addition, if depth*MW < 4125 we have an increase of C1 associated with an increase of permeability and the opposite effect if depth*MW>4125



```
Analysis of Variance Table

Model 1: logK ~ CPOR + VSHALE
Model 2: logK ~ C1 + VSHALE + CPOR + I(3000 * C1/(DEPT * MW)) + I(3000/DEPT)
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    290 154.867
2    287  72.479  3    82.387 108.74 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5.7: Permeability prediction ANOVA for the model with using CPOR and VShale, and the subset selection model showing the statistical imprtance of the addition mud gas features.

5.4 Petrel visualization

All the predictions were loaded to Petrel. This allowed to compare the results by well while displaying any logs to study more the causes of errors. Also, Petrel is the software used by geoscientist for the interpretation, so they could include these results in their interpretation workflows.
Figures 5.8, 5.9 and 5.10 show Petrel visualization of the prediction and the actual response using the best method for each of the three tasks

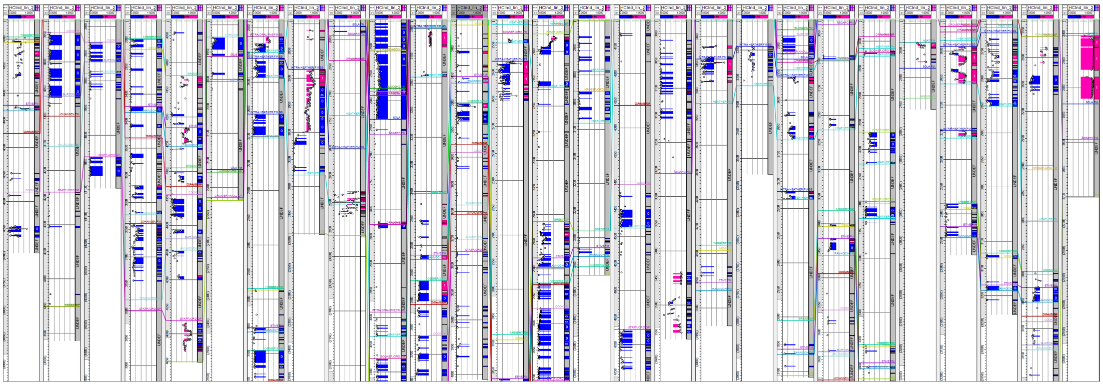


Figure 5.8: Task-1 Petrel visualisation of the classification results: for each wells two track are displayed (predicted HCInd probability from best model (track-1) and true HCind response (track-2)). Blue represent water and pink hydrocarbon. Results quality are moderate

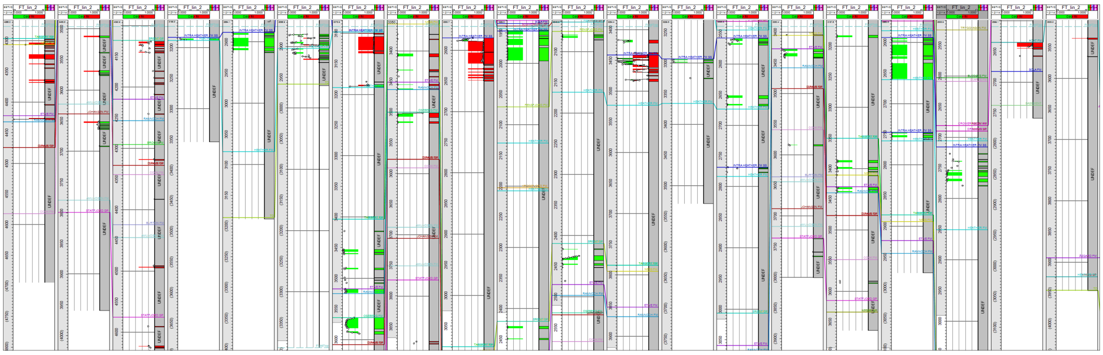


Figure 5.9: Task-2 Petrel visualisation of the classification results for each wells two track are displayed (predicted FluidType probability from best model (track-1) and true FluidType response (track-2)) showing excellent fit. Green represent oil and red gas

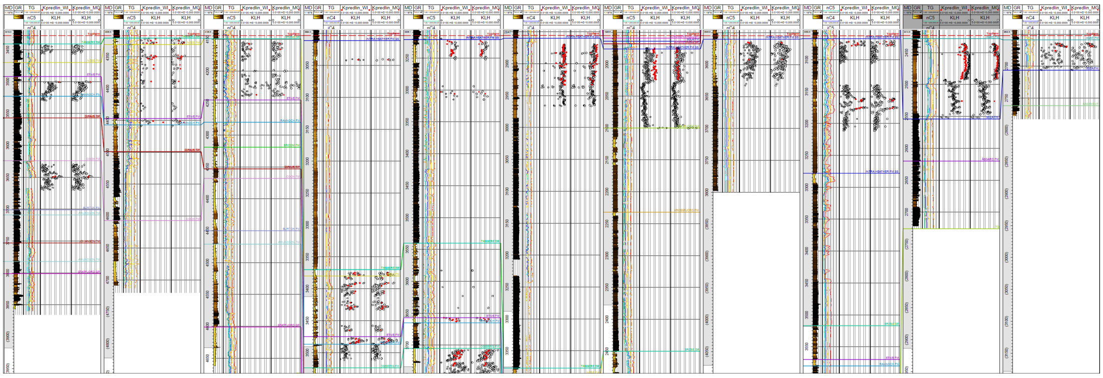


Figure 5.10: Task-3 Petrel visualisation of the regression results: for each well three track are displayed (Track1: GR log, Track-2: Real permeability (black points) and predicted permeability from standard approach of linear model with porosity and Vshale (red points), Track3: Real permeability (black points) and predicted permeability from linear model with porosity, Vshale, and mud gas data (red points) showing a better fit from the model with mud gas data in some wells, generally the results are comparable

