Prediction of Rate of Chemical Reaction using Computational Intelligence

Thesis submitted in partial fulfillment of the requirements for the award of the degree of

Master of Engineering

in

Computer Science and Engineering

submitted by

Abhishek Kapoor
(Reg no: 801532007)

Under Supervision of

Dr. V.P. Singh
Dr. Prashant Singh Rana

Thapar University
Patiala-147004, Punjab, India
June 2017

Candidate Declaration

I hereby certify that the work, which is being presented in the thesis, titled **Prediction of Rate of Chemical Reaction using Computational Intelligence**, in partial fulfillment of the requirements for the award of the degree of **Master of Engineering** and submitted to Thapar University is an authentic record of my own work carried out under the supervision of **Dr. V.P. Singh** and **Dr. Prashant Singh Rana**. I have also cited the reference about the text(s)/figure(s)/table(s) from where they have been taken.

The matter presented in this thesis has not been submitted elsewhere for the award of any other degree or diploma from any institution.

Abhishek Kapoor

This is to certify that the above statement made by the candidate is correct to the best of our knowledge.

(Dr. V.P. Singh)
Assistant Professor
CSED

(Dr. Prashant Singh Rana)
Assistant Professor
CSED

Acknowledgements

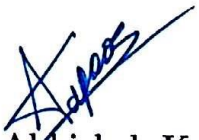
First, I would like to express my deep gratitude towards my supervisors **Dr. V.P. Singh and Dr. Prashant Singh Rana** for their invaluable advice and encouragement at every step of my Master's program. Without their unfailing support and belief in me, this thesis would not have been possible. Their contribution to this thesis goes well beyond their role as an academic supervisor and includes constant support on a personal level without which this journey may never have been completed, and for this, I am truly grateful. They are great mentors for my life as well.

I am also thankful to **Dr. Ashutosh Mishra**, P.G. Coordinator, for the motivation and inspiration that triggered me for the thesis work. He has always been supportive and provided us required information at regular intervals.

I would like to acknowledge **Dr. Maninder Singh**, Head, Computer Science and Engineering Department, Thapar University for setting good standards for his students and providing all the help and facilities that were essential throughout the journey. Your encouragement time and again has helped students to achieve the set goals.

I will be failing in my duty if I do not express my gratitude towards **Dr. S.S. Bhatia**, Dean of Academic Affairs, Thapar University for making provisions of infrastructure such as library facilities, computer labs equipped with Internet facilities, immensely useful for the learners to equip themselves with the latest knowledge in the field. I would also like to thank **Dr. Niyati Baliyan**, visiting Assistant Professor, Thapar University for supporting me in my two years of journey in Thapar University.

Above all, thanks to the Almighty, my family and friends for always being there for me and staying calm at times required.



Abhishek Kapoor

Abstract

Rate of chemical reaction is one of the most important thing in chemical kinetics. Rate of reaction determines how fast or slow a reaction is taking place. Traditionally, Arrhenius equation is used to know rate of chemical reaction at different temperature and concentration. However, finding values of parameters of Arrhenius equation for a reaction can be costly as well as time consuming.

The work presented in this thesis mainly focuses on predicting value of rate of reaction using machine learning techniques. The objective of this thesis is to find optimum parameters for rate of reaction from easily measurable features of chemical reaction, such as temperature, pressure, density and concentration of reactants involved in a chemical reaction.

First of all, we collected data by simulating chemical reaction using Cantera. We choose 5 different reaction and simulated them in Cantera. For simulating these 5 reaction, we created an input file which have data about these 5 chemical reaction and then this input file was given to Cantera. Then we varied temperature, pressure of environment and changed the amount of reactants being used for chemical reaction then we notice corresponding rate of reaction.

We applied various machine learning models on data obtained from simulation to predict rate of reaction and compared their performances with each other to find the best machine learning model. To check the robustness of best model, we used k-fold cross validation.

Keywords: Rate of Reaction Prediction, Cantera, Simulation, Machine Learning Models.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

GA	Genetic Algorithm
LM	Linear Model
R^2	Coefficient of Determination
RMSD	Root Mean Square Devation
RMSE	Root Mean Square Error
SVM	Support Vector Machine
RF	Random Forest
CFD	Computational Fluid Dynamics
GBM	Gradient Boosted Model
kNN	k-nearest neighbour
SDT	Single Decision Tree
QSPR	Quantitative Structure Reactivity Relationship
OECD	Organization for Economic Co-operation and Development
QSAR	Quantitative Structure Activity Relationship
PAC	Probability Approximately Correct
BST	Boosted Tree

Chapter 1

Introduction

Chemistry is concerned with change entirely, changes that we can easily observe in our daily lives and changes that are not so familiar to us. Most of us have seen a rusted piece of iron, some of us must have seen burning of coal or any other fuel, curdling of milk, etc. These are some chemical changes that we come across in our daily lives. Our knowledge of chemistry has so far enabled us to understand about different types of changes.

We are surrounded by objects that in some way or the other are products of a chemical reaction, the plastic bags we carry, the oil used for cooking, combustion of fuels in vehicles, steel, concrete, and many other chemicals or substances that are being used or consumed by us in enormous amounts. However for all these reactions, scientists across the globe were unable to predict the time required by any particular reaction to give the desired products. They were also unable to explain the parameters that affected this time. The parameters that governed the speed of a reaction were uncertain.

The process of transforming one group of substance to another is known as chemical reaction. In a chemical reaction, there is no change in nuclei of atoms, the change only occurs in position of electrons and in the bonds among atoms[1]. Reaction involving radioactive and unstable elements are studied under different subbranch of chemistry which is known as Nuclear chemistry. In nuclear chemistry nuclei of atoms also change in chemical reaction.

Substances present in environment before the start of a chemical reaction are known as reactants. A chemical reaction can make one or more substance as product and these products have different properties than that of reactants. A reaction can have many steps and each step is known an elementary reaction. A chemical equation is used to represent a chemical reaction. Chemical equation give us information about reactants, products and also the intermediate products if any formed during reaction.

Rate of a chemical reaction is different at different temperatures and at pressures. Usually, reaction occurs at much higher pase at higher temperature as we have got more thermal energy at higher temperature and this thermal energy makes

it possible for molecules to get enough energy to break existing bonds. Direction of reaction can either be forward or reverse and reaction will continue until state of equilibrium is achieved. Spontaneous reaction are those which move only in forward direction, i.e, they do not require any input energy to occur, on the other hand, reaction which need some energy from environment are known as non-spontaneous reaction.

We can measure reaction rate by measuring speed at which regents are consumed, or the rate at which products are being formed. Rates of chemical reactions mainly depend on the nature of the reactants, temperature, presence of a catalyst, and concentration. We can measure rate of a reaction by following two ways:

1. By measuring the rate at which reactants are being consumed.
2. By measuring the rate at which end products are being formed.

Depending on reaction we are studying, we choose the best method. It is easier to compute the difference in amount of product formed in chemical reaction after some unit of time for some reaction and for other reaction is easier to compute the change in quantity of reagents that has been used up in chemical reaction in a unit of time. Equation 1.1 defines the rate of reaction.

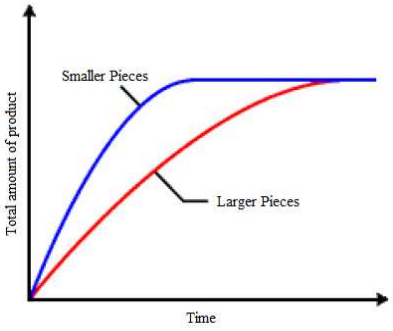
$$r = \frac{a}{t} \tag{1.1}$$

Where,

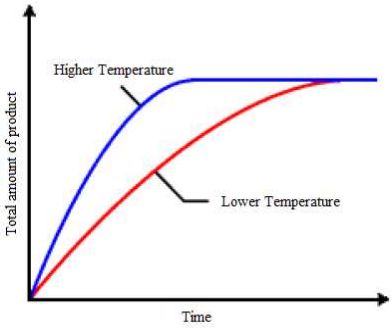
- a is amount of reactant used
- t is time taken
- r is rate of reaction

The Figure 1.1 shows the difference how the rate of reaction is affected by various factors like size of pieces of reactants, different temperatures and concentrations. If line is more steep, then it means rate of reaction is high and as the line becomes horizontal then it means the reaction has stopped. At the starting phase of reaction the rate is highest as then concentration of reactant is highest at that time. All reaction in Figure 1.1 justifies this statement.

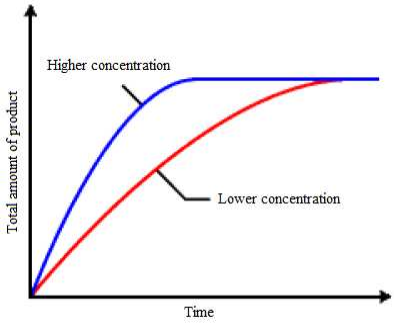
Chemical reaction is nothing but some sequence of some events that are taking place as reactants are transformed into product(s). Every event in a chemical reaction constitutes basically a step that can be seen as the breaking-up of a molecule ("dissociation") into many simpler units or a coming-together of discrete particles ("collision").



(a) Reactant Size



(b) Temperature



(c) Concentration

Figure 1.1: Effect of various parameters on rate of reaction

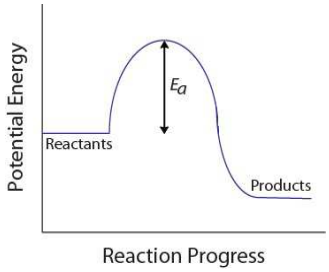


Figure 1.2: Activation Energy

Collision theory of reactivity states that when molecules of reactant "effectively collide" then chemical reaction occur. Meaning of "effective collision" is that the molecules should be oriented correctly so that the old bond can be broken and new bonds can be formed such that new product can be formed[2]. Molecules must have some kinetic energy during collision so that effective collision can occur. Such minimum energy required by molecule is also known as activation energy. Figure 1.2 shows graph of activation energy. Once energy reaches to activation energy, reactants start converting into products. Activation energy plays a very important role in rate of reaction as when activation energy is higher then rate of reaction is less as lot of energy is required to reach to level of activation energy.

Kinetic theory of gases[3] tells us that there will be only one event for 1000 binary

collisions in which three molecules simultaneously come together. This is the reason that termolecular processes are very rare. Collision between 4 molecules are so unlikely to happen that we have never observed any four-way collisions in an elementary reaction. Consider a simple bimolecular step

$$X + Y \rightarrow \text{end product(s)}$$

X and Y will react if they are able to come enough close that they can break some of their existing bonds and facilitate creation of any new bond which are required in products, such encounters are knows as collisions.

