sustainability

MDPI

Article

Discovering Energy Consumption Patterns with Unsupervised Machine Learning for Canadian In Situ Oil Sands Operations

Minxing Si [1,2], Ling Bai [3] and Ke Du [1,*]

[1] Department of Mechanical and Manufacturing Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada; minxing.si@ucalgary.ca
[2] Tetra Tech Canada Inc., 140 Quarry Park Blvd Suite 110, Calgary, AB T2C 3G3, Canada
[3] VL Energy Ltd., 208 Kincora Pt NW, Calgary, AB T3R 0A5, Canada; ling.bai@vlenergy.ca
* Correspondence: kddu@ucalgary.ca

Abstract: Canada's in situ oil sands can help meet the global oil demand. Because of the energy-intensive extraction processes, in situ oil sands operations also play a critical role in meeting the global carbon budget. The steam oil ratio (SOR) is an indicator used to measure energy efficiency and assess greenhouse gas (GHG) emissions in the in situ oil sands industry. A low SOR indicates an extraction process that is more energy efficient and less carbon intensive. In this study, we applied machine learning methods for data-driven discovery to a public database, Petrinex, containing operating data from 2015 to 2019 extracted from over 35 million records for 20 in situ oil sands extraction operations. Two unsupervised machine learning methods, including clustering and association rules, showed that the cyclic steam stimulation (CSS) recovery method was less efficient than the steam-assisted gravity drainage (SAGD) recovery method. Chi-square tests showed a statistically significant association between the CSS recovery method and high SOR ($p < 0.005$). Two association rules suggested that the occurrence of non-condensable gas (NCG) co-injection produced a low SOR. Chi-square tests on the two rules identified a statistically significant relationship between gas co-injection and low SOR ($p < 0.005$). Association rules also indicated that there was no association between the production regions and SORs. For future in situ oil sands development, decision-makers should consider SAGD as the preferred method because it is less carbon intensive. Existing in situ oil sands projects and future development should explore the possibility of NCG co-injection with steam to reduce steam consumption and consequently reduce GHG emissions from the extraction processes.

Keywords: in situ oil sands; data mining; Petrinex; k-means; unsupervised machine learning; clustering

check for updates

Citation: Si, M.; Bai, L.; Du, K. Discovering Energy Consumption Patterns with Unsupervised Machine Learning for Canadian In Situ Oil Sands Operations. *Sustainability* **2021**, *13*, 1968. https://doi.org/10.3390/su13041968

Received: 17 January 2021
Accepted: 8 February 2021
Published: 11 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

To keep the average global temperature rise below 2 °C, a third of global oil reserves have to remain undeveloped [1]. In 2019, Canada was the fourth largest oil producer, contributing 5% to the global oil production [2], and had the third largest proven oil reserves (following Venezuela and Saudi Arabia) with over 167 billion barrels (bbls) [3]. Canada plays a critical role in meeting the global carbon budget. Masnadi et al. [4] reported that Canada was the fourth highest carbon-intensive upstream oil producer in the world, after Algeria, Venezuela, and Cameroon. This is because over half of the oil production in Canada comes from an unconventional oil resource called oil sands.

Oil sands account for 64% of Canada's oil production and 98% of Canada's oil reserves [5]. Oil sands is a mixture of sand, water, clay, and heavy oil. The heavy oil separated from the oil sands is called bitumen, which contains particulate organic material, hydrocarbons, associated metals, and sulphur compounds [6]. A solid at room temperature, bitumen is the most viscous hydrocarbon [7]. Almost all oil sands reserves in Canada are concentrated in the Athabasca, Cold Lake, and Peace River regions in Northern Alberta. In situ oil sands extraction is one of two methods used to recover bitumen from oil sands.

The other extraction method is surface mining. Surface mining for bitumen extraction will become less economic than in situ recovery [1]. Eighty-one percent of the oil sands reserves in Canada need to be recovered by in situ oil sands extraction [5]. Therefore, in the long term, in situ recovery will be the primary method for future oil sands development. Two in situ recovery methods are commercially used: steam-assisted gravity drainage (SAGD) and cyclic steam stimulation (CSS). Both of these methods inject steam into reservoirs to reduce the bitumen's viscosity, which allows the bitumen emulsion to be pumped to the surface for oil/water separation and further processing [8].

Canada's oil sands development received broad criticism for its environmental impacts, such as high energy use [9], high greenhouse gas (GHG) emissions [10], reduced water quality [6], increased land disturbance [11], and increased fresh water use [12]. Among these, carbon emissions have been a focal point, both nationally and internationally. Pipeline development, such as Keystone XL, was rejected, citing climate change concerns [13]. The oil sands industry has contributed to Canada's economic opportunities. More than 400,000 people are employed by the oil sands industry and its related sectors [14]. Steam oil ratio (SOR) is an indicator used to measure energy efficiency and assess GHG emissions in the in situ oil sands industry. It measures the amount of steam injected into oil reservoirs and the amount of oil retrieved from underground reservoirs. A low SOR indicates that relatively little steam is required to produce a barrel of oil, which indicates that the extraction process is more energy efficient and less GHG-emission-intensive because most GHG emissions from in situ oil sands extractions are from steam generation. To reduce the use of steam, non-condensable gas (NCG), such as produced gas (mainly methane from oil-producing wells), is co-injected with steam by some operators. Solvents, such as hexane (C6), pentane (C5), and butane (C4), were also tested to dilute bitumen instead of steam [15]. Owing to the high cost and slow bitumen recovery rate of solvents, solvent-based methods have not been commercialised [16]. However, solvent co-injection with steam has been used by some in situ oil sands operators [17].

To control GHG emissions, the Government of Alberta implemented an emissions cap of 100 megatonnes of GHG emissions per year from oil sands extraction in 2017 [18]. Approximately one billion barrels of crude oil were produced from oil sands and contributed 70 megatonnes of GHG emissions in 2019 [19]. While balancing economic benefits and GHG emissions, decision makers face challenges to decide which development projects should take priority and which reserves should remain underground.

In this study, we aimed to discover which recovery method and which region had low SORs, low energy consumption, and consequently lower GHG emissions for in situ oil sands extraction by characterising patterns of emissions and fuel use data. The discovered patterns can provide information to decision makers for reviewing and approving new project applications while maximising the economic benefits and meeting the emissions cap of 100 megatonnes GHG emissions per year from oil sands extraction. We applied knowledge discovery in databases (KDD) to Petrinex, Canada's Petroleum Information Network, and used data mining techniques (specifically, unsupervised machine learning algorithms) to discover patterns from 20 in situ oil sands extraction schemes. Petrinex provides information for collecting royalties and facilitates commercial activities, such as production accounting. A detailed explanation regarding the data warehouse is provided by Alberta Energy Regulator (AER)'s Manual 011 [20]. The unsupervised machine learning algorithms used were clustering and association rules. KDD via unsupervised machine learning techniques has been widely researched and used in a range of applications in various industries [21]. For example, Lv [22] developed segmentation rules for batch process monitoring using the k-means clustering algorithm. Independent component analysis has been used for fault diagnosis and detection in industrial processes [23,24], and data clustering was used in chemical processes to detect faults on a separation tower [25]. In the oil industry, machine learning techniques were used to predict pressure, volume, and temperature (PVT) properties of crude oil [26,27], crude oil price [28,29], and enhanced oil recovery [30,31]. In oil sands operations, machine learning methods were applied to analyse

incident reports and increase process safety [32,33], and predict crude oil production from in situ oil sands extraction [34,35]. To the best of our knowledge, data mining techniques have not been applied to the Canadian oil and gas data warehouse or, more broadly, to any oil and gas data warehouse.

