REPUBLIC OF TÜRKİYE

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

ESTIMATION OF THE CRITICAL PROPERTIES OF COMPOUNDS USING MACHINE LEARNING ALGORITHMS

Roua ALHALOUSH

MASTER OF SCIENCE THESIS

Department of Chemical Engineering

Chemical Engineering Program

Supervisor

Prof. Dr. Hasan SADIKOĞLU

July, 2023

REPUBLIC OF TÜRKİYE

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

ESTIMATION OF THE CRITICAL PROPERTIES OF COMPOUNDS USING MACHINE LEARNING ALGORITHMS

A thesis submitted by Roua ALHALOUSH in partial fulfillment of the requirements for the degree of MASTER OF SCIENCE is approved by the committee on 25.07.2023 in the Department of Chemical Engineering, Chemical Engineering Program.

Prof. Dr. Hasan SADIKOĞLU
Yıldız Technical University
Supervisor

Approved By the Examining Committee

Prof. Dr. Hasan SADIKOĞLU, Supervisor
Yildiz Technical University _____

Assoc. Prof. Dr. Nil ACARALI, Member
Yildiz Technical University _____

Assist. Prof. Dr. Özgün YÜCEL, Member
Gebze Technical University _____

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of Estimation Of The Critical Properties Of Compounds Using Machine Learning Algorithms supervised by my supervisor, Prof. Dr. Hasan SADIKOĞLU. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Roua ALHALOUSH

Signature

Dedicated to my family

ACKNOWLEDGEMENTS

I would like to express my gratitude to Prof. Dr. Hasan SADIKOĞLU and to Res. Asst. Mert Akın İNSEL who supported me through all stages of my work on this thesis. I would like to thank all Yıldız Technical University Chemical Engineering Department faculty members and professors who helped me gain my chemical engineering knowledge and skills.

I owe my deepest and most sincere thanks to my dear family, who have not spared any support throughout my entire education life and supported me in every moment of my life. My dear mother Samar ALHALOUSH, and my brother Mohamed ALHALOUSH thank you for being the source of motivation that helped me move forward in all aspects of life.

Roua ALHALOUSH

TABLE OF CONTENTS

LIST OF SYMBOLS vii
LIST OF ABBREVIATIONS viii
LIST OF FIGURES ix
LIST OF TABLES x
ABSTRACT xi
ÖZET xiii
1 INTRODUCTION 15
1.1 Literature Review 15
1.2 Objective of the Thesis 18
1.3 Original Contribution 19
2 THEORY 20
2.1 Group Contribution Method 20
2.2 Lydersen Method 20
3 MACHINE LEARNING 24
3.1 Data in Machine Learning 24
3.2 Types of Machine Learning 25
4 METHODOLOGY 30
4.1 Data Collection and Selection Method 30
4.2 Used Software 31
4.3 Applied models 31
4.4 Evaluation method 39
5 RESULTS 46
5.1 The Results Obtained by Applying the Lydersen Method 46
5.2 Evaluation of Machine Learning Pergormances in Estimating The Critical Temprature 48
5.3 Evaluation of Machine Learning Pergormances in Estimating The Critical Pressure 51
5.4 Evaluation of Machine Learning Pergormances in Estimating The Critical Volume 54
5.5 Comparison Between Results 58
5.6 Pearson correalation 61
5.7 Feature Selection 64

5.8 Comparison with other studies. 66
5.9 Conclusion 66
REFERENCES **68**
A LYDERSEN CONSTANTS FOR ALCOHOL COMPUNDS **74**
PUBLICATIONS FROM THE THESIS **79**

LIST OF SYMBOLS

R^2	Coefficient of Correlation
P_c	Critical Pressure
T_c	Critical Temperature
V_c	Critical Volume
M_w	Molecular Weight
T_b	Normat Boiling Temperature
ΔP	Pressure Contribution
ΔT	Temprature Contribution
ΔV	Volume Contribution

LIST OF ABBREVIATIONS

ARE	Average Relative Error
CV	Cross Validation
DT	Decision Tree
GCM	Group Contribution Method
GPR	Gaussian Process Regression
KNN	K-Nearest Neighbor
MAE	Mean Absolute Errors
ML	Machine Learning
MMC	Maximum Margins Classifier
NN	Neural Network
RF	Random Forests
RMSE	Root of Mean Squared Errors
SSR	Sum of Squared Residuals
SST	Total Sum of Squares
SVC	Support Vector Classifier
SVM	Support Vector Machines

LIST OF FIGURES

Figure 1.1 Temperature-pressure phase diagram [78]. 15
Figure 1.2 Hydrogen bond between alcohol atoms [13]. 18
Figure 2.1 The molecular structure of 1,2,3-benzenetriol. 22
Figure 3.1 The steps of the machine learning process. 24
Figure 3.2 Types of machine learning algorithms with model examples. 29
Figure 4.1 Decision tree structure. 33
Figure 4.2 Neural Network structure[79]. 39
Figure 4.3 Cross validation method [80]. 40
Figure 4.4 The algorithm of Bayesian optimization [73]. 44
Figure 5.1 The literature values vs. estimated values of critical pressure. 47
Figure 5.2 The literature values vs. estimated values of critical temperature. 47
Figure 5.3 The literature values vs. estimated values of critical volume. 47
Figure 5.4 The Literature vs. predicted value of Tc for the SVM model. 49
Figure 5.5 The literature vs. predicted value of Tc for the GPR model. 49
Figure 5.6 The literature vs. predicted value of Pc for the GPR model. 52
Figure 5.7 The literature vs. predicted value of Pc for the NN model. 53
Figure 5.8 The literature vs. predicted value of Vc for the GPR model. 56
Figure 5.9 The literature vs. predicted value of Vc for the linear regression model. 56
Figure 5.10 The literature vs. predicted value of Vc for the stepwise linear regression model. 56
Figure 5.11 The literature vs. predicted value of Vc for the NN model. 57
Figure 5.12 The pearson correlation for input variables. 61
Figure 5.13 $\sum\Delta P$ graphed against $\sum\Delta V$. 62
Figure 5.14 M_w graphed against $\sum\Delta V$. 63
Figure 5.15 Mw graphed against $\sum\Delta P$. 63
Figure 5.16 Feature selection with F-test algorithm for the critical temperature. 64
Figure 5.17 Feature selection with F-test algorithm for the critical pressure. 65
Figure 5.18 Feature selection with F-test algorithm for the critical volume 65

LIST OF TABLES

Table 1.1 Critical properties of two-carbon organic compounds [12]. 17

Table 2.1 Group contributions for the Lydersen method[9]. 21

Table 2.2 Lydersen constants calculation for 1,2,3-benzentriol. 23

Table 2.3 Estimated vs. literature values of critical properties of 1,2,3-benzentriol. 23

Table 4.1 Description of the used data. 30

Table 4.2 The range of atoms number in used compounds. 30

Table 4.3 The range of values of the critical properties. 31

Table 5.1 Evaluation of the estimations of the Lydersen method. 46

Table 5.2 Evaluation of machine learning performance in predicting Tc. 48

Table 5.3 Training time and optimal hyperparameters for Tc estimation. 50

Table 5.4 Evaluation of machine learning performance in predicting Pc. 51

Table 5.5 Training time and optimal hyperparameters for Pc estimation. 53

Table 5.6 Evaluation of machine learning performance in predicting Vc. 55

Table 5.7 Training time and optimal hyperparameters for Vc estimation. 57

Table 5.8 Comparison between the Tc prediction accuracy of Lydersen method and ML models. 59

Table 5.9 Comparison between the Pc prediction accuracy of Lydersen method and ML models. 59

Table 5.10 Comparison between the Vc prediction accuracy of Lydersen method and ML models. 60

Table 5.11 outliers analysis 61

Table 5.12 Comparison between this study and other studies in the literature 66

ABSTRACT

Estimation Of The Critical Properties Of Compounds Using Machine Learning Algorithms

Roua ALHALOUSH

Department of Chemical Engineering

Chemical Engineering Program

Master of Science Thesis

Supervisor: Prof. Dr. Hasan SADIKOĞLU

Critical properties are considered a cornerstone in operating and designing industrial units. Obtaining these properties experimentally requires excessive consumption of energy and time. Thus, Group Contribution Methods are used for estimating the critical properties depending on the structure of chemical compounds. Lydersen method is one of the most successful methods developed in this scope. Many researchers used this method as a basis to develop improved models for the estimation of critical properties. Recently, machine learning algorithms were employed to enhance the estimation of different properties of chemical compounds. In this study, various machine learning algorithms (Linear Regression, SVM, DT, Ensembles, GPR, and NN) were used to estimate the critical properties of alcohol compounds. In this scope, Lydersen method was applied to 150 alcohol compounds to obtain Lydersen constants which were used to construct the machine-learning models. R^2, MSE, RMSE, and MAE were calculated to evaluate the accuracy of the machine learning models. 10-fold cross-validation method was used to test the validity of the models. Pearson correlation was performed to understand the relationship between input variables. Also, feature

selection was performed to understand the importance of the input variables in building the machine learning models. The estimation accuracy of the Lydersen method was compared to that of the machine learning models. Tc, Pc, and Vc were best estimated by the SVM model, GPR model, and GPR model, respectively. These models were found to yield the best performance with R^2 values of 0.9842, 0.9685, and 0.9964, respectively. In the estimation of Tc, Pc, and Vc using Lydersen method the R^2 values were found as 0.9759, 0.9318, and 0.9943, respectively. The obtained results were compared to the literature. The results obtained indicate that machine learning algorithms are powerful tools to generate robust models for property estimation.

Keywords: Critical properties, Lydersen method, group contribution method, machine learning algorithms.

.

ÖZET

Makine Öğrenmesi Algoritmalari Kullanilarak Bileşiklerin Kritik Özelliklerinin Tayini

Roua ALHALOUSH

Kimya Mühendisliği Anabilim Dalı

Yüksek Lisan Tezi

Danışman: Prof. Dr. Hasan SADIKOĞLU

Kritik özellikler, endüstriyel birimlerin işletilmesinde ve tasarlanmasında bir mihenk taşı olarak kabul edilir. Bu özellikleri deneysel olarak elde etmek aşırı enerji ve zaman tüketimi gerektirir. Bu nedenle, kimyasal bileşiklerin yapısına bağlı olarak kritik özelliklerin tahmin edilmesinde Grup Katkı Yöntemleri kullanılmaktadır. Lydersen yöntemi bu kapsamda geliştirilen en başarılı yöntemlerden biridir. Birçok araştırmacı, kritik özelliklerin tahmini için geliştirilmiş modeller geliştirmek için bu yöntemi temel olarak kullandı. Son zamanlarda, kimyasal bileşiklerin farklı özelliklerinin tahminini geliştirmek için makine öğrenimi algoritmaları kullanıldı. Bu çalışmada, alkol bileşiklerinin kritik özelliklerini tahmin etmek için çeşitli makine öğrenme algoritmaları (Lineer Regresyon, SVM, DT, Ensembles, GPR ve NN) kullanılmıştır. Bu kapsamda, makine öğrenme modellerini oluşturmak için kullanılan Lydersen sabitlerini elde etmek için 150 alkol bileşiğine Lydersen yöntemi uygulanmıştır. Makine öğrenimi modellerinin doğruluğunu değerlendirmek için R^2, MSE, RMSE ve MAE hesaplandı. Modellerin geçerliliğini test etmek için 10 kat çapraz doğrulama yöntemi kullanılmıştır. Girdi değişkenleri arasındaki ilişkiyi anlamak için Pearson

korelasyonu yapıldı. Ayrıca, makine öğrenmesi modellerinin oluşturulmasında girdi değişkenlerinin önemini anlamak için özellik seçimi yapılmıştır. Lydersen yönteminin tahmin doğruluğu, makine öğrenimi modellerininkiyle karşılaştırıldı. Tc, Pc ve Vc sırasıyla SVM modeli, GPR modeli ve GPR modeli tarafından en iyi şekilde tahmin edilmiştir. Bu modellerin sırasıyla 0,9842, 0,9685 ve 0,9964 R^2 değerleri ile en iyi performansı verdiği bulunmuştur. Lydersen yöntemi kullanılarak Tc, Pc ve Vc tahmininde R^2 değerleri sırasıyla 0.9759, 0.9318 ve 0.9943 olarak bulunmuştur. Elde edilen sonuçlar literatür ile karşılaştırıldı. Elde edilen sonuçlar, makine öğrenimi algoritmalarının, özellik tahmini için sağlam modeller oluşturmak için güçlü araçlar olduğunu göstermektedir.

Anahtar Kelimeler: Kritik özellikler, Lydersen yöntemi, grup katkı yöntemi, makine öğrenimi algoritmaları.

1
INTRODUCTION

1.1 Literature Review

One of the basic roles of a chemical engineer is to design and operate chemical industrial units. To achieve this role a comprehensive knowledge of the chemical compounds that will be utilized in industrial production is needed. This knowledge includes the accurate determination of the chemical and physical properties of each compound such as vapor pressure, heat capacity, density, the heat of vaporization, and many other properties. most of these properties are generally obtained through the usage of the theorem of corresponding states[1]. This theorem suggests that all fluids diverge from the ideal gas behavior approximately to the same degree when compared at the same reduced pressure and reduced temperature. Obtaining the reduced pressure and temperature depend on the critical properties of the compound[2]. The critical properties of a chemical compound are the properties of the compound at its critical point. The critical point can be identified as the endpoint that terminates the phase equilibrium curve as illustrated in (Figure 1.1)[3].

The critical temperature T_c is the minimum temperature value at which a fluid cannot be liquified despite the amount of pressure applied. Similarly, the critical pressure P_c is the minimum value of pressure required to liquefy a fluid at its critical temperature. While the critical volume V_c can be identified as the volume of one mol of a substance at its critical point[4].

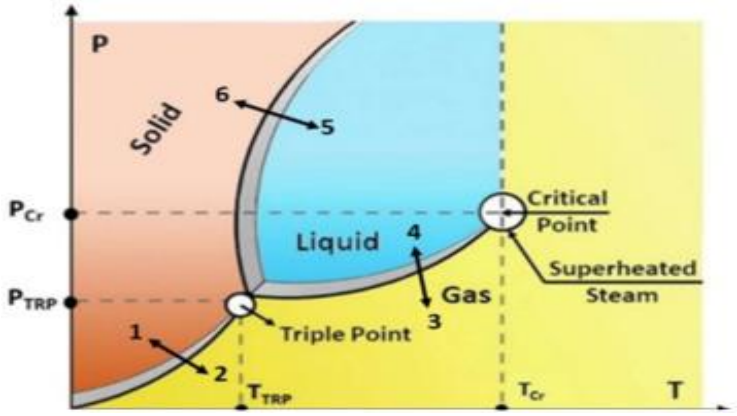


Figure 1.1 Temperature-pressure phase diagram [78].

Obtaining critical properties experimentally is very difficult and even impossible in some cases. The intensive need for energy and time rises the cost of applying experiments needed to obtain critical properties. In addition, large molecules become unstable and degrade under higher temperatures needed in such experiments[5], [6]. Due to the prementioned reasons non-experimental approaches were developed to estimate the critical properties of chemical compounds. The estimation approaches depend on the Group Contribution Method (GCM). The group contribution method depends on the simple principle of dividing each chemical compound into small parts. These parts can be atoms, bonds, or functional groups. Each part has a specific value known as a contribution. The contributions are measured using experimentally obtained data. By gathering the contributions of all the parts constituting the molecule different properties of the compound can be estimated[7].

The Lydersen method developed by Lydersen in 1955 is a group contribution method that is known to be one of the first effective methods used to estimate the critical properties of various chemical compounds. This method represents the base that many researchers used to develop enhanced estimation models such as Joback[8], Klincewicz[9], Gani-Constantinou[10], and others[6].

Recently researchers started to take advantage of the massive development that occurred in the machine learning field. Throughout the past few years, researchers applied machine learning methods to develop enhanced models for property estimation. The use of these developed models helped in improving the accuracy of the estimated data, which in turn raised the reliability of such data in designing and operating industrial units.

In this study, the ability of machine learning algorithms to provide enhanced models for the estimation of critical properties will be investigated. The performance of these models will be evaluated by comparing their estimation accuracy with the estimation accuracy of the Lydersen method. To perform this investigation a particular group of chemical compounds must be chosen. Alcohol compounds were selected to perform this study. The reasons for choosing alcohol compounds to be the subject of this study will be fully understood through the following parts.

Alcohols are defined as a group of organic compounds that contain at least one hydroxyl group (-OH) connected to a carbon atom within their structure. Alcohol compounds are considered to be among the most popular organic compounds encountered in chemical industries. The application fields of these compounds are endless. The alcohol compounds can be used in the industrial manufacturing of cleaning agents, fuels, drugs, dyes, textiles, and cosmetics. On a laboratory scale, alcohols are considered to be among the most frequently used solvents. They are frequently utilized in the synthesis of new chemical compounds[11].