In a gas, collision frequency between molecules of X and molecules of Y is directly proportional to concentration of X and Y; doubling the concentration of X will result in doubling the frequency of collisions between X and Y, and if we double concentration of X we will get the same effect. If all collisions in a chemical reaction are leading to product(s) only, then rate of bimolecular process will be first order in X and Y, or second order overall:

$$\text{rate} = k[X][Y]$$

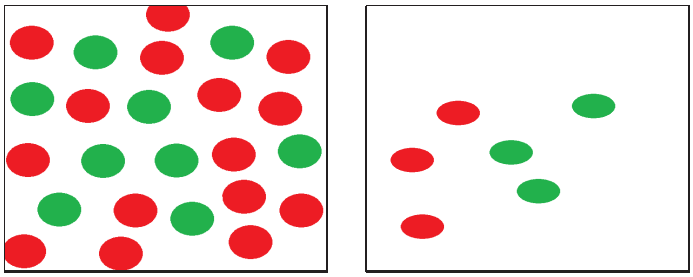
We know that chemical reactions occurs at higher phase at higher temperatures. Everyone knows that milk becomes sour at much more speed when stored outside refrigerator at room temperature, butter goes rancid more quickly in summer than in winter, and eggs at sea level hard-boil more rapidly as compared to mountains. For this same reason, cold-blooded animals, such as amphibians, reptiles, and fishes become sluggish in cold weather.

It makes sense why reaction occurs more rapidly at higher temperature. Thermal energy relates direction to motion at the molecular level with the increase in temperature, molecules move at much higher speed and collide more vigorously as the temperature rises, this increases the probability of bond cleavages and rearrangements as described above.

1.1 Factors That Affect the Rate of a Chemical Reaction

1.1.1 Concentration of Reactants

When we increase the concentration of chemicals, then there are more molecules per unit area and probability of collision among these molecules increase signif-



(a) High Concentration (b) Low Concentration

Figure 1.3: Effect of concentration on rate of reaction

icantly. Similarly, as concentration of product increases then rate of reaction decreases as there are less molecules of reactants. In gas, pressure is used as a measure of concentration.

Figure 1.3 shows when concentration is high, then there is very high probability that the elements will collide whereas when concentration is low, then there is very low probability that the elements will collide.

1.1.2 Temperature

When we increase temperature of environment in which reaction is taking place then molecules involved in that reaction get more thermal energy. As energy in the system increases more and more collision starts breaking activation energy. Therefore, we can say mostly that when we increase temperature then rate of relation also increases. It has been observed that when we increase temperature by 10 C then rate of reaction doubles. There are some reactions for which when we increase temperature the rate of reaction also increases but only up to a threshold level of temperature when temperature crosses this threshold then rate of reaction starts decreasing.

1.1.3 Medium or State of Matter

Another factor which affects the rate of reaction is state of reactants. Medium of reactants affects the rate of reaction. Rate of reaction will be affected if medium is aqueous, solid, gas. Same chemical reactions will occur at different rates when state of chemical will be changed. In both liquid as well as in solid, rate of reaction is greatly affected by surface area. Size and shape of reactants in solid also affect the rate of reactions.

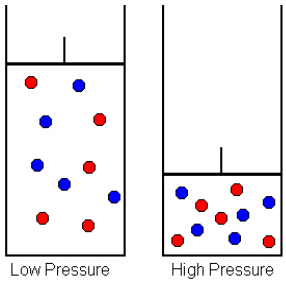


Figure 1.4: Effect of pressure on rate of reaction

1.1.4 Presence of Catalysts and Competitors

For a chemical reaction, there are some chemicals, which when added in lower the rate of reaction. These are called catalysts and are not consumed in a chemical reaction. When we add catalyst, then number of collisions in a reaction increase dramatically or bonds in reactant molecule become weak or it changes orientation of reactant molecule such that it becomes easy for reactants to react. When we add catalyst in a reaction, the rate of reaction usually increases.

1.1.5 Pressure

When we apply more pressure in environment then molecules of reactant come together and hence probability of collisions, among different reactant increases. As we increase pressure, the rate of reaction also increases.

1.1.6 Mixing

When we mix two reactants then it facilitates them to react at much higher pase and hence our rate of reaction increases.

1.2 Arrhenius Equation

By 1890, humans knew that by increasing temperatures we can speedup some chemical reaction, mostly rate becomes double when temperature rises by 10°C, but no one knows the clear reason why was this happening. Finally, the Swedish chemist Svante Arrhenius, in 1899, combined the Boltzmann distribution law and concepts of activation energy into a relationship[4] which is one of most significant relationships in physical chemistry:

Arrhenius equation is given in equation 4.1 as:

$$k = Ae^{-Ea/(RT)} \tag{1.2}$$

Where,

- R = the universal gas constant.
- T = absolute temperature.
- k = rate constant.
- E_a = activation energy.
- A = pre-exponential factor

Currently, Arrhenius equation is best seen as an empirical relationship. Arrhenius equation can be employed to model creep rates, population of crystal vacancies, diffusion coefficient's temperature variation, and such other thermally-induced reactions/processes. Arrhenius equation supports generalization of reason why at room temperature when we increase temperature by 10°C then for many chemical reaction rate of reaction just doubles. In Arrhenius equation, A is known as pre-exponential factor. It is a constant value that differs with different chemical reaction. Value of A is determined with help of experiments. It basically tells us how many collision are taking place in a chemical reaction in a unit time. In chemistry, activation energy usually denoted by E_a was first introduced by Svante Arrhenius in 1889. Activation energy describes the minimum amount of energy which should be available with reactants in a chemical system so that reactants can start changing into products. The most common units for measuring activation energy is kilo calories per mole or kiloJoules per mole (kJ/mol). As we increase temperature, then rate of reaction also increases, however, there are some cases in which when we increase temperature then rate of reaction decreases. In such cases, to fit rate constant in Arrhenius expression we have to follow an approximately exponential relationship, this results in negative value of activation energy.

1.3 Rate Equation

For any chemical reaction

$$n_1\ X + n_2\ Y \rightarrow n_3\ E + n_4\ F,$$

we use rate law or rate equation to relate the rate of reaction to concentration of reactants. Generalized form of rate equation is:

$$r = k(T)[X]^a[Y]^b \tag{1.3}$$

Reaction's rate constant depends on many factors and not only on temperature.

In equation 1.3, a and b are reaction orders and k(T) is rate constant sometimes also known as rate coefcient. k(T) depends on all those factors except concentration which affect rate of reaction. Concentration is taken into account explicitly for determining rate of reaction. We can find reaction orders from reaction mechanism. In multi-step reaction, detailed mechanism determines the rate equation but for single step reactions the reaction order is equal to stoichiometric coefficients of reactants.

When concentration of reaction involved in reaction is one unit then the rate of reaction we will get will be defined as rate constant. This is the reason that sometimes rate constant is also known as specific rate constant. Rate constant's characteristics are as follows:

- Rate constant is a measure of the rate of reaction. Greater is the value of the rate constant, faster is the reaction.
- For each particular value of temperature, there is one particular value of rate constant in each reaction.
- Even for same reaction, rate constant's value changes along with temperatures.
- In any reaction, reactant's concentration does not have any effect on rate constant.
- Order of reaction determines the units of rate constant.

For an elementary reaction, stoichiometric coefficient and the order are both equal to number of molecules participating(molecularity). There is no effect of concentration of reactant on speed of reaction when the reaction is of zero order. In 1_st order reactions, rate of reaction is directly proportional to concentration of 1 reactant and in first order reaction if we double the amount of reactant then reaction rate will get doubled. In second order reactions, rate of reaction is proportional to the product of concentration of two reactants and in second order reaction if we double the amount of reactants, then reaction rate quadruples. If reaction rate

depends on concentration of three molecules, then it is third order reaction.

$$2\ H_2(g) + 2\ NO(g) \rightarrow N_2(g) + 2\ H_2O(g)$$

For above reaction if we compute its rate equation, then it will be :

$$r = k[H_2][NO]^2$$

For some reactions, stoichiometric coefficients are not reflected in experimental rate equation. Like one can see in above reaction the stoichiometric coefficients of both reactants are 2 yet overall reaction is of third order.

To explain reaction rate, a mechanism which consists of number of elementary steps is used. Not every step affects the reaction rate; usually the slowest step determines overall rate of reaction. For example:

1. $2\ NO(g) \rightarrow N_2O_2(g)$ (fast equilibrium)
2. $N_2O_2 + H_2 \rightarrow N_2O + H_2O$ (slow)
3. $N_2O + H_2 \rightarrow N_2 + H_2O$ (fast)

In above mechanism reactions 1 and 3 are much faster than the reaction 2, so in this example rate determining step is reaction number 2 which is a bimolecular reaction. Here 2_nd reaction is a second order reaction and rate equation of 2_nd reaction is:

$$r = k_2[H_2][N_2O_2] \tag{1.4}$$

where, k_2 is rate constant. In the above example, one can observe product of 1_{st} reaction, i.e, N_2O_2 is an intermediate product which is very unstable and concentration of N_2O_2 depends on first step's equilibrium, and is given below:

$$[N_2O_2] = K_1[NO]^2 \tag{1.5}$$

in above equation K_1 is equilibrium constant of 1_{st} reaction. When we substitute equation 1.4 in equation 1.5, we get rate equation as

$$r = k_2K_1[H_2][NO]^2$$

If we assume $k = k_2K_1$ then this equation will be equivalent with equation 1.3. Rate equation is used as a mechanism to predict chemical rate of reaction in agreement with experiment.

One can wonder why second molecule of H_2 is not there in rate equation. This is because after the second rate determining step, third step is a very rapid step, so it has very minimum effect on overall rate of reaction.

1.4 Research Motivation

The reaction rate or rate of reaction is defined as how rapidly or slowly a reaction is taking place for e.g., iron rusting is a slow reaction under earth's atmosphere that may take years, on other hand, it takes just few seconds to complete combustion of cellulose in fire. Another example can be combustion of methane, a very quick reaction, whereas, fermentation of sugar into ethyl alcohol and carbon dioxide can take several days to complete. Formation of diamond is also a chemical reaction which takes many years to occur.

Finding rate of reaction is always hard as already available techniques are highly costly and they are also very time consuming. We can use machine learning to predict rate of reaction, however there is no on shelf dataset available for predicting the chemical rate of reaction.

1.5 Thesis Organization

The thesis is organized into 7 chapters. A brief outline is given below:

- **Chapter 1**: This chapter introduces the basics of chemical reactions and then describes different factors that affect the reaction. Further this chapter discuss about Arrhenius equation which defines relationship between rate of reaction and temperature. This chapter also discusses what rate equation is and how it is used to find rate of reaction with suitable example.
- **Chapter 2**: This chapter presents literature review of available work which has been already done in the domain of finding rate of reaction. This chapter discusses various developed approaches to find rate of reaction and this chapter also discusses some literature about Cantera. Further this chapter discusses various machine learning approaches.
- **Chapter 3**: This chapter defines the research problem. Further it discusses what are current gaps in the domain of research are and then it discusses what are objectives of this thesis.

