iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** II **Month of publication:** February 2020

DOI: http://doi.org/10.22214/ijraset.2020.2113

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 8 Issue II Feb 2020- Available at www.ijraset.com

Estimation of Iron using Multiple Linear Regression Models

Mohammed Azam Sayeed[1], Noor Ayesha[2], Mahadeva B K[3], Shanavaz H[4]
[1]*Department of Computer sciences, SET Jain University Bangalore -562112*
[3,4]*Department of Basic Sciences, SET Jain University Bangalore -562112*

Abstract: ***This paper proposes an alternative methodology to the colorimetric estimation of Iron. The importance of Iron is crucial for both Plants and human health, which makes colorimetric estimating iron having many applications in the real world. The alternative methodology replaces the dependence on expensive, non-portable, electrical colorimeter involving manual pre-calibration with the overhead of storing samples to shift to using of data-based approach with Linear regression model which makes the procedure more convenient and approachable to estimating Elements concentration based on the color complex generated on reaction with corresponding reagents for assisting researchers, laboratory technicians, and farmers.***
Keywords: Iron, Ferric, DNA, hemoglobin, anemia, cytochrome, mitochondria, colorimeter, colorimetry, FAS, KCNS, Nitric acid, Multiple Linear regression, Feature extraction, feature selection, Heatmap, correlation, scatter matrix, Linear regression algorithm

I. IMPORTANCE OF IRON IN NATURE

[1]The Iron is an essential micronutrient for almost all living organisms because it plays a critical role in metabolic processes such as DNA synthesis, respiration, and photosynthesis. Further, many metabolic pathways are activated by iron, and it is a prosthetic group constituent of many enzymes. An imbalance between the solubility of iron in soil and the demand for iron by the plant are the primary causes of iron chlorosis. Although abundant in most well-aerated soils, the biological activity of iron is low because it primarily forms highly insoluble ferric compounds at neutral pH levels. Iron plays a significant role in various physiological and biochemical pathways in biological systems. In plants, iron is involved in the synthesis of chlorophyll, and it is essential for the maintenance of chloroplast structure and function. It serves as a component of many vital enzymes such as cytochromes of the electron transport chain, and it is thus required for a wide range of biological functions. In Humans Iron is an essential element for almost all living organisms as it participates in a wide variety of metabolic processes, including oxygen transport, deoxyribonucleic acid (DNA) synthesis, and electron transport. However, as iron can form free radicals, its concentration in body tissues must be tightly regulated because, in excessive amounts, it can lead to tissue damage. Disorders of iron metabolism are among the most common diseases of humans and encompass a broad spectrum of diseases with diverse clinical manifestations, ranging from anemia to iron overload, and possibly to neurodegenerative diseases.

Heavy metals are environmental pollutants, and their toxicity is a problem of increasing significance for ecological, nutritional, evolutionary, and environmental reasons. Only a few Heavy metals (Fe, Cu, and Zn) are known to be essential for plants and animals (Wintz et al., 2002).

A. Effects of Iron on Plant Growth and Health

[1]Iron is the third most limiting nutrient for plant growth and metabolism, primarily due to the low solubility of the oxidized ferric form in aerobic environments (Zuo and Zhang, 2011; Samaranayke et al., 2012). Iron deficiency is a common nutritional disorder in many crop plants, resulting in poor yields and reduced nutritional quality. In plants, iron is involved in chlorophyll synthesis, and it is essential for the maintenance of chloroplast structure and function. In aerobic soils, iron is predominantly found in the Fe +3 form, mainly as a constituent of oxyhydroxide polymers with extremely low solubility. In most cases, this form does not sufficiently meet plant needs. The visual symptoms of inadequate iron nutrition in higher plants are interveinal chlorosis of young leaves and stunted root growth. When the insoluble ferric (Fe 3+) form is reduced, it is converted to a ferrous form in the soil and is then absorbed by plants. a critical component of proteins and enzymes, iron plays a significant role in basic biological processes such as photosynthesis, chlorophyll synthesis, respiration, nitrogen fixation, uptake mechanisms (Kim and Rees, 1992), and DNA synthesis, through the action of the ribonucleotide reductase (Reichard, 1993). It is also an active cofactor of many enzymes that are necessary for plant hormone syntheses, such as ethylene, lipoxygenase, 1-aminocyclopropane acid-1-carboxylic oxidase (Siedow, 1991), or abscisic acid.