Since alcohol compounds are of this degree of importance their properties were investigated thoroughly. One of the noticeable facts about alcohol compounds is that they tend to have higher values of critical properties than most of the other compounds that involve the same amount of carbon atoms within their structure. This can be seen clearly in (Table 1.1) where the critical properties of chemical compounds from different organic groups containing two carbon atoms within their structure are presented.

Table 1.1 Critical properties of two-carbon organic compounds [12].

Organic group	**Name of compound**	**Chemical structure**	**M_w**	**T_C**	**P_C**	**V_C**
Alkane	ethane	C_2H_6	30.07	305.42	48.8	147.9
Alkene	ethylene	C_2H_4	28.054	282.36	50.32	129.1
Alkyne	acetylene	C_2H_2	26.038	308.32	61.39	113
Amine	ethylamine	C_2H_7N	45.085	456.15	56.24	182
Alcohol	ethyl alcohol	C_2H_6O	46.069	516.25	63.84	166.9
Ether	dimethyl ether	C_2H_6O	46.069	400.1	53.7	170
Aldehyde	acetaldehyde	C_2H_4O	44.053	461	55.5	157
Thiol	ethyl mercaptan	C_2H_6S	62.136	499.15	54.9	207
Alkyl halide	ethyl chloride	C_2H_5Cl	64.514	460.35	52.69	200

Table 1.1 Critical properties of two-carbon organic compounds [12] (contineud).

Carboxylic acid	acetic acid	$C_2H_4O_2$	60.053	592.71	57.86	171
Ester	methyl formate	$C_2H_4O_2$	60.053	487.2	59.98	172

From the table, it is obvious that alcohols and carboxylic acids have the highest values for their critical properties. This is caused by the presence of the hydroxyl group in the structure of these compounds. The hydroxyl group is extremely polar which causes the molecules of these compounds to be attached to each other with a hydrogen bond which is illustrated in (Figure 1.2)[13]. The presence of hydrogen bonds increases the energy needed to perform experiments to obtain the critical properties. thus, there is a greater need to develop enhanced models for the estimation of the critical properties of these compounds. The data size available for alcohol compounds in the literature was larger than that of the carboxylic acid compounds. For these reasons, the critical properties of alcohol compounds were selected to be the subject of this study.

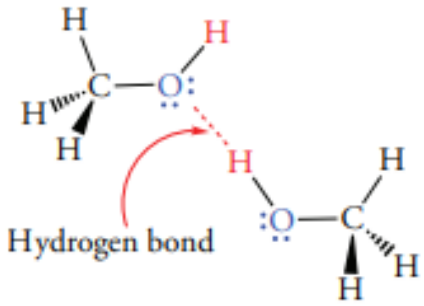


Figure 1.2 Hydrogen bond between alcohol atoms [13].

1.2 Objective of the Thesis

The aim of this study is to investigate the ability of machine learning algorithms to generate enhanced models for the estimation of critical properties. The generated models will be based on the group contribution method known as the Lydersen method. In this scope, the Lydersen method was applied to 150 alcohol compounds

to calculate Lydersen constants which were utilized as the basis of the machine learning models. The machine learning algorithms were applied using MATLAB software. The utilized machine learning algorithms in this study are Linear Regression, SVM, DT, Ensembles, GPR, and NN algorithms. A comparison between the performances of different machine learning algorithms was made. Also, a comparison between the performance of machine learning models and the performance of the original Lydersen method was done. Additionally, a comparison between the results obtained in this study and the results of other studies found in the literature is held.

1.3 Original Contribution

In this study, Lydersen constants for 150 alcohol compounds were obtained. Different machine-learning algorithms were applied to generate enhanced models for the estimation of critical properties The algorithms with the best performance were identified. The generated models are expected to yield more accurate results in comparison with the original Lydersen method. This study appears to be the only study that focuses on alcohol compounds which are known to be one of the organic groups with the highest values of critical properties.

2
THEORY

In this chapter, the main concepts used in the construction of this study will be discussed. For this sake, Group Contribution Method (GCM) will be explained in detail. Then, the Lydersen method which is the building block for the machine learning models utilized in this study. An explanatory example of the calculations of Lydersen method will be demonstrated.

2.1 Group Contribution Method

Group Contribution Method (GCM) is an approach used to estimate the values of thermodynamical properties. This method depends on the fact that all chemical compounds are formed from common building units. These units can be taken as atoms and bonds. Also, in a more complex form, the building units can be taken as functional groups. According to the group contribution method, each unit contributes with a specific value to a particular property, and this contribution will be the same in any chemical compound. This means that the properties of any chemical compound can be estimated depending on its molecular structure without the need to set up costly experiments[14]. This method can be described mathematically through equation (2.1).

$$X = f(X_0 + \sum_i n_i R_i) \tag{2.1}$$

Where X is the property that will be estimated, f is a simple function, X_0 is a constant, n_i is the number of groups i in the molecule, and R_i is the contribution of group i. This general equation represents the mathematical implication of the GCM's concept, which was used by researchers to create models for investigating various properties such as the Lydersen method that represents the base of our study and will be discussed in detail in the subsequent section.

2.2 Lydersen Method

The Lydersen method founded by Lydersen in 1955 is a group contribution method used to estimate the values of critical properties for different chemical compounds. Within this method, 43 different functional groups were specified with different

contributions. The critical properties can be estimated by dividing the chemical compound into its constituent functional groups [15]. The contribution of each functional group can be determined from (Table 2.1).

Table 2.1 Group contributions for the Lydersen method[9].

Group type	Group	*ΔT*	*ΔP*	*ΔV*	Group type	Group	*ΔT*	*ΔP*	*ΔV*
Non-ring group contribution	−CH3	0.02	0.227	55	Oxygen group contribution	−O− (non-ring)	0.021	0.16	20
	−CH2−	0.02	0.227	55		−O− (ring)	0.014	0.12	8
	>CH−	0.012	0.210	51		>C=O (non-ring)	0.04	0.29	60
	>C<	0.00	0.210	41		>C=O (ring)	0.033	0.2	50
	=CH2	0.018	0.198	45		O=CH− (Aldehyde)	0.048	0.33	73
	=CH−	0.018	0.198	45		−COOH (acid)	0.085	0.4	80
	=C<	0	0.198	36		−COO− (ester)	0.047	0.47	80
	=C=	0	0.198	36		=O (except as above)	0.02	0.12	11
	≡CH	0.005	0.153	36	Nitrogen group contribution	−NH2	0.031	0.095	28
	≡C−	0.005	0.153	36		>NH (non-ring)	0.031	0.135	37
Ring group contribution	−CH2−	0.013	0.184	44.5		>NH (ring)	0.024	0.09	27
	>CH−	0.012	0.192	46		>N− (non-ring)	0.014	0.17	42
	>C<	-0.007	0.154	31		−N= (non-ring)	0.007	0.13	32
	=CH−	0.011	0.154	37		−CN	0.06	0.36	80
	=C<	0.011	0.154	37		−NO2	0.055	0.42	78
Halogen group contributio	−F	0.018	0.224	18	Sulfur group contribution	−SH	0.015	0.27	55
	−Cl	0.017	0.32	49		−S− (non-ring)	0.015	0.27	55
	−Br	0.01	0.5	70		−S− (ring)	0.008	0.24	45
	−I	0.012	0.83	95		=S	0.003	0.24	47
Oxygen group contribution	−OH (alcohol)	0.082	0.06	18					
	−OH (phenol)	0.031	-0.02	3					

The contributions are added together to get the summation of contributions. The summation of contributions value is then used in specific equations to predict the critical properties. The equations (2.2), (2.3), and (2.4) are the equations used to estimate the critical temperature T_c, critical pressure P_c, and critical volume V_c respectively in the Lydersen method [16].

$$T_c = \frac{T_b}{0.567+\sum n_i \Delta T_i - (\sum n_i \Delta T_i)^2} \quad (2.2)$$

Where T_b is the normal boiling temperature (K), n_i is the number of group i in the molecule, and ΔT_i is the contribution of group i to the total critical temperature T_c(K).

$$P_c = \frac{M}{(0.34+\sum n_i \Delta P_i)^2} \tag{2.3}$$

Where M is the molecular weight of the molecule (g/mol), n_i is the number of group i in the molecule, ΔP_i is the contribution of group i to the total critical pressure P_c (bar).

$$V_c = 40 + \sum n_i \Delta V_i \tag{2.4}$$

Where n_i is the number of group i in the molecule, ΔV_i is the contribution of group i to the total critical molar volume V_c(cm^3/mol).

2.2.1 Lydersen method calculation example

In this part the application of Lydersen method will be demonstrated by an example of calculating the critical properties of 1,2,3-benzenetriol shown in (Figure 2.1).

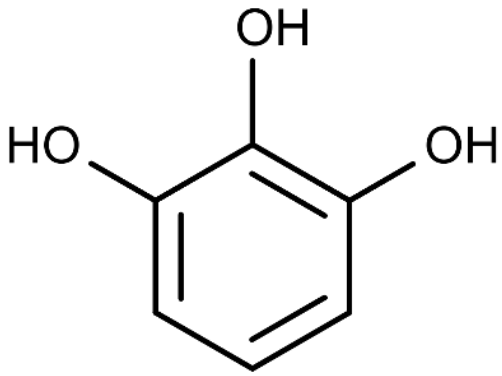


Figure 2.1 The molecular structure of 1,2,3-benzenetriol.

In the beginning, the compound will be divided into its constituent functional groups as shown in the (Table 2.2).

Table 2.2 Lydersen constants calculation for 1,2,3-benzentriol.

Group Type	Group	Number of Groups	ΔT	ΔP	ΔV
Ring group contribution	=CH-	3	0.033	0.462	111
	=C<	3	0.033	0.462	111
Oxygen group contribution	-OH (phenol)	3	0.093	-0.06	9
Summation	Σ		0.159	0.864	231

Given that the normal boiling point T_b of 1,2,3-benzenetriol is 581.85 (K) and its molecular weight Mw is 126.12 (g/mol) and using Lydersen equations (2.2), (2.3), and (2.4) the estimated values of the critical properties paired with the literature values are shown in the (Table2.3).

Table 2.3 Estimated vs. literature values of critical properties of 1,2,3-benzentriol.

Critical property	T_c	P_c	V_c
Units	K	bar	Cm^3/mol
Estimated value	830	87.0	271
Literature value	830	88.1	318

When we compare the values of the critical properties estimated by Lydersen method with the values obtained from the literature it can be concluded that even though Lydersen method provides proper estimations there is still room for improvement. This improvement can be performed using advanced computational techniques such as machine learning which is the main topic of this study. machine learning will be discussed in detail in the next chapter.

3
MACHINE LEARNING

Machine learning is a subsection of computer science that mainly focuses on the ways that a computer can follow to imitate the learning process of a human. In this scope, the computer uses different algorithms to discover a hidden pattern between a set of data. This hidden pattern resembles the mathematical relationship that best describes these data. By applying different algorithms, the computer is expected to find a model that fits the data and can be used either to predict or classify new data [17]. The process of machine learning consists of multiple steps that can be represented in (Figure 3.1). These steps will become clearer in the following sections.

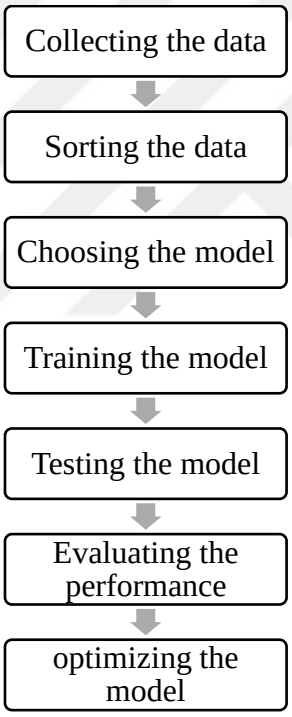


Figure 3.1 The steps of the machine learning process.

3.1 Data in Machine Learning

Similar to other computer science machine learning field mainly deals with data. For this reason, the process of applying machine learning for any purpose starts by collecting suitable data. The data collected must be reliable and sufficient to drive the learning process. Generally, the collected data are divided into three main sets:

- Training set which can be described as a sample taken from the data set in order to train the machine learning algorithms. It generally represents about 80% of the total data set.
- Test set which represents about 10% of the data set which is used to evaluate the performance of machine learning models.
- Validation set which is a sample taken from the data set that is used for tuning the hyperparameters of the algorithms[18].

The data set may include independent variables which can be called inputs or in machine learning terminology they are often referred to as features. Also, it may include dependent variables which can also be called as outputs, or as in machine learning terminology they are referred to as responses. when both input variables and the desired output are available in the data set the data point can be referred to as a labeled example. In contrast, when only input variables are available in the data set the data point is called an unlabeled example [19].

3.2 Types of Machine Learning

Depending on the data offered to the model machine learning algorithms can be classified into the following categories:

- Supervised machine learning
- Unsupervised machine learning
- Semi-supervised machine learning
- Reinforcement learning [20]

Knowing that the learning method used in this study is the supervised machine learning one, it will be discussed in detail in the following section while the other methods will be handled briefly.

3.2.1 Supervised Learning

Supervised machine learning requires training data that include both inputs which can be referred to as features and the desired output which also can be referred to as response. This means that this type of learning deals only with labeled examples.

This learning category is usually used to fit the data into a model that relates the inputs to the desired output [20]. The generated model can be used for two purposes:

1- Prediction: to predict the response for future input data.

2- Inference: to understand the relationship between the inputs and the output.

In this category of machine learning the training data can be seen as a collection of training examples. Training examples containing one or more inputs, and the desired output are represented mathematically by vectors. Similarly, the total set of training data is represented mathematically as a matrix [21].

As we mentioned earlier the main goal is to find the best model that can successfully describe the relationship between the features and the responses. This model is found by applying different algorithms that mainly work to optimize an objective function. The objective function is a function that measures the difference between the estimated data and the true data. This objective function is iteratively optimized to get the best-fitting model that is usually identified by distinguishing parameters and hyperparameters. The hyperparameters can be defined as some kind of free parameters that are not found by the algorithms. Instead, they are included as inputs by giving them an initial value. Then their values are specified or tuned using optimization methods over the validation set [22].

Variables can be classified as quantitative or qualitative variables. Quantitative variables are the variables that are related to a certain quantity and can take numerical values. Weight, height, the degree of temperature, and other different measurements are considered examples of quantitative variables. On the other hand, qualitative variables are the variables that are related to certain qualities and can take a value of a specific class or category. The condition of an appliance (on/off), the type of a device (A/B/C), and the status of a student in a certain class (failed/ passed) can be good examples of qualitative variables. Depending on whether the data set includes quantitative or qualitative variables the algorithms used for supervised machine learning can be categorized into two different classes: regression and classification [23].

3.2.1.1 Regression

Regression is a group of statistical methods used for numerical prediction, where the model used generally produces a continues-valued function that demonstrates

the relationship between a dependent variable and independent variables. The models used for regression can be called predictors [24]. There are many machine learning algorithms that are used for regression such as:

- Linear Regression
- polynomial Regression
- Support Vector Regression

3.2.1.2 Classification

Classification is a group of statistical processes used for class labeling the data. This means that it is used for identifying the class or category that the data belongs to. In this case, the models used for classification can be also called classifiers [25]. Some of the machine learning algorithms that are used for classification are as follows:

- Decision Tree
- Random Forests
- Support Vector Classification

The differentiation between classification and regression models is not always at this clarity. For example, the name of the logistic regression model implies that it is a regression model, but this is not the case. This model is actually used for classification by estimating the class probability of the data [26]. Also, some of the models can be used for both classification and regression purposes such as:

- K-Nearest Neighbor
- Support Vector Machine

The models used in this study will be discussed in detail in the methodology section.

3.2.2 Unsupervised Machine Learning

The unsupervised machine learning algorithms are characterized by the fact that the data offered to the models include only inputs. This indicates that this type of machine learning deals only with unlabeled examples. This type of machine learning is typically more complex than supervised learning. The fact that the response data do not exist makes the situation more challenging because the

supervision factor is missing [27]. Unsupervised machine learning can be used for data analysis by exploring the given data. This type of learning generally attempts to find a common structure in the given data. One of the methods used in this context is clustering. In clustering the algorithm is trying to determine the shared attributes between the data, in order to sort the data with similar attributes into subgroups called clusters. Each cluster contains data with common features while the data found in different clusters are considered to be dissimilar [28]. Some of the clustering methods are:

- Hierarchical clustering
- K-means clustering

3.2.3 Semi-supervised machine learning

It is a type of machine learning that lies between supervised and unsupervised machine learning methods. This type of machine learning becomes useful when the data set contains few labeled examples with a large number of unlabeled examples. This is the case when the input variables can be collected easily while the desired output variables need much more effort and can be extensively harder to collect. It has been proven that the accuracy of learning algorithms is highly improved when a small number of labeled examples is incorporated with the unlabeled examples [29]. Semi-supervised learning can be used for two purposes:

1. transductive learning: which refers to the use of learning algorithms to conclude the proper labels for the given unlabeled data.
2. inductive learning: which refers to the use of learning algorithms to estimate the correct relationship between inputs and outputs.