This study:

1. Assesses the impact of production regions and recovery methods on steam injection and oil production using clustering, unsupervised machine learning algorithms;
2. Evaluates whether production regions have a relationship with solution gas production by an unsupervised machine learning method, namely association rules;
3. Evaluates whether solvent co-injection with steam can reduce SORs and whether production regions have a relationship with solution gas production by an unsupervised machine learning method, namely association rules.

2. Materials and Methods

The KDD process is iterative, interactive, and includes the following main steps [36]:

1. Data selection: Relevant data are retrieved from the database, then a subset of data samples is selected to create a target dataset on which the discovery will be performed.
2. Data pre-processing: Outliers, inconsistent, or missing data are removed.
3. Data transformation: Appropriate data forms are created for mining. The task may consist of dimension reduction, data integration, and other steps.
4. Data mining or pattern discovery: Interesting patterns are extracted. Data mining is an essential step in the process of KDD [37]. Data mining tasks are generally grouped as predictive or descriptive. The predictive task builds a model to predict the future with methods such as correlation and regression. The descriptive task characterises properties of the data with methods such as clustering, identifying frequent patterns, and understanding associations.
5. Interpretation and evaluation: The mined patterns are interpreted and evaluated (commonly with pattern visualisation techniques).

KDD is a computational process for finding useful knowledge from a large amount of data [38]. For this study, we followed the standard KDD steps described above using the data science library Pandas (version 0.25.3) and the programming language Python (version 3.73) (Figure 1).

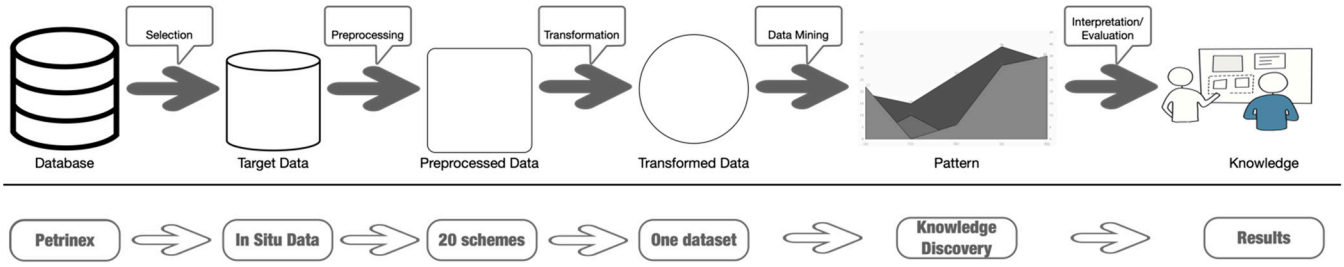

Figure 1. Knowledge discovery in database processes.

2.1. Data Selection

The monthly operating data obtained from Petrinex [39] contain mandatory reports on monthly activities from oil and gas licensees or operators in Alberta to the AER [40]. The monthly data from 2015 to 2019 were then tabulated into one dataset with 29 columns and over 35 million rows. The 29 columns provided information such as facility location, facility operator, well status, facility activity, and facility type. The 35 million rows contained monthly records for the entire oil and gas industry in Alberta. The monthly records included oil and solution gas production from oil batteries, fuel gas use, steam injection volumes, and NCG injection volumes.

The following procedures were performed to select data associated with in situ oil sands schemes:

1. Under the reporting facility types, battery (BT) and injection facility (IF) were selected.
2. Under the reporting facility subtypes, in situ oil sands and sulphur reporting at oil sands were selected.
3. BT and IF were linked by 11,000 well IDs provided in the Well to Facility Link Report [39]. The paired injection wells and producing wells for the scheme had the same well IDs. Depending on the stage of production, the number of wells for each scheme ranged from 100 to over 600 wells. The linked BT and IF IDs formed a dataset for in situ oil sands extraction schemes only, which was the target dataset in this study. The linked BT and IF IDs for each scheme are provided in the Supplementary Material.

The oil sands scheme included all BT and IF IDs associated with in situ oil sands extraction and excluded bitumen upgrading and producing wells. The BT is the facility that separates and measures products from producing wells. The IF is where steam is injected into the oil sands reservoir.

2.2. Data Preprocessing

In this study, 11 monthly records with oil production less than 5000 m^3 (approximately 100 bbl/day) were removed. These months had production interruptions such as the 2016 forest fire in Northern Alberta, or production started with volumes that were 5 to 10 times smaller than the following months. A detailed analysis of the data removal is provided in the Supplementary Material. In addition, MEG Energy's Christina Lake scheme did not have any fuel use data due to confidentiality. Therefore, this scheme was removed. With the exclusions removed, 20 in situ oil sands schemes with 1127 monthly records were populated for knowledge discovery (Table 1). The 20 schemes accounted for 82.4% of all in situ oil sands extractions in 2019 [41].

Table 1. In situ oil sands schemes. CSS: cyclic steam stimulation; SAGD: steam-assisted gravity drainage.

Operation (In-Text Reference) *	Operator	Scheme Name	Region	Recovery (Extraction) Method
IMOCL	Imperial Oil Resources	Cold Lake	Cold Lake	CSS
SUFB	Suncor Energy Inc.	Firebag	Athabasca	SAGD
CNRLWL	Canadian Natural Resources Limited (CNRL)	Wolf Lake, Primrose, and Burnt Lake	Cold Lake	CSS
CVECL	Cenovus Energy Inc.	Christina Lake	Athabasca	SAGD
CVEFC	Cenovus Energy Inc.	Foster Creek	Athabasca	SAGD
COPSM	ConocoPhillips Canada Resources Corp.	Surmont	Athabasca	SAGD
CNOOCLK	CNOOC Petroleum North America ULC	Long Lake	Athabasca	SAGD
HSESR	Husky Oil Operations Limited	Sunrise	Athabasca	SAGD
CNRLJF	Canadian Natural Resources Limited	Jackfish	Athabasca	SAGD
HSETL	Husky Oil Operations Limited	Tucker Lake	Cold Lake	SAGD
CNRLKB	CNRL	Kirby	Athabasca	SAGD
AOCLM	Athabasca Oil Corporation	Leismer	Athabasca	SAGD
SHAMR	PetroChina Canada Ltd.	Mackay River	Athabasca	SAGD
AOCHS	Athabasca Oil Corporation	Hangingstone	Athabasca	SAGD
PGFLB	Pengrowth Energy Corporation	Lindbergh	Cold Lake	SAGD
CNULPR	Canadian Natural Upgrading Limited	Peace River	Peace River	CSS
SUMR	Suncor Energy Inc.	Mackay River	Athabasca	SAGD
COGGD	Connacher Oil and Gas Limited	Great Divide	Athabasca	SAGD
OSUM	Osum Production Corp.	Orion	Cold Lake	SAGD
JCOS	Japan Canada Oil Sands Limited	Hangingstone	Athabasca	SAGD

* Operators are based on the 2019 Alberta Energy Regulator (AER) ST53 report. Assets may have changed ownership from 2015 to 2019.

2.3. Data Transformation

Of the 29 columns in the target dataset, 12 were removed. The removed columns contained information such as scheme locations and submission dates. A list of removed columns is provided in the Supplementary Material. Data in the target dataset were summarised to extract the operating parameters listed in Table 2 and for pattern discovery and unsupervised machine learning. The dataset had 1127 rows, each representing one monthly record. Of the 20 schemes, 13 had 60 monthly records, 3 schemes had 59 monthly records, and the remaining 4 schemes had 29 to 57 monthly records. The target dataset is provided in [42].