- **Chapter 4**: This chapter discusses about simulation results and dataset. This chapter defines how simulation was done and it has a brief description about Cantera which is an open source simulator which we used to simulate a chemical reaction and collect data. Further this chapter discusses about dataset, it discusses what is range of different attributes and what units we have used.
- **Chapter 5**: This chapter discuss about methodology followed and models used. This chapter has detailed discussion about what was the methodology which we used and further this chapter discusses about internal working of 5 best performing models.
- **Chapter 6**: In this chapter results of our research are discussed. Various models are compared in this chapter and to verify robustness of our model we have performed cross validation and the result of cross validation is also presented in this chapter. Further, our results are also discussed in this chapter.
- **Chapter 7**: This chapter summarizes the key findings and main contributions of the thesis and lists the possible future research directions.

Chapter 2

Literature Survey

L. Elliott et. al. [5] presented an approach in which they find rate coefficients of a reaction with help of limited species production data. First of all, they retrieved data about reaction by an inversion technique which was based upon nature inspired genetic algorithm, which is further based on principle of survival of fittest. Then they experimentally find reaction rate of coefficients for 3 different flames namely hydrogen,oxygen and nitrogen with highest accuracy possible. Then they take some data about production of species at different conditions and then they used genetic algorithm to find value of rate coefficients , such that Arrhenius equation is satisfied. Results of their paper show that process developed by them can predict the reaction coefficients with significant accuracy. Further, they concluded that this technique can be used where we do not have any information about rate coefficient and we do not have enough data then this technique which uses genetic algorithm can be used to predict the rate coefficient and further rate of reaction in particular environment.

W. Polifke et. al. [6] suggested a technique to derive rate coefficient from chemical data. They employed genetic algorithm to derive rate coefficient from data of heat release and species production. They used their method to determine rate coefficient of methane combustion. There technique can be used to predict rate of reaction and heat release in a chemical reaction by using detailed mechanism of chemical reaction. The technique developed by them can even be performed by a human who does not have knowledge about different chemicals or chemical kinetics at all, moreover, this technique needs very less amount of human effort. The basic aim of their study was to develop a technique by which their could find reaction coefficient for those chemical reactions of which, they had detailed mechanism and this technique should be able to predict rate of reaction as quickly as possible. They used CHEMKIN [7] package to simulate the rate of reaction and heat release and verify results of their experiments.

S. Vakalisa et. al. [8] used Cantera and Matlab for creating a multibox concept for thermodynamic model of biomass downdraft gasifiers. Biomass gasification is a complex process, influenced by various parameters. Numerous models have been

developed in order to describe the process of biomass gasification, a respectable amount of which concern the specific case of downdraft gasifiers. A traditional black-box modeling approach cannot reflect the wide range of operational conditions and the four major different process zones (i.e. drying, pyrolysis, oxidation, reduction) in a downdraft gasifier. In their research, they created and assessed the performance of the multi-box, by inserting real operating data (i.e. biomass composition, equivalent ratio, operation temperatures), collected during a monitoring activity that have been performed by the authors on a down draft gasier (design: Joos-Gasier), in North Italy. The values which were returned from the model were compared with the actual measurements and they found that there model was able to predict value with significant accuracy.

L. Tao [9] provided better simulation of combustion by coupling OpenFoaM [10] and Cantera. OpeamFOAM is computational fluid dynamics software and it is open source. In his research, tao he coupled Cantera and OpenFoam, and then he calculated the same H_2-O_2 combustion case with a solver coupled with Cantera and other solver which was not coupled with Cantera. Then he analysed the results by comparing different simulators and he concluded that best simulation was when he coupled OpenFoam with Cantera.

C. Togbe et. al. [11] used Cantera for modeling 1-pentanol oxidation, further they simulated laminar flame speeds and also they used Cantera to evaluate densities. To better understand characteristics of combustion of 1-pentanol better first of all they studied old data of 1-pentanol and then they presented new data about 1-pentanol's combustion. They gathered data about 1-penranol by performing experiments in two different environments. They used jet-stirred reactor also known as JSR to measure species concentration profiles at 10 atm. They measured concentration of species at different temperatures and also over different equivalence ratios. Further they measured 1-pentanol-air's flame speed at 1 atm and temperature of 423 K. For measuring flame speed, the equivalence ration was taken between in a range of 0.7 to 1.4. Further they modeled their experimental configuration in Cantera and performed oxidation of 1-pentanol. Additionally, when they analysed model developed in Cantera with experimental results they concluded that model they developed in Cantera showed very similar results with that of experimental result.

S. Khaitan et. al. [12] developed a method to get information about reaction mechanisms as well to analyse result by using Cantera. Further, they developed a technique which can be used to evaluate rate of reaction by using simulation in Cantera.

X. Li et. al. [13] developed a simulation in Cantera for getting value of rate of reaction of different chemical reaction involving ozone as a reactant and they tested their model at different temperature.

S. Gupta and N. Basant [14] in their study, developed QSRR (Quantitative Structure Reactivity Relationship) using single decision tree and decision tree boost for predicting the rate constants (k_{O3} and k_{SO4}) of reaction of O_3 and SO_4 with diverse organic chemicals in aqueous medium by following the OECD(Organization for Economic Co-operation and Development) guidelines for QSAR(Quantitative StructureActivity Relationship) analysis.

To know how much a substance or a chemical will pollute water we need to know aqueous rate constant of that substance or chemical. Aqueous rate constant is also written as K_{OH}. It is very difficult to find K_{OH} for each chemical by doing various experiments. Such excrements are not only time consuming but are costly as well so X. Luo et. al. [15] in their study developed a Quantitative StructureActivity Relationship (QSAR) model to predict the aqueous rate constant. They tried to develop a model which is more generalizable and can be applied to many chemicals. They tested their QSAR model by using the guidelines which are given by OECD. Their model was able to predict K_{OH} with satisfactory accuracy Model developed by them can be used for those chemicals too, which are not studied well.

P. Baldi et al[16] developed a system using machine learning which was able to predict the outcome of multistep reactions. They developed a rule based model for predicting such multi step reactions. Further they tried to generalise his model such that model can be used to predict wide number of reactions under various different conditions.

2.1 Machine Learning Approaches

Learning is nothing but inferring knowledge from the information we gathered in past. In humans learning starts from the minute we are born and this process of learning continues till the end of life and in this duration humans try to gather as much knowledge as possible and then try to learn from that knowledge which was gained from various experiences.

Artificial Intelligence(AI) tries to simulate process of learning which happens in not only humans but also in other living things, in lifeless machine. Artificial Intelligence enables the machines to perform the task with highest amount of precision as well as accuracy given to them without needing human interference. Machine

learning is a subfield of AI and the main area in which machine learning works is to develop new algorithms and as well as understand and evaluate algorithms which enable the machine to learn. These days in industries, machine learning is one of the most popular area of interest/work. Machine learning tries to bring other fields like brain modeling,human psychology and statistics together to build an intelligent system. Neural network which are inspired from working of brain are used widely in machine learning for learning from data. Machine learning uses analysis of data and machine learning algorithms too use analysis skills, thus statistics plays a very important part in machine learning. When a computer is used to solve or deal with a particular task is then that task is known as task domain or sometimes also referred to as knowledge base. Information that is produced by or gotten from the task constitutes its knowledge base. To represent knowledge base we use numerical, discrete value, relational literals and Boolean or sometimes their combination is also used. Input-output pairs are used to represent knowledge base, here input given to task is input and results which we get frrom that task is output. Knowledge base's data can be used to classify output for a given input. Knowledge base is not enough to know the internal working of a task but it is enough for classifying a given input to some output. As when we have a lot of information it is next to impossible for humans to get information from it, machine learning, on the other hand can easily do this. With the help of more data a computational model is made which can represent that task with significant accuracy without knowing internal working of that task. An algorithm can use computational model to predict output for some unabsorbed input for that particular task. The computational model can be of any type it can be simply some rules, a formula, or some mathematical operations which when applied to input give an output.

Every machine learning algorithm uses different technique to make computational model from knowledge base but goal of every machine learning algorithm is to infer knowledge from knowledge base.

To learn about a process, machine learning algorithms need dataset. Dataset have data regarding, which outputs was given for a particular input Each input has some attributes in it, which tell us about properties of that particular input, an input can have two attribute or sometimes, there can be thousands of such attributes. An attribute can either be continuous or discrete. Discrete attributes as the name implies have distinct values such as shape of an object can either be square, rectangle etc; on the other hand continuous have numeric values such as area of shape. Every dataset has some input and some output attribute. Input is basically given to the learning algorithm and the goal of learning algorithm is to

map the given input to the output corresponding to that particular task.

It is assumed in machine learning that values of input and output are inter-dependent. Input attributes given in dataset are known as features in machine learning. The computational model can also be thought as of a function which simply maps our input to an output.

Machine learning has many applications these days, for example, we can train a computational model from emails such that computation model can learn to distinguish between important and spam email. Once a computational model is trained, then that model can be used to keep important mails in one folder and spam on other.

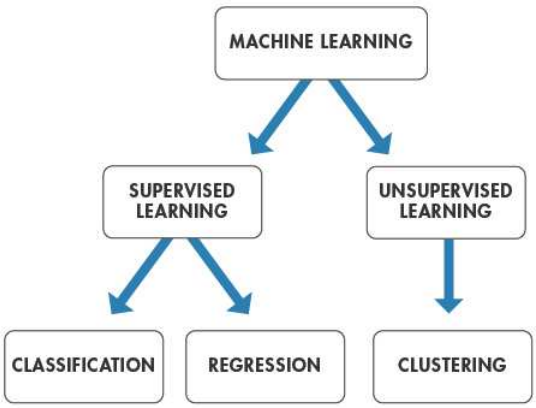


Figure 2.1: Machine learning categorization

Figure 2.1 shows the categorization of machine learning approaches. It is divided into two categories i.e. supervised learning and unsupervised learning. Supervised learning is classified into (i) Classification - predict discrete valued output (ii) Regression - predict continuous valued output. Clustering is unsupervised learning. In supervised learning a computational model is trained to map input given in knowledge base to corresponding output. In supervised learning labels are given along data, on other hand, in unsupervised learning we are not given labels along with input. An example of supervised learning is to classify whether an email is important or not or for predicting how many runs a cricketer will make in his next match and examples on unsupervised learning includes clustering of documents.

Classification algorithms are those which are used to classify in which class a given input belongs to. Here classes are discrete in nature and some set of rules or a model can be made to classify a given input to a particular class. Input given to a classification algorithm can be either discrete or it can be continuous, but

there are some algorithms which only take input in form of discrete attributes but output of such algorithms are always discrete.

Regression algorithms create predictions based on some mathematical operations or some equation creates a model which gives a continuous value as output while taking some input. Here input can either be discrete or continuous.

Table 2.1 shows the recently developed classification and regression models. Implementation of all these models is available in R.