There are many methods that can be used for semi-supervised learning, but the most common methods are the following ones:

- Generative models
- Transductive support vector machines
- Graph-based methods like: Laplacian regularization [30].

3.2.4 Reinforcement learning

Reinforcement learning is a branch of machine learning that relies on a reward system. This type of learning is considered to be a dynamic model, where sequential decisions are made. In this type of learning the machine is expected to make intelligent decisions in order to maximize the reward collected from a certain environment [31]. The details of this learning method are beyond the scope of this study. (Figure 3.2) summarizes the different methods of machine learning.

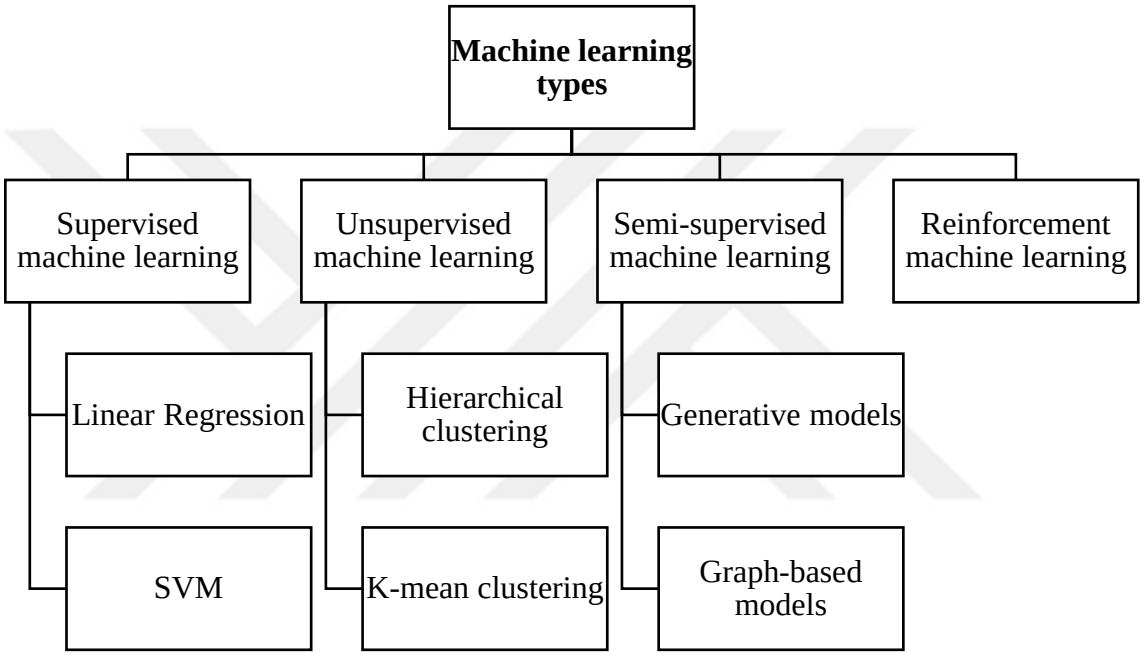


Figure 3.2 Types of machine learning algorithms with model examples.

4
METHODOLOGY

4.1 Data Collection and Selection Method

In this study, 150 alcohol compounds were analyzed. The literature data used for Lydersen constants calculations and for the evaluation of the success of the generated model in this study were collected from Yaws' Handbook of Thermodynamic and Physical Properties of Chemical Compounds [12]. The description of the data set used in this study can be illustrated in the (Table 4.1). The composition of alcohol compounds used in this study varies according to (Table 4.2). The range values of each of the critical properties used in this study are presented in the (Table 4.3).

Table 4.1 Description of the used data.

Factor	Name	Unit	Description
M_W	Molecular weight	g/mol	input
T_b	Normal boiling temperature	K	input
Tc	Critical temperature	K	output
Pc	Critical pressure	Bar	output
Vc	Critical volume	cm^3/mol	output
ΣΔT	Summation of temperature contribution	K	input
ΣΔP	Summation of pressure contribution	Bar	input
ΣΔV	Summation of volume contribution	cm^3/mol	input

Table 4.2 The range of atoms number in used compounds.

The atom type	Min atom number	Max atom number
C	1	24
O	1	6
H	4	42
N	1	1

Table 4.3 The range of values of the critical properties.

Property	Unit	Max. value	Min. value
Tc	K	506.2	959
Pc	Bar	11.4	88.1
Vc	cm^3/mol	117.8	1220

4.2 Used Software

For the calculations of Lydersen's constants Excel program was used. For the application of machine learning algorithms, MATLAB program was used.

4.3 Applied models

In this section, the machine learning models used in this study will be discussed in detail.

4.3.1 Linear Regression

Linear regression is one of the well-known machine learning models. Despite its simplicity linear regression model succeeded to maintain a place among the most used models in machine learning and many other statistical disciplines. This model is considered to be one of the supervised learning models since it requires the availability of both input and output variables. This model suggests that there is a linear relationship between the input and the output variables. It is generally used for estimating the relationship between one or more input variables and the desired output. In case there is only one input variable combined with the desired output this type of regression is called simple linear regression. On the other hand, if there is more than one input variable combined with the desired output it is called multiple linear regression[32].

Simple linear regression depends on measuring the change that occurs in the dependent output variable as a result of a unit change in the independent input variable. This can be mathematically presented by the equation (4.1).

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{4.1}$$

Where Y is the dependent output variable, β_0 is the intercept, β_1 is the slope, X is the independent input variable, and ϵ is the error term. $\beta_0$and β_1 are two unknown constants that represent the model coefficients, or in machine learning terminology they are called the model parameters[33].

For multiple linear regression, the previous equation can be adjusted into the equation (4.2).

$$Y = \beta_0 + \sum_{i=1}^{N} \beta_i X_i + \epsilon_i \quad (4.2)$$

Where Y is the dependent output variable, β_0 is the intercept coefficient, β_i is ith the regression coefficient, X is the independent input variable, and ϵ is the error term [34].

Using the training data set the regression coefficients can be estimated. These estimated parameters define the preliminary linear function that will be tested later in the test set. The found linear model can be then used to predict future response values for new data [35].

4.3.2 Decision Tree

The Decision Tree models (DT) are machine learning models that are known for their simplicity and interpretability. These models are considered to be non-parametric supervised machine learning models. In decision tree models, the main data set is divided into smaller sets called regions. The division of the data set is done based on particular rules called splitting rules. Each region contains a group of data with specific similarities[36]. The decision tree model can be used for both regression and classification. Regression trees are used when the input and output variables are considered to be quantitative. While classification trees are used when the input and output variables are considered to be qualitative. These models can be used to predict future output values according to the region that the future input belongs to. For instance, in regression tree models the mean value of the output variables in a specific region is considered to be the predicted output for any future inputs that fall into the same region [37]. Similarly, in classification tree models the predicted output is taken as the common class in the specific region that the future inputs belong to. The performance of regression trees is measured using MSE evaluation metrics. For the classification trees, the performance is evaluated by measuring the impurity in the data using the entropy function [38].

The structure of the decision tree is built based on a top-down algorithm. The first node in the flowchart of the tree is called the root node which represents a single region that all data belongs to. the root node is divided into two internal nodes depending on a suitable splitting rule. The splitting of internal nodes is repeated until the terminal nodes are reached. Each terminal node indicates a smaller region that contains a subset of the data. The nodes are connected together using segments called branches [39]. A visual representation of the decision tree structure is shown in (Figure 4.1).

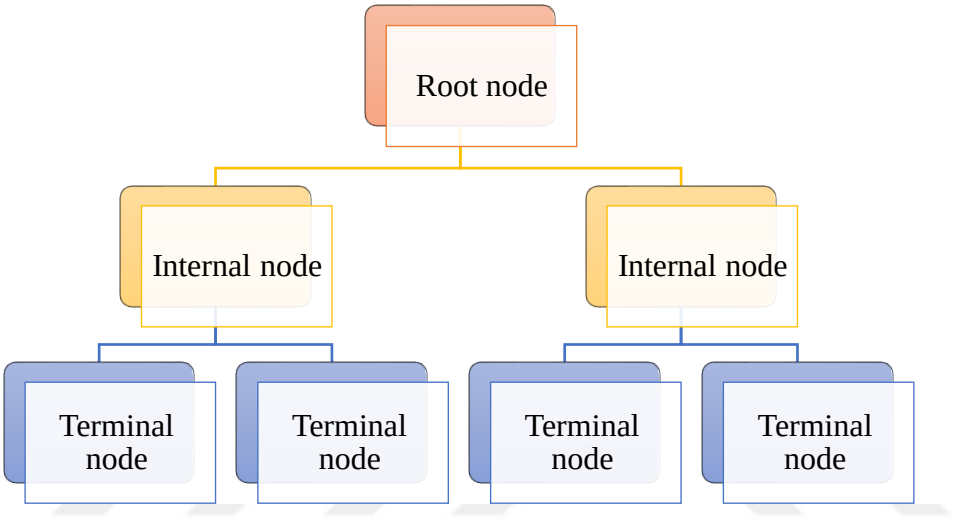


Figure 4.1 Decision tree structure.

4.3.3 Support Vector Machines (SVM)

Support vector machines are supervised machine learning models. These models are considered to be highly accurate models and that is what makes them among the most used machine learning models. Although support vector machine models can be used for both classification and regression analysis, they use slightly different algorithms for each purpose. The main concept of support vector machine models is to separate the data space into two distinct subspaces. This concept makes the support vector machines a binary classification model by structure[40].

The support vector machines can be thought of as an extension of the idea of a simple classifier called a Maximum Margin Classifier (MMC) also, known as hard margins. The concept of this classifier is that the separation of the data space is done by using a linear hyperplane. The hyperplane is a subspace that has a dimension

that is less by one unit than the dimension of the total data space. This hyperplane is combined with two margins that are placed on the opposite side of the hyperplane [41].

For classification, the margins are placed above the nearest data points to the hyperplane. These data points are then called support vectors. The support vectors are the factors that determine the position of the hyperplane and affect its orientation. The classification is achieved by maximizing the distance between the two margins to make the wider separation space possible. For regression, the margins are placed in such a way that the space contained between them holds all the data. The function that represents the hyperplane is then used to estimate the response values for future inputs.

Despite the simplicity of the application of the Maximum Margin Classifier idea, it fails to perform well in most data sets. There are two reasons for that:

1- The data are overlapped which makes it impossible to separate them using a linear hyperplane.
2- The existence of outliers because the structure of MMC makes it sensitive to outliers [42][.

This obstacle has been overcome through extending the idea of MMC by tolerating the misclassify of a few data to get a better performance for the whole data set. This is done by adding a slack variable that allows some data to fall on the wrong side of the margins or the hyperplane in some cases. This misclassifying will increase the error in the model and to solve this problem a hyperparameter called C, the cost of misclassifying is added to the objective function. A trade-off between the width of the margins and the total error is done to choose the best C. After the addition of C and the slack variable to the MMC, the resulting model is called Support Vector Classifier (SVC) which is also known as soft margins [43].

In other cases, the data set is linearly inseparable. Such data require a nonlinear hyperplane to divide the space, which makes both SVC and MMC inapplicable. This hurdle can be overcome by transferring the data into a higher dimensional space, finding a suitable hyperplane to separate the data, and then projecting them back to their original dimension. This process is computationally expensive, so a trick called the Kernel trick is used instead [44].

The kernel trick is a method that is used to find a nonlinear hyperplane to separate the data without the need to move to a space with a higher dimension. It depends on a simple concept of using the Kernel function which works by quantifying the similarities of two data. The combination of the Kernel function with the SVC is called Support Vector Machines. The kernel function has multiple types which are:

1- Linear Kernel
2- The Gaussian RBF Kernel (Radial Basis Function)
3- Polynomial Kernel
4- Sigmoid Kernel [46].

4.3.4 Gaussian Process Regression

Gaussian Process Regression (GPR) are non-parametric supervised machine learning algorithms. These algorithms gained popularity due to their ability to make good predictions even with small datasets. The name of the model suggests that it can only be used for regression analysis, but in fact, the model can also be used for classification [46]. The Gaussian Process Regression models are considered to be probabilistic models. This means that these models do not give an exact prediction value but instead, it gives a probability distribution of what the output value could be. This makes GPR models different from other models that give a single output as a prediction, which reflects the uncertainty in the given predictions[47]. Gaussian Process is entirely determined by two functions:

- The mean function m(x),
- The covariance function k(x, x')

So, we can express the Gaussian Process mathematically by the equation (4.3).

$$f(x) \sim GP(m(x), k(x, x')) \tag{4.3}$$

Usually, the mean function is taken to be zero, which means that the covariance function defines the Gaussian Process function's behavior [48]. The generally used covariance functions are listed as follows:

- Constant
- Linear
- White Gaussian noise

- Squared exponential
- Periodic
- Rational quadratic [47].

4.3.5 Tree-Based Ensemble Methods

Ensemble methods are commonly used supervised machine learning algorithms. These methods are generally used with the aim of raising the accuracy of the models that are considered to have weak prediction ability. They work by combining the outcomes of multiple models which results in better prediction ability. Ensemble methods can be used for both classification and regression analysis[49]. In regression analysis, ensemble methods give us the final prediction by taking the average prediction of all individual model predations. On the other hand, in classification, the final class prediction is performed according to the majority vote of classes predicted by individual models [50]. Ensemble methods can be categorized into two groups according to the type of models used in constructing the ensemble:

1- Heterogenous Ensemble methods works with models that are of different kinds. For example, different models like DT, SVM, KNN, or any other kind of model can work together in heterogeneous ensemble methods. The main concept behind this group of ensemble methods is that various models are used with the same data set. Each model predicts its own outcomes and at the end, the predictions of all models are gathered to predict the final outcome with enhanced accuracy.

2- Homogenous Ensemble methods uses a single type of model to predict the outcomes. In these methods, the data set is altered multiple times to get different variations of the same set. Then the selected model is used repeatedly to predict outcomes for each variation of the data set [51].

Tree-Based Ensemble Methods are ensemble methods that use the Decision Tree (TD) model as the basic prediction algorithm. Boosting, Bagging, and Random

Forests algorithms are the most popular ensemble methods. These algorithms are going to be discussed in the following sections.

4.3.5.1 Bagging (Bootstrap Aggregation)

Bootstrap Aggregation also known as Bagging is an ensemble learning algorithm that aims to lower the value of the variance of a certain model which subsequently reduces the total error and enhances the prediction ability of a model. The main idea of this algorithm that it derives different variations from the data set to obtain multiple subsets. The subsets are created by taking different samples from the data set with replacements. Then a Decision Tree (TD) model is applied to each sample to get predictions. By combining the predictions of each decision tree, the final prediction model is obtained [52]. This is mathematically expressed by the equation (4.4).

$$f(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \tag{4.4}$$

Where B is the number of samples, f_b is the decision tree model for a sample b, x is the future input variable, and $f(x)$ is the bagging model used to make future predictions [53].

4.3.5.2 Random Forests (RF)

Random Forests (RF) is an ensemble learning algorithm that provides an enhancement to the bagging algorithm explained previously. This enhancement is obtained by decorrelating the decision trees. The decorrelation of the decision trees causes the covariance between the trees to reduce which in turn reduces the total variance of the ensemble method. The reduction in the variance results in a reduction in the error and enhances the prediction ability of the model Decorrelating the decision trees is achieved by taking a random subset of the input variables at each split. Thus, the decision trees experience a different variety of features and the correlation between them is reduced [54].

4.3.5.3 Boosting

Is another ensemble learning algorithm that includes the building of multiple decision trees to enhance the prediction ability. The boosting works in a different way than bagging and random forests algorithms. In this method the inputs are fitted to a decision tree then the outcomes are divided into two parts. The first part is the

variables that were fitted in a good way to the model which are passed as a final output. The second part is the data that were not fitted well by the first decision tree which are referred to as residuals. These residuals are further taken to the next step where they will be fitted into another decision tree and the resulting outcomes are going to experience the same process. This process is repeated until the model is improved and the error is reduced. This way the model performs better in predicting new outcomes [55].

4.3.6 Neural Networks (NN)

Neural Networks (NN) are machine learning algorithms that are considered to be the heart of deep learning. The Neural Networks algorithms are based on a structural imitation of the neural networks of the human brain. The structure of neural networks is consisted of multiple layers each of which includes points or nodes called neurons. The points are connected together to form the network. What is meant by connection here is actually the reciprocal transmission of signals between the points through connection channels also known as edges [56]. In the simplest model of the neural network, the network is made of three layers which are:

1- Input layer: the layer that receives the input variables.

2- Hidden layer: the layer that processes the inputs and converts them to the outputs by applying computational operations.

3- Output layer: the layer that produces the final output variables [57].