Table 2. Operating parameters retrieved from the data warehouse.

Operating Parameters	Units	Selection Method
Fuel Use	10^3 m^3	ActivityID column select FUEL ProductID column select GAS
Flare Volume	10^3 m^3	ActivityID column select FLARE ProductID column select GAS
Vented Gas Volume	10^3 m^3	ActivityID column select VENT ProductID column select GAS
Oil Production Volume	m^3	ActivityID column select PROD ProductID column select OIL
Steam Injection Volume	m^3	ActivityID column select INJ ProductID column select STEAM
Gas Injection Volume	10^3 m^3	ActivityID column select INJ ProductID column select GAS
Solution Gas Volume	10^3 m^3	ActivityID column select PROD ProductID column select GAS
Other Solvent Injection Volume	m^3	ActivityID column select INJ ProductID column select C3-SP, COND, etc.

Monthly SORs were calculated by dividing injection steam volumes by oil volumes and then cross-checked against AER ST53 statistical reports to ensure BT and IF were linked correctly and other parameters were appropriately extracted for each scheme. In Petrinex, steam quantity is reported in m^3 of cold water equivalent at a temperature of 15 °C, and fuel gas quantity is reported in 10^3 m^3 at 15 °C and 101.325 kPa absolute pressure.

2.4. Data Mining

Two unsupervised machine learning techniques were used: clustering and association rules. Unsupervised learning is used to discover the underlying patterns within the data to learn more about it. Unsupervised learning was conducted using the R programming language. The k-means algorithm was executed using an R function called kmeans. The association rule algorithm was implemented using a package in R called arules.

2.4.1. Clustering

Cluster analysis splits data into groups based on a similarity measure and is used to explore hidden patterns [43]. In this study, we used a k-means algorithm with the Euclidean distance similarity metric. We divided monthly production volumes and steam injection volumes into k clusters based on the distance to the centroid of a cluster, with the objective of maximising the similarity within groups and minimising the similarity between groups [44]. The algorithm aims to minimise the Euclidean distances of all points with their nearest cluster centres by minimising the within-cluster sum of squared errors (SSE).

By clustering, we analysed how oil production responded to steam injection. Minimising steam injection quantities is the key to reducing GHG emissions from in situ oil sands

extraction. We also examined how production regions and recovery methods influenced oil production and steam injection. The steam injection and oil production data were normalised using the z standardisation method before being fed into the algorithm. The number of clusters (k) was selected based on the rule suggested by Harigan [45]. The rule uses the intuition that when clusters are well separated by K* being the right number of clusters, then:

- For K < K*, a (K + 1) cluster partition should be the K cluster partition with one of its clusters split into two. This would significantly decrease the total within-cluster variation (W_K);
- For K > K*, both the K and (K + 1) cluster partitions will be equal to the right cluster partition with some of the right clusters split randomly, so that W_K and W_{K+1} are not significantly different.

2.4.2. Association Rule

Association rules are used to identify sets of items that frequently occur together in a dataset. It is a popular unsupervised machine learning technique for market basket analyses, writer evaluations, medical diagnoses, etc. [43].

In this study, we evaluated whether co-injections were associated with low SORs and if production regions were associated with high solution gas oil ratios (SGORs). The association rule had three parameters: support, confidence, and lift [46,47]. In this context, the association rule can be written as:

$$X \Rightarrow Y\ [support,\ confident,\ lift].$$

Support measures how frequently X and Y happen together and is expressed as:

$$support\ \{X\ \Rightarrow\ Y\ \} = \frac{Number\ of\ months\ containing\ both\ X\ and\ Y}{m},$$

where X is the co-injection or production region, Y is low SOR or high SGOR, and m is the number of months in the entire dataset, which was 1127 in this study.

Confidence is the conditional probability that Y is true under the condition of X and expressed as:

$$Confidence\ \{X\ \Rightarrow\ Y\ \} = \frac{Number\ of\ months\ containing\ both\ X\ and\ Y}{Months\ containing\ X} = P(Y\,|X).$$

Lift is used to measure the correlation between X and Y and is written as:

$$Lift\ \{X\ \Rightarrow\ Y\ \} = \frac{Confidence\ \{X\ \Rightarrow\ Y\ \}}{Percentage\ of\ months\ containing\ Y},$$

when $Lift < 1$, X is negatively correlated with Y. When $Lift > 1$, X is positively correlated with Y, and when $Lift = 1$, X and Y are independent.

For association rule mining, two parameters need to be defined: the minimum support threshold (min_sup) and the minimum confidence threshold (min_conf). In this study, we set the min_sup to 10% to ensure that at least two schemes had co-injection or a high SGOR with at least 96 monthly records. The min_sup threshold also filtered out some injection activities that might not have been intended to recover bitumen. We set the min_conf to 80% to ensure a high $P(Y \mid X)$.

We categorised SORs, solvent co-injection volumes, and SGORs based on the median values. The criteria used are presented in Table 3. The solvents co-injected with steam by the 20 selected schemes were gas (mainly methane), natural gas condensate, and propane (C3).

Table 3. Criteria used for association rule mining. SOR: steam oil ratio; NCG: natural gas condensate; SGOR: solution gas oil ratio; C3: propane.

Rule $X \Rightarrow Y$	Criteria	Categorisation
SOR (Y)	SOR $\geq$ median	High SOR
	SOR < median	Low SOR
NCG/condensate/C3 injection (X)	Injection volume $\geq$ median	With co-injection
	Injection volume < median	Without co-injection
SGOR (Y)	SGOR $\geq$ median	High SGOR
	SGOR < median	Low SGOR
Production region (X)	Athabasca, Cold Lake, and Peace River	

The cut-off values (medians) are presented in Table 4.

Table 4. Cut-off values.

Production Indicators	Cut-Off Values (Median)
NGC co-injection volume	1456 10^3 m^3
SOR	3.31
SGOR	0.01444 10^3 m^3 solution gas/m^3 of oil

2.5. Interpretation and Evaluation

The uncovered patterns were visualised and are presented in the Results section. A chi-square test for independence was used to assess the statistical significance level of the dependence between the antecedent (X) and the consequent (Y) in an association rule (X $\Rightarrow$ Y) [48,49]. The null hypothesis and an alternative hypothesis for the chi-square test are:

- H_o: The antecedent (X) and the consequent (Y) are independent.
- H_a: The antecedent (X) and the consequent (Y) are not independent.

3. Results

3.1. Clustering

There were 1127 monthly records grouped into nine clusters based on steam injection and oil production. Among the nine clusters, clusters 4 and 9 were the least efficient, with more steam injection used per unit of bitumen produced (compared to the average). Cluster 2 was the most efficient, with the lowest steam injection per unit of bitumen produced (Figure 2).

3.2. Association Rule and Chi-Square Test

We tested 23 rules to determine whether solvent co-injection with steam, recovery methods, and production regions impacted SORs and SGOR. The results of the association rules are presented in Table 5. Among the 23 rules, Rules 1, 5, 11, 17, and 22 met the criteria of support, confidence, and lift, indicating the antecedent itemset implies the consequent itemset.

Chi-square tests were conducted on Rules 1, 5, 11, 17, and 22 for statistical significance. The Pearson *p*-values from the chi-square tests for all five rules were less than 0.05; therefore, we rejected the null hypothesis and concluded that there was a statistical association between the antecedent itemset and the consequent itemset.