Table 2.1: Some machine learning packages in R

S.No	Model	Model Type	Method	Package	Tuning Parameter
1	ada	Classification	ada	ada	maxdepth, iter, nu
2	avNNet	Dual Use	avNNet	caret	decay, size, bag
3	bag	Dual Use	bag	caret	vars
4	bdk	Dual Use	bdk	kohonen	xweight,topo,xdim,ydim
5	blackboost	Dual Use	blackboost	mboost	maxdepth, mstop
6	Boruta	Dual Use	Boruta	Boruta	mtry
7	bstTree	Dual Use	bstTree	bst	maxdepth, nu, mstop
8	C5.0	Classification	C5.0	C50	winnow, trials, model
9	cforest	Dual Use	cforest	party	mtry
10	ctree	Dual Use	ctree	party	mincriterion
11	cubist	Regression	cubist	Cubist	committees, neighbors
12	enet	Regression	enet	elasticnet	lambda, fraction
13	foba	Regression	foba	foba	lambda, k
14	GAMens	Classification	GAMens	GAMens	fusion, iter, rsm_size
15	gamLoess	Dual Use	gamLoess	gam	degree,span
16	gbm	Dual Use	gbm	gbm	trees, shrinkage,depth
17	gcvEarth	Dual Use	gcvEarth	earth	degree
18	glm	Dual Use	glm	stats	None
19	icr	Regression	icr	caret	n.comp
20	J48	Classification	J48	RWeka	C
21	JRip	Classification	JRip	RWeka	NumOpt
22	knn	Dual Use	knn	caret	k
23	lars	Regression	lars	lars	fraction
24	lda	Classification	lda	MASS	None
25	leapSeq	Regression	leapSeq	leaps	nvmax
26	Linda	Classification	Linda	rrcov	None
27	lm	Regression	lm	stats	None
28	logforest	Classification	logforest	LogForest	None
29	M5	Regression	M5	RWeka	smoothed,pruned,rules
30	nb	Classification	nb	klaR	usekernel, fL

Chapter 3

Problem Formulation

By nature, chemistry is concerned with change. Chemical reactions convert substances with well-defined properties into other materials with different properties. Much of our study of chemical reactions is concerned with the formation of new substances from a given set of reactants. However, it is equally important to understand how rapidly chemical reactions occur.

The rates of reactions span an enormous range, from those that are complete within fractions of seconds, such as certain explosions, to those that take thousands or even millions of years, such as the formation of diamonds or other minerals in earth's crust. The area of chemistry that is concerned with the speeds, or rates, of reactions is called chemical kinetics. Chemical kinetics is a subject of broad importance. It relates, for example, to how quickly a medicine is able to work, to whether the formation and depletion of ozone in the upper atmosphere are in balance, and to industrial problems such as the development of catalysts to synthesize new materials.

There are many factors which affect rate of reaction which are listed below:

- Concentration
- Pressure
- Temperature
- Presence of catalyst
- Physical state of the reactants and surface area

Currently to find rate of reaction for a particular environment we use Arrhenius equation. However, for using Arrhenius equation we need value of parameters and finding value of parameters of Arrhenius equation is very expensive and time consuming. There is need to develop a technique by which we can find rate of reaction which is less time consuming as well as cheap. In this research, we tried to use machine approach to predict rate of reaction.

3.1 Research Gaps

Following are the gaps that are identified during literature survey:

1. The already present techniques involve finding Arrhenius parameters and then use Arrhenius equation to find rate of reaction but these techniques are very costly and time consuming [13].
2. There is need to develop a technique which can predict rate of reaction with minimum number of attributes about reaction [14].
3. There is no on-the-shelf dataset available which can be used for training model [15].
4. No one has applied machine learning techniques to find rate of reaction yet [16] .
5. To find chemical rate of reaction by using physio chemical properties of reactants need to be studied.

3.2 Research Objectives

The following research objectives are formulated:

1. To collect data from simulation and use this data for model training.
2. To create and analyse various machine learning algorithms on obtained dataset and identify best model which is able to find rate of reaction most accurately.
3. To check the robustness of our selected model.

Chapter 4

Simulation and Dataset

These days inorder to understand complex biochemical processes, simulation and modeling is being widely used. Reaction prediction has the potential to increase our understanding of biochemical catalysis and metabolism [17].

4.1 Cantera

Cantera 2.3.0 [18] is an is open-source software that can enhance problem solving in the areas of kinetics, thermodynamics and transport phenomena. Various components can be inserted in Cantera like gas mixtures, specific reactor designs, kinetics mechanisms, reactivity of surfaces and other relevant files. Cantera being open-source software, thus it allows a wide range of interference from the user. Moreover, Cantera can run on various environments like Matlab, FORTRAN, Python and C++. In our case the model is developed in a Python-Cantera environment although in the future it could be the case that Cantera will be also applied by the authors in other interfaces in order to process more complex mechanisms. Cantera may apply both stoichiometric and non-stoichiometric methods in order to calculate the thermodynamic equilibrium. The method of element potential minimization (non-stoichiometric) and the Villars Cruise Smith algorithm (stoichiometric) can be applied depending on the preference of the user. Cantera can be used to get rate of progress of chemical reaction at different temperatures and pressures.

Cantera uses modified Arrhenius equation for getting value of rate constant.

$$k = AT^{n}e^{-Ea/(RT)} \tag{4.1}$$

In above equation if we will put n(which is a contant) = 0 then it will become same as original Arrhenius equation. Typically value of n lies between - 1 and 1. Various predictions for n are derived using theoretical analyses. It has been pointed out that “it is not feasible to establish, on the basis of temperature studies of the rate constant, whether the predicted T dependence of the pre-exponential factor(A) is

observed experimentally". However, if we can get some evidence, from theory or from experiment then we can perform incisive tests of Arrhenius law.

4.1.1 Input File

The most important part for a simulating a chemical reaction in Cantera is to develop an input file. Input file contains nearly everything which is important for Cantera to work. It contains a lot of information about chemical being used, such as name of chemical, its composition etc. Input file has directives and entries, both of them have a syntax similar to a function. An entry has information about object, that object can be anything like a phase, species or a reaction. Directives have information which explains about the parameters of entries, for example, how different errors are being handled by entries can be found in its directive.

Input file of Cantera follows the same syntax as that of Python which make it easier to learn.

Every entry in Cantera have a field in which one can assign values to. for example a species entry will be written as

species(name='C12', atoms='C:12')

There are some fields which are mandatory for every entry and if that field is not filled then error message will be shown.

4.1.1.1 Species Name

The name filed of species can have any character provided that character is not a reserved character of XML. Generally, we use - or + signs to indicate charge of that particular molecule, but this is not mandatory. For example,

name = 'CH4 (singlet)'

name = 'H2O'

name = 'water'

name = 'arg_2 +'

4.1.1.2 Elemental Composition

Entry of atom also specifies the composition of elements, for example:

atoms = “H:2 O:1” # H20

atoms = “H:2, O:” # H2O with optional comma

atoms = “Y:1 Ba:2 Cu:3 O:6.5” # stoichiometric YBCO

atoms = “” # an empty site represented

atoms = “Ne:1 E:-2” # Ne++

There are several ways by which we can specify a species in liquid solution, or species, or surfaces however there is only one way by which we can represent species in gaseous form as the composition of gas is well defined. On other hand like in liquid solution a species can be defined both without and with water molecules including in the solvation cage. On other hand when we talk about surface species we are allowed to even skip the field of atom completely, however when we skip the atom field than it means that it has nothing in composition and it is an empty surface site.

Vacancies which are present in solid are also represented in input file. To represent a charge vacancy we define it to be composed of electrons for example:

species(name = ‘Yttria-Stabilized Zirconia-oxygen-vacancy’,

atoms = ‘O:0, E:2’,

...,) Here in above, atom number is optional and if it is 0 then this implies that element is not present there.

Number of atoms will always be a whole number for any element, the only exception to this is some special element.

4.1.1.3 Thermodynamic Properties

Thermodynamic properties which are appropriate for a particular interface or phase are represented using the ideal interface and phase entries. Although each one may use different expressions to compute the properties, they all require thermodynamic property information for the individual species. For the phase types implemented at present, the properties needed are:

- The molar heat capacity at constant pressure $c^0p(T)$ for a range of temperatures and a reference pressure P0;
- The molar enthalpy $h^(T0, P0)$ at P0 and a reference temperature T0 ;
- The absolute molar entropy $s^(T0, P0)$at (T0,P0)

4.1.1.4 Species Transport Coefficients

Some coefficients which can tell how each species is affecting transport properties of phase are required by transport property models. An entry that specifies species specific coefficients can be given in transport field.

Parameters which are required by ideal gas transport property model are given by the entry type which is gas transport and currently gas transport is only entry type available. In gas transport, entry value of field as well as their units are at par with the transport database parameters[19]. The data which is given in the transport database is directly used in the gas transport entry without even converting units. The numeric fields in transport coefficient should only have numeric values, no string is allowed.

4.1.1.5 Thermodynamic Property Models

There are different type of parameterization implemented in different types of entry for describing heat capacity. It is not important that each species in a system will use same type of parameterization and in fact each species in a system can use different type of parameterization for heat capacity. Currently, several types are implemented which provide species properties appropriate for models of pure compounds, ideal solutions and ideal gas mixtures.

4.1.1.6 The NASA 7-Coefficient Polynomial Parameterization

To solve problem related to combustion, it is very important that one has knowledge about thermochemical properties such as enthalpy, heat capacity and entropy. In 1961 Sanford Gordon and Frank Zeleznik found a new technique by which more than one property can be found with a single regression polynomial. They named that technique as NASA 7 which was first published by Gordon and Zeleznik [20] in 1962 and then by McBride et. al. [21]in 1963.

The NASA 7-coefficient polynomial parameterization is used to compute the species reference-state thermodynamic properties $\hat{c}_p^0(T)$, $\hat{h}^0(T)$ and $\hat{s}^0(T)$.

The NASA parameterization represents $\hat{c}_p^0(T)$ with a fourth-order polynomial:

$\frac{c_p^0(T)}{R} = a_0+a_1T+a_2T^2+a_3T^3+a_4T^4$

$\frac{h^0(T)}{RT}=a_0 + \frac{a_1}{2}T + \frac{a_2}{3}T^2 + \frac{a_3}{4}T^3 + \frac{a_4}{5}T^4 + \frac{a_5}{T}$

$\frac{s^0(T)}{R}=a_0 \ln T + a_1T + \frac{a_2}{2}T^2 + \frac{a_3}{3}T^3 + \frac{a_4}{4}T^4 + a_6$

NASA 7 uses only 7 coefficients for each temperature region, however, the NASA polynomial form with 7 coefficients is now outdated. The NASA 7 polynomial form is used by some of popular programs such as Chemkin. Chemkin is not compatible with the most recent form which for each temperature region uses 9 coefficients.