The neurons of the input layer are connected to the neurons of the hidden layer. Each of the input variables is distributed into a neuron in the input layer. A numerical value called weight is assigned to each of the edges connecting the neurons. The inputs are processed with the weights and are fed to the hidden layer. Then a bias value is added at the hidden layer and fed into an activation function. The output of the activation function determines if a particular neuron will be activated. The activated neurons transmit the variables into the next layer, this process is called as forward propagation. When the variables reach the output layer

the variable with the highest value is taken as the desired output [58]. This process is illustrated in (Figure 4.2).

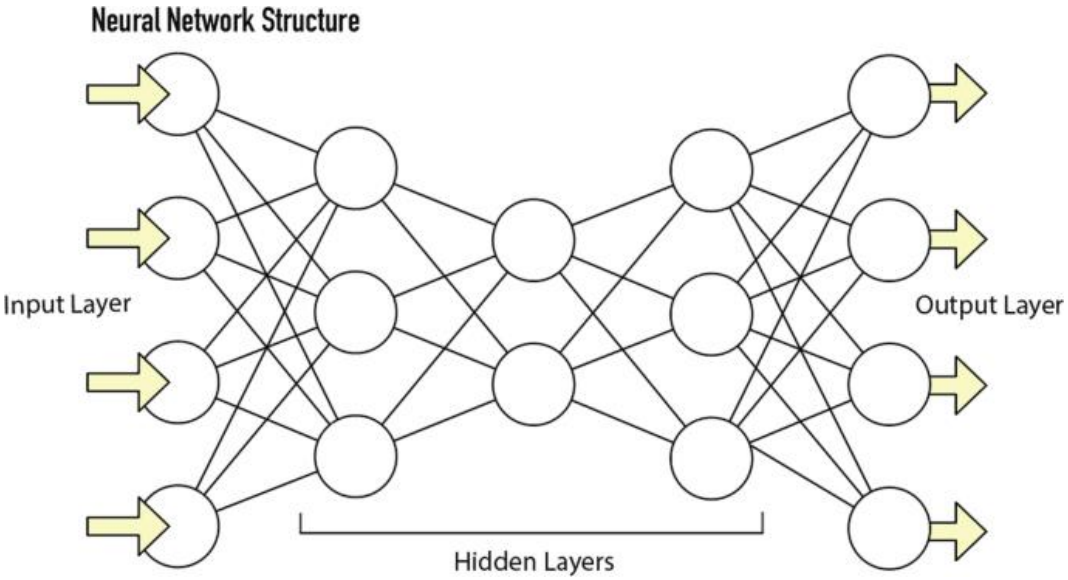


Figure 4.2 Neural Network structure[79].

4.4 Evaluation method

Evaluation methods are the methods used to evaluate the performance of a machine learning model. Machine learning models are fitted to the data in the training set. To evaluate their performance the trained models are tested in the test set. The evaluation is performed by measuring the estimation error made by the model. The error is measured using different evaluation metrics like R^2, MSE, RMSE, and MAE these metrics will be discussed in detail in the following sections. Sometimes, the dataset is not large enough so dividing it into two sets for training and testing will result in a poorly trained model that yields weak predictions. In these cases, validation methods are used to solve the problem. There are many validation methods that could be used in machine learning but the most popular one is Cross Validation method which will be discussed furtherly in the next section.

4.4.1 Cross-Validation Method

The cross-validation method is a validation method used in machine learning problems. This method follows a hold-out technique that enables it to deal with small datasets [59]. In this technique instead of dividing the original data set into a training set and a test set the whole data set is used for both training and testing the

model. This is achieved by dividing the data set into roughly equal non-overlapping parts. Then a single part is held out for testing the model while the other parts are used for training it. The performance of the model is evaluated by measuring the error in the test part. After obtaining the error of that part using evaluation metrics another part will be held out for testing and the process will be repeated over all the parts. each part will yield an error and at the end, the average of all the calculated errors will be reported as the final error of the model [60]. Cross validation method illustrated in (Figure 4.3) can be summarized by the following algorithm:

1- Divide the data set into k equal parts also called folds
2- Hold out the k^{th} part for testing
3- Train the model over the remaining parts
4- Evaluate the performance of the trained model using the held-out k^{th} part
5- Repeat the steps from 2 to 4 on each part
6- Combine the performance evaluation results for all parts [61].



Figure 4.3 Cross validation method [80].

4.4.2 Evaluation Metrics

4.4.2.1 Mean Squared Error

The Mean Squared Error (MSE) is a statistical equation used to evaluate the performance of regression models. This equation is considered to be a measure of the model's error which in our study can be defined as the difference between the output variable estimated using the regression model and the actual output value obtained from the literature or obtained experimentally [61]. The mathematical expression of MSE is presented in the equation (4.5).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \quad (4.5)$$

Where N is the number of data, Y_i is the actual output value of the data i, and $\hat{Y}_i$ is the estimated output value of the data i. The MSE is considered to be a loss function which means that the main goal in regression models is to minimize the value of this function to obtain a better fit of the data which yields better-estimating performance [62].

4.4.2.2 Root Mean Squared Error

The Root Mean Squared Error (RMSE) is one of the evaluation metrics used in this study. It is used to measure the residuals which are the difference between the actual output value and the estimated output value for each data point. it is taken as the squared root of MSE, and it is represented mathematically by the equation (4.6).

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2} \quad (4.6)$$

Where N is the number of data, Y_i is the actual output value of the data i, and $\hat{Y}_i$ is the estimated output value of the data i [63].

4.4.2.3 Mean Absolute Error

The Mean Absolute Error (MAE) is a statistical approach used to measure the absolute value of the difference between the actual output value and the estimated output value. The MAE which often referred to as the sum of absolute errors can be represented mathematically by the equation (4.7).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|(Y_i - \hat{Y}_i)\right| \quad (4.7)$$

Where N is the number of data, Y_i is the actual output value of the data i, and $\hat{Y}_i$ is the estimated output value of the data i [64].

4.4.2.4 The Coefficient of Correlation R^2

The coefficient of correlation R^2 is a statistical evaluation metric used to measure the performance of regression models that are usually used for prediction. It measures the accuracy of the fitted model. The values of R^2 ranges from [0 - 1], 1 being the perfect fit. The closer the value of R^2 gets to 1 the more accurate model is obtained. The value of R^2 can be calculated using the equation (4.8).

$$R^2 = 1 - \frac{SSR}{SST} \tag{4.8}$$

Where SSR is the Sum of Squared residuals and SST is the Total Sum of Squares. These two values can be calculated using equations (4.9) and (4.10), respectively.

$$SSR = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 \tag{4.9}$$

Where N is the number of data, Y_i is the actual output value of the data i, and $\hat{Y}_i$ is the estimated output value of the data i.

$$SST = \sum_{i=1}^{N} (Y_i - \overline{Y}_i)^2 \tag{4.10}$$

Where N is the number of data, Y_i is the actual output value of the data i, and $\overline{Y}_i$ is the mean value of the actual output of the data i [65].

4.4.2.5 Average Relative Error ARE

The Average Relative Error (ARE) is a measure of the difference between the actual output value and the estimated output value it can be calculated using the equation (4.11).

$$ARE(\%) = \frac{1}{N} \left| \frac{\hat{Y}_i - Y_i}{Y_i} \right| * 100 \tag{4.11}$$

Where N is the number of data, Y_i is the actual output value of the data i, and $\hat{Y}_i$ is the estimated output value of the data i [66].

4.4.3 Tuning of Hyperparameters

When a machine learning model is being trained the hyperparameters of the model must be given initial values in order to complete the training process. However, the values of the hyperparameters must be tuned later in the process to get the optimal values that enable the model to give the best performance possible [67]. There are a couple of methods that are generally used to find the optimal hyperparameters of the objective function such as:

- Manual Search
- Random Search
- Grid Search
- Bayesian optimization

The manual search, random search, and grid search are easily applicable methods that basically depend on the examination of various combinations of hyperparameters to get to the optimal combination but each in its unique way. Unfortunately, despite the simplicity of these methods they cannot be applicable to all cases especially when there is an enormous number of hyperparameter combinations that require a lot of computational work. in these cases, the Bayesian optimization method becomes very beneficial [68]. This method will be furtherly discussed since it is the method used in this study to estimate the hyperparameters of machine learning models.

4.4.3.1 Bayesian Optimization

Bayesian optimization is an optimization approach used to find the optimal hyperparameters for a machine learning model. This optimization is very useful in optimizing functions that demand excessive computational work. it is also very beneficial in evaluating black-box functions [69]. The main concept of this approach is that rather than making a big effort to find the hyperparameters using the objective function itself, it is possible to constitute a probability model of the objective function and utilize it to get to the optimal hyperparameters [70]. This is done through a certain algorithm that can be explained in a simplified way through the following steps:

1- Use a limited number of samples to build a surrogate model of choice that approximates the objective function.

2- Use an acquisition function

3- find the maximum point of the acquisition function.

4- Use the found point as a new sample to modify the surrogate model

5- Repeat steps 3 and 4 at each iteration until the maximum number of iterations is reached.

6- Examine the found samples to find the global minimum and select it as the optimal solution for the optimization problem [71].

In the above algorithm, the surrogate model is a probability model of the objective function that is constructed using samples of the pairs of hyperparameters to the

objective function score. Usually, the Gaussian regression process is used as the surrogate model for Bayesian optimization. The acquisition function is a function constructed based on the surrogate model rather than the original objective function. The expected improvement function is the most common type of acquisition function generally used in Bayesian optimization [72]. This algorithm is illustrated in (Figure 4.4).

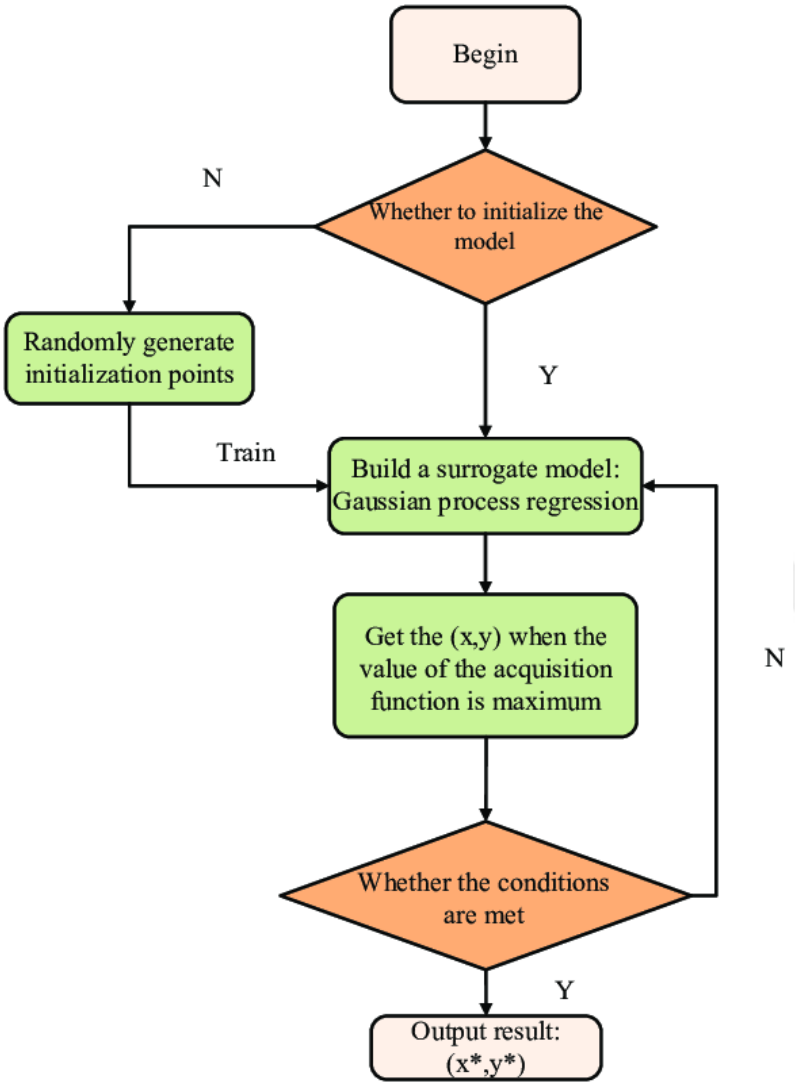


Figure 4.4 The algorithm of Bayesian optimization [73].

4.4.4 Pearson Correlation

Pearson correlation is a measure that is used to indicate whether two set of variables are related to each other. It can take values between -1 and 1. The closer the value of the correlation gets to 1 means that the two variables are strongly related to each

other. On the other hand, the closer it gets to -1 means that the variables are strongly inversely proportional. A value close to 0 indicates that the variables are independent from each other [74].

4.4.5 Feature Selection

Feature selection in machine learning is a group of methods that is usually used to understand how important an input variable can be in building the machine learning models. The application of these methods helps us to avoid the increase in model's complexity and enhances the overall accuracy of the model [26].

5
RESULTS

In this study, the critical properties of 150 alcohol compounds were estimated using different machine-learning models. The validation of machine learning models was fulfilled using a 10-fold cross-validation method. The performance of the models was evaluated using evaluation metrics. The results obtained for critical properties will be presented in the following sections. Also, a comparison between the results obtained by Lydersen method and the results obtained by machine learning algorithms will be discussed. Pearson correlation was performed to understand the relationship between the inputs. Also, feature selection was performed to understand the importance of each of the input features in building the machine learning models.

5.1 The Results Obtained by Applying the Lydersen Method

The Lydersen method was used to estimate the critical properties of 150 alcohol compounds. The critical properties estimated using the Lydersen method were compared to the values of the critical properties of the same compounds found in the literature. The performance of the Lydersen method was then measured using different evaluation methods. In (Table 5.1) the results of the evaluation metrics used are presented.

Table 5.1 Evaluation of the estimations of the Lydersen method.

	RMSE	MSE	R^2	MAE
T_c	13.6938	187.5192	0.9759	7.2244
P_c	3.9243	15.4003	0.9318	1.5737
V_c	20.2195	408.8284	0.9943	12.6520

The results in the table demonstrate that the Lydersen method is a relatively successful method to obtain critical properties. The graphs of the literature values against the estimated values for critical pressure, critical temperature, and critical

volume are shown in (Figure 5.1), (Figure 5.2), and (Figure 5.3), respectively. It can be observed from the graphs that the Lydersen method relatively succeeded to predict the values of the critical properties. The presence of outliers indicates that there is room for improving its performance by applying machine learning algorithms.

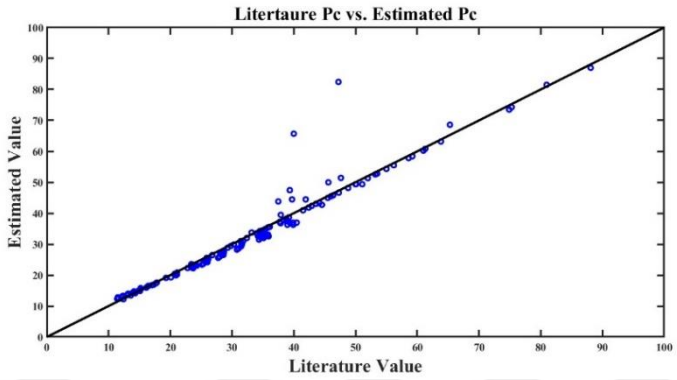


Figure 5.1 The literature values vs. estimated values of critical pressure.

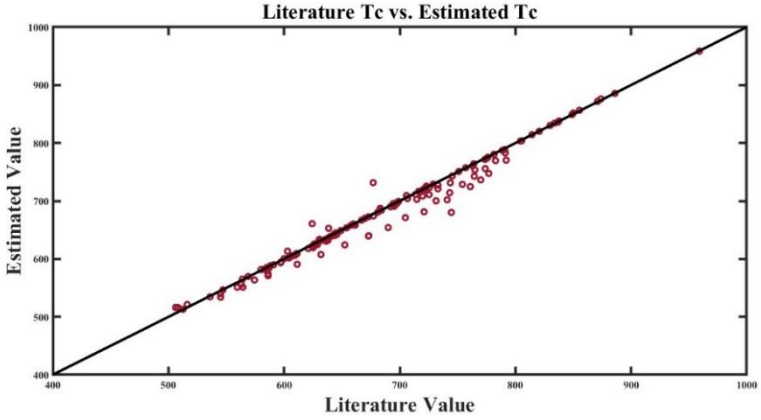


Figure 5.2 The literature values vs. estimated values of critical temperature.

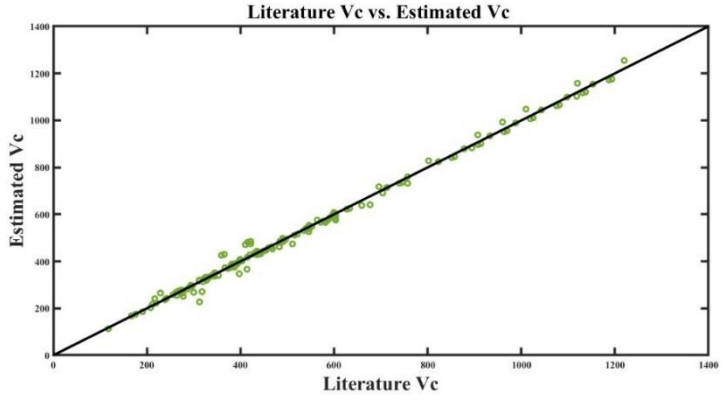


Figure 5.3 The literature values vs. estimated values of critical volume.