5.5 General comments

The linear models had outperformed the nonlinear models. This is probably due to the existence of true linear physical relation between the features and the response as we know for example that log of permeability has a linear relation to the porosity and that rate and permeability has a linear relation.

We think that due to some limitations of the mud gas data2.4.3, the data suffers from experiment bias, a systematic error due to different tools used and measurements from well to well that affected more the nonlinear models. In fact, the results from simple randomly sampled cross validation or validation set were different and much better, than from LOWOCV, specially using nonlinear methods. Due to this bias, the non linear methods as they are by construction more complex, tend more easily to overfit the data by trying to explain this systematic noise than the linear models. The low number of data points enhances further this problem.
For example, figure5.11 shows that for the neural network model the data fits very well both training and validations set, but not the LOWOCV. In fact, The mse of the validation set keep decreasing after 50 iterations and have similar level to the training set mse, while the mse from the LOWOCV is more noisy, starts to increase again after only 20 iterations and its level is not comparable to the mse level of the training set. In fact, this could explained by the fact that quickly the network start to learn patterns only relevant for the training and validation set, but not for the LOWOCV. It is starting

to overfit and learning from the systematic noise that is present in both validation and training set but not if we consider a complete new well as sed by the LOWOCV.

To solve this issue further work, need to be done on sampling and on the features, may be trying to remove background trends as done in time series. We could also implement parameters optimization as grid search on the LOWOCV technique to better optimize the parameters of the models. . . .

