A MACHINE LEARNING DATA-DRIVEN MODEL FOR BHP ESTIMATION IN MULTI-FRACTURED HORIZONTAL SHALE OIL WELLS

Agustín Terminiello[1], Matías Nasca[2], Juan Filipich[3], Donald Mc Intyre[4], Pablo Crespo[5]

1: Pan American Energy SC. ATerminiello@pan-energy.com

2: Pan American Energy SC. MNasca@pan-energy.com

3: Pan American Energy SC. JFilipich@pan-energy.com

4: Pan American Energy SC. DMcIntyre@pan-energy.com

5: Pan American Energy SC. PCrespo@pan-energy.com

Palabras clave: Bottom Hole Pressure, Vertical Lift Performance Curve, Machine Learning

ABSTRACT

Una correcta estimación de la presión de fondo del pozo es crítica para una adecuada evaluación de cualquier tipo de reservorio Shale no convencional, ya que provee los datos de entrada necesarios para la construcción de gráficos diagnóstico de producción (como, por ejemplo, el gráfico de Flujo Lineal y el gráfico Log-Log). Asimismo, permite el análisis de transientes de caudal (RTA) y de los modelos de simulación numérica. Los sensores de presión de fondo permanentes no suelen ser comunes en campos maduros en donde haya culminado la fase exploratoria y, por esta razón, el ingeniero de reservorios debe emplear diferentes métodos para estimar la presión de fondo a partir de datos de producción y de presión de boca de pozo. Además, se deben correr gradientes dinámicos frecuentemente para lograr estimaciones de pérdida de carga a lo largo de la cañería. Para el caso de pozos con sistema de elevación artificial, este cálculo puede no ser del todo simple y, por lo tanto, información adicional acerca del sistema de bombeo será necesaria para la estimación de la presión de fondo.

El objetivo del presente trabajo es construir un modelo de Machine Learning que permita predecir la presión de fondo en pozos horizontales fracturados de la Fm. Vaca Muerta en Argentina. Las variables de entrada para el entrenamiento del algoritmo incluyen la ubicación geográfica del pozo y del yacimiento, la zona navegada en el reservorio, la historia de producción, los parámetros del sistema de elevación artificial, el diseño de completación y propiedades del fluido de reservorio.

INTRODUCTION

The bottom hole flowing pressure (BHP) for any oil well can be obtained either from a direct measurement, or from an indirect calculation from production rates and well head pressure (WHP). For the former case, a down hole permanent gauge must be attached to the tubing to measure the bottom hole pressure versus time in a continuous trend for the given depth where the pressure sensor has been set. For the latter, a pressure drop equation has to be solved in order to relate well head and bottom hole pressures using surface production rates, fluid properties and tubular geometry.

From the first law of thermodynamics, the following equation can be obtained to calculate the pressure drop for a fluid flowing between two points (from A to B) along a tubing of length L, internal diameter D, and height Δh:

$$\Delta P = P_A - P_B = g\delta\Delta h + \frac{\delta}{2}\Delta v^2 + f_D\frac{L}{2D}\delta v^2$$

where:

ΔP is the pressure drop generated when the fluid goes from point A to point B

g is the gravitational acceleration

δ is the fluid density

v is fluid velocity along the tubular

f_D is the Darcy friction factor that depends on the Reynolds Number (N_{Re}) and the tubular roughness.

The first term of the equation represents the gravitational pressure drop that is experienced in the wellbore as the fluid ascends from the reservoir to the well head. For the case of horizontal wells, Δh can be considered from the landing point to the well head. The second term stands for the pressure drop due to kinetic energy loss which in most cases can be considered negligible for oil wells. Finally, the third term is the Darcy-Weisbach equation and it represents the pressure drop associated to friction between the fluid and the tubular walls. This term involves the Darcy friction factor which can be estimated using empirical correlations. The fluid velocity is involved in this term as well as the tubular inner diameter which together with the fluid properties (oil density and viscosity) will determine the Reynolds number and therefore the flow regime which in most cases will be turbulent (N_{Re}>4000).

It is noteworthy that the given equation is only valid for a single phase liquid flow system where no gas nor water flows along the tubing. For a multi-fractured horizontal well, this is only the case where the well head pressure is above the bubble point and so there is no gas liberation inside the wellbore. Additionally, the water rate must also be zero (or very close to zero) for the flow rate to be considered "single-phase". Therefore, this should be the case of a horizontal well where the flowback back period has ceased, the water cut is negligible and so, a single phase flows from the reservoir to the well head.

It should be pointed out that for many shale reservoirs the flowback period can be very extensive and water cut values remain as high as 10 or 20 per cent for some months since the time the well was put on production. Additionally, as the reservoir is produced and depleted, the well head pressure will go below the bubble point and gas will be liberated at some point on its way to the surface. As a result, a multiphase flow model will be needed to predict the bottom hole pressure of the well. This model will have to account for oil, gas, and water flow along the tubing as well as for the non-equal phase velocities and the slippage effect (or liquid holdup).

During multiphase flow inside the tubing, the lighter phase will move faster than the denser phase which will "slip down" towards the bottom of the wellbore. This phenomenon results in non-equal phase velocities as well as in two different flow directions inside the well. The denser phase will start going up from the reservoir towards the well head together with the lighter phase until some point in the tubing where it "slips down" and inverts its flow direction towards the bottom. The determination of the value of the liquid holdup is critical for pressure drop calculations in multiphase models and it can be obtained from slippage correlations.

Multiphase models which consider the slippage effect are called separated flow models and start by determining the flow regime that exists inside the tubular. Later on, the slippage effect will be quantified using specific correlations according to the selected flow regime.

The Duns and Ross flow map easily identifies four types of flow regimes in gas-liquid flows which depend on the individual velocities of each phase: bubble flow, plug flow, slug flow and mist flow. As the gas velocity increases for a given liquid velocity, the flow regime progresses from bubble flow to plug, slug, and finally to mist flow where gas becomes the continuous phase with liquid droplets entrained in it.

A great disadvantage of separated flow models is the great difficult to code them and deploy them on computer programs or production dashboards as most flow correlations are presented in graphic form. A common procedure to estimate the bottom hole pressure of a well where no down hole permanent gauge is available, is to build a vertical lift performance curve (VLP) that can make accurately predictions based on production rates and well head pressure. The first step would be to run a dynamic gradient (DG) in the wellbore in order to measure pressure at different depths inside the tubing. This results can later on be loaded on a production software where different multiphase correlations can be used to match the real data and obtain a VLP curve for the given wellbore. Finally, using these lift curves, well head pressures can be converted into bottom hole pressures using as input variable: production rates, PVT properties and well geometry. Each well will have its own VLP and so dynamic gradients would have to be run periodically in all the wells of the field.

Nevertheless, this process can be time-consuming specially when the BHP wants to be estimated for quick diagnostic analysis on production dashboards for different wells from different fields.

The objective of this work is to train a Machine Learning model that can easily be deployed on production dashboards as well as on reservoir modelling software for the automatic estimation of the BHP taking its input from in-house production data bases and other real time surface measurements. The novelty of this solution is the simplicity to compute and deploy, and for that reason, to serve as a handy tool for reservoir and production engineers in their early-time diagnostic analysis and production monitoring.