5.2 Evaluation of Machine Learning Pergormances in Estimating The Critical Temprature

Various machine-learning algorithms were used to predict the critical temperature of alcohol compounds in this study. The performance of the models was evaluated with different evaluation metrics. In (Table 5.2) the results obtained for each evaluation metric are presented.

Table 5.2 Evaluation of machine learning performance in predicting Tc.

Model Type	RMSE	MSE	R^2	MAE
Linear Regression	14.1995	201.6251	0.9742	10.6603
Linear Regression	16.3774	268.2199	0.9657	9.0997
Linear Regression	14.2054	201.7922	0.9742	10.1177
Stepwise Linear Regression	13.6399	186.0468	0.9762	8.5842
Tree	23.6586	559.7281	0.9284	17.6742
SVM	11.1151	123.5452	0.9842	6.3156
Gaussian Process Regression	11.2072	125.6016	0.9839	6.6627
Kernel	86.1050	7414.0765	0.0511	69.7499
Kernel	44.3061	1963.0322	0.7488	29.6906
Ensemble	21.9918	483.6381	0.9381	15.7310
Neural Network	51.4130	2643.2921	0.6617	36.2888

According to the results shown in the table, it can be clearly seen that the models with the best performance in estimating the critical temperature values were the

Support Vector Machines (SVM) model with RMSE= 11.1151, MSE=123.5452, MAE= 6.3156, and R^2 = 0.9842, and the Gaussian process regression (GPR) with RMSE=11.2072, MSE=125.6016, R^2=0.9839, and MAE=6.6627. (Figure 5.4) and (Figure 5.5) illustrates the literature values of critical temperature graphed against the predicted values produced by the SVM model, and GPR model respectively.

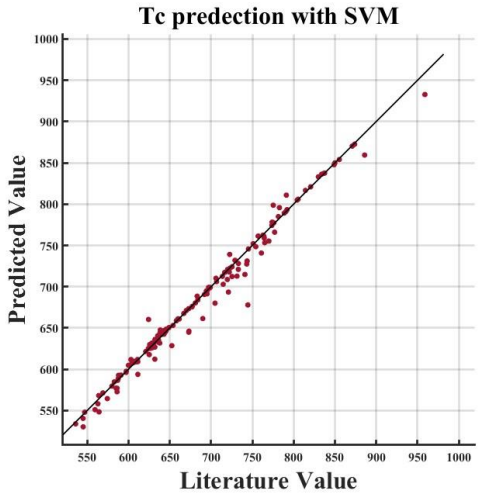


Figure 5.4 The Literature vs. predicted value of Tc for the SVM model.

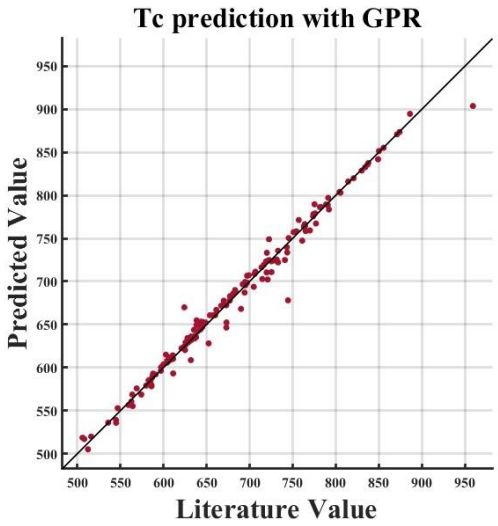


Figure 5.5 The literature vs. predicted value of Tc for the GPR model.

The optimal hyperparameters and training time for the machine learning models used for critical temperature estimation are presented in (Table 5.3).

Table 5.3 Training time and optimal hyperparameters for Tc estimation.

Model	Training time(sec)	Optimal hyperparameters
Linear Regression	2.2065	Terms: Linear; Robust option: Off
Linear Regression	2.1425	Terms: Interactions; Robust option: Off
Linear Regression	2.9826	Terms: Linear; Robust option: On
Stepwise Linear Regression	11.501	Initial terms: Linear; Upper bound on terms: Interactions; Maximum number of steps: 1000
Tree	73.459	Surrogate decision splits: Off; Optimized Hyperparameters; Minimum leaf size: 1;
SVM	233.71	Kernel scale: 1; Optimized Hyperparameters; Kernel function: Linear; Box constraint: 12.2313; Epsilon: 1.2231; Standardize data: Yes;
Gaussian Process Regression	201.93	Signal standard deviation: 10.6578; Optimize numeric parameters: Yes; ; Optimized Hyperparameters; Basis function: Constant; Kernel function: Nonisotropic Rational Quadratic; Kernel scale: 84.5996; Sigma: 10.6578; Standardize data: Yes;
Kernel	3.523	Learner: SVM; Number of expansion dimensions: Auto; Regularization strength (Lambda): Auto; Kernel scale: Auto; Epsilon: Auto; Iteration limit: 1000
Kernel	5.2343	Learner: Least Squares Kernel; Number of expansion dimensions: Auto; Regularization strength (Lambda): Auto; Kernel scale: Auto; Iteration limit: 1000

Table 5.3 Training time and optimal hyperparameters for Tc estimation (continued).

Ensemble	525.9	; Optimized Hyperparameters; Ensemble method: Bag; Minimum leaf size: 8; Number of learners: 30; Number of predictors to sample: 1;
Neural Network	569.5	Iteration limit: 1000; ; Optimized Hyperparameters; Number of fully connected layers: 3; Activation: ReLU; Regularization strength (Lambda): 0; Standardize data: Yes; First layer size: 10; Second layer size: 10; Third layer size: 10;

5.3 Evaluation of Machine Learning Pergormances in Estimating The Critical Pressure

The critical pressure values of various alcohol compounds were estimated using machine-learning algorithms in this study. Multiple evaluation metrics were used to evaluate the performance of each model. In (Table 5.4) the evaluation metrics results for the estimation of critical pressure obtained for each model are summarized.

Table 5.4 Evaluation of machine learning performance in predicting Pc.

Model Type	RMSE	MSE	R^2	MAE
Linear Regression	7.161965	51.29374	0.776154	5.191756
Linear Regression	9.55322	91.26401	0.601724	3.589971
Linear Regression	7.562933	57.19796	0.750388	4.615379
Stepwise Linear Regression	7.688031	59.10582	0.742063	3.307497
Tree	5.619468	31.57842	0.862192	3.31255

Table 5.4 Evaluation of machine learning performance in predicting Pc (continued).

SVM	4.837848	23.40478	0.897862	2.427041
Gaussian Process Regression	2.687156	7.220805	0.968488	1.377517
Kernel	12.55812	157.7063	0.31177	8.107197
Kernel	8.420451	70.904	0.690575	4.303623
Ensemble	4.904907	24.05812	0.895011	2.323244
Neural Network	3.528076	12.44732	0.94568	1.673407

According to the results shown in the table, it can be clearly seen that the models with the best performance in estimating the critical pressure values was the Gaussian Process Regression GPR model with RMSE= 2.687156, MSE=7.220805, MAE= 1.377517, and R^2 = 0.968488. also, Neural Networks NN with RMSE=3.528076, MSE=12.44732, R^2=0.94568, and MAE=1.673407. (Figure 5.6) and (Figure 5.7) illustrates the literature values of critical pressure graphed against the predicted values produced by the GPR model and NN model respectively.

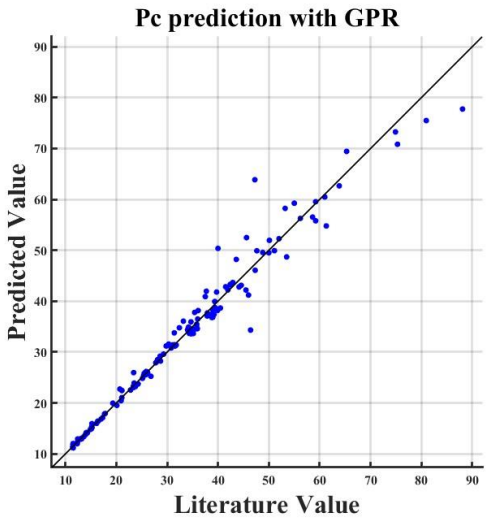


Figure 5.6 The literature vs. predicted value of Pc for the GPR model.

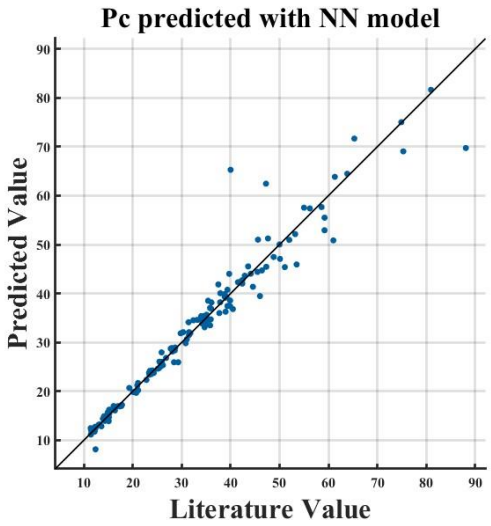


Figure 5.7 The literature vs. predicted value of Pc for the NN model.

The optimal hyperparameters and training time for the machine learning models used for critical pressure estimation are presented in (Table 5.5).

Table 5.5 Training time and optimal hyperparameters for Pc estimation.

Model	Training time(sec)	Optimal hyperparameters
Linear Regression	7.998531	Terms: Linear; Robust option: Off
Linear Regression	6.465572	Terms: Interactions; Robust option: Off
Linear Regression	5.82708	Terms: Linear; Robust option: On
Stepwise Linear Regression	10.22741	Initial terms: Linear; Upper bound on terms: Interactions; Maximum number of steps: 1000
Tree	73.84641	Surrogate decision splits: Off; ; Optimized Hyperparameters; Minimum leaf size: 1;
SVM	270.5283	Kernel scale: 1; ; Optimized Hyperparameters; Kernel function: Linear; Box constraint: 12.2313; Epsilon: 1.2231; Standardize data: Yes;

Table 5.5 Training time and optimal hyperparameters for Pc estimation (continued).

Gaussian Process Regression	138.4467	Signal standard deviation: 10.6578; Optimize numeric parameters: Yes; ; Optimized Hyperparameters; Basis function: Constant; Kernel function: Nonisotropic Rational Quadratic; Kernel scale: 84.5996; Sigma: 10.6578; Standardize data: Yes;
Kernel	7.606753	Learner: SVM; Number of expansion dimensions: Auto; Regularization strength (Lambda): Auto; Kernel scale: Auto; Epsilon: Auto; Iteration limit: 1000
Kernel	6.045319	Learner: Least Squares Kernel; Number of expansion dimensions: Auto; Regularization strength (Lambda): Auto; Kernel scale: Auto; Iteration limit: 1000
Ensemble	234.0193	; Optimized Hyperparameters; Ensemble method: Bag; Minimum leaf size: 8; Number of learners: 30; Number of predictors to sample: 1;
Neural Network	462.0245	Iteration limit: 1000; ; Optimized Hyperparameters; Number of fully connected layers: 3; Activation: ReLU; Regularization strength (Lambda): 0; Standardize data: Yes; First layer size: 10; Second layer size: 10; Third layer size: 10;

5.4 Evaluation of Machine Learning Pergormances in Estimating The Critical Volume

Various machine-learning algorithms were used to predict the critical volume of alcohol compounds in this study. The performance of the models was evaluated with different evaluation metrics. In (Table 5.6) the evaluation metrics results for the estimation of critical volume obtained for each model are summarized.

Table 5.6 Evaluation of machine learning performance in predicting Vc.

Model Type	RMSE	MSE	R^2	MAE
Linear Regression	20.03008	401.2042	0.994397	13.02917
Linear Regression	44.85432	2011.91	0.971901	17.36886
Linear Regression	21.12409	446.2274	0.993768	11.86253
Stepwise Linear Regression	20.1475	405.9218	0.994331	12.9369
Tree	32.19543	1036.546	0.985523	22.58299
SVM	23.2903	542.4379	0.992424	17.86443
Gaussian Process Regression	16.01828	256.5853	0.996416	10.99362
Kernel	274.9112	75576.17	-0.05554	198.6456
Kernel	113.192	12812.42	0.821055	63.78985
Ensemble	21.89346	479.3238	0.993306	16.01408
Neural Network	20.03008	401.2041	0.994397	13.02917

According to the results shown in the table, it can be clearly seen that the model with the best performance in estimating the critical volume values was the Gaussian Process Regression GPR model with RMSE= 16.01828, MSE=256.5853, MAE= 10.99362, and R^2 = 0.996416. Additionally, Linear regression, stepwise linear regression, and Neural Network NN models were observed to provide good estimation accuracy. (Figure 5.8), (Figure 5.9), (Figure 5.10) ,and (Figure 5.11) illustrates the literature values of critical volume graphed against the predicted values produced by the GPR, linear regression, stepwise linear regression, and NN models respectively.

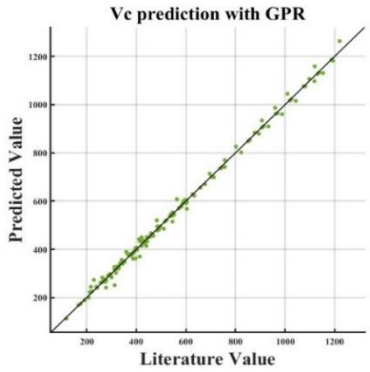


Figure 5.8 The literature vs. predicted value of Vc for the GPR model.

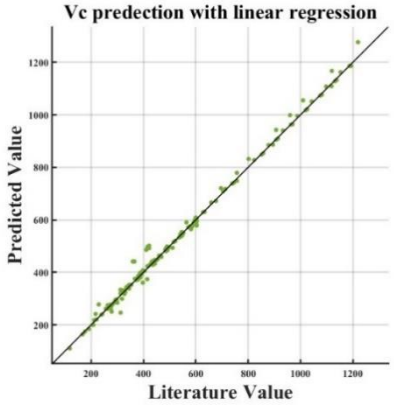


Figure 5.9 The literature vs. predicted value of Vc for the linear regression model.

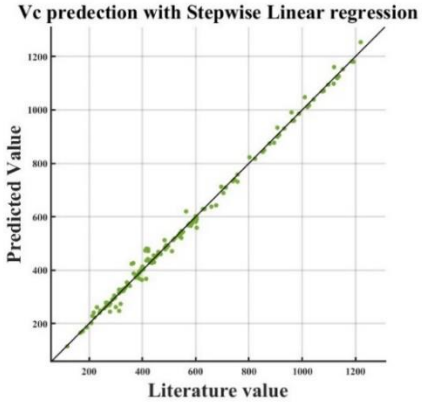


Figure 5.10 The literature vs. predicted value of Vc for the stepwise linear regression model.

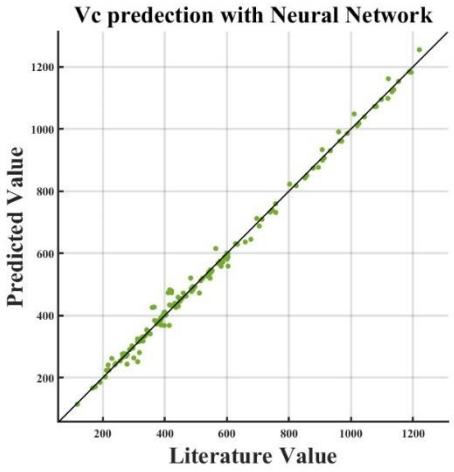


Figure 5.11 The literature vs. predicted value of Vc for the NN model.

The optimal hyperparameters and training time for the machine learning models used for critical volume estimation are presented in (Table 5.7)

Table 5.7 Training time and optimal hyperparameters for Vc estimation.

Model	Training time(sec)	Optimal hyperparameters
Linear Regression	2.318626	Terms: Linear; Robust option: Off
Linear Regression	1.143311	Terms: Interactions; Robust option: Off
Linear Regression	2.201102	Terms: Linear; Robust option: On
Stepwise Linear Regression	5.210274	Initial terms: Linear; Upper bound on terms: Interactions; Maximum number of steps: 1000
Tree	54.54287	Surrogate decision splits: Off; ; Optimized Hyperparameters; Minimum leaf size: 1;
SVM	239.1611	Kernel scale: 1; ; Optimized Hyperparameters; Kernel function: Linear; Box constraint: 232.765; Epsilon: 23.2765; Standardize data: Yes;

Table 5.7 Training time and optimal hyperparameters for Vc estimation (continued).