A NASA parameterization is defined by an embedded NASA entry. Very often, two NASA parameterizations are used for two contiguous temperature ranges. This can be specified by assigning the thermo field of the species entry as a sequence of two NASA entries:

species(name = “H2”,

atoms = “ H:2 ”,

thermo = (

NASA([100.00, 900.00], 2.344331120E+00, 7.980520750E-03,

-1.947815100E-05, 2.015720940E-08,-7.376117610E-12,

-9.179351730E+02, 6.830102380E-01),

NASA([900.00, 3000.00], [3.337279200E+00, -4.940247310E-05, 4.994567780E-07, -1.795663940E-10,2.002553760E-14, -9.501589220E+02, -3.205023310E+00])))

4.1.1.7 The NASA 9-Coefficient Polynomial Parameterization

After NASA 7, an extension of NASA 7 was developed which is known as the NASA 9-coefficient polynomial parameterization which uses 9 coefficients instead of just 7 and these two new coefficients are used in each temperature region, and also used when we have multiple regions in temperature range.

The NASA 9 parameterization represents the specie’s thermodynamic properties with the following equations:

C0p(T)R=a0T2+a1T1+a2+a3T+a4T2+a5T3+a6T4

H0(T)RT=a0T2+a1lnTT+a2+a32T+a43T2+a54T3+a65T4+a7T

s0(T)R=a02T2a1T1+a2lnT+a3T+a42T2+a53T3+a64T4+a8

An example of H20 in which its three different temperature regions can be represented like:

species(name=‘H2O’,

atoms='H:2 O:1',

thermo=(NASA9([100.00, 900.00],

[3.823578150E+04, -7.124641460E+02, 1.268013678E+00,

3.134567891E-03, -3.231268190E-07, -6.123762611E-10,

3.123512341E-13, -3.145671581E+04, -5.168816141E+00]),

NASA9([900.00, 5900.00],

[2.314145131E+05, -2.135146151E+03, 7.334131213E+00,

-8.114257101E-05, 3.671413413E-09, -2.125841223E-12,

1.236999901E-16, -2.812698172E+04, -1.5235891732E+01]),

NASA9([5900.00, 100000.00],

[-2.134113146E+09, 2.121212572E+06, -1.246824242E+02,

4.134141312E-02, -1.231142847E-06, 7.0134131101E-11,

-8.842351500E-16, -8.01332111E+06, 2.131171123E+03])),)

It should be noted that in above example the coefficients are dummy ones which are just meant so that reader can understand how NASA 9 works. There is a tool named as NASA thermoBuild which is used to collect thermodynamic properties for different species, one can easily generate an input file along with thermodynamic properties of that species.

For simulation, we first created an input file for Cantera. We obtained data about chemical reaction from GRI 3.0 [22]. Then we feed this input file in Cantera and simulated different chemical reaction with different mole fraction of chosen chemical species. To generate data, random values of temperature, pressure and concentration of reactants was taken and these values were given to Cantera to get net production rate. The range of temperature, pressure and concentration of reactants is shown in Table 4.1.

Table 4.1: Description of data

Parameter	Lower bound	Upper bound
Temperature(Kelvin)	200	5000
Pressure(Pascal)	8000	1000000
Concentration of reactant(moles)	0	10

4.2 Dataset

4.2.1 Features

We noted down following properties of chemical reaction during simulation :

1. Density
2. Temperature
3. Pressure
4. Concentration of 1^{st} element
5. Concentration of 2^{nd} element
6. Activation energy of chemical reaction
7. Frequency factor(A)
8. Value of n

Density is a physical property of matter, as each element and compound has a unique density associated with it. Density is defined in a qualitative manner as the measure of the relative "heaviness" of objects with a constant volume. For gases, the density may vary with the number of gas molecules in a constant volume. Density in gas is very different from that of solids. In solids, density is relationship between mass of compound and volume, i.e, how much space it takes, whereas, in gases it is more complex and mainly depends on pressure and temperature. Density in solids and liquid, is also responsive to pressure and temperature but this response is very less. When a substance is more dense then it means the molecules in substance are near and possibility of collisions between molecules increases. Particles can only react when they collide. If you heat a substance, the particles move faster and hence collide more frequently. That will speed up the rate of reaction. Temperature also affects the rate of chemical reactions as temperature increases the molecules begin to travel faster and hence probability of collision between molecules increases.

Pressure also affects the chemical rate of reaction, when we increase the pressure then concentration of reactants will increase and hence speed of chemical reaction will increase. In our experimental reaction, we considered only two reactants and the concentration of both were recorded in dataset.

As we increase concentration of chemicals involved in reaction the rate of reaction also increases. Activation energy, A(frequency factor) and constant n are unique values for every chemical reaction. The value of all 3 attributes are taken from GRI 3.0 [22]. The unit of rate of reaction in our dataset is kmol/m^3/s. Activation energy is the threshold energy that reactants need to possess before the formation products. If the threshold value is not met, the products are not formed. Hence, we can say that, increase in activation energy decreases product formation and hence, the overall rate of reaction.

Chapter 5

Methodology and Models

5.1 Approach

The methodology followed is shown in Figure 5.1. We take 5 chemical reactions and simulate them in Cantera to get the respective value of reaction rate. The chemical reactions which we choose are the following:

- $CO + O\ (+M) \leftrightarrow CO_2(+M)$
- $CO + O_2 \leftrightarrow CO_2 + O$
- $C_2H_2 + OH \leftrightarrow CH_3 + CO$
- $CO + OH \leftrightarrow CO_2 + H$
- $CH_2O + O_2 \leftrightarrow HO_2 + HCO$

We created separate dataset for different chemical reactions from the combined dataset which we obtained from simulation.In data cleansing, we removed activation energy, A(frequency factor) and n from our dataset, these values were not varying. After data cleansing we were left with 5 different attributes.we trained our models with 70 % data and the remaining was used as test data to verify our results.

5.2 Model Evaluation

There are many ways by which we can measure the performance of prediction. Some methods perform better in some application, while others perform better in some other application. Thus, depending on the application we choose method to evaluate our prediction. A brief discussion on some of methods which are used to evaluate performance of prediction is given in following subsections

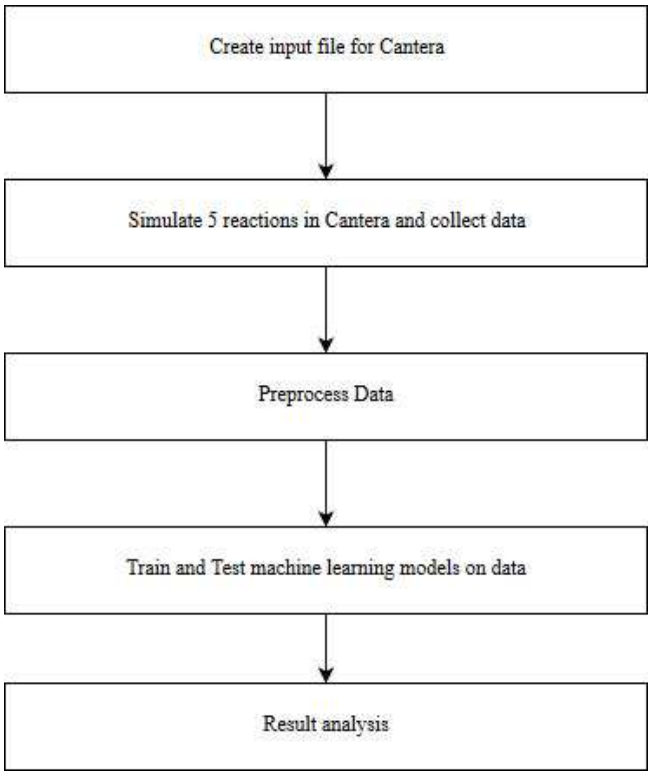


Figure 5.1: Methodology

5.2.1 RMSE

One of the most popular techniques to measure performance of regression model is RMSE(Root Mean Square Error) [23]. However, this method can be used only when we compare models whose errors are measured in the same units. To calculate RMSE we use following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}} \tag{5.1}$$

where, a_i is actual target p_i is predicted target n is the total number of instances.

5.2.2 Correlation (r)

Correlation [23] is a statistical technique which describes to what extent are two variables related to each other. With the help of correlation, we can find how much predicted values are related with actual values. Two variables are considered as highly and positively correlated if value of correlation tends to 1 and if value of correlation is tending towards 0 then it means there is no relation between two

variables. Correlation is defined as follows

$$r = \frac{\sum_{i=1}^{n}(x-\overline{x})(y-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x-\overline{x})^2 \sum_{i=1}^{n}(y-\overline{y})^2}} \tag{5.2}$$

where x is the actual value, y is the predicted value, $\overline{x}$ is the mean of the all actual values, $\overline{y}$ is the mean of the all predicted values, and n is the number of instances.

5.2.3 Coefficient of determination (R^2)

The coefficient of determination (R^2) [23] summarizes the explanatory power of the regression model. R^2 describes the proportion of variance of the dependent variable explained by the regression model. If the regression model is perfect then R^2 is 1 and if the regression model is a total failure then R^2 is zero i.e., no variance is explained by regression. The coefficient of determination is computed by taking the square of r (i.e. correlation). It is defined as follows:

$$R^2 = r \times r \tag{5.3}$$

5.2.4 Accuracy

The accuracy is calculated as percentage deviation of predicted target with actual target with acceptable error [23].

$$\begin{aligned} Accuracy &= \tfrac{100}{n} \textstyle\sum_{i=1}^{n} q_i \\ q_i &= \begin{cases} 1 & \text{if } abs(p_i - a_i) \leq err \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{5.4}$$

where a is actual target, p is predicted target, err is the acceptable error, and n is the total number of instances.

5.2.5 k fold Cross Validation

To check robustness of our model we use k-fold cross validation[24]. In this technique, dataset is randomly partitioned into k equal sized samples and out of these k samples, k-1 samples are used for training and remaining 1 sample is used for testing our result and this process is repeated k times (the folds) such that each

and every k sample is used for testing once. The k results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once.

5.3 Models

5.3.1 Random Forest

Random Forest is one of the most popular machine learning algorithm which was created by Breiman [25]. Random Forest is nothing but a group of many simple decision trees and all these trees are able to predict the outcome for any input. These trees are able to predict to which class a particular input belongs, if our problem is of classification, and if problem is of regression, these trees are able to predict a continuous number. In case of classification each tree in random forest votes for a particular class and the class which has most votes is given as output for that particular input, on the other hand, in regression output of every tree is averaged to obtain the output for that particular input.

Random Forest can be seen as ensemble of many simple decision trees. Ensembling of many decision trees in random forest has shown dramatic improvement in performance of model. Random Forest is also able to overcome the issue of overfitting, which is one of biggest problem in single decision tree. During training of model each decision tree in model is trained on random subset of features of training data. Another ensemble technique, bagging, selects random sub samples of training data and trains model on them but random forest is different from bagging as here we are not only choosing random samples of training data but we are choosing random sample of features as well. Random forest with help of multiple decision tree are much more generalised when compared to single decision tree as there is very less chance of overfitting. Random Forest can also be used to rank features. The idea of feature selection using random forest was given in the original paper of random forest itself [25].