The model was trained using a certain group of variable that will be discussed on the next section. As a result, the BHP can be expressed as a function "f" which represents the calibrated

Machine Learning algorithm:

$$BHP = f(x_1, x_2, x_3 \dots x_n)$$

The variables x_1, x_2, x_3,…,x_n represent the necessary input for the Machine Learning predictive model as well as the set of variables with which this algorithm has originally been trained.

TRAINING DATA SET PREPARATION

To build the Training Data Set for the Machine Learning algorithm, a group of 32 multi-fractured horizontal wells of the Vaca Muerta formation were selected.

The selection criteria were the following:

- Well type: multi-fractured horizontal.
- Production fluid: oil.
- Reservoir: Vaca Muerta formation (Neuquina Basin, Argentina).
- In case of a flowing well: at least one dynamic gradient run on the well.
- In case of a well with an artificial lift system (ALS): either one dynamic gradient run on the well during the flowing period and/or one dynamic gradient after the ALS has been installed.

The following table shows all 32 selected wells:

Well	Field	Landing Zone	Type of ALS installed	# DG during flowing period	# DG during ALS period
#1	A	Z1	None	2	
#2	A	Z1	None	2	
#3	A	Z1	None	1	
#4	A	Z1	None	1	
#5	A	Z1	None	1	
#6	A	Z2	None	1	
#7	A	Z1	None	1	
#8	A	Z1	None	1	
#9	A	Z1	None	1	
#10	B	Z1	None	1	
#11	B	Z1	None	1	
#12	B	Z1	None	1	
#13	C	Z1	None	1	
#14	C	Z1	None	1	
#15	D	Z2	Gas Lift	3	2
#16	D	Z1	Gas Lift	3	3
#17	D	Z1	Gas Lift	6	1
#18	D	Z1	Gas Lift	x	2
#19	D	Z1	Rod Pumping System	2	5
#20	D	Z1	Gas Lift	x	2
#21	D	Z1	Gas Lift	4	4
#22	D	Z1	Gas Lift	2	3
#23	D	Z1	Gas Lift	2	5
#24	D	Z1	Gas Lift	1	6
#25	D	Z2	Rod Pumping System	3	x
#26	D	Z2	Gas Lift	4	5
#27	D	Z2	Gas Lift	1	3
#28	D	Z2	Rod Pumping System	3	x
#29	D	Z1	Gas Lift	4	6
#30	D	Z1	Gas Lift	x	4
#31	D	Z1	Gas Lift	4	5
#32	D	Z1	Gas Lift	6	1

Table 1. The 32 multi-fractured horizontal wells of the sample.

As it can be observed on table 1, wells have been selected from 4 different oil fields (A, B, C and D). All wells were drilled in the Vaca Muerta formation and navigate either the landing zone Z1 or Z2. While Z1 is located towards the base of the formation, Z2 is located at an average vertical distance of 80 m over Z1.

There are 14 flowing wells and 18 wells where an artificial lift system (either Rod Pumping or Gas Lift) has been installed. All flowing wells have at least one dynamic gradient while wells with artificial lift have at least one dynamic gradient during the flowing period and/or one dynamic gradient after the ALS has been installed.

Dynamic gradients run during the flowing period of each well were uploaded to a production software together with the reservoir fluid PVT properties and tubular geometry. Each field and each landing zone have their own reservoir fluid type and PVT analysis. In all cases, a multiphase production correlation was found to successfully match real data and hence, a vertical lift performance curve was generated for each well. For the case of wells with more than one dynamic gradient, the production correlation with the best match to real data was used for the analysis.

The following figure shows the dynamic gradient for Well #3 together with the production correlation of Beggs and Brill that was chosen to best match the real data:

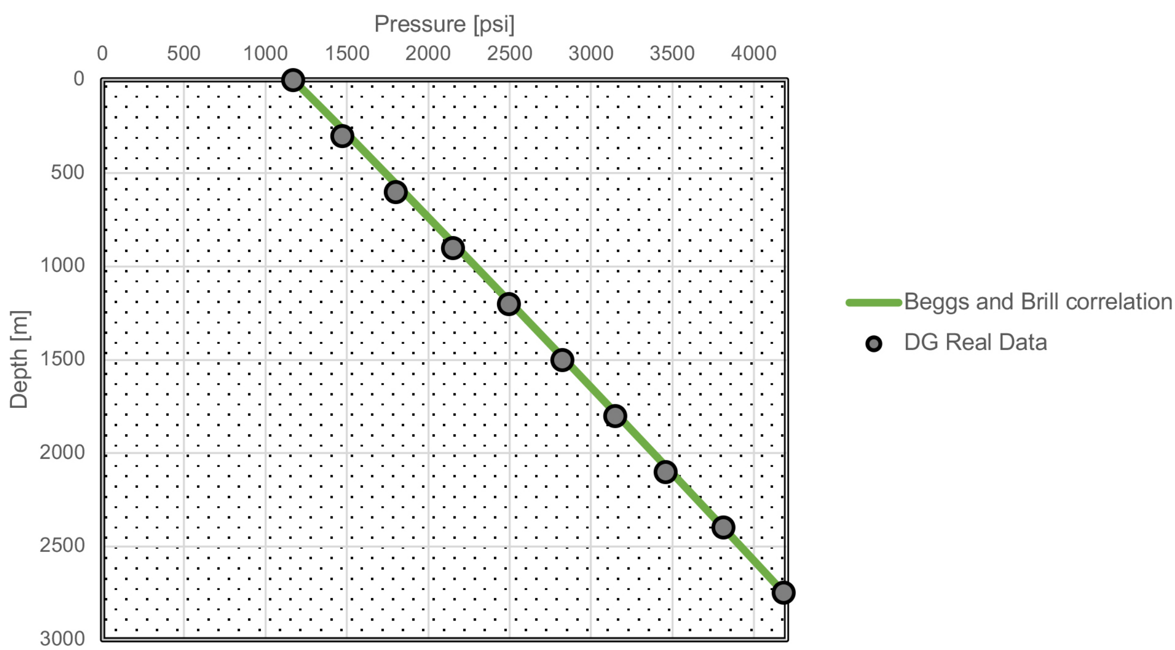

Figure 1. Dynamic gradient for Well #3 and production correlation match.