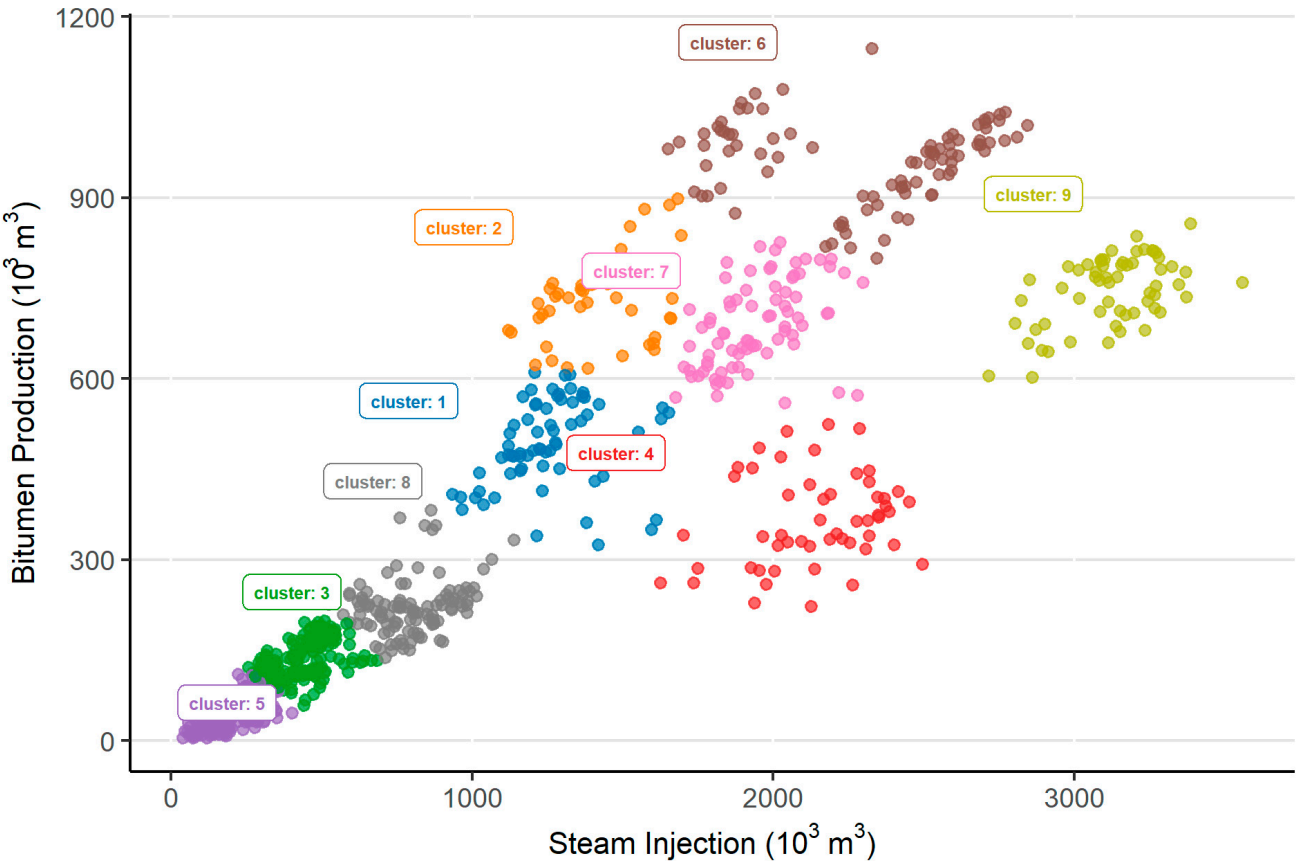

Figure 2. Cluster analysis based on monthly steam injection and monthly bitumen production.

Table 5. Association rule results.

Rule ID	Antecedent (X)	Consequent (Y)	Support	Confidence	Lift
1	With solvent co-injection	Low SOR	19%	93%	1.9
2	Without solvent co-injection	Low SOR	31%	39%	0.8
3	Method = CSS	Low SOR	0%	1%	0.0
4	Method = SAGD	Low SOR	50%	57%	1.2
5	Method = CSS	High SOR	16%	99%	2.0
6	Method = SAGD	High SOR	34%	40%	0.8
7	Region = Athabasca	Low SOR	43%	63%	1.3
8	Region = Cold Lake	Low SOR	7%	26%	0.5
9	Region = Peace River	Low SOR	0%	2%	0.0
10	Method = SAGD, without solvent co-injection	Low SOR	37%	48%	0.8
11	Method = SAGD, with solvent co-injection	Low SOR	22%	93%	1.6
12	Method = CCS, without solvent co-injection	Low SOR	1%	1%	0.0
13	Method = SAGD, with solvent co-injection, Region = Athabasca	Low SOR	93%	93%	0.4
14	Method = SAGD, with solvent co-injection, Region = Cold Lake	Low SOR	0.4%	100%	0.4
15	Method = SAGD, without solvent co-injection, Region = Athabasca	Low SOR	38%	50%	0.6
16	Method = SAGD, without solvent co-injection, Region = Cold Lake	Low SOR	11%	43%	0.5
17	Method = CSS	High SGOR	16%	100%	2.0
18	Method = SAGD	High SGOR	34%	39%	0.8
19	Method = SAGD, with solvent co-injection	High SGOR	14%	60%	1.5
20	Method = SAGD, without solvent co-injection	High SGOR	26%	34%	0.9
21	Without solvent co-injection, region = Athabasca	High SGOR	12%	20%	1.7
22	Without solvent co-injection, region = Cold Lake	High SGOR	29%	87%	3.0
23	Without solvent co-injection, region = Peace River	High SGOR	7%	100%	15.0

4. Discussion

4.1. Efficiency of Recovery Methods

Twenty schemes were clustered into nine groups based on steam injection and oil production (Figure 3). CNULPR using the CSS recovery method and other SAGD schemes was grouped into cluster 5, which had the lowest overall oil production and steam injection volumes. This pattern indicated that the CSS method shared similar characteristics with the SAGD method when production volume was low. The maximum oil production under cluster 5 was 110,468 m^3/month.