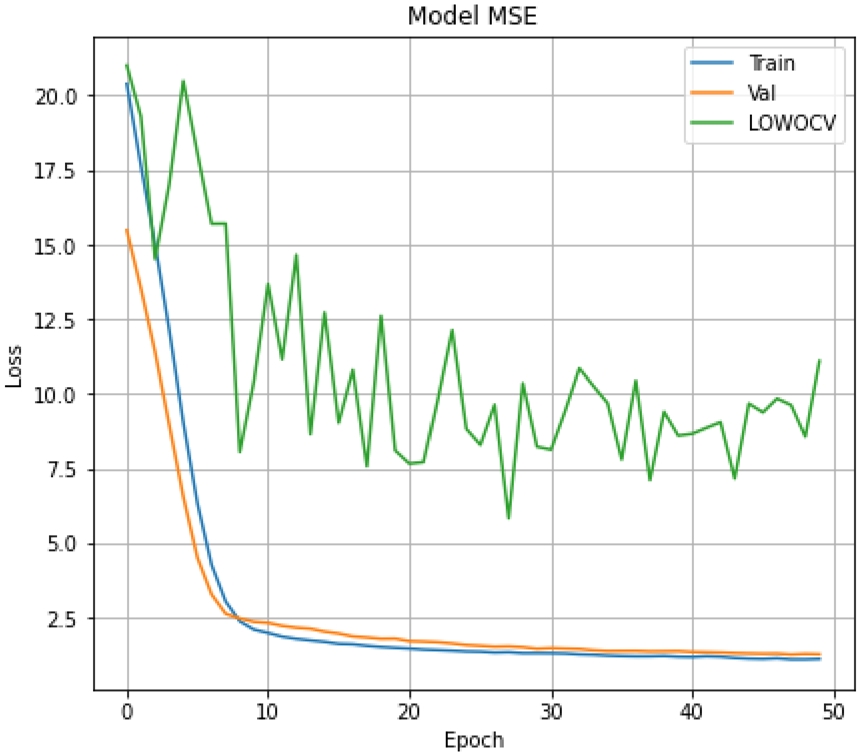


Figure 5.11: Evolution of mse with epochs using the neural network model for the permeability prediction for train set, validation set and LOWCV showing that the data fits very well both training and validations set, but not the LOWOCV.The mse of the validation set keep decreasing after 50 iterations and have similar level to the training set mse, while the mse from the LOWOCV is more noisy and starts to increase again after only 20 iterations.

5.6 Limitations and potential improvements

As shown in this chapter, we succeeded to have an excellent prediction in task-2 (oil and gas classification) and task-3 permeability prediction. Task-1 (water and hydrocarbon classification) prediction had moderate results. We made some inference and explanations on mud gas features and how it relates to the response variable. However, we noticed some limitations:

- Number of sample points and wells are small making difficult to generalise these findings, a way forward is using more data and more wells,
- Various sampling rates, background trends, scale and experiment bias in the data are affecting the different models making learning real and relevant physical behavior more difficult. Collecting more data and generating more standardized features between wells may improve the learning capacity of the models. We can consider removing background trend, normalize features, calibrating with drilling parameters and mud type ...
- Due to the systematic bias 4.15, hyperparameters tuning on validation set or cross validation set was not optimal for optimizing the models as these hyperparameters will be affected by this noise. A better way would be to use the LOWOCV method to optimize these hyperparameters instead and implement a grid search.
- Causality assessment and evaluation of the different model results allowed to determine more features that could improve the models as reservoir pressure for permeability prediction. A way forward is to collect these data and update the models with these features.
- Traditional mud gas systems is limited to a high resolution (C3) quantities and to (C5) with lower confidence. However, advanced mud gas systems allow now for more accurate quantification of gas quantities from Methane (C1) to Octane (C8) which would unlock more potential from mud gas in the future.

Chapter 6

Conclusion

Mud gas data are continuous measurements of the different compounds of the gas released from the formation while drilling. Due to the difficulties of manual interpretation of these data, their use has always been limited.

The focus of this study was to use big data analytics to extract insights from mud gas data that helps petroleum system analysis. We specially focused on three predictive tasks:

1. Identifying hydrocarbon bearing zones in the reservoir rock
2. Identifying fluid type (Gas or Oil) in hydrocarbon bearing zones
3. Estimating permeability in hydrocarbon bearing zones

In this study, we succeeded in generating robust and stable ML models for these three predictive tasks. The reservoir fluid prediction (task-1 and task-2) would allow assessment of the rock fluid type while drilling allowing to decrease cost and to have better decision on data acquisition program. The permeability prediction task would allow better prediction and understanding of permeability, allowing ultimately for better informed development decision and decrease data acquisition costs (core, pressure points).