It can be observed that the match is good for all points, in particular for the one at z: 2794 m which corresponds to the deepest point in the well where pressure has been measured. The production correlation of Beggs and Brill was then used to generate a VLP curve for the well and use it to convert well head pressures into bottom hole pressures after the whole production history of the well was uploaded to the software. Figure 2 shows the well head pressure of Well #3 together with the curve of bottom hole pressure that was generated using the VLP:

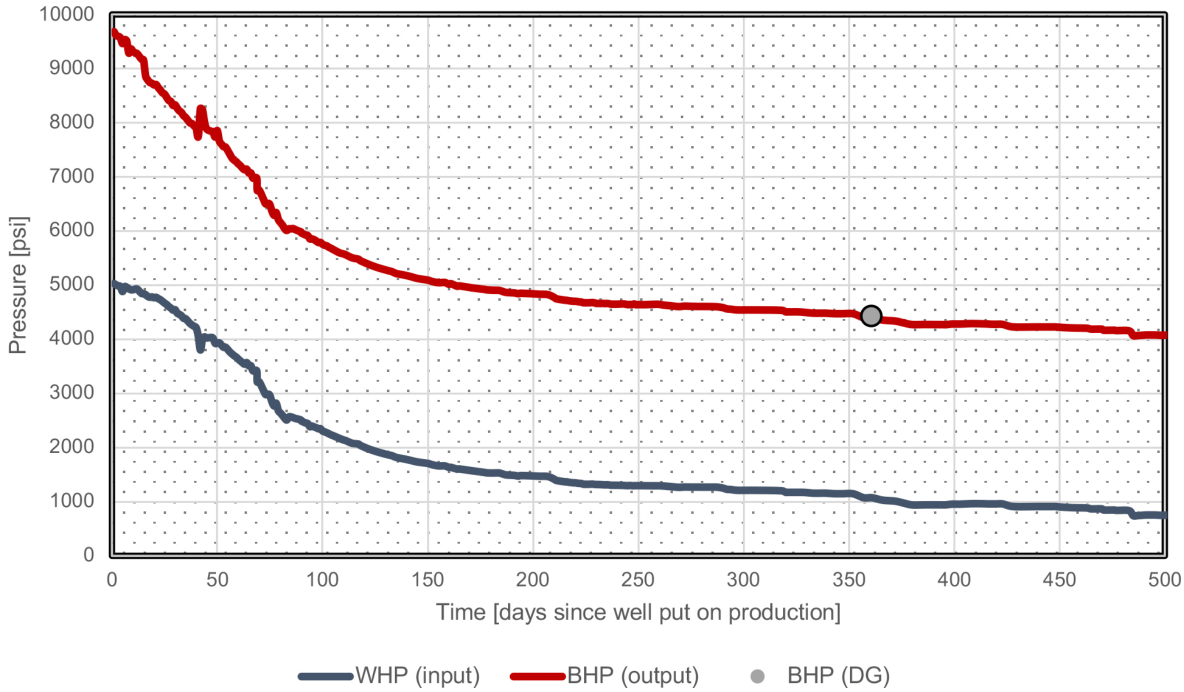

Figure 2. WHP and BHP for Well #3 (flowing well).

The graph also shows that for the day 373 since the well was put on production, there is a direct measurement of pressure of 4420 psi. These values correspond to the BHP estimated from the direct measurement in the dynamic gradient at z: 2794 m.

For the dynamic gradients run in wells after the ALS has been installed, the pressure measured at the landing zone depth was taken as a particular direct measurement of the BHP associated to a specific date. These dynamic gradients were neither uploaded to production software nor matched with correlations as most of them do not work properly after the flowing period of the well has ceased.

Except for Wells #18, #20 and #30 which do not have dynamic gradients during the flowing period, a VLP was generated for all the other 29 wells of the sample. As it was explained before, these VLPs were then used to convert WHP to BHP for each well since the day production started until present day or until the date when the ALS has been installed. This is to say that the VLPs can only predict the BHP accurately during the flowing period of the well. It has been found that these VLPs overestimated the BHP once the rod pumping system or gas lift were used to produce the well.

Figure 3 shows the WHP and BHP versus time for Well #16. It can be observed that for the flowing period there is a continuous trend of BHP versus time which has been estimated from WHP and a lift curve. There are 3 dynamic gradients for this period. For this well, the first dynamic gradient was the one used to build the VLP and convert WHP to BHP. In the day 873 since the well was put on production, an artificial lift system was installed. From this date onwards,

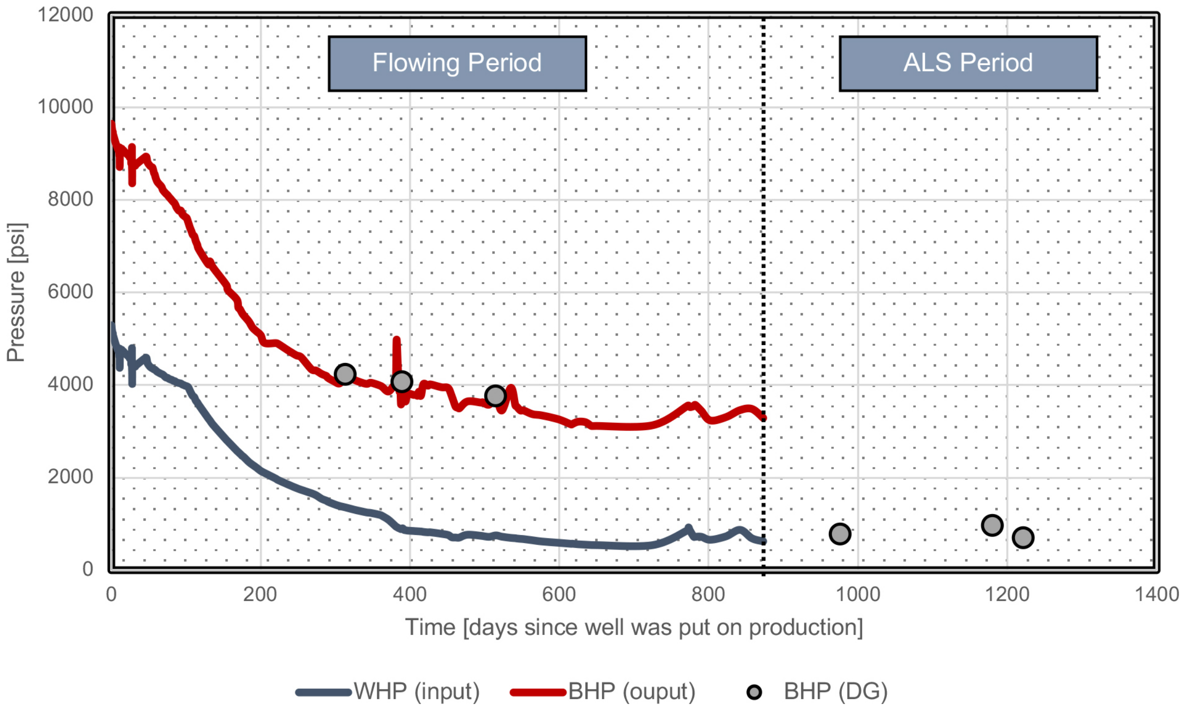

Figure 3. WHP and BHP for Well #16 (ALS well).

the VLP is no longer used to estimate BHP. Nonetheless, for the ALS period there are 3 dynamic gradients and therefore, there are 3 particular direct measurements of the BHP of the well.

So far it has been explained how the BHP (target variable) has been estimated for each case: either from VLPs or from direct measurements of dynamic gradients. So far so good. The rest of the training data set for the Machine Learning algorithm consists of input variables. These variables are divided in 4 groups:

- Georeferencing variables
- Production variables
- Wellbore variables
- Reservoir fluid variables

The first group sets the well in place (both in surface and in subsurface). There are 2 variables in this group: the Field name (A, B, C or D) and the Landing Zone (Z1 or Z2).