Figure 5.2: Representation of Random Forest

5.3.2 BST (Boosted Tree)

There are many techniques by which we can ensemble models to enhance performance of models, two of them are bagging and boosting. In boosting, several weak learners are bring together to become a very strong learner. Now one can think what actually is weak learner? Well, weak learner is one who has slightly better accuracy than totally random chance. The idea of boosting was built upon Leslie Valiants Probability Approximately Correct (PAC) or work on distribution free learning, which was basically a framework for analysing the machine learning problems. The basic hypothesis for boosting was to filter out those observations which are handled by weak learner with high accuracy and with rest of data focusing on creating a new weak leaner which can learn some more observation from it with high accuracy. It may then filter out those observation which new model can handle with high accuracy and with the remaining data creating another model and this process will continue until there is no change in accuracy by adding new models.

A boosted decision tree can be created by combining many trees with the help of boosting. In boosted tree, every tree depends on the tree which are already created before it, new tree tries to fit the residuals of prior tree and hence minimizing error. Thus, by doing this boosting helps tree to increase accuracy of model.

When in dataset features are correlated then usually decision trees give high amount of accuracy. If features in dataset are not correlated i.e the have high degree of entropy then training a tree from dataset will not work and that tree will have very less accuracy. In AdaBoost we create decision tree which have only single split and that is why such decision tree are called as decision stumps due to its shortness. To create trees, Adaboost gives different weights to different observations, observations which are hard to classify have got more weight and observation which can easily be classified have less weight. Now weak learner which we are creating in Adaboost have more focus in classifying the difficult observation rather then the simple one which are already classified. Algorithms like Adaboost were put in a framework of statistics by Brieman who called them as ARCing(Adaptive Reweighting and Combining) algorithms. All steps in an ARCing algorithms involves minimization of weights and then classifier and weighted input are recomputed for further steps.

The framework for ARCing algorithms was developed further by Friedman and he named them as GBM (Gradient Boosting Machines). Later, he referred them as gradient boosting or sometimes gradient boosted tree.

In the language of statistics boosting is problem of numerical optimization in which objective is to bring loss of model to minimum levels by putting more and more weak learners using gradient descent like procedure.

Boosted algorithms are also referred to as stage wise additive model, as in such algorithms we are adding only one weak learner at a time and the already added models are not changing at all in other steps.

To expand boosting beyond classification the generalization of boosted models was done which enables boosted algorithm to use differential loss functions and hence boosted models can work with regression problems as well on multi class classification.

For boosting, weak learners should not be too weak or too strong. This can be achieved by putting some constraint on tree. A general idea is that if we put more constraints on tree then we will need more number of trees or its vice versa, i.e, if there are less constraints on a tree then we will need fewer trees. Some of constraints which can be applied on trees are as follows:

- Number of trees:- if we keep adding more and more trees then out model can overfit. This can be stopped by adding new trees only till performance of model is increasing, if it stop increasing then immediately stop.
- Tree depth:- when height of tree is increased then it become more complex

and can also overfit thats why shorter trees are more preferred. We should keep height between 4 to 8 to get better results.

- Number of leaves or number of nodes:- one can apply constraint on this too to avoid overfitting.
- Number of observations per split:- This can also be used as a constraint. In this we put minimum number of training data at a node before we can split node.
- Minimimum improvement to loss:- is a constraint on the improvement of any split added to a tree.

5.3.3 kNN (k nearest neighbours)

kNN stand for k nearest neighbour, it is one of non parametric algorithms for learning. The meaning of non parametric is that kNN basically does not assume any underlying data distribution. Being non parametric can be very useful as data in the real world does not follow any distribution(e.g. linearly separable, gaussian mixtures etc) then in such a dataset, algorithm which is non parametric can be useful . kNN along with non parametric algorithm is also a lazy algorithm i.e. any generalization is not done by taking any training data points or we can say that in kNN there is not a separate training phase or it is very negligible to be called a separate phase which makes training of data extremely fast. When we say there is no generalization or there is lack of generalization then it means that all training data is used. To be exact it means that when we are testing data then all training points are needed but on other hand when we see other algorithms like SVM one the curve is made in SVM then all training points can be ignored and prediction will depend only on that curve where as in kNN you need all the training points every time.

In kNN there is negligible cost in training but cost of testing phase is extremely high, Here cost doesn't mean money but here cost is in terms of money as well as time. kNN needs more memory as it stores all training data in memory. More time is needed in kNN as we have to consider all data points and have to find distance of point to all other points.

5.3.4 GBM (Gradient Boosting Machines)

GBM or Gradient Boosting Machines are one of most powerful machine learning algorithms. GBM's have shown high success in recent times for both classification and regressing dataset. In GBM's new models are fit during learning technique to reduce error and increase error of new model made[26]. The model which are fit during learning process are weak learners. The main idea on which this algorithm work is to make a new model from ensemble models such that the new model's loss function is highly correlated with the negative gradient of ensemble models. There can be any any loss function being used but to understand process better lets assume regular square error is being used as loss function, in this case learning algorithm will do consecutive error-fitting. A researcher can use any loss function in learning process, there are wide range of error functions which are available right now and a researcher if he wants can make his own loss function for a particular task.

GBMs are very highly flexible and they can be customized for a different datasets differently. Very high degree of freedom is there in GBM's during model design and also different loss function can be chosen by doing trial and error. Due to so many different features of GBM's they are used widely with greater success.

5.3.5 M5 (Model Tree)

In model trees, instead of simple nodes at leaf node there are decision trees. They were originally proposed by Quinlan. An open-source implementation, called M50, was presented in [27] and has proven to be successful in many practical applications. The inclusion of linear regression models rather than constant predictors at the leaf nodes is essential: standard regression trees with constant predictors produce much less accurate predictions than model trees [28]. M5 trees are grown using the standard top-down approach for growing decision trees. Once an unpruned tree has been grown, multiple linear regression models are placed at each node of the tree. Following this, the tree is pruned, potentially replacing large subtrees by a single linear regression model. Finally, the linear regression models along the path from the root node of the tree to each leaf node are combined into a single linear regression model, at the leaf node, using a smoothing process that produces a linear combination of linear regression models. For further details, see [27, 28]

Chapter 6

Results and Discussion

Some chemical reactions occur more rapidly then others. In our dataset the range of our 5 chosen chemical reaction is shown in Table 6.1.

Table 6.1: Range of chemical rate of reactions

Chemical equation	Range of reaction rate $\times 10^4$(kmol/m^3/s)
CO + O (+M) $\leftrightarrow$ CO_2(+M)	[0,1.9773]
CO + O_2 $\leftrightarrow$ CO_2 + O	[0,0.2853]
C_2H_2 + OH $\leftrightarrow$ CH_3 + CO	[0,4.9999]
CO + OH $\leftrightarrow$ CO_2 + H	[0,212.7785]
CH_2O + O_2 $\leftrightarrow$ HO_2 + HCO	[0,24.9139]

Table 6.2: Parameters for chemical reaction

Chemical Equation	Rate Parameter		
	A	n	E_a(cal/mol)
CO + O (+M) $\leftrightarrow$ CO_2(+M)	1.8E+07	0	2385
CO + O_2 $\leftrightarrow$ CO_2 + O	2.5E+09	0	47800
C_2H_2 + OH $\leftrightarrow$ CH_3 + CO	4.83E-07	4	-2000
CO + OH $\leftrightarrow$ CO_2 + H	4.76E+04	1.228	70
CH_2O + O_2 $\leftrightarrow$ HO_2 + HCO	1.0E+11	0	40000

We have taken accepted error as 1% of range of rate of reaction. Parameter for chemical reaction used in input file of Cantera are given in Table 6.2. Values in Table 6.2 are taken from GRI 3.0 [22].The results reported in GRI 3.0 are products of computational and experimental research sponsored by the Gas Research Institute

We trained 20 machine learning models on our data. Here we provided results of 5 best models which we found. We run these 5 models on dataset having data of 5 different chemical reaction. Parameters of model which we trained are given in Table 6.3. We used R for training and prediction. To measure the performance of models we used correlation,R square,RMSE,Accuracy. Results of 5 chosen reaction is shown in Tables 6.4 to 6.8

Table 6.3: Machine learning models

Model name	Parameters
Random Forest [29]	ntree=800, mtry=2
Bst [30]	mstop = 150, maxdepth = 5,nu = 0.1
GBM [31]	shrinkage=0.1,n.minobsinnode=10,n.trees = 150, interaction.depth = 3, shrinkage =0.1 and n.minobsinnode = 10
KNN[32]	k=5
M5[33]	pruned = 'Yes', smoothed = 'No', rules = 'No'

Table 6.4: Results for CO + OH $\leftrightarrow$ CO_2 + H

Model name	Correlation(r)	R^2	RMSE	Accuracy
M5	0.973	0.948	15004.86	70.58
Random Forest	0.988	0.978	12739.62	71.68
Gbm	0.983	0.967	19536.19	51.75
knn	0.876	0.767	40632.67	50.45
Bst Tree	0.990	0.982	9088.52	80.84

Table 6.5: Results for CO + O (+M) $\leftrightarrow$ CO_2(+M)

Model name	Correlation(r)	R^2	RMSE	Accuracy
M5	0.964	0.930	233.28	72.08
Random Forest	0.991	0.983	141.81	78.84
Gbm	0.983	0.967	230.30	66.01
knn	0.902	0.815	485.38	57.85
Bst Tree	0.995	0.991	105.21	83.87

Table 6.6: Results for CO + O_2 $\leftrightarrow$ CO_2 + O

Model name	Correlation(r)	R^2	RMSE	Accuracy
M5	0.978	0.957	29.66	73.52
Random Forest	0.990	0.981122	19.84	78.73
Gbm	0.980	0.960	35.24	63.82
knn	0.905	0.820	70.39	59.82
Bst Tree	0.996	0.992	14.21	82.37

6.1 k fold Cross Validation

We observed from our results that Bst tree outperform all other models and is very closely followed by random forest.