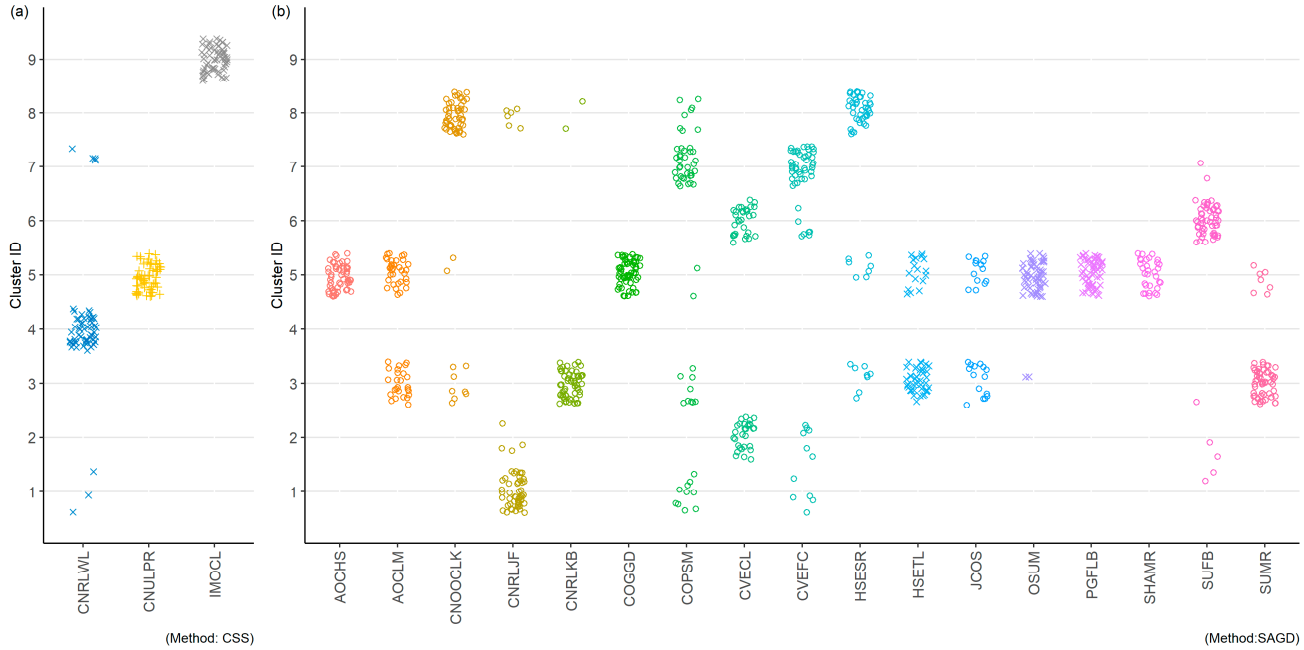

Figure 3. Clustering results: x represents the Cold Lake region; + represents the Peace River region; O represents the Athabasca region. (**a**) includes schemes using the CSS recovery method. (**b**) includes schemes using the SAGD recovery method.

The other two CSS schemes, IMOCL and CNRLWL, were different from SAGD. IMOCL had a steady operation in 2015–2019, and all 60 monthly data points were clustered together and formed independent cluster 4. Fifty-three out of 60 monthly data points for CNRLWL were grouped into cluster 4. Clusters 4 and 9 injected more steam to generate similar oil production in comparison to other clusters that were SAGD schemes. This pattern indicated that the CSS method might be less efficient than the SAGD method when the schemes proceed toward maturity. Rule 5 in Table 5 and the subsequent chi-square test also indicated that the CSS method has higher SOR and is less efficient with the rule: $\{\text{Method} = \text{CSS}\} \Rightarrow \{\text{High SOR}\}$ (support : 16%, confidence : 99%, lift : 2.0).

The HSETL, OSUM, and PGFLB schemes are located in the Cold Lake region. They were grouped together with the schemes in the Athabasca region, which implied that different regions might not have an impact on the oil and steam interaction.

4.2. Solvent Co-Injection with Steam

Solvent co-injection with steam to improve heavy oil recovery efficiency was first reported in the 1960s [50] and has been successfully used in California for producing and transporting heavy crude oil [51]. The solvents used in the 20 selected schemes were gas (mainly methane), C3, and natural gas condensate. The CNOOCLK, COGGD, and IMOCL schemes injected condensate between 2015 and 2019. The injection volumes per month were 683 m^3 for CNOOCLK and 740 m^3 for COGGD. The IMOCL scheme had a monthly average condensate injection of 9827 m^3; large monthly volumes (greater than 10,000 m^3) were injected from June 2017 to January 2019. By December 2019, the

condensate injection by IMOCL was stopped. Only CVEFC injected C3 at an average of 2868 m^3/month from January 2018 to December 2019. The weighted average of the SOR for CVEFC increased by 8% from 2.56 to 2.77 m^3/m^3 when comparing before and after C3 co-injection. However, these co-injection activities did not meet the min_sup threshold of 10%. Only gas co-injection met both min_sup and min_conf thresholds.

For gas co-injection, we used the median value (1456 10^3 m^3) of gas injection volume as a cut-off. Six schemes were considered co-injection schemes. Three schemes, CVECL, CVEFC, and SUFB, continuously injected gas between 2015 and 2019 for 60 months. The weighted average SOR of these three schemes was 2.36 m^3/m^3; it was 45% lower than the weighted average SOR for the 14 schemes without gas co-injection that were fully operational between 2015 and 2019 (Figure 4). The CNRLJF and COPSM schemes began gas co-injection in mid-2016 and early 2017, respectively. The SHAMR scheme was a new operation that began in June 2017; gas co-injection started in September 2018. The weighted average SOR of SHAMR was two times higher than the weighted average SOR of CVECL, CVEFC, and SUFB.

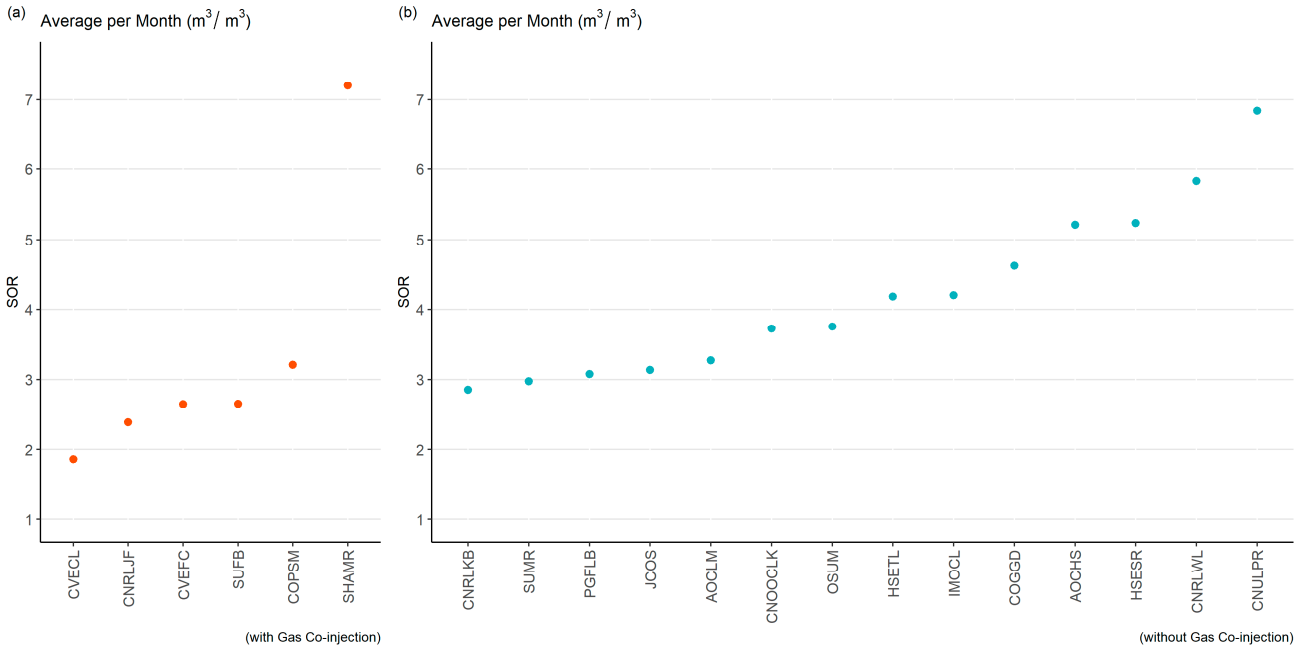

Figure 4. Co-injection effects on the steam oil ratio. (**a**) includes schemes with gas co-injection. (**b**) includes schemes without gas co-injection.