Multiple ML models were used for the different prediction tasks including Linear models with subset selection, Lasso, Ridge, Random Forest, Boosting, SVM and Neural Network. Parameters' tuning was done to optimize these models. The linear model after subset selection showed the best performance in the three tasks. It showed a moderately good prediction for the classification of water and hydrocarbon samples, a very good prediction of distinguishing between oil and gas, and a very good performance on permeability prediction, better than conventional approach of using just a poro-perm relationship. The linear models outperforming the nonlinear models were attributed to the existence of true linear physical relation between the features and the response as we know for example that log of permeability has a linear relation to the porosity and that rate and permeability has a linear relation. Also, we think due to the small number of wells (also data points) that the data suffers from experiment bias, a systematic error due to different tools, measurements and drilling parameters from well to well that affected more the nonlinear models. In fact, very low error on test set was observed using random sample training and test set split.

To reduce the effect of this systematic error on choosing the best model, we developed a new validation procedure consisting of leaving one well out cross validation allowed to have more realistic assessment of the model metrics as this would remove sampling and experiment bias from the validation set.

As the linear model with best or stepwise subset selection showed the best performance in the three tasks, it also allowed some inference. Despite the high collinearity of the mud gas, the subset selection method allowed to eliminate multiple features enhancing at some degree the interpretability of the relations. Thus, for the second classification task for example, we validated the importance of mud gas ratios for the prediction of fluid type as established first by Pixler 1969[22]. Thus, an increase of the proportion of the heavy carbon components increases the probability of oil rather than gas. In addition, we noticed other features such as mud weight and depth which are important for the prediction. For the permeability prediction, some of the generated features such as C1/MW and C1/(DEPT*MW) gave an interesting improvement to the prediction. These new features were considered as corrected gas rates to the mud weight and depth, and needs further investigation.

The different predictions were then exported for each well and imported to the main geological interpretation software ‘Petrel’, allowing for better visualization of the results. In addition, as the best model for the three tasks were linear prediction with features selection, we were able to easily export the model to petrel through writing the formula in the software which would allow to test the results on more wells.

In this project, we showed that data analytics and machine learning unlocked some of the potential of mud gas data that previously had very limited uses by predicting some fluid and reservoir properties. But we are still scratching the surface.

Multiple ways forward exist. We can try to generate standardized features from mud gas data that would be less affected by bias for example removing gas background trend from the mud gas logs and correcting with drilling parameters and mud types. We can also use more features such as reservoir pressure (Pr) which could be important for predicting permeability. We can collect and use more data from more wells and handle better missed values. In addition, we can improve the parameter tuning of the models by implementing a LOWOCV with grid search to generate more accurate models. To simplify visualization, handling of the data and allow geoscientists to take real time decisions, a web base application could be developed to run directly the models in Petrel or TechLog. Moreover, other properties could be used a response variable as GOR, porosity,...

Furthermore, new advances in mud gas systems allow now for more accurate quantification of gas quantities from Methane (C1) to Octane (C8) which would unlock more potential from mud gas in the future.

Bibliography

[1] F. Anifowose and M. Mezghani. Mud gas data could reveal a wealth of reservoir information. *Journal of petroleum technology*, October 2021. url: https://jpt.spe.org/mud-gas-data-could-reveal-a-wealth-of-reservoir-information.

[2] G.E. Archie. The electrical resistivity log as an aid in determining some reservoir characteristics. *Society of Petroleum Engineers*, December 1942.

[3] P. Bhandari. Random and systematic error | differences, sources examples, may 7,2021. url: https://www.scribbr.com/methodology/random-vs-systematic-error/.

[4] B. Biju-Duval. *sedimentary geology: Sedimentary Basins, Depositional Environments, Petroleum Formation.* Technip, 2002.

[5] L.M. Blumberg. *Gas Chromatography.* Elsevier, 9 edition, 2012. doi: https://doi.org/10.1016/B978-0-12-385540-4.00002-X.

[6] Francois Chollet et al. Keras, 2015.

[7] J. Craig and F. Quagliaroli. The oil gas upstream cycle: Exploration activity. *EPJ Web of Conferences*, 246:00008, 01 2020.

[8] T. Darling. *Well Logging and Formation Evaluation.* Elsevier, 1 edition, February 2005.

[9] J.W. Emerson, W.A. Green, B. Schloerke, J. Crowley, D. Cook, H. Hofmann, and H Wicham. The generalized pairs plot. *Journal of Computational and Graphical Statistics*, 55:79–91, 2012.