The second group consists of all those dynamic variables which are measured in the separator and/or in the well head. In this group there are 4 variables: oil production rate, gas production rate, water production rate and the well head pressure.

The third group of "Wellbore variables" consists of: the horizontal well lateral length, the number of fracture stages and the landing point true vertical depth.

Finally, the fourth group consists of four variables which characterize the reservoir fluid type. These variables are oil and gas density, oil viscosity and bubble point pressure. These variables can be obtained for example, from a PVT analysis.

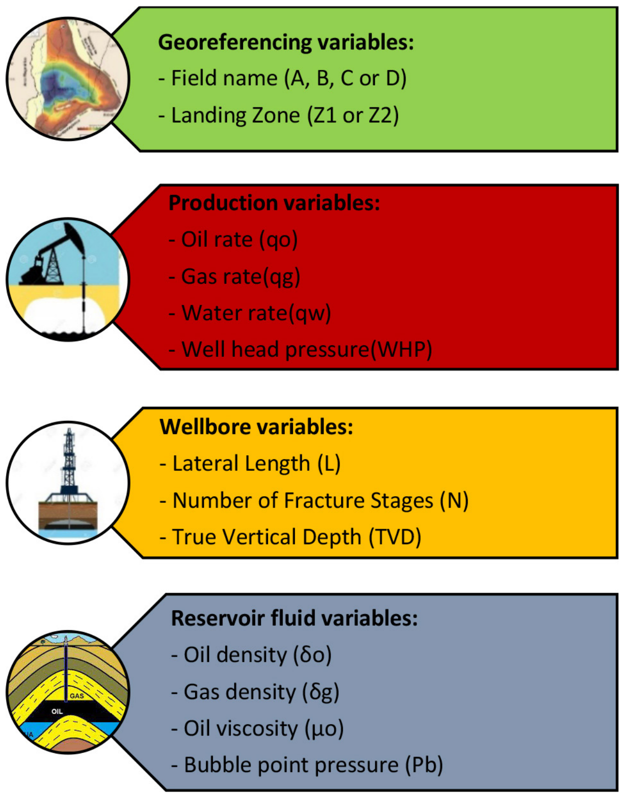

Figure 4. Input variables for the Machine Learning model training process.

The input variables for the training process of the Machine Learning model are summarized in the figure below:

A final comment should be made on how the Training Data Set file was made for each well of the sample.

As it was stated before, for the flowing period of each well, a continuous curve of BHP versus time was generated using a vertical lift performance curve. Each VLP was built from a production correlation that best matched the real data of a dynamic gradient. BHP predictions were made from the day that each well was put on production until present day or until the day when an ALS has been installed. This was made for all the wells of the sample except for Wells #18, #20 and #30 which do not have a dynamic gradient during their flowing period.

Table 2 shows the Training Data Set of Well #27 for BHP prediction before ALS. It can be seen that all input variables previously described are included as well as the target variable: the BHP. Additionally, the ALS start date has also been added. For Well #27, a Gas Lift system was installed on the day #425 since production started. The Training Data Set for this well contains BHP estimations from the first day of production to the day #425 when the ALS was installed. All the wells of the sample (except for #18, #20 and #30) have a Training Data Set as the one shown in table 2 for their flowing period.

Georeferencing		Production					Wellbore			Reservorio fluid					
Field name	Landing zone	t [d]	qo [m3/d]	qg [Mm3/d]	qw [m3/d]	WHP [psia]	Lateral Lenght [m]	#Fracture Stages	TVD [m]	γ_o [°API]	γ_g [SG]	μ_o [cp]	Pb [psi]	ALS start day	BHP [psi]
D	Z2	0.1	0.0	0.0	20.4	4564	2000	27	2815	47.9	0.81	0.42	4378	425	9143
D	Z2	0.1	0.0	0.0	24.2	5220	2000	27	2815	47.9	0.81	0.42	4378	425	9809
D	Z2	0.2	0.0	0.0	28.3	5217	2000	27	2815	47.9	0.81	0.42	4378	425	9805
D	Z2	0.2	0.0	0.0	28.1	5220	2000	27	2815	47.9	0.81	0.42	4378	425	9809
D	Z2	0.3	0.0	0.0	26.6	5219	2000	27	2815	47.9	0.81	0.42	4378	425	9808
D	Z2	0.3	0.0	0.0	27.4	5220	2000	27	2815	47.9	0.81	0.42	4378	425	9809
D	Z2	0.4	0.0	0.0	29.3	5216	2000	27	2815	47.9	0.81	0.42	4378	425	9804
D	Z2	0.4	0.0	0.0	27.1	5214	2000	27	2815	47.9	0.81	0.42	4378	425	9802
D	Z2	0.4	0.0	0.0	27.8	5197	2000	27	2815	47.9	0.81	0.42	4378	425	9785
D	Z2	0.5	0.0	0.0	26.6	5187	2000	27	2815	47.9	0.81	0.42	4378	425	9775
D	Z2	0.5	0.0	0.0	28.8	5190	2000	27	2815	47.9	0.81	0.42	4378	425	9779
D	Z2	0.6	0.0	0.0	28.1	5189	2000	27	2815	47.9	0.81	0.42	4378	425	9777
D	Z2	0.6	0.0	0.0	27.6	5190	2000	27	2815	47.9	0.81	0.42	4378	425	9778
D	Z2	0.6	0.0	0.0	25.7	5190	2000	27	2815	47.9	0.81	0.42	4378	425	9778
D	Z2	0.7	0.0	0.0	27.4	5188	2000	27	2815	47.9	0.81	0.42	4378	425	9776
D	Z2	0.7	0.0	0.0	27.1	5189	2000	27	2815	47.9	0.81	0.42	4378	425	9777
D	Z2	0.8	0.0	0.0	26.2	5190	2000	27	2815	47.9	0.81	0.42	4378	425	9778
D	Z2	0.8	0.0	0.0	27.6	5189	2000	27	2815	47.9	0.81	0.42	4378	425	9777
D	Z2	0.9	0.0	0.0	28.8	5188	2000	27	2815	47.9	0.81	0.42	4378	425	9776
D	Z2	0.9	0.0	0.0	28.1	5190	2000	27	2815	47.9	0.81	0.42	4378	425	9778
D	Z2	0.9	0.0	0.0	26.4	5188	2000	27	2815	47.9	0.81	0.42	4378	425	9776
D	Z2	1.0	0.0	0.0	27.6	5189	2000	27	2815	47.9	0.81	0.42	4378	425	9777
D	Z2	1.0	0.0	0.0	28.1	5187	2000	27	2815	47.9	0.81	0.42	4378	425	9775
D	Z2	1.1	0.0	0.0	26.6	5186	2000	27	2815	47.9	0.81	0.42	4378	425	9774
D	Z2	1.1	0.0	0.0	27.8	5190	2000	27	2815	47.9	0.81	0.42	4378	425	9778

Table 2. Training Data Set for Well #27 (for BHP prediction before ALS).