The two association rules and chi-square tests suggested that the occurrence of gas co-injection implied a low SOR, including $\{\text{Gas Co} - \text{injection}\} \Rightarrow \{\text{Low SOR}\}$ $(\text{support} : 19\%, \text{confidence} : 93\%, \text{lift} : 1.9)$ and $\{\text{Method} = \text{SAGD, and Gas Co} - \text{injection}\} \Rightarrow \{\text{Low SOR}\}$ $(\text{support} : 22\%, \text{confidence} : 93\%, \text{lift} : 1.6)$. The distribution of SORs is provided in Figure 5.

4.3. Solution Gas and Production Region

On average, between 2015 and 2019, in situ oil sands extractions produced 21 m^3 of solution gas/1 m^3 of bitumen, with a median of 14 m^3/m^3. The Peace River region only had one scheme: CNULPR. The arithmetic mean of the SGOR for CNULPR was 81 m^3/m^3. The arithmetic mean of the SGOR for the Cold Lake region was 36 m^3/m^3, and for the Athabasca region, it was 11 m^3/m^3. Although the schemes in the Cold Lake region had higher SGORs (Figure 6), none of these schemes used gas co-injection (Figure 5).

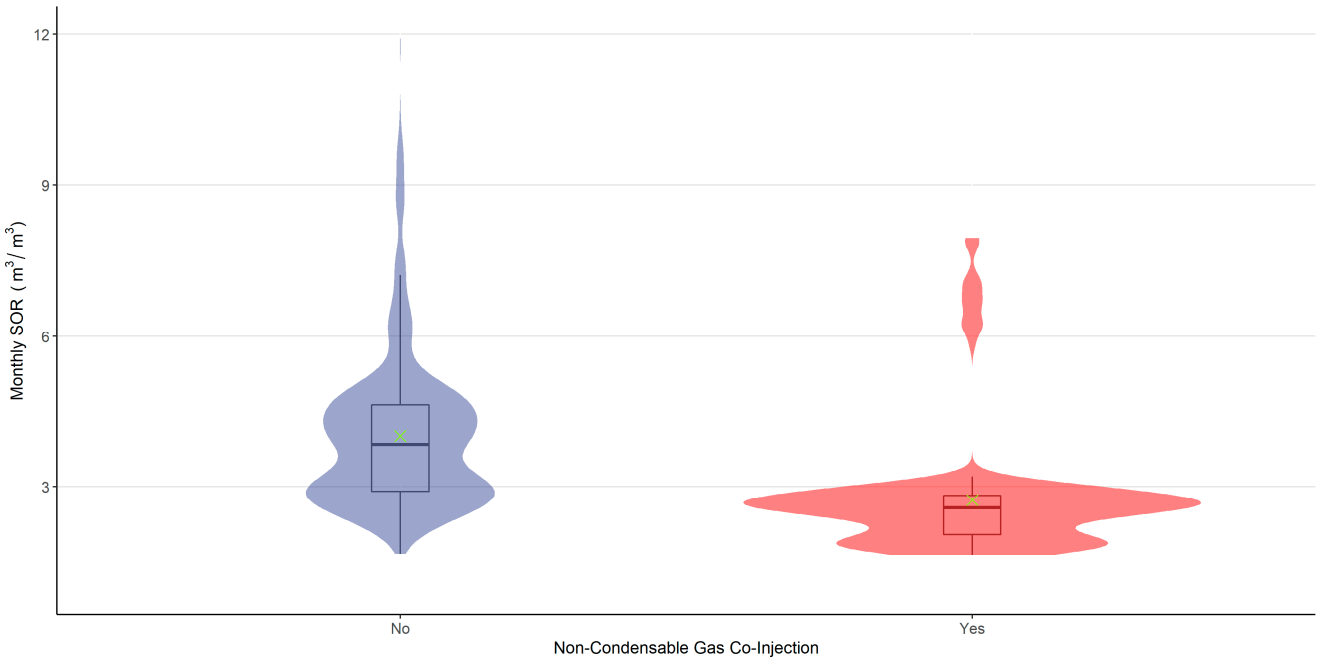

Figure 5. Monthly SORs between schemes without gas co-injection and schemes with gas co-injection. X represents the arithmetic mean of the monthly SORs. The upper and lower bars in the box indicate the 25th and 75th percentile values, respectively. The middle bar in the box indicates the median. The whiskers extend to values within 1.5 times of the interquartile range (IQR). The violin plot describes the distribution of the monthly SORs using a density curve. The width of each curve represents the frequency of SORs.

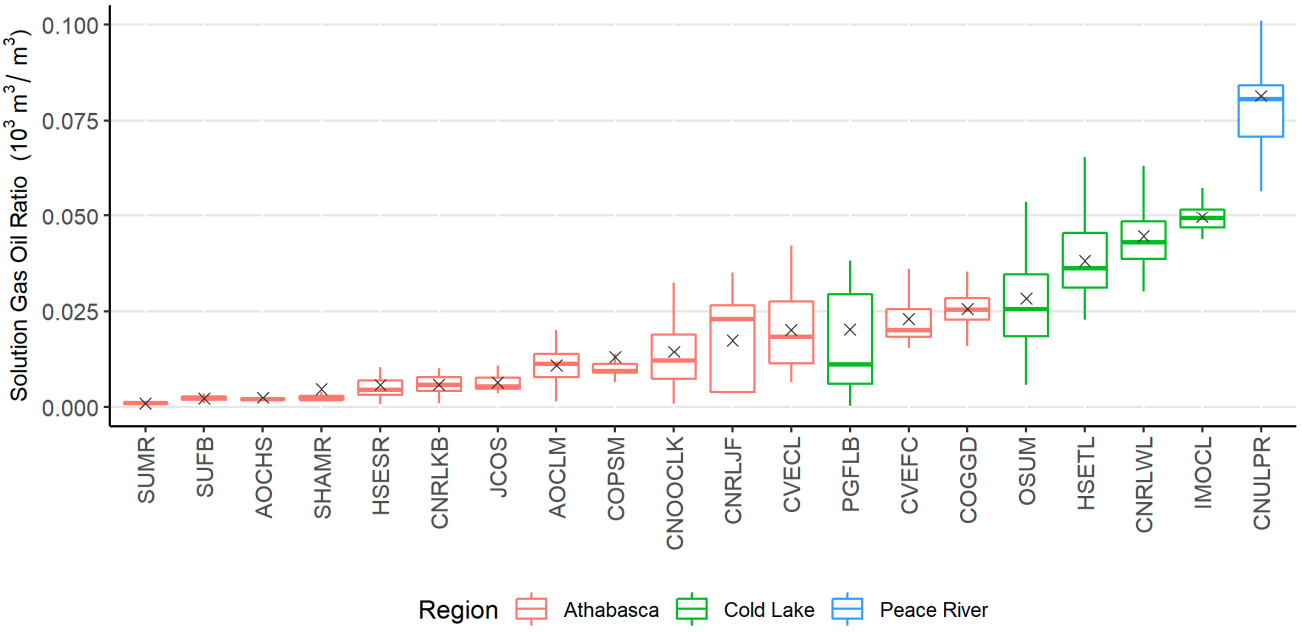

Figure 6. Average solution gas produced between 2016 and 2019 by in situ production regions. See Figure 5 for an explanation of the boxplot. The boxplot is based on monthly data. The X symbol in the box represents the arithmetic mean.

Rule 22 and its chi-square test also suggested that there was a strong relationship between the Cold Lake production region and a high SGOR, with 29% support, 87% confidence, and 3.0 lift.