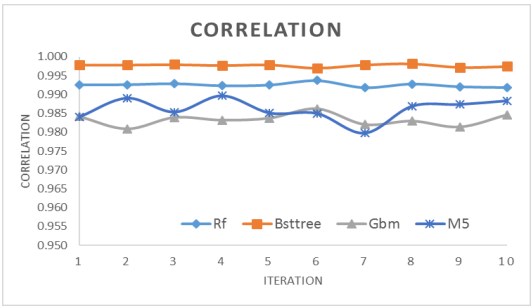
In this work, repeated K fold cross validation is used that describes the consis-

Table 6.7: Results for $CH_2O + O_2 \leftrightarrow HO_2 + HCO$

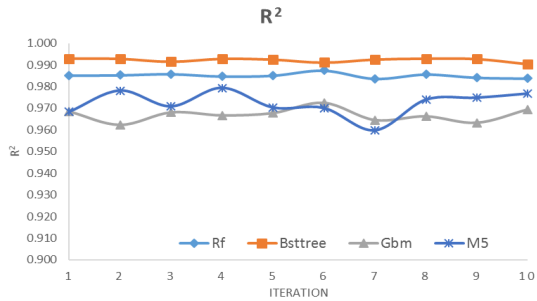
Model name	Correlation(r)	R^2	RMSE	Accuracy
M5	0.982	0.965	2698.71	76.31
Random Forest	0.987	0.975	2189.77	78.24
Gbm	0.980	0.962	3592.14	66.7
knn	0.925	0.856	6264.20	60.18
Bst Tree	0.995	0.991	1470.33	85.21

Table 6.8: Results for $C_2H_2 + OH \leftrightarrow CH_3 + CO$

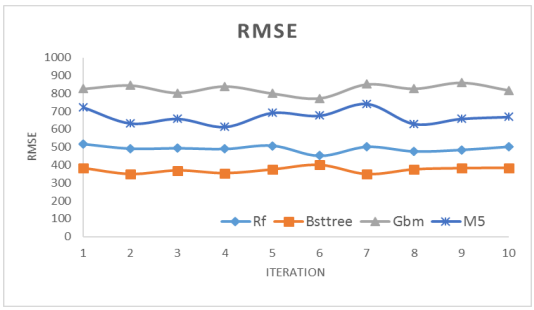
Model name	Correlation(r)	R^2	RMSE	Accuracy
M5	0.986	0.973	629.80	66.48
Random Forest	0.993	0.98	452.87	74.08
Gbm	0.986	0.972	771.63	57.45
knn	0.935	0.874	1406.22	50.95
Bst Tree	0.996	0.992	349.08	81.24



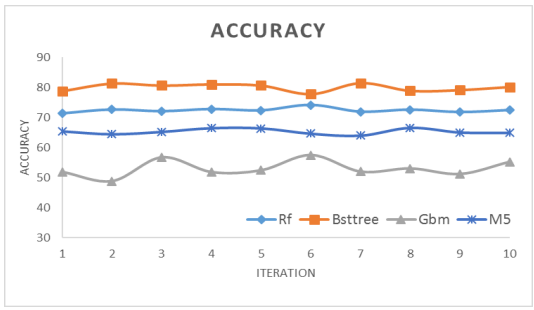
(a) Correlation



(b) R^2



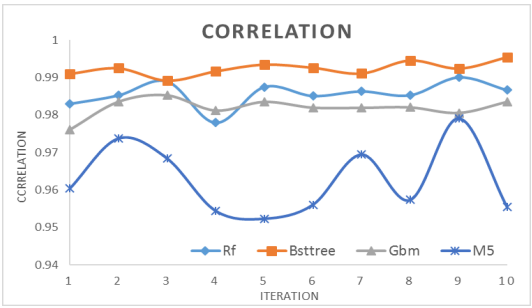
(c) RMSE



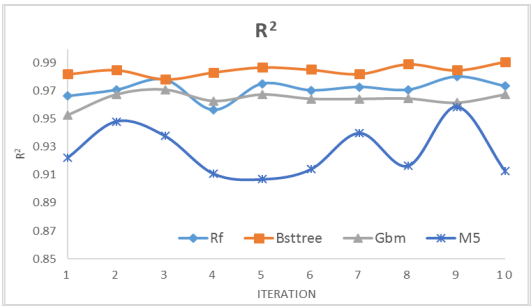
(d) Accuracy

Figure 6.1: Cross validation of $C_2H_2 + OH \leftrightarrow CH_3 + CO$

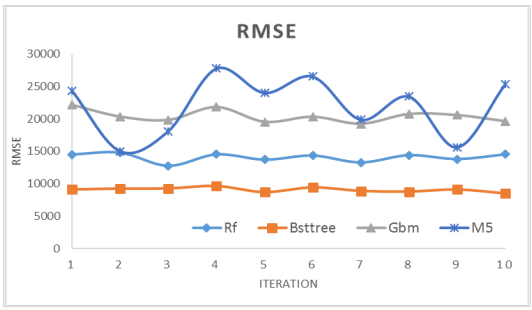
tency in the accuracy which means proposed model is not effected from these problems.Results of cross validation are shown in Figures 6.1 to 6.5. The result of cross validation conclude that the proposed model is free from overfitted/underfitted/biased issues.



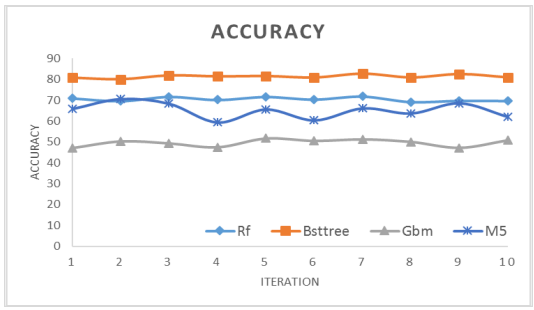
(a) Correlation



(b) R^2

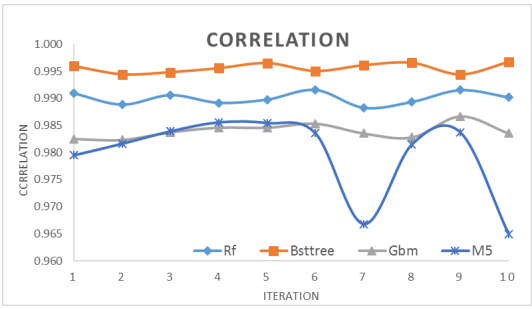


(c) RMSE

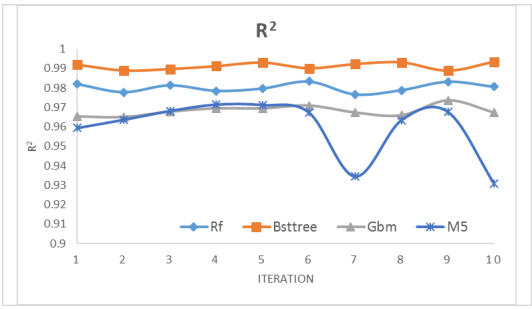


(d) Accuracy

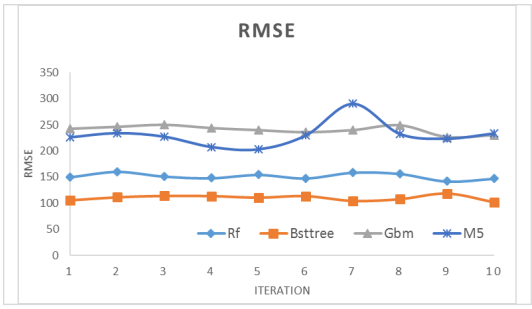
Figure 6.2: Cross validation of CO + OH $\leftrightarrow$ CO_2 + H



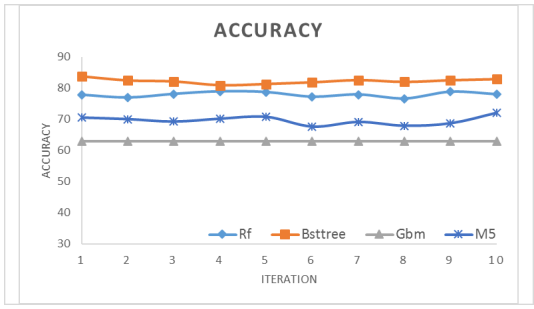
(a) Correlation



(b) R^2

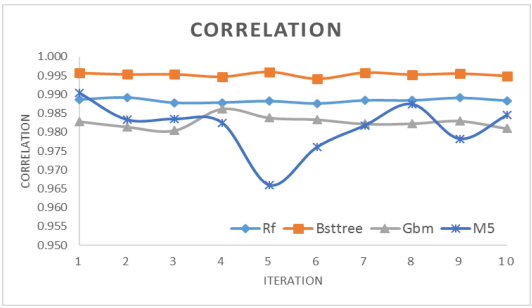


(c) RMSE

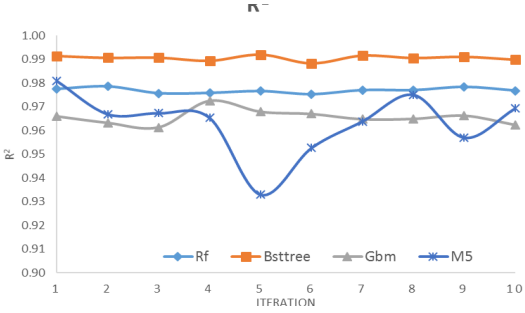


(d) Accuracy

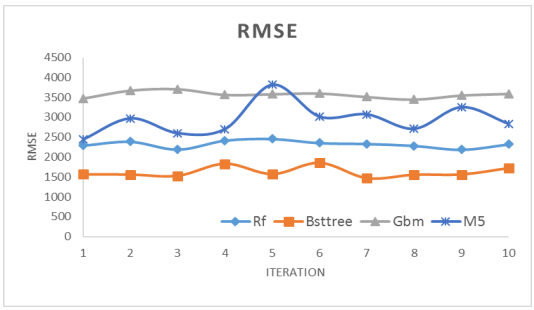
Figure 6.3: Cross validation of CO + O (+M) $\leftrightarrow$ CO_2(+M)



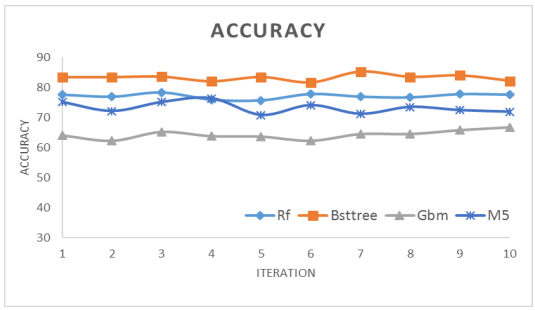
(a) Correlation



(b) R^2

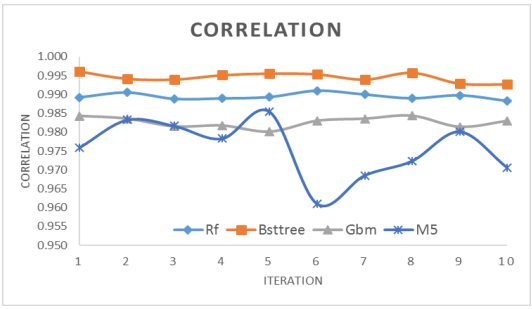


(c) RMSE

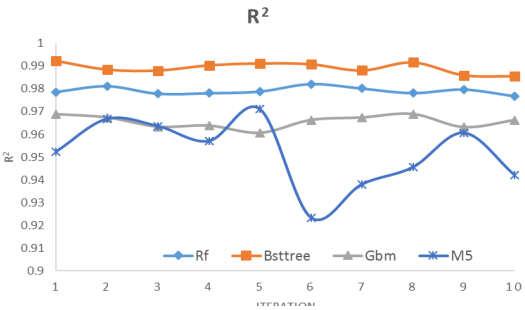


(d) Accuracy

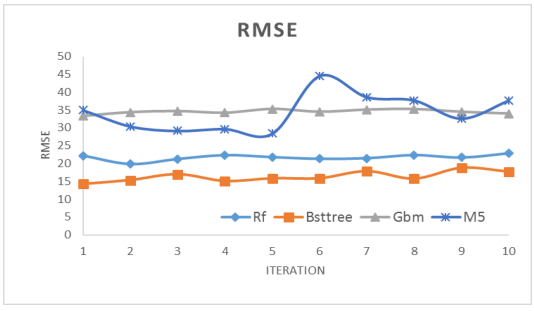
Figure 6.4: Cross validation of $CH_2O + O_2 \leftrightarrow HO_2 + HCO$