For BHP predictions after the artificial lift system has been installed in the well, a different Training Data Set was generated. For the ALS period there is no continuous curve of BHP versus time as there is for the flowing period. Instead, for this case, the Training Data Set consists of particular measurements of BHP estimated from dynamic gradients (as shown in table 3). It should be noted that no variables associated to the ALS are included in the Training Data Set (the algorithm is only told the exact date when the flowing period begins and ends for each well). All the wells from field D are included on this data set except for Wells #25 and #28 which do not have dynamic gradients in their ALS period.

While there is a separated Training Data Set file for each well for their flowing period, the Training Data Set for the ALS period comprises all the wells in one single file. Each line of the Training Data Set includes one value of BHP estimated from a dynamic gradient run in a well during the ALS period. There are 57 lines on this file as there is a total of 57 dynamic gradients run during the ALS period.

The following table shows the Training Data Set for BHP predictions after ALS:

	Georeferencing		Production					Wellbore			Reservorio fluid				
Well name	Field name	Landing zone	t [d]	qo [m3/d]	qg [Mm3/d]	qw [m3/d]	WHP [psia]	Lateral Lenght [m]	#Fracture Stages	TVD [m]	δo [°API]	δg [SG]	μo [cp]	Pb [psi]	BHP [psi]
Well #15	D	Z2		31.5	7.2	5.2	158	2500	33	2745	47.9	0.81	0.42	4378	757
Well #15	D	Z2		26.0	5.1	5.2	128	2500	33	2745	47.9	0.81	0.42	4378	759
Well #16	D	Z1		8.3	1.9	6.4	129	1500	20	2820	47.9	0.81	0.42	4378	963
Well #16	D	Z1		12.1	4.8	7.5	152	1500	20	2820	47.9	0.81	0.42	4378	776
Well #16	D	Z1		4.4	2.0	7.7	118	1500	20	2820	47.9	0.81	0.42	4378	693
Well #18	D	Z1		8.3	1.9	6.4	180	1547	25	2792	47.9	0.81	0.42	4378	940
Well #18	D	Z1		11.5	1.2	6.8	175	1547	25	2792	47.9	0.81	0.42	4378	947
Well #20	D	Z1		1.0	1.5	1.1	141	1500	18	2790	47.9	0.81	0.42	4378	443
Well #20	D	Z1		1.4	1.3	2.2	125	1500	18	2790	47.9	0.81	0.42	4378	1650
Well #21	D	Z1		9.0	0.7	1.5	145	2000	23	2741	47.9	0.81	0.42	4378	660
Well #21	D	Z1		12.3	2.5	4.2	149	2000	23	2741	47.9	0.81	0.42	4378	676
Well #21	D	Z1		19.7	3.5	4.0	151	2000	23	2741	47.9	0.81	0.42	4378	941
Well #21	D	Z1		2.2	3.5	4.8	141	2000	23	2741	47.9	0.81	0.42	4378	1436
Well #22	D	Z1		7.8	2.8	2.6	155	2000	27	2780	47.9	0.81	0.42	4387	1427
Well #22	D	Z1		5.4	5.7	4.8	173	2000	27	2780	47.9	0.81	0.42	4387	605
Well #22	D	Z1		10.9	3.4	8.6	138	2000	27	2780	47.9	0.81	0.42	4387	1111
Well #23	D	Z1		8.7	1.5	7.2	149	1767	24	2792	47.9	0.81	0.42	4387	1053
Well #23	D	Z1		12.7	5.7	8.3	191	1767	24	2792	47.9	0.81	0.42	4387	854
Well #23	D	Z1		9.6	1.5	8.8	144	1767	24	2792	47.9	0.81	0.42	4387	1129
Well #23	D	Z1		11.9	0.6	6.8	165	1767	24	2792	47.9	0.81	0.42	4387	1033
Well #23	D	Z1		13.3	5.2	13.4	139	1767	24	2792	47.9	0.81	0.42	4387	970
Well #24	D	Z1		3.7	2.4	4.0	138	1200	16	2807	47.9	0.81	0.42	4378	888
Well #24	D	Z1		4.5	2.9	3.5	138	1200	16	2807	47.9	0.81	0.42	4378	500
Well #24	D	Z1		3.0	1.3	3.2	147	1200	16	2807	47.9	0.81	0.42	4378	457
Well #24	D	Z1		7.2	11.7	4.2	127	1200	16	2807	47.9	0.81	0.42	4378	1397
Well #24	D	Z1		9.4	2.2	7.0	109	1200	16	2807	47.9	0.81	0.42	4378	1200
Well #24	D	Z1		22.4	6.6	14.7	147	1200	16	2807	47.9	0.81	0.42	4378	1309

Table 3. Training Data Set for BHP prediction after ALS.

Both Training Data Sets (before and after the ALS) were used to train the Machine Learning algorithm to accurately predict the BHP of any oil producer multi-fractured horizontal well of the Vaca Muerta Formation taking as input all the group of variables previously described.

MULTIVARIATE ANALYSIS

Multivariate analysis was performed on this work to understand on the first place the relationships between the input variables and the target variable and their relevance in the performance of the Machine Learning algorithm.

By definition, the Pearson Correlation Coefficient (PCC) offers a measurement of the linear correlation between two variables. For this example, let us consider two hypothetical variables A and B.

$$PCC\ (A,B) = \frac{cov\ (A,B)}{\sigma_A \cdot \sigma_B}$$

where:

$cov\ (A,B)$ is the covariance between A and B.

σ_A and σ_B are the standard deviations of A and B respectively.

If PCC equals 0, then there is no linear relationship between the variables A and B.

On the opposite, if PCC equals 1 or -1 there is a perfect linear relationship between both variables.

Early multivariate analysis is part of the preliminary "exploratory" process of the model development workflow as it is used to understand the relationships between variables: if one variable increases or decreases, so does the other one? If they both increase or decrease, do they both do it at the same rate?

When moving on to the model building phase it is desirable to avoid highly correlated variables in order to avoid multicollinearity.