5. Conclusions

In this study, machine learning methods for data-driven discovery were applied to a public database, Petrinex, containing operating data from 2015 to 2019 that were extracted from over 35 million records for 20 in situ oil sands extraction schemes. The use of clustering and association rules and two unsupervised machine learning methods implied that: (1) the CSS recovery method was less efficient than SAGD recovery as schemes proceed toward maturity (Rule 5); (2) gas co-injection resulted in low SORs (Rules 1 and 11); and (3) the Cold Lake region had higher SGOR compared to the two other regions (Rule 22). The procedures and analyses introduced in this study for the two unsupervised machine learning algorithms can be applied to any database in any country for data-driven pattern discovery.

The chi-square test carried out on Rule 5 $\{\text{Method} = \text{CSS}\} \Rightarrow \{\text{High SOR}\}$ $(\text{support}: 16\%, \text{confidence}: 99\%, \text{lift}: 2.0)$ showed that there was a significant association between the CSS recovery method and high SOR ($p < 0.005$). SAGD recovery might be the preferred method for decision makers to consider in the future for in situ oil sands development projects because the SAGD method is less GHG-emission-intensive. By choosing SAGD recovery as a preferred method, the economic benefits are maximised while GHG emissions are minimised.

Rule 1 $\{\text{Gas Co} - \text{injection}\} \Rightarrow \{\text{Low SOR}\}$ $(\text{support}: 19\%, \text{confidence}: 93\%, \text{lift}: 1.9)$ and Rule 11 $\{\text{Method} = \text{SAGD, and Gas Co} - \text{injection}\} \Rightarrow \{\text{Low SOR}\}$ $(\text{support}: 22\%, \text{confidence}: 93\%, \text{lift}: 1.6)$, shown in Table 5, suggested that the occurrence of gas co-injection implied a low SOR. Chi-square tests on Rules 1 and 11 showed that there was a statistically significant relationship between gas co-injection and low SOR ($p < 0.005$). The association rules also indicated that there were no associations between the production regions and SORs.

SORs are also affected by other factors, such as operational efficiency and equipment maintenance. The application of the SAGD method and gas co-injection alone may not result in low SORs. Existing in situ oil sands projects and future developments should explore the possibility of gas co-injection with steam to reduce steam consumption and consequently reduce GHG emissions from the extraction processes.

Supplementary Materials: The following are available online at https://www.mdpi.com/2071-1050/13/4/1968/s1, Table S1: Battery and Injection Facility IDs for each scheme, Table S2: Number of monthly data used in the study, Table S3: Removed monthly data, Table S4: Removed Columns, Table S5: Criteria used for Association Rule, Table S6: Summary of Gas Co-injection.

Author Contributions: Conceptualisation, M.S. and L.B.; methodology, M.S.; software, M.S.; formal analysis, M.S.; investigation, M.S.; resources, K.D.; writing—original draft preparation, M.S.; writing—review and editing, K.D.; visualisation, M.S.; supervision K.D.; project administration, K.D. and L.B.; funding acquisition, K.D. and L.B. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada (fund number CRDPJ535813-18) and by Mitacs through the Mitacs Accelerate program (fund number IT18400).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository. The data presented in this study are openly available in Mendeley Data at 10.17632/8ngkgz69zb.4.

Acknowledgments: The authors would like to thank the Environment and Water group and Jessica Coles from Tetra Tech Canada Inc. for editing the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The sponsors had no role in the design, execution, interpretation, or writing of the study.

References

1. McGlade, C.; Ekins, P. The geographical distribution of fossil fuels unused when limiting global warming to 2 °C. *Nature* **2015**, *517*, 187–190. [CrossRef]
2. U.S. Energy Information Administration (EIA). What Countries Are the Top Producers and Consumers of Oil? 2020. Available online: https://www.eia.gov/tools/faqs/faq.php?id=709&t=6 (accessed on 25 September 2020).
3. U.S. Energy Information Administration (EIA). International Energy Statistics, Crude Oil Including Lease Condensate 2019. Available online: https://www.eia.gov/international/data/world/petroleum-and-other-liquids/annual-crude-and-lease-condensate-reserves?pd=5&p=0000000000000000000008&u=0&f=A&v=mapbubble&a=-&i=none&vo=value&vb=106&t=C&g=00000000000000000000000000000000000000000000000000001&l=249-ruvvvvvfvtvnvv1vrvvvvfvvvvvvfvvvou20evvvvvvvvvvnvvvs0008&s=315532800000&e=1577836800000&ev=true (accessed on 3 December 2019).
4. Masnadi, M.S.; El-Houjeiri, H.; Schunack, D.; Li, Y.; Englander, J.G.; Badahdah, A.; Monfort, J.-C.; Anderson, J.E.; Wallington, T.J.; Bergerson, J.A.; et al. Global carbon intensity of crude oil production. *Science* **2018**, *361*, 851–853. [CrossRef]
5. Natural Resources Canada. Crude-Oil-Facts 2019. Available online: https://www.nrcan.gc.ca/science-data/data-analysis/energy-data-analysis/energy-facts/crude-oil-facts/20064 (accessed on 3 December 2019).
6. Alexander, A.; Chambers, P. Assessment of seven Canadian rivers in relation to stages in oil sands industrial development, 1972–2010. *Environ. Rev.* **2016**, *24*, 484–494. [CrossRef]
7. Banerjee, D. *Oil Sands, Heavy Oil, & Bitumen: From Recovery to Refinery*; PennWell Corporation: Tulsa, OK, USA, 2012.
8. Giove, A.; Sciarrabba, T. *In-Situ Bitumen Extraction*; Oil Sands Mag: Calgary, AB, Canada, 2019.
9. Brandt, A.R.; Englander, J.; Bharadwaj, S. The energy efficiency of oil sands extraction: Energy return ratios from 1970 to 2010. *Energy* **2013**, *55*, 693–702. [CrossRef]
10. Englander, J.G.; Brandt, A.R.; Elgowainy, A.; Cai, H.; Han, J.; Yeh, S.; Wang, M.Q. Oil Sands Energy Intensity Assessment Using Facility-Level Data. *Energy Fuels* **2015**, *29*, 5204–5212. [CrossRef]
11. Jordaan, S.M.; Keith, D.W.; Stelfox, B. Quantifying land use of oil sands production: A life cycle perspective. *Environ. Res. Lett.* **2009**, *4*, 024004. [CrossRef]
12. Jordaan, S.M. Land and Water Impacts of Oil Sands Production in Alberta. *Environ. Sci. Technol.* **2012**, *46*, 3611–3617. [CrossRef] [PubMed]
13. Goldenberg, S. Keystone XL Pipeline: Obama Rejects Controversial Project. 2012. Available online: https://www.theguardian.com/environment/2012/jan/18/obama-administration-rejects-keystone-xl-pipeline (accessed on 3 December 2019).
14. Natural Resources Canada. Oil Sands: Economic Contributions 2016. Available online: https://www.nrcan.gc.ca/energy/publications/18756 (accessed on 28 September 2020).
15. Zhang, Y.; Hu, J.; Zhang, Q. Simulation Study of CO2 Huff-n-Puff in Tight Oil Reservoirs Considering Molecular Diffusion and Adsorption. *Energies* **2019**, *12*, 2136. [CrossRef]
16. Keshavarz, M. Analytical Modeling of Steam Injection and Steam-Solvent Co-Injection for Bitumen and Heavy Oil Recovery with Parallel Horizontal Wells. Ph.D. Thesis, University of Calgary, Calgary, AB, Canada, April 2019.
17. Cenovus Energy. Cenovus Uses Solvents to Improve Its SAGD 2020. Available online: https://www.cenovus.com/technology/solvents.html (accessed on 11 August 2020).
18. Oil Sands Advisory Group. *The Oil Sands Advisory Group ("OSAG") Recommendations on Implementation of the Oil Sands Emissions Limit Established by the Alberta Climate Leadership Plan ("ACLP"): Executive Summary*; Oil Sands Advisory Group: Edmonton, AB, Canada, 2017.
19. Government of Alberta. Capping Oil Sands Emissions 2020. Available online: https://www.alberta.ca/climate-oilsands-emissions.aspx (accessed on 30 September 2020).
20. Alberta Energy Regulator. *How to Submit Volumetric Data to the AER*; Alberta Energy Regulato: Calgary, AB, Canada, 2019.
21. Ge, Z.; Song, Z.; Ding, S.X.; Huang, B. Data Mining and Analytics in the Process Industry: The Role of Machine Learning. *IEEE Access* **2017**, *5*, 20590–20616. [CrossRef]
22. Lv, Z.; Yan, X.; Jiang, Q. Batch process monitoring based on just-in-time learning and multiple-subspace principal component analysis. *Chemom. Intell. Lab. Syst.* **2014**, *137*, 128–139. [CrossRef]
23. Liu, Y.; Zhang, G. Scale-sifting multiscale nonlinear process quality monitoring and fault detection. *Can. J. Chem. Eng.* **2015**, *93*, 1416–1425. [CrossRef]
24. Yu, H.; Khan, F.; Garaniya, V. Nonlinear Gaussian Belief Network based fault diagnosis for industrial processes. *J. Process. Control.* **2015**, *35*, 178–200. [CrossRef]
25. Thomas, M.C.; Zhu, W.; Romagnoli, J.A. Data mining and clustering in chemical process databases for monitoring and knowledge discovery. *J. Process. Control.* **2018**, *67*, 160–175. [CrossRef]
26. Shokrollahi, A.; Tatar, A.; Safari, H. On accurate determination of PVT properties in crude oil systems: Committee machine intelligent system modeling approach. *J. Taiwan Inst. Chem. Eng.* **2015**, *55*, 17–26. [CrossRef]
27. Ramirez, A.M.; Valle, G.A.; Romero, F.; Jaimes, M. Prediction of PVT Properties in Crude Oil Using Machine Learning Techniques MLT. *SPE Lat. Am. Caribb. Pet. Eng. Conf.* **2017**. [CrossRef]
28. An, J.; Mikhaylov, A.; Moiseev, N. Oil price predictors: Machine learning approach. *Int. J. Energy Econ. Policy* **2019**, *9*, 1–6. [CrossRef]