Gaussian Process Regression	202.8109	Signal standard deviation: 189.204; Optimize numeric parameters: Yes; ; Optimized Hyperparameters; Basis function: Constant; Kernel function: Nonisotropic Rational Quadratic; Kernel scale: 84.5996; Sigma: 189.204; Standardize data: Yes;
Kernel	2.776785	Learner: SVM; Number of expansion dimensions: Auto; Regularization strength (Lambda): Auto; Kernel scale: Auto; Epsilon: Auto; Iteration limit: 1000
Kernel	2.598434	Learner: Least Squares Kernel; Number of expansion dimensions: Auto; Regularization strength (Lambda): Auto; Kernel scale: Auto; Iteration limit: 1000
Ensemble	429.6577	; Optimized Hyperparameters; Ensemble method: Bag; Minimum leaf size: 8; Number of learners: 30; Number of predictors to sample: 1
Neural Network	598.0305	Iteration limit: 1000; ; Optimized Hyperparameters; Number of fully connected layers: 3; Activation: ReLU; Regularization strength (Lambda): 0; Standardize data: Yes; First layer size: 10; Second layer size: 10; Third layer size: 10;

5.5 Comparison Between Results

In this section, a comparison between the results obtained using the Lydersen method and using the machine learning algorithm will be held. In addition, a comparison between the results obtained between the results of this study and other studies conducted in the same field will be presented.

5.5.1 The Comparison of Critical Temperature Results

In (Table 5.8) The results of the evaluation metrics for both the Lydersen method and the machine learning models which were selected to be the best models for predicting the critical temperature values.

Table 5.8 Comparison between the Tc prediction accuracy of Lydersen method and ML models.

	RMSE	MSE	R^2	MAE
Lydersen Method	13.6938	187.5192	0.9759	7.2244
SVM	11.1151	123.5452	0.9842	6.3156
Gaussian Process Regression	11.2072	125.6016	0.9839	6.6627

As the results show in the previous table, both the SVM and the GPR models generated by machine learning algorithms delivered better results in estimating the critical temperature than Lydersen method.

5.5.2 The Comparison of Critical Pressure Results

The results of the evaluation metrics that measure the performance of both the Lydersen method and the machine learning models are presented in (Table 5.9) Considering that the Gaussian Prosses Regression GPR model and Neural Network NN were selected as the models with the best performance in estimating the critical pressure values.

Table 5.9 Comparison between the Pc prediction accuracy of Lydersen method and ML models.

	RMSE	MSE	R^2	MAE
Lydersen Method	3.9243	15.4003	0.9318	1.5737
GPR	2.687156	7.220805	0.96845	1.377517
Neural Network	3.528076	12.44732	0.94568	1.673407

As the results show in the previous table, both the GPR model and NN model generated by machine learning algorithms delivered better results in estimating the critical pressure than Lydersen method.

5.5.3 The Comparison of Critical Volume Results

The performance evaluation results of the critical volume obtained using the Lydersen method and using the Gaussian Process Regression GPR model, Linear regression model, stepwise linear regression model, and neural network model are presented in (Table 5.10).

Table 5.10 Comparison between the Vc prediction accuracy of Lydersen method and ML models.

	RMSE	MSE	R^2	MAE
Lydersen Method	20.2195	408.8284	0.9943	12.6520
GPR	16.01828	256.5853	0.9964	10.9934
Linear Regression	20.03008	401.2042	0.994397	13.02917
Stepwise Linear Regression	20.1475	405.9218	0.994331	12.9369
Neural Network	20.03008	401.2041	0.994397	13.02917

As the results show in the previous table, the GPR, Linear regression, stepwise linear regression, and Neural network NN models generated by machine learning algorithms delivered better results in estimating the critical volume than Lydersen method.

5.5.4 Outliers Analysis

In this section, the performance of machine learning algorithms will be evaluated by reviewing the estimation results of some data points. These points were found to be outliers when estimated by the Lydersen method. In (Table 5.11) the

estimation results of the outliers are presented. It can be clearly seen that the usage of machine learning algorithms improved the estimation accuracy of these points.

Table 5.11 outliers analysis

Property	Compound Name	Literature Value	Lydersen Estimation	ML-Estimation
P_c (bar)	Glycerol	40	65.596	50.3751
P_c (bar)	Cyclohexanol	37.49	43.812	40.891
T_c (K)	1-eicosanol	792	770.379	793.244
T_c (K)	1-haptanol	631.9	607.598	612.173
V_c (cm^3/mol)	3-isobutylphenol	415.5	481	433.25
V_c (cm^3/mol)	pyrocatechol	300	268	284.35

5.6 Pearson correalation

Pearson correlation was used to understand the relationship between the features and the results were obtained as in (Figure5.12).

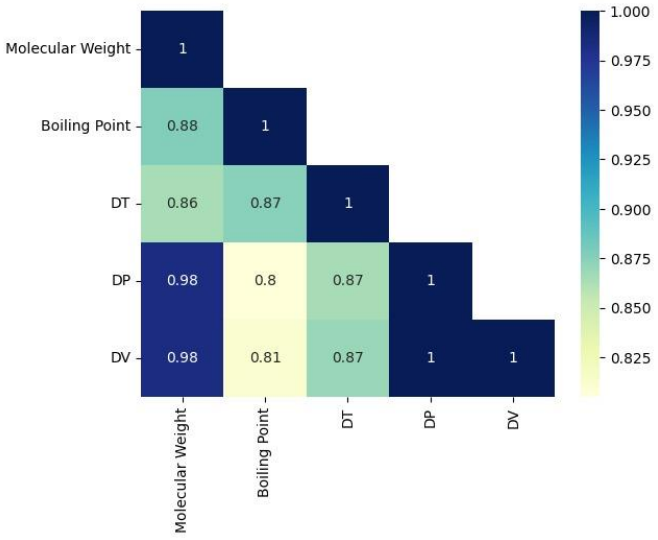


Figure 5.12 The pearson correlation for input variables.

From (Figure 5.12) it can be seen that the summation of pressure contribution ∑ΔP and the summation of temperature contribution ∑ΔV are correlated with R=1 which means that any of these features can be used in the construction of the model. A graph was constructed to fully understand the relationship between these two features and from (Figure 5.13) it can be seen that these two features are multiples of each other.

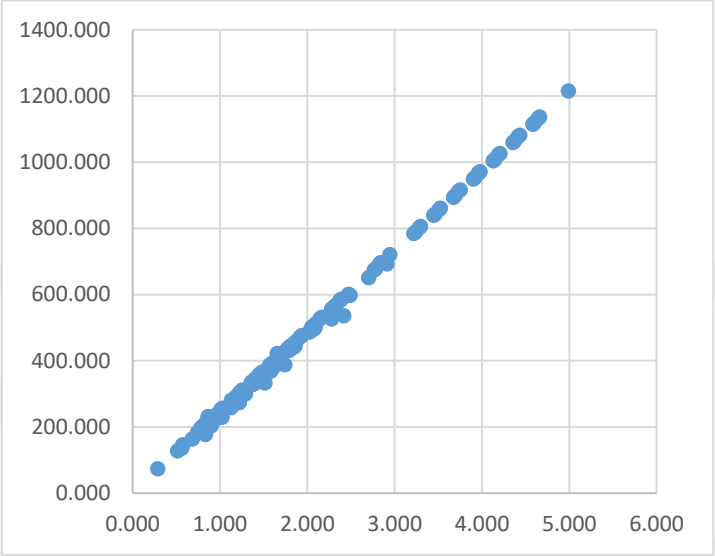


Figure 5.13 ∑ΔP graphed against ∑ΔV.

Also, it can be seen that the molecular weight M_w is correlated with both the summation of volume contribution ∑ΔV and the summation of pressure contribution ∑ΔP are related with R=0.98. This is predictable since as the molecular weight increases the summation of volume contribution ∑ΔV will increase because of the increase in the number of the molecule's constituents which will cause the occupied volume to increase. (Figure 5.14) represent the relationship between the ∑ΔV and the M_w.

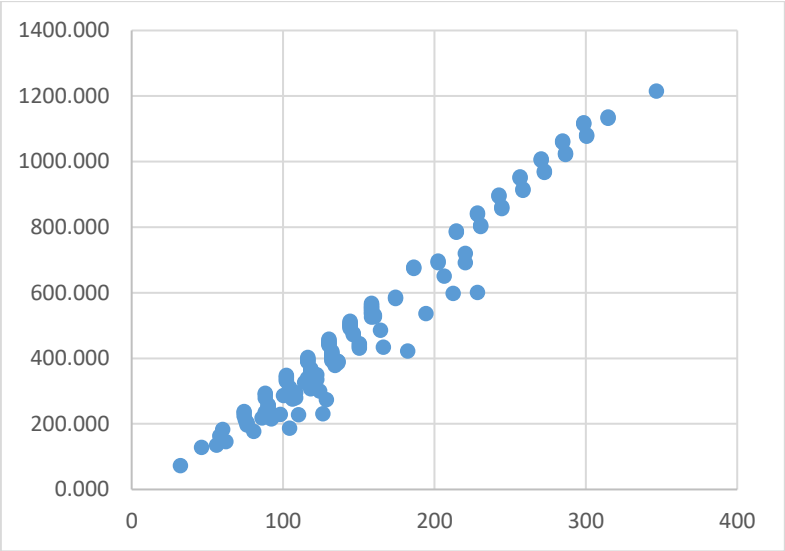


Figure 5.14 M_w graphed against $\sum\Delta V$.

The M_w is correlated with the $\sum\Delta P$ Since the $\sum\Delta V$ is correlated with $\sum\Delta P$ with an R=1. This means that if there is a feature that is correlated with one of them it will definitely be correlated with the other one by the same degree. (Figure 5.15) represent the relationship between the $\sum\Delta P$ and the M_w.

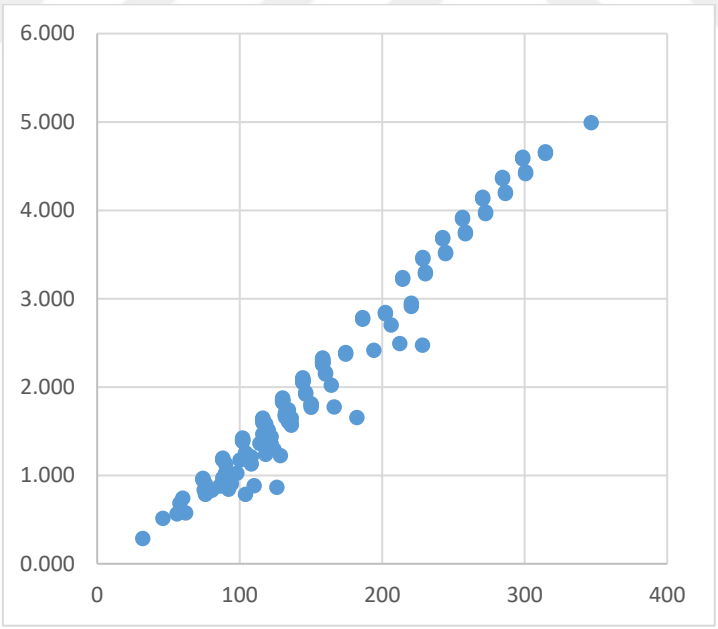


Figure 5.15 Mw graphed against $\sum\Delta P$.

Similarly, in (Figure 5.12) it can be seen that the summation of temperature contribution $\sum\Delta T$ is correlated with both $\sum\Delta P$ and $\sum\Delta V$ with the same value of R=0.87.

5.7 Feature Selection

In this study, feature selection was made to understand which were the most important features in the estimation of the critical properties. The feature selection was performed by applying the F-test algorithm for each of the critical properties.

5.7.1 Feature Selection for the Estimation of Critical Temperature

In equation (2.2) it can be seen that the Lydersen method uses only 2 features in the estimation of the critical temperature Tc which are the normal boiling point Tb and the summation of temperature contributions ∑ΔT. Machine learning algorithms use 5 features for the estimation of the same property which are Tb, Mw, ∑ΔT, ∑ΔP, ∑ΔV. As shown in (Figure 5.16) the most important feature in the estimation of critical temperature was found to be the normal boiling point Tb. It can be seen that the machine learning models utilized the other features with a noticeable importance degree. This enhanced the estimation accuracy of the machine learning algorithms over Lydersen method.

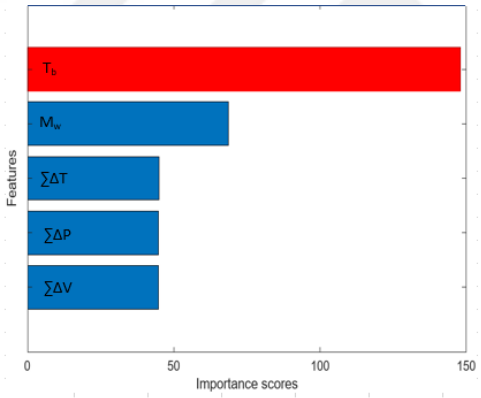


Figure 5.16 Feature selection with F-test algorithm for the critical temperature.

5.7.2 Feature Selection for the Estimation of Critical Pressure

In equation (2.3) the Lydersen method uses only 2 features in the estimation of the critical pressure Pc which are the molecular weight Mw and the summation of pressure contributions ∑ΔP. Machine learning algorithms use 5 features for the estimation of the same property which are Tb, Mw, ∑ΔT, ∑ΔP, ∑ΔV. As shown in (Figure 5.17) the most important feature in the estimation of critical pressure was found to be the summation of pressure contributions ∑ΔP. It can be seen that the machine learning models utilized the other features with a noticeable importance

degree. This enhanced the estimation accuracy of the machine learning algorithms over Lydersen method.

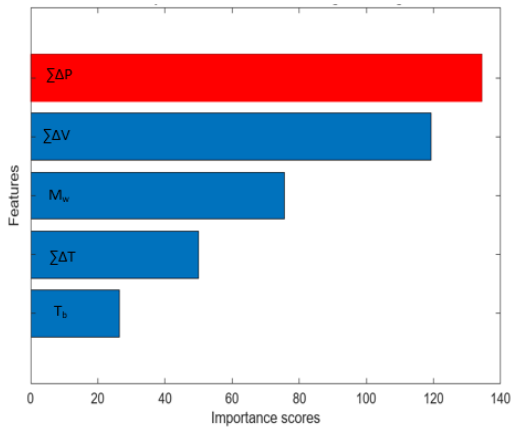


Figure 5.17 Feature selection with F-test algorithm for the critical pressure.

5.7.3 Feature Selection for the Estimation of Critical Volume

In equation (2.4) the Lydersen method uses only one feature in the estimation of the critical volume Vc which is the summation of volume contributions ∑ΔV. Machine learning algorithms use 5 features for the estimation of the same property which are Tb, Mw, ∑ΔT, ∑ΔP, ∑ΔV. As shown in (Figure 5.18) the most important feature in the estimation of critical pressure was found to be the summation of volume contributions ∑ΔV. It can be seen that the machine learning models utilized the other features with a noticeable importance degree. This enhanced the estimation accuracy of the machine learning algorithms over Lydersen method.

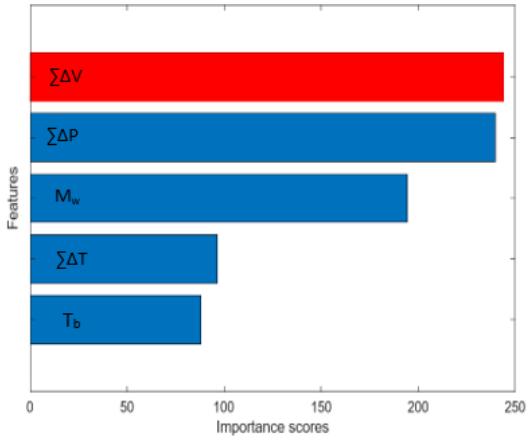


Figure 5.18 Feature selection with F-test algorithm for the critical volume.

5.8 Comparison with other studies.

In this part, the evaluation results obtained in this study are compared with the results of other studies from the literature that hold similar objectives. The results are shown in (Table 5.12). Taking into consideration the differences between the type of materials, algorithms, and number of data and comparing the results of R^2 and ARE metrics the comparison table shows that this study provides sufficiently good results.

Table 5.12 Comparison between this study and other studies in the literature

Ref.	Investigated properties			Data size			Validation metrics		
	T_c	P_c	V_c	T_c	P_c	V_c	T_c	P_c	V_c
[66]	☒	☒	☒	763	508	1743	AAE=7.79, ARE=1.11%	AAE=178 ARE=6.98 %	AAE=0.008, ARE=2.66%
[75]	☒	☐	☐	155	-	-	R^2 =0.9143, ARE=3.99%	-	-
[76]	☒	☒	☐	1078	1078	-	MAE=2.8, RMSE=5.2, R^2=0.99	MAE=0.34 RMSE=0.69 R^2=0.979	-
[77]	☒	☐	☐	800	-	-	R^2=0.9	-	-
[78]	☒	☒	☒	1792	1726	1801	AAE=23.77 ARE=5.29%	AAE=3.18, ARE=8.29%	AAE=19.92, ARE=6.15%
Our study	☒	☒	☒	150	150	150	RMSE= 11.115, MSE= 123.5, R^2= 0.9842 MAE= 6.316 ARE= 0.7699%	RMSE= 2.687 MSE= 7.221 R^2= 0.9685 MAE= 1.3775 ARE= 1.791%	RMSE= 16.0 MSE= 256.6 R^2= 0.996 MAE= 10.99 ARE= 0.081%

.