The following figure shows the PCC for all the input and target variables:

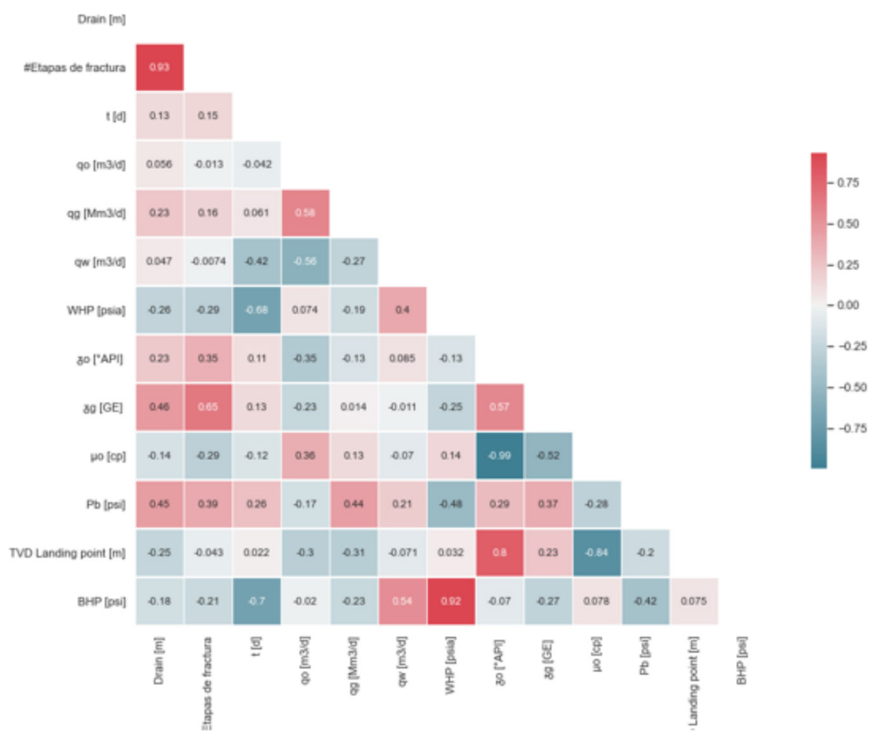

Figure 5. PCC for the all the variables of this work (input and target variables)

If we focus on the last line (towards the base of the triangle), we see the PCC between each input variable and the BHP (target variable). It should be pointed out that the PCC between the WHP and the BHP is 0,92 (very close to 1) which is associated to a high correlated behavior between these two variables.

MODEL DEVELOPMENT

The Machine Learning model development can be separated in three different stages:

- Data Preprocess
- Model Building
- Model Validation

Data Preprocess

At this stage, the data sets previously described were used to train the Machine Learning model.

Before the training process, features (i.e. different combinations of the input variables) were engineered and selected. This is done at this stage since some features may be irrelevant or less significant to the dependent variable. Therefore, their unnecessary inclusion in the model might lead to inaccurate or less reliable predictions and in an undesirable increase in the model complexity.

On the one hand, data quality assurance & quality control was conducted to exclude any outlier or neglect bad values from the overall Training Data Set due to operational or mechanical problems encountered in the wellbore while rates and well head pressures were being measured. For example, frac hits or temporary shut-ins may result in higher well head pressures that cannot be well represented in the model. On the other hand, based on domain expertise knowledge and categorical encoding methods, the data was enriched with new features.

Finally, by applying different features selection techniques (based-on filter, wrapper, and embedded methods) a smaller set of features was found which resulted in a better prediction power.

Model Building

Within the model building process multiple approaches were tested always starting simple and adding complexity as required to minimize the prediction error. Among the models trained for this testing exercise there were linear regression models, decisions trees based-on boosting and bagging methods, and stacked models.

After running these models with different types of features sets, the best combination was an average output of XGBoost and CatBoost Regression models. Both algorithms are based on a "boosting" technique which consists of adding up the contributions of multiple weak learner models. At each iteration, the solution space where the previous learner had the highest errors is then emphasized for the next learner to focus on. Therefore, after a series of iterations, the original "weak" learners become "strong" learners.

This boosting learning procedure is represented on figure 6. As it can be observed, the function "$f(x)$" which predicts the BHP can be represented as a linear combination of the terms "$\alpha_t.h_t(x)$", where "α_t" stands for the weight of each weak learner "$h_t(x)$".

At this stage, the resulting ensemble model provided good predictability. However, a refined tuning was later on performed in order to maximize the prediction accuracy.

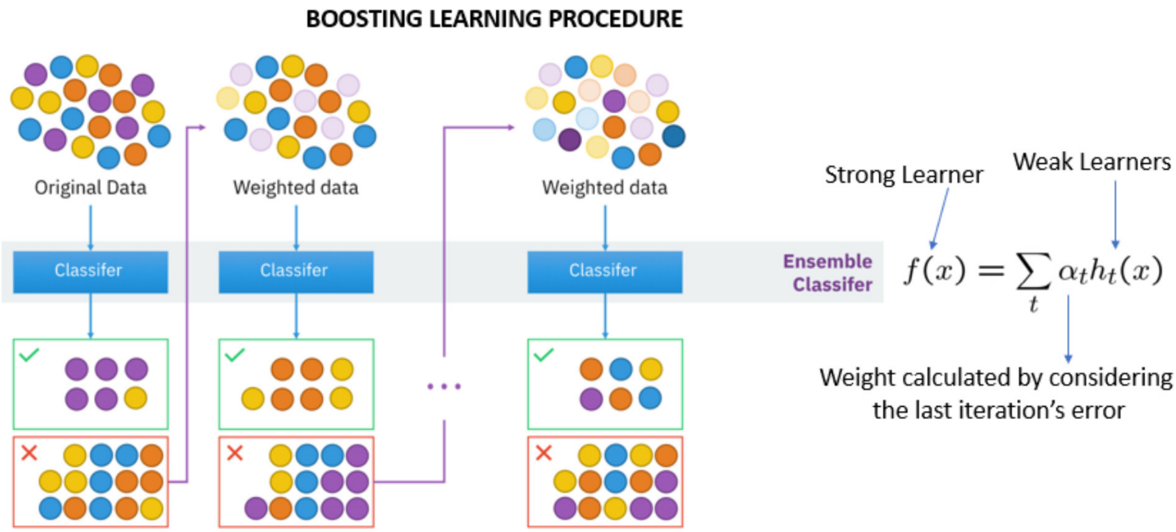

Figure 6. Conceptual representation of the Boosting Learning Procedure.

The final Machine Learning model that can accurately predict the Bottom Hole Pressure of the well from a given set of input variables (x_1, x_2, x_3,…, x_n) can be represented as follows:

$$BHP = f(x_1, x_2, x_3, \dots, x_n) = \frac{XGBoost_1 + CatBoost_2}{2}$$

The BHP is the mean value of the outputs of two XGBoost and CatBoost independent models.

Model Validation

The Coefficient of Determination (R^2), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as prediction accuracy measurements were the statistical metrics used to evaluate the performance of the model. That is, estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

There are various sorts of validation techniques on a limited data sample. On this study, a variant of a cross-validation method was selected as it is simple to understand and generally results in a less biased or less optimistic estimate of the model skill in comparison with other methods (such as a simple train/test split, also known as “hold-out” validation technique).

The selected validation technique of the model was a variant of the K-Fold method which will be described on the next section.

BLIND TEST

Three blind tests methods were carried out to assess the accuracy of the BHP predictive model as part of the model validation stage. These methods were the following:

- K-Fold Method
- BHP comparison between a well with a down hole permanent gauge vs the BHP estimated with the Machine Learning algorithm
- Other wells not used on the training of the algorithm with Dynamic Gradients, and therefore, with direct measurements of the BHP for a given point in time

K-Fold Method

The K-Fold method is a cross-validation process that is used to evaluate the performance of Machine Learning models for a given data sample. The parameter "K" represents the number of groups in which the data sample will be split into. Under a cross-validation method, models are evaluated by training numerous Machine Learning models on subsets of the available input data and evaluating them on the matching subset of the data.