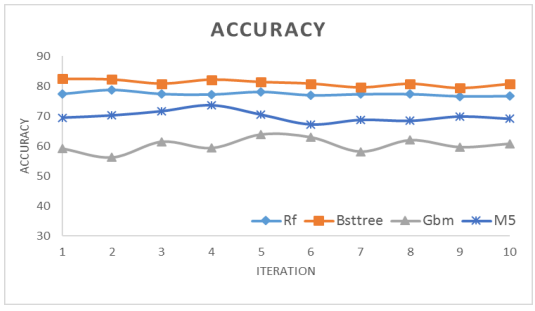
(a) Correlation



(b) R^2



(c) RMSE



(d) Accuracy

Figure 6.5: Cross validation of $CO + O_2 \leftrightarrow CO_2 + O$

Chapter 7

Conclusion and Future scope

This chapter is the concluding part of the thesis and also proposes some suggestions about how present work can be extended. Section 7.1 brings out the overall conclusions of the research work carried out in this thesis and in section 7.3 suggestions regarding the future research directions and possible extensions of the work presented in the thesis are made.

7.1 Conclusion

In this thesis we developed a machine learning approach to predict rate of chemical reaction. Finding parameters of Arrhenius equation can by costly as well as time consuming so proposed approach can be used as it don't need any such value. This approach can be effectively applied, where, we don't have enough data about the chemical species, which are being used, in our chemical reaction like, when we create a new drug, we don't have any previous data about our drug so we can predict with this approach how much rapidly drug will dissolve in human body in various situations like when a human have fever. This approach is much faster and cheaper when compared to previous approach.

7.2 Thesis Contributions

1. A machine learning technique to predict chemical rate of reaction is purposed.
2. We used Cantera to simulate chemical reaction to generate the dataset.
3. In this thesis, we have performed a comparative study of models on obtained dataset and tried to find which model works better, to predict chemical rate of reaction.
4. We used k cross validation method to check robustness of our models.

7.3 Future Scope

- In this thesis, ten machine learning models are used for predicting chemical rate of reaction. New machine learning approaches are available and they need to be explored for accurate and fast predictions.
- Physio chemical properties of reactants can be used to measure new features and improve the accuracy.

References

[1] V. Gold, K. Loening, A. McNaught, and P. Shemi, "Iupac compendium of chemical terminology," *Blackwell Science, Oxford*, 1997.

[2] I. Chorkendorff and J. W. Niemantsverdriet, *Concepts of modern catalysis and kinetics*. John Wiley & Sons, 2006.

[3] E. H. Kennard, "Kinetic theory of gases," 1938.

[4] K. J. Laidler, "The development of the arrhenius equation," *J. Chem. Educ*, vol. 61, no. 6, p. 494, 1984.

[5] S. Harris, L. Elliott, D. Ingham, M. Pourkashanian, and C. Wilson, "The optimisation of reaction rate parameters for chemical kinetic modelling of combustion using genetic algorithms," *Computer methods in applied mechanics and engineering*, vol. 190, no. 8, pp. 1065–1090, 2000.

[6] W. Polifke, W. Geng, and K. Döbbeling, "Optimization of rate coefficients for simplified reaction mechanisms with genetic algorithms," *Combustion and Flame*, vol. 113, pp. 119–134, apr 1998.

[7] R. J. Kee, F. M. Rupley, and J. A. Miller, "Chemkin-ii: A fortran chemical kinetics package for the analysis of gas-phase chemical kinetics," tech. rep., Sandia National Labs., Livermore, CA (USA), 1989.

[8] S. Vakalisa, D. Prandoa, F. Patuzzia, and M. Baratieria, "Thermodynamic modeling of biomass downdraft gasifiers: Introduction to the" multi-box" concept," 2014.

[9] L. Tao and Z. Dongmei, "Numeric simulation and analysis of h_2 o_2 premixed combustion based on OpenFOAM," in *2012 IEEE Symposium on Robotics and Applications (ISRA)*, Institute of Electrical and Electronics Engineers (IEEE), jun 2012.

[10] H. Jasak, A. Jemcov, Z. Tukovic, *et al.*, "Openfoam: A c++ library for complex physics simulations," in *International workshop on coupled methods in numerical dynamics*, vol. 1000, pp. 1–20, IUC Dubrovnik, Croatia, 2007.

[11] C. Togbé, F. Halter, F. Foucher, C. Mounaim-Rousselle, and P. Dagaut, "Experimental and detailed kinetic modeling study of 1-pentanol oxidation in a jsr and combustion in a bomb," *Proceedings of the Combustion Institute*, vol. 33, no. 1, pp. 367–374, 2011.

[12] S. K. Khaitan and M. Raju, "Dynamic simulation of air storage-based gas turbine plants," *International Journal of Energy Research*, vol. 37, pp. 558–569, nov 2011.

[13] X. Li, W. Zhao, J. Li, J. Jiang, J. Chen, and J. Chen, "Development of a model for predicting reaction rate constants of organic chemicals with ozone at different temperatures," *Chemosphere*, vol. 92, pp. 1029–1034, aug 2013.

[14] S. Gupta and N. Basant, "Modeling the reactivity of ozone and sulphate radicals towards organic chemicals in water using machine learning approaches," *RSC Adv.*, vol. 6, no. 110, pp. 108448–108457, 2016.

[15] X. Luo, X. Yang, X. Qiao, Y. Wang, J. Chen, X. Wei, and W. J. G. M. Peijnenburg, "Development of a qsar model for predicting aqueous reaction rate constants of organic chemicals with hydroxyl radicals," *Environ. Sci.: Processes Impacts*, vol. 19, pp. 350–356, 2017.

[16] M. A. Kayala and P. F. Baldi, "A machine learning approach to predict chemical reactions," in *Advances in Neural Information Processing Systems 24* (J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, eds.), pp. 747–755, Curran Associates, Inc., 2011.

[17] P. Rydberg, D. E. Gloriam, J. Zaretzki, C. Breneman, and L. Olsen, "SMARTCyp: A 2d method for prediction of cytochrome p450-mediated drug metabolism," *ACS Medicinal Chemistry Letters*, vol. 1, pp. 96–100, jun 2010.

[18] D. G. Goodwin, H. K. Moffat, and R. L. Speth, "Cantera: An object-oriented software toolkit for chemical kinetics, thermodynamics, and transport processes," 2016. Version 2.2.1.

[19] R. J. Kee, G. Dixon-Lewis, J. Warnatz, M. E. Coltrin, and J. A. Miller, "A fortran computer code package for the evaluation of gas-phase multicomponent transport properties," *Sandia National Laboratories Report SAND86-8246*, vol. 13, pp. 80401–1887, 1986.

[20] F. J. Zeleznik and S. Gordon, *A General IBM 704 or 7090 Computer Program for computation of chemical equilibrium compositions, rocket performance, and Chapman-Jouguet dentonations: Frank J. Zeleznik and Sanford Gordon.* National Aeronautics and Space Administration, 1962.

[21] B. McBride, S. Heimel, J. Eblers, and S. Gordon, "Thermodynamic properties to 6000," *K for*, vol. 210, 1963.

[22] G. P. Smith, D. M. Golden, M. Frenklach, N. W. Moriarty, B. Eiteneer, M. Goldenberg, C. T. Bowman, R. K. Hanson, S. Song, W. C. Gardiner, Jr., V. V. Lissianski, and Z. Qin, "Gri mech 3.0."

[23] P. S. Rana, H. Sharma, M. Bhattacharya, and A. Shukla, "Quality assessment of modeled protein structure using physicochemical properties," *Journal of bioinformatics and computational biology*, vol. 13, no. 02, p. 1550005, 2015.

[24] P. Pantola, A. Bala, and P. S. Rana, "Consensus based ensemble model for spam detection," in *Advances in Computing, Communications and Informat-*

ics (ICACCI), 2015 International Conference on, pp. 1724–1727, IEEE, 2015.

[25] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[26] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[27] Y. Wang, I. H. Witten, M. van Someren, and G. Widmer, "Inducing models trees for continuous classes," in *Proceedings of the Poster Papers of the European Conference on Machine Learning, Department of Computer Science, University of Waikato, New Zeland*, 1997.

[28] J. R. Quinlan *et al.*, "Learning with continuous classes," in *5th Australian joint conference on artificial intelligence*, vol. 92, pp. 343–348, Singapore, 1992.

[29] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[30] Z. Wang, T. Hothorn, and M. Z. Wang, "Package bst,"

[31] G. Ridgeway *et al.*, "gbm: Generalized boosted regression models," *R package version*, vol. 1, no. 3, p. 55, 2006.

[32] M. Kuhn, "Caret package," *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, 2008.

[33] G. Holmes, M. Hall, and E. Frank, "Generating rule sets from model trees," in *Twelfth Australian Joint Conference on Artificial Intelligence*, pp. 1–12, Springer, 1999.

List of Publications

1. Abhishek Kapoor, Akash Shrivastava, Jagmeet Kaur, Prashant Singh Rana and V.P. Singh, *"Predicting Rate of Chemical Reaction Using Machine Learning Techniques"*, Computational Biology and Chemistry, Elsevier. [IF=1.04] [Communicated]

thesis

ORIGINALITY REPORT

%12	%11	%5	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.thapar.edu:8080 Internet Source	%1
2	drjosephryan.com Internet Source	%1
3	Rana, Prashant Singh, Harish Sharma, Mahua Bhattacharya, and Anupam Shukla. "Quality assessment of modeled protein structure using physicochemical properties", Journal of Bioinformatics and Computational Biology, 2014. Publication	%1
4	www.aresinstitute.org Internet Source	%1
5	www.inderscience.com Internet Source	%1
6	www.cantera.org Internet Source	%1
7	www.oppapers.com Internet Source	%1
8	etd.gatech.edu Internet Source	<%1