Iron is a major component of plant redox systems that are used in cytochrome essentially acts as an electron carrier in the respiratory chain and Mitochondria contain a large number of metalloproteins that require iron to carry out their function.
Iron deficiency-induced chlorosis is a major problem in plants, and it affects both yield and crop quality resulting in growth retardation; reduction in leaf size; deepening of green leaf color (particularly in the youngest leaves); reddening or purpling of stems and older leaves; wilting of shoots; yellowing and dieback of oldest leaves (especially from the tips or margins); brown or black speckles or larger necrotic patches on leaves; blackening of leaf tips and stem bases; stiffening of stems; root stunting (particularly of adventitious roots); lack of root branching; root flaccidity; root blackening (particularly of the apices); and formation of precipitates on roots (Snowden and Wheeler, 1993). On the other hand, Iron toxicity can promote the formation of reactive oxygen-based radicals, which can damage vital cellular constituents (e.g., membranes) by lipid peroxidation. Bronzing (coalesced tissue necrosis), acidity, and/or blackening of the roots are symptoms of plants exposed to above-optimal iron levels. Typically, approximately 80% of iron is found in photosynthetic cells where it is essential for the biosynthesis of cytochromes and other heme molecules, including chlorophyll, the electron transport system, and the construction of Fe-S clusters (Briat et al., 2007; Hansch and Mendel, 2009) Also Too much or too little light can quickly stress a plant, which makes them more prone to disease, pests, and premature death. However, finding optimal lighting for your plant can take some trial and error experiments. Crops may either need direct bright sunlight or indirect low light based on plants' physiological requirements. Thus it makes it essential to determine the amount of Iron to avoid deficiency or toxicity of iron Fe3+ for optimal health of the plant.

B. Effects of Iron on Human Health

[2]In the human body, iron mainly exists in complex forms bound to protein (hemoprotein) as heme compounds (hemoglobin or myoglobin), heme enzymes, or nonheme compounds (flavin-iron enzymes, transferring, and ferritin).The body requires iron for the synthesis of its oxygen transport proteins, in particular hemoglobin and myoglobin, and for the formation of heme enzymes and other iron-containing enzymes involved in electron transfer and oxidation-reductions. Almost two-thirds of the body iron is found in the hemoglobin present in circulating erythrocytes, 25% is contained in a readily mobilizable iron store, and the remaining 15% is bound to myoglobin in muscle tissue and a variety of enzymes involved in the oxidative metabolism and many other cell functions.
The physical state of iron entering the duodenum greatly influences its absorption. At physiological pH, ferrous iron (Fe^{+2}) is rapidly oxidized to the insoluble ferric (Fe^{+3}) form. Gastric acid lowers the pH in the proximal duodenum reducing Fe^{+3} in the intestinal lumen by ferric reductases, thus allowing the subsequent transport of Fe^{+2} across the apical membrane of enterocytes. This enhances the solubility and uptake of ferric iron. Iron deficiency is defined as a condition in which there are no mobilizable iron stores and in which signs of a compromised supply of iron to tissues, including the erythron, are noted. Iron deficiency can exist with or without anemia. Some functional changes may occur in the absence of anemia, but the most functional deficits appear to occur with the development of anemia. Even mild and moderate forms of iron deficiency anemia can be associated with functional impairments affecting cognitive development, immunity mechanisms, and work capacity. Iron deficiency during pregnancy is associated with a variety of adverse outcomes for both mother and infant, including increased risk of sepsis, maternal mortality, perinatal mortality, and low birth weight. Iron deficiency and anemia also reduce learning ability and are associated with increased rates of morbidity. A nutritional iron deficiency arises when physiological requirements cannot be met by iron absorption from the diet. Dietary iron bioavailability is low in populations consuming monotonous plant-based diets with little meat. The plasma or serum pool of iron is the fraction of all iron in the body that circulates bound primarily to transferrin. The classical ways of estimating the level of iron in the plasma or serum include measuring the total iron content per unit volume in µg/dL;

C. Motivation to estimate Iron in the sample:

1) Determine the Iron concentration in soil and water samples assist to increase crop health for better yield and can avoid iron toxicity in the environment, which generally forms as hard water which could potentially destroy crop fields due to salinity and also can cause health issues on consumption by animals and humans alike.
2) Nutritional iron deficiency can be discovered by determining Ferric Ions in serum to check for diseases such as anemia variants which is a crippling issue especially in developing countries where children are affected the most through malnutrition.
3) Estimate the hardness of the water sample by inferring from the Ferric ion concentration in the water sample. Also, it helps to administer precision Ferric amount supplements to patients to curate to biological needs.
4) Iron being one of the most necessary minerals for plant health extensively documented in research, could be key to precision agriculture could be adopted by custom created fertilizers to cater to specific plant and the environment in consideration for maximum profits for farmers.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 8 Issue II Feb 2020- Available at www.ijraset.com

II. COLORIMETRIC ESTIMATION OF IRON

A colorimeter is a device that is used in Colorimetry. It refers to a device that helps specific solutions to absorb a particular wavelength of light. The colorimeter is usually used to measure the concentration of a known solute in a given solution with the help of the Beer-Lambert law.

A. Principle of Colorimeter

It is a photometric technique which states that when a beam of incident light of intensity Io passes through a solution, the following occur:

A part of it is reflected which is denoted as Ir

A part of it is absorbed which is denoted as Ia

Rest of the light is transmitted and is denoted as It

Therefore, Io = Ir + Ia + It

To determine Ia the measurement of Io and It is sufficient, therefore, Ir is eliminated. The amount of light reflected is kept constant to measure Io and It.