5.9 Conclusion

In this study, critical properties estimation was performed using different machine learning algorithms. The estimations were obtained for 150 different alcohol compounds. The utilized machine learning models were constructed on the basis of Lydersen constants. A 10-fold cross-validation method was used to validate the

performance of the models. The Bayesian optimization was used for tuning the hyperparameters of the models. Pearson correlation was utilized to understand the relationship between the input variables. Feature selection was performed to determine the importance of each of the input variables in building the machine-learning models. For comparison's sake, the Lydersen method was also utilized to estimate the critical properties of the compounds. After evaluating the results obtained using both the Lydersen method and machine learning algorithms it was observed that the machine learning models performed better at the estimation of the critical properties. SVM and GPR models were found to perform better at the estimation of the critical temperature than the original Lydersen method. GPR and NN models provided enhanced accuracy for the estimation of critical pressure when compared to the Lydersen method. Machine learning models performed exceptionally well at the estimation of critical volume since each of the GPR, NN, linear regression, and stepwise linear regression models provided better estimation accuracy than the Lydersen method. The data points that were found as outliers when estimated by Lydersen method were proven to be estimated with enhanced accuracy by the machine learning algorithms. It can be concluded that machine learning models were proven to be more successful in the estimation of critical properties than the original Lydersen method. When compared to similar studies in the literature this study appears to be the only study that was made on alcohol compounds which are known to be one of the organic groups with the highest values of critical properties. The models provided in this study are reported to be successful models for the estimation of critical properties of alcohol compounds with a number of carbon atoms that ranges between 1 to 24. The calculated Lydersen constants for 150 alcohol compounds are provided at the end of this study. Researchers can utilize these data for their future studies on the estimation of critical properties. This study can be enhanced by the use of chemical compounds from different organic groups using a larger amount of data. Also, the study can be expanded by the addition of other critical properties estimation methods such as Joback and Gani-Constantinou or other methods. Other artificial intelligence and deep learning algorithms may be utilized for the sake of comparison and improvement. The results of this study indicate that machine learning algorithms are powerful tools to generate robust models for property estimation. Thus, researchers can invest in developing new models using machine learning algorithms for the estimation of other properties.

REFERENCES

[1] Y. Nannoolal, J. Rarey, and D. Ramjugernath, "Estimation of pure component properties: Part 2. Estimation of critical property data by group contribution," Fluid Phase Equilib, vol. 252, no. 1–2, pp. 1–27, Mar. 2007, doi: 10.1016/J.FLUID.2006.11.014.

[2] J. W. Tester and M. Modell, "Thermodynamics and Its Applications," Book, vol. 3rd Editio, p. 936, 1997.

[3] A. R. Honerkamp-Smith, S. L. Veatch, and S. L. Keller, "An introduction to critical points for biophysicists; observations of compositional heterogeneity in lipid membranes," Biochimica et Biophysica Acta (BBA) - Biomembranes, vol. 1788, no. 1, pp. 53–63, Jan. 2009, doi: 10.1016/J.BBAMEM.2008.09.010.

[4] R. M. Stephenson and S. Malanowski, "Handbook of the Thermodynamics of Organic Compounds," Handbook of the Thermodynamics of Organic Compounds, 1987, doi: 10.1007/978-94-009-3173-2.

[5] C. F. Spencer and T. E. Daubert, "A critical evaluation of methods for the prediction of critical properties of hydrocarbons," AIChE Journal, vol. 19, no. 3, pp. 482–486, May 1973, doi: 10.1002/AIC.690190310.

[6] B. E. Poling et al., Properties of Gases and Liquids. McGraw-Hill Education, 2001. Accessed: Jan. 30, 2023. [Online]. Available: https://www.accessengineeringlibrary.com/content/book/9780070116825

[7] Z. Kolska, M. Zabransky, and A. Randov, "Group Contribution Methods for Estimation of Selected Physico-Chemical Properties of Organic Compounds," Thermodynamics - Fundamentals and Its Application in Science, Oct. 2012, doi: 10.5772/49998.

[8] K. G. Joback and R. C. Reid, "Estimation of Pure-Component Properties from Group-Contributions," Chem Eng Commun, vol. 57, no. 1–6, pp. 233–243, Jul. 1987, doi: 10.1080/00986448708960487.

[9] K. M. Klincewicz and R. C. Reid, "Estimation of critical properties with group contribution methods," AIChE Journal, vol. 30, no. 1, pp. 137–142, 1984, doi: 10.1002/AIC.690300119.

[10] L. Constantinou and R. Gani, "New group contribution method for estimating properties of pure compounds," AIChE Journal, vol. 40, no. 10, pp. 1697–1710, 1994, doi: 10.1002/AIC.690401011/TITLE/NEW_GROUP_CONTRIBUTION_METHOD_FOR_ESTIMATING_PROPERTIES_OF_PURE_COMPOUNDS.

[11] J. Falbe, H. Bahrmann, W. Lipps, D. Mayer, and G. D. Frey, "Alcohols, Aliphatic," Ullmann's Encyclopedia of Industrial Chemistry, Jan. 2013, doi: 10.1002/14356007.A01_279.PUB2.

[12] C. L. Yaws, "Yaws' handbook of thermodynamic and physical properties of chemical compounds : physical, thermodynamic and transport properties for 5,000 organic chemical compounds," 2003.

[13] R. J. Ouellette and J. D. Rawn, “Organic Chemistry - Structure, Mechanism, and Synthesis,” Organic Chemistry, pp. 315–342, 2014.

[14] L. Constantinou and R. Gani, “New group contribution method for estimating properties of pure compounds,” AIChE Journal, vol. 40, no. 10, pp. 1697–1710, Oct. 1994, doi: 10.1002/AIC.690401011.

[15] S. S. Godavarthy, R. L. Robinson, and K. A. M. Gasem, “Improved structure–property relationship models for prediction of critical properties,” Fluid Phase Equilib, vol. 264, no. 1–2, pp. 122–136, Mar. 2008, doi: 10.1016/J.FLUID.2007.11.003.

[16] J. O. Valderrama and P. A. Robles, “Critical properties, normal boiling temperatures, and acentric factors of fifty ionic liquids,” Ind Eng Chem Res, vol. 46, no. 4, pp. 1338–1344, Feb. 2007, doi: 10.1021/IE0603058/ASSET/IMAGES/MEDIUM/IE0603058E00015.GIF.

[17] A. I. Khan and S. Al-Habsi, “Machine Learning in Computer Vision,” Procedia Comput Sci, vol. 167, pp. 1444–1451, Jan. 2020, doi: 10.1016/J.PROCS.2020.03.355.

[18] J. Han and M. Kamber, “Mining Sequence Patterns in Transactional Databases,” Data Mining: Concepts and Techniques, pp. 498–512, 2011, Accessed: Jun. 19, 2023. [Online]. Available: http://medcontent.metapress.com/index/A65RM03P4874243N.pdf

[19] A. Jung, “Machine Learning,” 2022, doi: 10.1007/978-981-16-8193-6.

[20] T. Jo, “Machine Learning Foundations,” Machine Learning Foundations, 2021, doi: 10.1007/978-3-030-65900-4.

[21] G. Bounds, “Supervised Learning,” Encyclopedia of Machine Learning and Data Mining, pp. 1213–1214, 2017, doi: 10.1007/978-1-4899-7687-1_803.

[22] T. Hastie, J. Friedman, and R. Tibshirani, “The Elements of Statistical Learning,” 2001, doi: 10.1007/978-0-387-21606-5.

[23] N. R. Chrisman, “Rethinking Levels of Measurement for Cartography,” Cartogr Geogr Inf Sci, vol. 25, no. 4, pp. 231–242, 1998, doi: 10.1559/152304098782383043.

[24] K. A. Venkatesh, K. Mohanasundaram, and V. Pothyachi, Regression tasks for machine learning. Elsevier, 2023. doi: 10.1016/b978-0-323-91776-6.00009-9.

[25] K. A. Ross, “Classification,” Encyclopedia of Database Systems, pp. 426–432, 2018, doi: 10.1007/978-1-4614-8265-9_552.

[26] G. James, D. Witten, T. Hastie, and R. Tibshirani, “An Introduction to Statistical Learning,” 2021, doi: 10.1007/978-1-0716-1418-1.

[27] M. E. Celebi and K. Aydin, “Unsupervised learning algorithms,” Unsupervised Learning Algorithms, pp. 1–558, Jan. 2016, doi: 10.1007/978-3-319-24211-8/COVER.

[28] J. Van Ryzin and University of Wisconsin--Madison. Mathematics Research Center., “Classification and clustering : proceedings of an advanced seminar conducted by the Mathematics Research Center, the University of Wisconsin--Madison, May 3-5, 1976,” p. 467.

[29] X. Zhu and A. B. Goldberg, “Introduction to Semi-Supervised Learning,” Introduction to Semi-Supervised Learning, 2009, doi: 10.1007/978-3-031-01548-9.

[30] “Semi-supervised Learning,” Encyclopedia of the Sciences of Learning, pp. 3036–3036, 2012, doi: 10.1007/978-1-4419-1428-6_2402.

[31] M. Van Otterlo and M. Wiering, “Reinforcement learning and markov decision processes,” Adaptation, Learning, and Optimization, vol. 12, pp. 3–42, 2012, doi: 10.1007/978-3-642-27645-3_1.

[32] D. L. Mohr, W. J. Wilson, and R. J. Freund, Statistical Methods. Elsevier, 2021. doi: 10.1016/B978-0-12-823043-5.00015-1.

[33] X. Gang Su, “Linear regression analysis: Theory and computing,” Linear Regression Analysis: Theory and Computing, pp. 1–328, Jan. 2009, doi: 10.1142/6986.

[34] A. C. Rencher and F. C. William, “Methods of multivariate analysis: Third edition,” Methods of Multivariate Analysis: Third Edition, pp. 1–758, Aug. 2012, doi: 10.1002/9781118391686.

[35] D. J. Olive, “Linear regression,” Linear Regression, pp. 1–494, Apr. 2017, doi: 10.1007/978-3-319-55252-1/COVER.

[36] L. Rokach and O. Maimon, “Data Mining with Decision Trees: Theory and Applications, 2nd Edition,” Data Mining with Decision Trees: Theory and Applications, 2nd Edition, vol. 81, pp. 1–305, Jan. 2014, doi: 10.1142/9097.

[37] B. Everitt and S. Rabe-Hesketh, “Regression Trees,” pp. 327–344, 2001, doi: 10.1007/978-1-4757-3285-6_16.

[38] T. R. Shultz et al., “Classification Tree,” Encyclopedia of Machine Learning, pp. 171–171, 2011, doi: 10.1007/978-0-387-30164-8_114.

[39] T. Landers and R. L. Rosenberg, “Decision Tree,” Encyclopedia of Database Systems, vol. 6, no. 1, pp. 765–765, 2009, doi: 10.1007/978-0-387-39940-9_2445.

[40] R. A. Berk, “Statistical Learning as a Regression Problem,” pp. 1–48, 2008, doi: 10.1007/978-0-387-77501-2_1.

[41] C. Cortes, V. Vapnik, and L. Saitta, “Support-vector networks,” Machine Learning 1995 20:3, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[42] Ingo. Steinwart and Andreas. Christmann, “Support vector machines,” p. 601, 2008.

[43] “Data Science, Analytics and Machine Learning with R,” Data Science, Analytics and Machine Learning with R, 2023, doi: 10.1016/C2020-0-01649-X.

[44] T. Hofmann, B. Schölkopf, and A. J. Smola, “Kernel methods in machine learning,” https://doi.org/10.1214/009053607000000677, vol. 36, no. 3, pp. 1171–1220, Jun. 2008, doi: 10.1214/009053607000000677.

[45] K.-L. Du and M. N. S. Swamy, “Kernel Methods,” Neural Networks and Statistical Learning, pp. 569–592, 2019, doi: 10.1007/978-1-4471-7452-3_20.

[46] C. E. Rasmussen, “Gaussian Processes in machine learning,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3176, pp. 63–71, 2004, doi: 10.1007/978-3-540-28650-9_4.

[47] N. Quadrianto, K. Kersting, and Z. Xu, “Gaussian Process,” pp. 535–548, 2017, doi: 10.1007/978-1-4899-7687-1_108.

[48] W. Menke, “Environmental Data Analysis with MatLab or Python: Principles, Applications, and Prospects, Third Edition,” Environmental Data Analysis with MatLab or Python: Principles, Applications, and Prospects, Third Edition, pp. 1–444, Jan. 2022, doi: 10.1016/C2021-0-01049-X.

[49] A. Kumar and M. Jain, “Ensemble Learning for AI Developers,” Ensemble Learning for AI Developers, 2020, doi: 10.1007/978-1-4842-5940-5.

[50] D. Opitz and R. Maclin, “Popular Ensemble Methods: An Empirical Study,” Journal of Artificial Intelligence Research, vol. 11, pp. 169–198, Aug. 1999, doi: 10.1613/JAIR.614.

[51] R. Polikar, “Ensemble based systems in decision making,” IEEE Circuits and Systems Magazine, vol. 6, no. 3, pp. 21–44, 2006, doi: 10.1109/MCAS.2006.1688199.

[52] L. Breiman, “Bagging predictors,” Mach Learn, vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1007/BF00058655/METRICS.

[53] C. S. Nam, J. Y. Jung, and S. Lee, “Human-Centered Artificial Intelligence: Research and Applications,” Human-Centered Artificial Intelligence: Research and Applications, pp. 1–295, Jan. 2022, doi: 10.1016/C2020-0-02460-6.

[54] V. K. Ayyadevara, “Pro Machine Learning Algorithms,” Pro Machine Learning Algorithms, 2018, doi: 10.1007/978-1-4842-3564-5.

[55] Z. H. Zhou, “Ensemble methods: Foundations and algorithms,” Ensemble Methods: Foundations and Algorithms, pp. 1–218, Jan. 2012, doi: 10.1201/B12207/ENSEMBLE-METHODS-ZHI-HUA-ZHOU.

[56] G. Dreyfus, “Neural networks: Methodology and applications,” Neural Networks: Methodology and Applications, pp. 1–497, 2005, doi: 10.1007/3-540-28847-3/COVER.

[57] R. Rojas, “Neural Networks,” 1996, doi: 10.1007/978-3-642-61068-4.

[58] Y. Chen, L. Li, W. Li, Q. Guo, Z. Du, and Z. Xu, “AI Computing Systems,” AI Computing Systems, pp. 1–427, Jan. 2024, doi: 10.1016/C2021-0-02950-3.

[59] M. Stone, “Cross-Validatory Choice and Assessment of Statistical Predictions,” Journal of the Royal Statistical Society: Series B (Methodological), vol. 36, no. 2, pp. 111–133, Jan. 1974, doi: 10.1111/J.2517-6161.1974.TB00994.X.

[60] S. Eslamian and F. Eslamian, Handbook of Hydroinformatics Volume I: Classic Soft-Computing Techniques Edited. Elsevier, 2023. Accessed: Jun. 21, 2023. [Online]. Available: http://www.sciencedirect.com:5070/book/9780128212851/handbook-of-hydroinformatics

[61] "Theory of Point Estimation," Theory of Point Estimation, 1998, doi: 10.1007/B98854.

[62] J. Shao, "Mathematical Statistics," 2003, doi: 10.1007/B97553.

[63] H.-P. Kriegel, P. Kunath, and M. Renz, "Root-Mean-Square Error," Encyclopedia of GIS, pp. 1794–1794, 2017, doi: 10.1007/978-3-319-17885-1_101126.

[64] J. Fürnkranz, "Mean Absolute Error," Encyclopedia of Machine Learning and Data Mining, pp. 806–806, 2017, doi: 10.1007/978-1-4899-7687-1_953.

[65] N. R. Draper and H. Smith, "R-square statistic," Applied regression analysis, p. 33, 1998.

[66] S. Y. Hwang and J. W. Kang, "Group Contribution-Based Graph Convolution Network: Pure Property Estimation Model," Int J Thermophys, vol. 43, no. 9, Sep. 2022, doi: 10.1007/S10765-022-03060-7.

[67] R. Egger, Ed., "Applied Data Science in Tourism," 2022, doi: 10.1007/978-3-030-88389-8.

[68] "Hyperparameter Tuning for Machine and Deep Learning with R," Hyperparameter Tuning for Machine and Deep Learning with R, 2023, doi: 10.1007/978-981-19-5170-1.

[69] T. Agrawal, "Bayesian Optimization," Hyperparameter Optimization in Machine Learning, pp. 81–108, 2021, doi: 10.1007/978-1-4842-6579-6_4.