29. Gumus, M.; Kiran, M.S. Crude oil price forecasting using XGBoost. In Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 5–8 October 2017; pp. 1100–1103. [CrossRef]
30. You, J.; Ampomah, W.; Sun, Q.; Kutsienyo, E.J.; Balch, R.S.; Dai, Z.; Cather, M.; Zhang, X. Machine learning based co-optimization of carbon dioxide sequestration and oil recovery in CO2-EOR project. *J. Clean. Prod.* **2020**, *260*, 120866. [CrossRef]
31. Krasnov, F.; Glavnov, N.; Sitnikov, A. A Machine Learning Approach to Enhanced Oil Recovery Prediction. Available online: https://www.researchgate.net/publication/321976616_A_Machine_Learning_Approach_to_Enhanced_Oil_Recovery_Prediction (accessed on 3 December 2019).
32. Kurian, D.; Sattari, F.; Lefsrud, L.; Ma, Y. Using machine learning and keyword analysis to analyze incidents and reduce risk in oil sands operations. *Saf. Sci.* **2020**, *130*, 104873. [CrossRef]
33. Kurian, D.; Ma, Y.; Lefsrud, L.; Sattari, F. Seeing the forest and the trees: Using machine learning to categorize and analyze incident reports for Alberta oil sands operators. *J. Loss Prev. Process. Ind.* **2020**, *64*, 104069. [CrossRef]
34. Esmaeili, S.; Sarma, H.K.; Harding, T.; Maini, B.B. Effect of Temperature on Bitumen/Water Relative Permeability in Oil Sands. *Energy Fuels* **2020**, *34*, 12314–12329. [CrossRef]
35. Li, C.; Jan, N.M.; Huang, B. Data analytics for oil sands subcool prediction—A comparative study of machine learning algorithms. *IFAC-Pap* **2018**, *51*, 886–891. [CrossRef]
36. Gullo, F. From Patterns in Data to Knowledge Discovery: What Data Mining Can Do. *Phys. Procedia* **2015**, *62*, 18–22. [CrossRef]
37. Han, J.; Pei, J.; Kamber, M. *Data Mining: Concepts and Techniques*; Elsevier: Waltham, MA, USA, 2011.
38. Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. *Advances in Knowledge Discovery and Data Mining*; American Association for Artificial Intelligence: Menlo Park, CA, USA, 1996.
39. Petrinex. Petrinex Public Data. Petrinex Public Data Web Portal 2019. Available online: https://www.petrinex.gov.ab.ca/PublicData (accessed on 2 December 2019).
40. Government of Alberta. Oil and Gas Conservation Act, Oil Gas Conserv Act Oil Gas Conserv Rules Alta Regul 1511971 Amend Alta Regul 172019 2019. Available online: http://www.qp.alberta.ca/documents/Regs/1971_151.pdf (accessed on 17 November 2019).
41. Alberta Energy Regulator. ST53: Alberta In Situ Oil Sands Production Summary 2020. Available online: https://www.aer.ca/providing-information/data-and-reports/statistical-reports/st53.html (accessed on 15 April 2020).
42. Si, M. Data Mining and Unsupervised Machine Learning in Canadian In Situ Oil Sands Database for Knowledge Discovery and Carbon Cost Analysis. Available online: https://data.mendeley.com/datasets/8ngkgz69zb/3 (accessed on 3 December 2019).
43. Rokach, L.; Maimon, O. Clustering Methods. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O., Rokach, L., Eds.; Springer: New York, NY, USA, 2005; pp. 321–352. [CrossRef]
44. Guha, S.; Mishra, N. Clustering Data Streams. In *Data Stream Manag*; Garofalakis, M., Gehrke, J., Rastogi, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2016; pp. 169–187. [CrossRef]
45. Hartigan, J.A. *Clustering Algorithms*; Wiley: New York, NY, USA, 1975.
46. Bagui, S.; Dhar, P.C. Positive and negative association rule mining in Hadoop's MapReduce environment. *J. Big Data* **2019**, *6*, 75. [CrossRef]
47. Li, Y.; Wang, J.; Duan, L.; Bai, T.; Wang, X.; Zhang, Y.; Qin, G. Association Rule-Based Feature Mining for Automated Fault Diagnosis of Rolling Bearing. *Shock. Vib.* **2019**, *2019*, 1–12. [CrossRef]
48. Brin, S.; Motwani, R.; Silverstein, C. Beyond market baskets: Generalizing association rules to correlations. In Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, May 1997; pp. 265–276.
49. Silverstein, C.; Brin, S.; Motwani, R. Beyond Market Baskets: Generalizing Association Rules to Dependence Rules. *Data Min. Knowl. Discov.* **1998**, *2*, 39–68. [CrossRef]
50. Farouq, A. *Application of Solvent Slugs in Thermal Recovery Operations*; 1965. Available online: https://www.osti.gov/biblio/6685144 (accessed on 15 April 2020).
51. Lechtenberg, H.J.; Gates, G.L.; Caraway, W.H.; Baptist, O.C. *Field Study of Viscous Oil Production by Solvent Stimulation Wilmington*; Bureau of Mines: Washington, DC, USA, 1972.