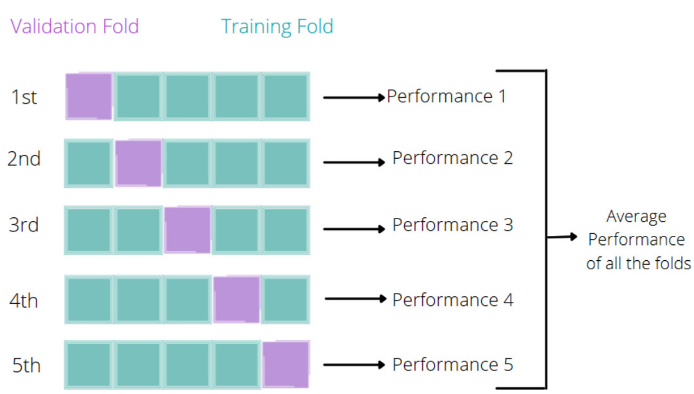

Figure 7. Overview of Cross Validation techniques.

Basically, this approach is used to avoid the overfitting or fluctuations in the training data that are selected and learned as concepts of the model. A more demanding approach to the K-Fold method in which the cross-validation process is repeated "n" times with different random splits of the sample, is called "double cross-validation" like the one used on this work.

The K-Fold method was applied in all the 32 wells of the sample. The 80% of the total data was used for Train, while the remaining 20% was used as Test using the cross validation technique previously described.

The results of the K-Fold method are shown on table 4:

Error Metric	Train Data	Test Data
Average R^2	0.999959	0.998210
Average RMSE	10.220	66.756
Average MAE	6.797	14.786
Standard Deviation R^2	0.000002	0.000657
Standard Deviation RMSE	0.196	12.271
Standard Deviation MAE	0.123	1.027

Table 4. K-Fold Blind Test results using K= 5 and n = 10.

Considering the magnitude order of the target variable, 66.756 psi as the average RMSE of the test data can be considered acceptable.

Well with Down Hole Permanent Gauge

A well with a down hole permanent gauge was used as blind test for the predictive model. The BHP was measured at reservoir depth in a continuous trend for 140 days.

It should be noted that this well was not part of the Training Data Sets of the model.

Well information:

- Horizontal well
- Lateral Length: 1200 m
- Number of fracture stages: 15
- Reservoir: Vaca Muerta Formation
- Landing Zone: Z1
- No ALS installed on the well

PVT properties as well as the reservoir depth were also known. All the necessary variables were input in the model and the BHP was estimated for the same 140 days where the down hole permanent gauge has been operating. Figure 8 shows the BHP estimated with the Machine Learning algorithm (red trend) and the BHP measured with the down hole permanent gauge (yellow trend). For both cases, the BHP has been normalized using the initial reservoir pressure as reference value.

As it can be observed the BHP estimation is very good and both trends lie over each other.

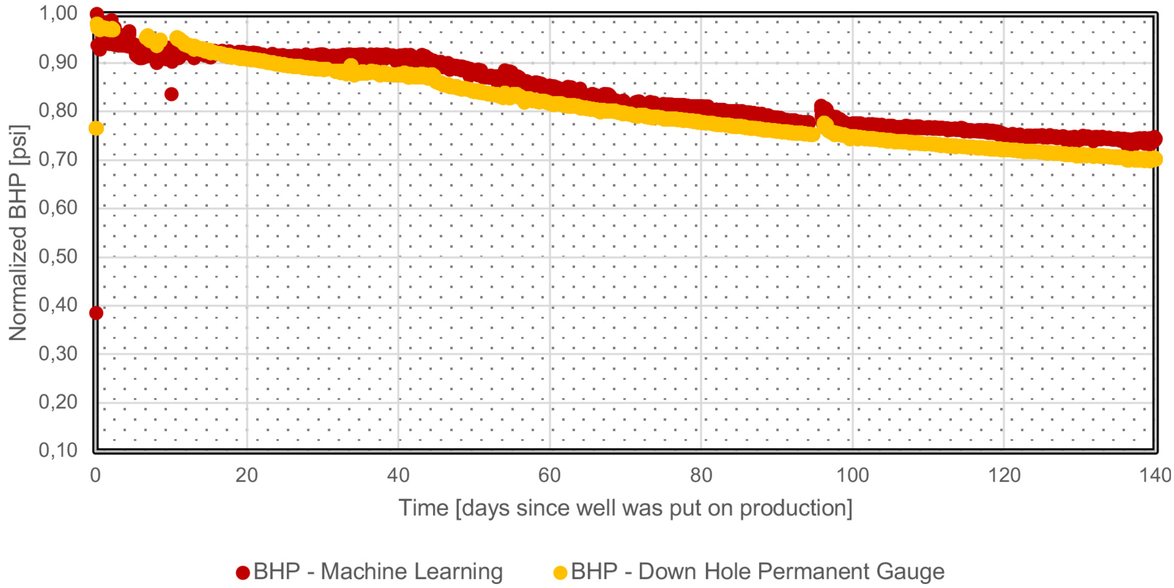

Figure 8. BHP estimated with the Machine Learning algorithm (red trend) and the BHP measured with the down hole permanent gauge (yellow trend) versus time.

Other Wells with Dynamic Gradients

The oil fields A, B, C and D have other wells that were not used during the training of the predictive Machine Learning model and that have dynamic gradients that were run at different moments of the wells′ lives.

These dynamic gradients offer a direct measurement of pressure very close to the reservoir depth and the well′s landing point. The deepest pressure measurement of the dynamic gradient is then extrapolated to the well′s TVD using the fluid average density obtained in the dynamic gradient lowest point.

The result of this procedure is a very good estimation of the BHP for a given point in time (the date when the dynamic gradient has been run in the well). These BHP estimations with dynamic gradients were used as an alternative blind test of the Machine Learning model.

The following figure shows a well from field A which has been used as Blind Test of the model. Originally, the well was produced through a casing of 5.5 in until the day 620 when a tubing of 2.875 in was installed. All the necessary variables were input in the model and the BHP was predicted as shown in the figure (blue trend). Two dynamic gradients were run in the well and were used to estimate the BHP at two different points in time. These direct measurements were plotted on the same graph (yellow points). As it can be observed, the model prediction is very close to the BHP estimated with dynamic gradients for both tubular diameters (casing and tubing). It should be noted that this well has no ALS installed.

The average absolute percentage error of the Machine Learning model prediction (vs the BHP estimation with dynamic gradients) for this well is only 2% which is considered acceptable.