[70] P. Liu, "Bayesian Optimization," 2023, doi: 10.1007/978-1-4842-9063-7.

[71] M. Pelikan, "Hierarchical Bayesian Optimization Algorithm," vol. 170, 2005, doi: 10.1007/B10910.

[72] F. Archetti and A. Candelieri, "Bayesian Optimization and Data Science," 2019, doi: 10.1007/978-3-030-24494-1.

[73] Y. Li, Y. Zhang, and Y. Cai, "A new hyper-parameter optimization method for power load forecast based on recurrent neural networks," Algorithms, vol. 14, no. 6, 2021, doi: 10.3390/A14060163.

[74] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," Springer Topics in Signal Processing, vol. 2, pp. 1–4, 2009, doi: 10.1007/978-3-642-00296-0_5/COVER.

[75] Y. Que, S. Ren, Z. Hu, and J. Ren, "Machine Learning Prediction of Critical Temperature of Organic Refrigerants by Molecular Topology," Processes, vol. 10, no. 3, p. 577, Mar. 2022, doi: 10.3390/PR10030577/S1.

[76] M. R. Dobbelaere, Y. Ureel, F. H. Vermeire, L. Tomme, C. V. Stevens, and K. M. Van Geem, "Machine Learning for Physicochemical Property Prediction of Complex Hydrocarbon Mixtures," Ind Eng Chem Res, vol. 61,

no. 24, pp. 8581–8594, Jun. 2022, doi: 10.1021/ACS.IECR.2C00442/SUPPL_FILE/IE2C00442_SI_001.PDF.

[77] V. Mann, K. Brito, R. Gani, and V. Venkatasubramanian, “Hybrid, Interpretable Machine Learning for Thermodynamic Property Estimation using Grammar2vec for Molecular Representation,” Fluid Phase Equilib, vol. 561, p. 113531, Oct. 2022, doi: 10.1016/J.FLUID.2022.113531.

[78] Y. Su, Z. Wang, S. Jin, W. Shen, J. Ren, and M. R. Eden, “An architecture of deep learning in QSPR modeling for the prediction of critical properties using molecular signatures,” AIChE Journal, vol. 65, no. 9, p. e16678, Sep. 2019, doi: 10.1002/AIC.16678.

[79] A. Ganesh and N. Urs, “Deep Learning: Why Neural Networks Are State of the Art,” pp. 31–46, 2022, doi: 10.1007/978-3-030-84729-6_2.

[80] Y. Wang, M. Khodadadzadeh, and R. Zurita-Milla, “Spatial+: A new cross-validation method to evaluate geospatial machine learning models,” International Journal of Applied Earth Observation and Geoinformation, vol. 121, p. 103364, Jul. 2023, doi: 10.1016/J.JAG.2023.103364.

A
LYDERSEN CONSTANTS FOR ALCOHOL COMPUNDS

No	Name	Formula	ΣΔT	ΣΔP	ΣΔV
1	allyl alcohol	C_3H_6O	0.138	0.683	163.000
2	3-amino-1-propanol	C_3H_9NO	0.173	0.836	211.000
3	2-aminoethoxyethanol	$C_4H_{11}NO_2$	0.214	1.223	286.000
4	1,2,3-benzenetriol	$C_6H_6O_3$	0.159	0.864	231.000
5	benzyl alcohol	C_7H_8O	0.168	1.211	295
6	bisphenol a	$C_{15}H_{16}O_2$	0.234	2.472	601
7	1,4-butanediol	$C_4H_{10}O_2$	0.244	1.028	256
8	1,3-butanediol	$C_4H_{10}O_2$	0.236	1.011	252
9	*sec*-butanol	$C_4H_{10}O$	0.154	0.951	234
10	*tert*-butanol	$C_4H_{10}O$	0.142	0.951	224
11	butanol	$C_4H_{10}O$	0.162	0.968	238
12	*trans*-2-butene-1,4-diol	$C_4H_8O_2$	0.24	0.97	236
13	2-butoxyethanol	$C_6H_{14}O_2$	0.223	1.582	368
14	4-butylphenol	$C_{10}H_{14}O$	0.177	1.812	445
15	2-butyne-1,4-diol	$C_4H_6O_2$	0.214	0.88	218
16	2-chloroethanol	C_2H_5ClO	0.139	0.834	177
17	*m*-chlorophenol	C_6H_5ClO	0.114	1.224	274
18	*m*-cresol	C_7H_8O	0.117	1.131	280
19	*p*-cumylphenol	$C_{15}H_{16}O$	0.203	2.492	598
20	cyclohexanol	$C_6H_{12}O$	0.159	1.172	286.5
21	1,2-decanediol	$C_{10}H_{22}O_2$	0.356	2.373	582
22	1,10-decanediol	$C_{10}H_{22}O_2$	0.364	2.39	586
23	1-decanol	$C_{10}H_{22}O$	0.282	2.33	568
24	2-decanol	$C_{10}H_{22}O$	0.274	2.313	564
25	diacetone alcohol	$C_6H_{12}O_2$	0.202	1.468	339
26	2,2-diethyl-1,3-propanediol	$C_7H_{16}O_2$	0.284	1.692	407
27	diethylene glycol	$C_4H_{10}O_3$	0.265	1.188	276
28	2,3-dimethyl-1,2-butanediol	$C_6H_{14}O_2$	0.256	1.448	348
29	4,4-dimethyl-1,2-pentanediol	$C_7H_{16}O_2$	0.276	1.675	403

30	2,2-dimethyl-1,3-propanediol	$C_5H_{12}O_2$	0.204	0.784	187
31	2,2-dimethyl-1,4-butanediol	$C_6H_{14}O_2$	0.264	1.465	352
32	6,6-dimethyl-1-heptanol	$C_9H_{20}O$	0.242	2.086	499
33	4,6-dimethyl-1-heptanol	$C_9H_{20}O$	0.246	2.069	505
34	4,5-dimethyl-1-octanol	$C_{10}H_{22}O$	0.266	2.296	560
35	4,4-dimethyl-1-pentanol	$C_7H_{16}O$	0.202	1.632	389
36	2,3-dimethyl-2,3-butanediol	$C_6H_{14}O_2$	0.244	1.448	338
37	2,3-dimethyl-2,3-pentanediol	$C_7H_{16}O_2$	0.264	1.675	393
38	2,4-dimethyl-2,3-pentanediol	$C_7H_{16}O_2$	0.268	1.658	399
39	2,3-dimethyl-2-butanol	$C_6H_{14}O$	0.174	1.388	330
40	3,3-dimethyl-2-ethyl-1-butanol	$C_8H_{18}O$	0.214	1.842	440
41	5,5-dimethyl-2-ethyl-1-hexanol	$C_{10}H_{22}O$	0.254	2.296	550
42	2,4-dimethyl-2-ethyl-1-pentanol	$C_9H_{20}O$	0.234	2.069	495
43	3,4-dimethyl-2-ethylphenol	$C_{10}H_{14}O$	0.17	1.769	434.5
44	4,6-dimethyl-2-heptanol	$C_9H_{20}O$	0.238	2.052	501
45	4,5-dimethyl-2-hexanol	$C_8H_{18}O$	0.218	1.825	446
46	3,7-dimethyl-2-octanol	$C_{10}H_{22}O$	0.258	2.279	556
47	3,4-dimethyl-2-pentanol	$C_7H_{16}O$	0.198	1.598	391
48	2,3-dimethyl-2-tert-butyl-1-butanol	$C_{10}H_{22}O$	0.234	2.279	536
49	4,4-dimethyl-3-ethyl-2-pentanol	$C_9H_{20}O$	0.226	2.052	491
50	4,4-dimethyl-3-isopropyl-1-pentanol	$C_{10}H_{22}O$	0.246	2.279	546
51	3,4-dimethyl-3-isopropyl-2-pentanol	$C_{10}H_{22}O$	0.238	2.262	542
52	dinonylphenol	$C_{24}H_{42}O$	0.457	4.99	1215
53	dipropylene glycol	$C_6H_{14}O_3$	0.289	1.608	378
54	2,6-di-tert-butyl-p-cresol	$C_{15}H_{24}O$	0.237	2.913	692
55	1,4-dodecanediol	$C_{12}H_{26}O_2$	0.396	2.827	692
56	1,12-dodecanediol	$C_{12}H_{26}O_2$	0.404	2.844	696
57	1-dodecanol	$C_{12}H_{26}O$	0.322	2.784	678
58	2-dodecanol	$C_{12}H_{26}O$	0.314	2.767	674

59	1,20-eicosanediol	$C_{20}H_{42}O_2$	0.564	4.66	1136
60	1,2-eicosanediol	$C_{20}H_{42}O_2$	0.556	4.643	1132
61	2-eicosanol	$C_{20}H_{42}O$	0.474	4.583	1114
62	1-eicosanol	$C_{20}H_{42}O$	0.482	4.6	1118
63	2-ethoxyethanol	$C_4H_{10}O_2$	0.183	1.128	258
64	2-(2-ethoxyethoxy)ethanol	$C_6H_{14}O_3$	0.244	1.742	388
65	ethyl alcohol	C_2H_6O	0.122	0.514	128
66	2-ethyl-1,3-butanediol	$C_6H_{14}O_2$	0.256	1.238	307
67	2-ethyl-1,3-pentanediol	$C_7H_{16}O_2$	0.288	1.675	413
68	3-ethyl-2-pentanol	$C_7H_{16}O$	0.206	1.615	395
69	3-ethyl-3-hexanol	$C_8H_{18}O$	0.222	1.859	444
70	3-ethyl-3-octanol	$C_{10}H_{22}O$	0.18	2.253	536
71	ethylene glycol	$C_2H_6O_2$	0.204	0.574	146
72	*p*-ethylphenol	$C_8H_{10}O$	0.137	1.358	335
73	furfuryl alcohol	$C_5H_6O_2$	0.16	1.023	229
74	glycerol	$C_3H_8O_3$	0.298	0.844	215
75	guaiacol	$C_7H_8O_2$	0.138	1.291	300
76	1,4-heptadecanediol	$C_{17}H_{36}O_2$	0.496	3.962	967
77	1,17-heptadecanediol	$C_{17}H_{36}O_2$	0.504	3.979	971
78	2-heptadecanol	$C_{17}H_{36}O$	0.414	3.902	949
79	1-heptadecanol	$C_{17}H_{36}O$	0.422	3.919	953
80	1,6-heptanediol	$C_7H_{16}O_2$	0.296	1.692	417
81	1,7-heptanediol	$C_7H_{16}O_2$	0.304	1.709	421
82	1-heptanol	$C_7H_{16}O$	0.222	1.649	403
83	2-heptanol	$C_7H_{16}O$	0.214	1.632	399
84	1,16-hexadecanediol	$C_{16}H_{34}O_2$	0.484	3.752	916
85	1,2-hexadecanediol	$C_{16}H_{34}O_2$	0.476	3.735	912
86	1-hexadecanol	$C_{16}H_{34}O$	0.402	3.692	898
87	2-hexadecanol	$C_{16}H_{34}O$	0.394	3.675	894
88	2,3-hexanediol	$C_6H_{14}O_2$	0.268	1.448	358
89	1,4-hexanediol	$C_6H_{14}O_2$	0.276	1.465	362
90	1,6-hexanediol	$C_6H_{14}O_2$	0.284	1.482	366
91	1-hexanol	$C_6H_{14}O$	0.202	1.422	348
92	2-hexanol	$C_6H_{14}O$	0.194	1.405	344
93	3-isobutylphenol	$C_{10}H_{14}O$	0.169	1.795	441
94	isopropyl alcohol	C_3H_8O	0.142	0.741	183

95	2-isopropyl-1-pentanol	$C_8H_{18}O$	0.226	1.842	450
96	2-isopropylphenol	$C_9H_{12}O$	0.149	1.568	386
97	2-methoxyethanol	$C_3H_8O_2$	0.163	0.901	203
98	2-(2-methoxyethoxy)ethanol	$C_5H_{12}O_3$	0.224	1.515	333
99	methyl alcohol	CH_4O	0.102	0.287	73
100	3-methyl-1,2-butanediol	$C_5H_{12}O_2$	0.248	1.221	303
101	2-methyl-1,2-propanediol	$C_4H_{10}O_2$	0.224	1.011	242
102	6-methyl-1-octanol	$C_9H_{20}O$	0.254	2.086	509
103	4-methyl-2-ethylphenol	$C_9H_{12}O$	0.157	1.585	390
104	2-methyl-2-nonanol	$C_{10}H_{22}O$	0.262	2.313	554
105	2-methyl-2-pentanol	$C_6H_{14}O$	0.182	1.405	334
106	4-methyl-2-pentanol	$C_6H_{14}O$	0.186	1.388	340
107	1-methylcyclohexanol	$C_7H_{14}O$	0.16	1.361	326.5
108	neopentyl glycol	$C_5H_{12}O_2$	0.244	1.238	297
109	1,4-nonadecanediol	$C_{19}H_{40}O_2$	0.536	4.416	1077
110	1-nonadecanol	$C_{19}H_{40}O$	0.462	4.373	1063
111	2-nonadecanol	$C_{19}H_{40}O$	0.454	4.356	1059
112	1,19-nonadecanediol	$C_{19}H_{40}O_2$	0.544	4.433	1081
113	1,9-nonanediol	$C_9H_{20}O_2$	0.344	2.163	531
114	1,4-nonanediol	$C_9H_{20}O_2$	0.336	2.146	527
115	1-nonanol	$C_9H_{20}O$	0.262	2.103	513
116	nonylphenol	$C_{15}H_{24}O$	0.277	2.947	720
117	1,2-octadecanediol	$C_{18}H_{38}O_2$	0.516	4.189	1022
118	1,18-octadecanediol	$C_{18}H_{38}O_2$	0.524	4.206	1026
119	1-octadecanol	$C_{18}H_{38}O$	0.442	4.146	1008
120	2-octadecanol	$C_{18}H_{38}O$	0.434	4.129	1004
121	1,4-octanediol	$C_8H_{18}O_2$	0.316	1.919	472
122	1,8-octanediol	$C_8H_{18}O_2$	0.324	1.936	476
123	4-octanol	$C_8H_{18}O$	0.234	1.859	454
124	1-octanol	$C_8H_{18}O$	0.242	1.876	458
125	1,15-pentadecanediol	$C_{15}H_{32}O_2$	0.464	3.525	861
126	1,4-pentadecanediol	$C_{15}H_{32}O_2$	0.456	3.508	857
127	2-pentadecanol	$C_{15}H_{32}O$	0.374	3.448	839
128	1-pentadecanol	$C_{15}H_{32}O$	0.382	3.465	843
129	2,2,3,4,4-pentamethyl-3-pentanol	$C_{10}H_{22}O$	0.222	2.279	526

130	1,5-pentanediol	$C_5H_{12}O_2$	0.264	1.255	311
131	3-pentanol	$C_5H_{12}O$	0.174	1.178	289
132	1-pentanol	$C_5H_{12}O$	0.182	1.195	293
133	*tert*-pentyl-alcohol	$C_5H_{12}O$	0.162	1.178	279
134	phenol	C_6H_6O	0.097	0.904	225
135	2-phenyl-2-propanol	$C_9H_{12}O$	0.188	1.648	391
136	2-phenylethanol	$C_8H_{10}O$	0.188	1.438	350
137	1,3-propanediol	$C_3H_8O_2$	0.224	0.801	201
138	1,2-propanediol (propylene glycol)	$C_3H_8O_2$	0.216	0.784	197
139	propargyl alcohol	C_3H_4O	0.11	0.564	135
140	pyrocatechol	$C_6H_6O_2$	0.128	0.884	228
141	sorbitol	$C_6H_{14}O_6$	0.58	1.654	422
142	*p*-tert-amylphenol	$C_{11}H_{16}O$	0.177	2.022	486
143	*p*-tert-butylcatechol	$C_{10}H_{14}O_2$	0.188	1.775	434
144	2-tert-butylphenol	$C_{10}H_{14}O$	0.157	1.795	431
145	*p*-tert-octylphenol	$C_{14}H_{22}O$	0.237	2.703	651
146	1,4-tetradecanediol	$C_{14}H_{30}O_2$	0.436	3.281	802
147	1,14-tetradecanediol	$C_{14}H_{30}O_2$	0.444	3.298	806
148	2-tetradecanol	$C_{14}H_{30}O$	0.354	3.221	784
149	1-tetradecanol	$C_{14}H_{30}O$	0.362	3.238	788
150	tetraethylene glycol	$C_8H_{18}O_5$	0.387	2.416	536

PUBLICATIONS FROM THE THESIS

Conference Papers

1. R. ’ALHALOUSH, M. A. ’Insel, and H. ’SADIKOĞLU, “Assessment Of Lydersen Method In Estimation Of Critical Temperature, Pressure, And Volume Of Various Alcohols,” in *5th International Eurasian Conference on Biological and Chemical Sciences (EurasianBioChem 2022)* , Ankara, Turkey: www.EurasianBioChem.org, Oct. 2022, pp. 499–499.