[10] J. Erzinger, T. Wiersberg, and M. Zimmer. Real-time mud gas logging and sampling during drilling. *Geofluids*, 6(3):225–233. doi: https://doi.org/10.1111/j.1468-8123.2006.00152.x.

[11] N.P. Giovanni and J.L. Douglas. Advances in mud gas interpretation whilst drilling. *SPWLA 49th Annual Logging Symposium*, May 25-28, 2008.

[12] F. Gradstein, E. Anthonissen, H. Brunstad, M. Charnock, O. Hammer, T. A. Hellem, and K. Lervik. Norwegian offshore stratigraphic lexicon (norlex). *Newsletters on Stratigraphy*, 44:73–86, 10 2010.

[13] J.H. Harworth, M.P. Sellens, and A Whittaker. Interpretation of hydrocarbon shows using light (c1-c5)hydrocarbon gases from mudlog data. *AAPG Bulletin*, 69 (08):1305–1310, 1985. doi: https://doi.org/10.1306/AD462BDC-16F7-11D7-8645000102C1865D.

[14] Witten D. Hastie T. James, G. and R. Tibshirani. *An introduction to statistical learning*. Springer, 2 edition, 2013.

[15] kinverarity. Lasio. url: https://lasio.readthedocs.io/en/latest/.

[16] S. Liqiang, W. Feng, M. Jianhai, F. Guoqing, and Y. Hang. Quantitative calculation of gor of complex oil-gas-water systems with logging data: A case study of the yingdong oil/gas field in the qaidam basin. *Natural Gas Industry B*, 1:172–177, 2014.

[17] C. McPhee, J. Reed, and I. Zubizarreta. *Developments in Petroleum Science*, volume 64. Elsevier, 2015.

[18] S. Moshood. *Petroleum Engineering: Principles, Calculations, and Workflows*. American Geophysical Union, 1 edition, 2018. url: https://ebookcentral-proquest-com.ezproxy.uis.no/lib/uisbib/detail.action?docID=5528446.

[19] University of Oslo. Porosity logs. url: https://www.uio.no/studier/emner/matnat/geofag/GEO4250/v08/ undervisningsmateriale/Lectures/.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[22] B. O. Pixler. Formation evaluation by analysis of hydrocarbon ratios. *J Pet Technol , SPE-2254-PA*, 21:665–670, 1969. doi: http://dx.doi.org/10.2118/2254-PA.

[23] M. Rowe and D. Muirhead. Mud-gas extractor and detector comparison. *SPE-188068-MS*, April 2017. doi: https://doi.org/10.2118/188068-MS.

[24] O. Serra. *Fundamental of well-log interpretation 1. the acquisition of logging data*, volume 15A. Elsevier, 3 edition, 1988.

[25] Shlumberger, 1998. url: https://glossary.oilfield.slb.com/.

[26] Shlumberger, 1998. url: https://www.slb.com/resource-library/oilfield-review/defining-series/defining-permeability.

[27] P. Tan, M. Steinbach, A. Karpatne, and V. Kumar. *Introduction to data mining.* Pearson, 2 edition, 2019.

[28] S. Theodoridis. *Machine learning : a Bayesian and optimization perspective.* Elsevier, 2 edition, 2020.

[29] T. Yang, I. Areif, M. Nieman, and et al. A machine learning approach to predict gas and oil ratio based on advanced mud gas data. *SPE-195459-MS*, 2019. doi: https://doi.org/10.2118/195459-MS.

[30] T. Yang, R. Basquet, A. Callejon, Van R., Joost J., and Bartusiak B. hale pvt estimation based on readily available field data. *Unconventional Resources Technology Conference*, August 2014. doi: https://doi.org/10.15530/URTEC-2014-1884129.

[31] T. Yang, Yerkinkyzy G., Uleberg K., and Arief I.H. Predicting reservoir fluid properties from advanced mud gas data. *SPE-201635-PA, SPE Res Eval Eng*, 24 (02):358–366, October 2020. doi: https://doi.org/10.2118/201635-MS.

[32] D.S. Yeung, I. Cloete, D. Shi, and W.W.Y. Ng. *Sensitivity Analysis for Neural Networks.* Springer, 1 edition, 2009.