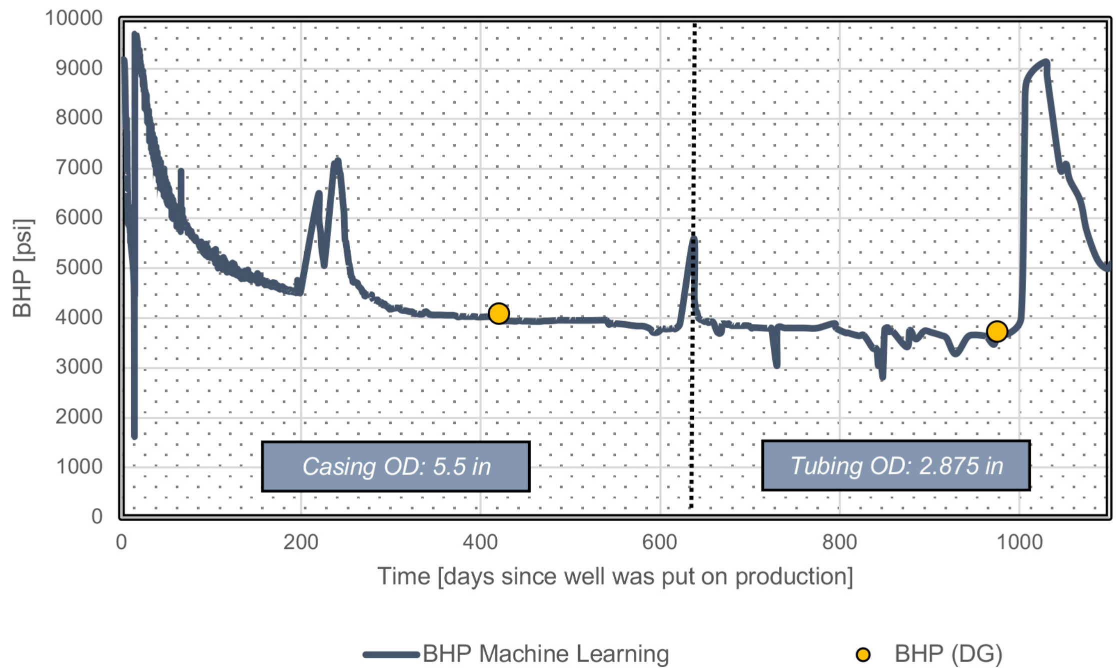

Figure 9. BHP vs time for a well of field A used as Blind Test.

Figure 10 shows the BHP predicted by the model for a well of field D together with two BHP estimations from dynamic gradients. For this case, a Gas Lift system was installed on the well 615 days after the put on production.

The average absolute percentage error of the Machine Learning model prediction (vs the BHP estimation with dynamic gradients) for this well is only 1.6%.

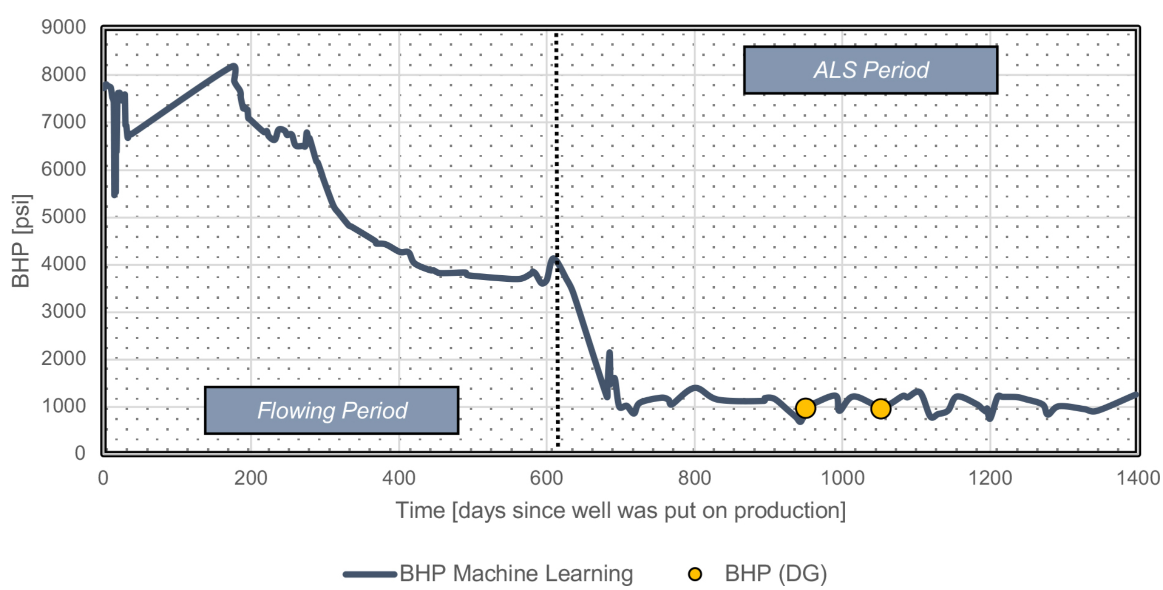

Figure 10. BHP vs time for a well of field D used as Blind Test.

FURTHER DISCUSSION

Supervised learning algorithms try to find hidden patterns among a big array of data, and therefore, relationships are stablished between the input features and the target variable. As it was done on this work, the chosen algorithm *learns* how to predict an output value from a given set of input variables. This learning process is a result of the previous training of the model and the continuous iterations of the boosting procedure.

Therefore, it should be noted that *"the model only learns what it has originally been taught"*. This is to say that, one should not expect the model to be able to predict the target variable from a completely different combination of input values (x_1, x_2, x_3,…, x_n) not used in the training process.

Moreover, the accuracy of the model is limited to the training universe conveyed by the Training Data Set. Machine Learning models should be re-trained over time as more data becomes available. This should result in more accurate predictions and lower estimation errors.

CONCLUSIONS

A Machine Learning model was built to predict the BHP of multi-fractured horizontal wells of the Vaca Muerta Formation. The novelty of this process relies on the simplicity to code and deploy on production dashboards as well as computer software.

Different input variables were used in the training process of the model. These variables can be divided in four groups: Georeferencing, production, wellbore, and reservoir fluid variables.

A XGBoost and CatBoost algorithms were selected to best predict the target variable.

Three different Blind Tests methods were carried out to assess the accuracy of the Machine Learning predictive model. All methods have shown successful results in relation to the performance of the algorithm.

The accuracy of the model is limited to the data universe in which it has originally been trained to predict the target variable.

ACKNOWLEDGEMENTS

The authors would like to thank Pan American Energy for the authorization to share this interdisciplinary and integrated work to build and calibrate a Machine Learning model that can accurately predict the bottom hole pressure of a multi-fractured horizontal well from a given set of production, reservoir, and well variables.

REFERENCES

Landset S, Khoshgoftaar TM, Richter AN, Hasanin T. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. J Big Data. 2015.

Kushmerick N, Weld DS, Doorenbos R. Wrapper induction for information extraction. Washington: University of Washington; 1997.

Tsymbal A, Pechenizkiy M, Cunningham P. Diversity in search strategies for ensemble feature selection. Inf Fusion. 2005.

Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. 10 Jun 2016.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, Andrey Gulin. CatBoost: unbiased boosting with categorical features. 20 Jan 2019.

M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society. Series B (Methodological).

Yagnik Pandya. 14 Feb 2021. Ensemble Methods in Machine Learning.

Eugenia Anello. K-Fold Cross Validation for Machine Learning Models. Published in Towards AI Feb 17, 2021.