1) *Beer's Law:* When a beam of monochromatic light is passed through a solution of a substance, the intensity of absorption increases exponentially as the concentration of the absorbing substance increases arithmetically.
   Equation: Log10 Io/It = asc where, as is absorbency index and c is the concentration of the solution.
2) *Lambert's Law:* When a beam of monochromatic light is passed through a medium, the intensity of absorption increases exponentially as the thickness of the absorbing material increases arithmetically

   Equation: A = log10 Io/It = asb Where A is called absorbance, as is the standard absorbance and b is the length/thickness of the solution, and transmittance is the reciprocal of absorbance A.

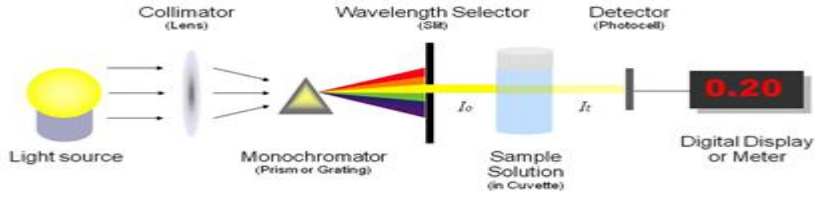


Fig. 1 Basic working of colorimeter

B. Working of Colorimeter

1) *Step 1:* Before starting the experiment it is important to calibrate the colorimeter. It is done by using the standard solutions of the known solute concentration that has to be determined. Fill the standard solutions in the cuvettes and place it in the cuvette holder of the colorimeter. adjust the knobs to set zero value for the blank solution with the right wavelength filter based on the blank solution.
2) *Step 2:* A light ray of a certain wavelength, which is specific for the assay is in the direction of the test solution. The light passes through a series of different lenses and filters. The colored light navigates with the help of lenses, and the filter helps to split a beam of light into different wavelengths allowing only the required wavelength to pass through it and reach the cuvette of the standard test solution.
3) *Step 3:* When the beam of light reaches cuvette, it is transmitted, reflected, and absorbed by the solution. The transmitted ray falls on the photodetector system where it measures the intensity of transmitted light. It converts it into the electrical signals and sends it to the galvanometer.
4) *Step 4:* The electrical signals measured by the galvanometer are displayed in the digital form.
5) *Step 5:* Formula to determine substance concentration in the test solution.

The relation between absorbance A, concentration c (expressed in mol dm-3) and path length t (expressed in cm) are given by Beer-Lambert's law. Beer-Lambert's law states that when a beam of monochromatic light is passed through a medium, the amount of light absorbed is directly proportional to the concentration of the solution and the path length(thickness) of the path of radiation through the transparent solution sample.

Equation: A = ∈cl, where ∈ is the molar extinction coefficient, c is the concentration, t is the path length and is constant for a given substance at a given wavelength. If t, the length is kept constant, then A=C. Hence a plot of absorbance against concentration gives a straight line.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 8 Issue II Feb 2020- Available at www.ijraset.com

C. Pros and Cons of Using a Colorimeter to Estimating Iron

1) Pros

a) It is a robust laboratory method for estimating the concentration of metal in the solution that generates distinct color based on its specific reagent reactions.

b) Chemical analysis through measurement of absorption of light radiation in the visible region of the spectrum (400-760 nm) concerning a known concentration of the substance is known as Colorimetry, Colorimeter correctly represents colorimetry and beer Lambert's Law.

c) Colorimeter gives accurate results on the amount of light absorbed using a photocell, can give accurate results even at low concentration of the solution upon comparison with titrimetry or gravimetry. Useful for estimation of metals like copper, iron, potassium, etc.

2) Cons

a) Adjustments should be made to the colorimeter on the blank sample to set zero using the two knobs first, and setting the right wavelength requires expertise on the chemical composition of the sample.

b) The device requires a power source and is not portable for on ground level sample collection testing for End users such as researchers, agricultural engineers.

c) It is difficult to collect and store sample in the cuvette, samples are generally discarded once values are recorded from the colorimeter

d) Traditional colorimeter devices are quite expensive for end-users especially farmers ranging from min price Rs 5200/Piece. Advance portable colorimeters are at an even higher min price range.

Fig. 2 Colorimeter calibration on balnk solution with correct filter

As we can see colorimetric estimation of Iron is a great methodology to determine Iron concentration in solution but due to its various Cons being a sizeable electric apparatus requiring prior adjustments makes it not a feasible solution to ground level research as samples collection and storage for in-depth analytics and research is difficult, there is an enormous data lost with colorimeter approach. To have a solution more reliable and feasible for end-users for research We will try to infer other analytical approaches that mimic colorimeter and much easier to use to avoid prior tuning, carrying apparatus to site or worrying about the sample storage.

D. Alternatives to Colorimetric Estimating Iron

1) Theory of Estimate Iron using FAS Solution: The method involves the reaction of Fe(III) with thiocyanate in an acidic medium to give intense red-colored complex which depends on the volume of Ferric in solution. Fe(II) doesn't give this reaction with thiocyanate. At high concentration react with the reagent thiocyanate KCNS (potassium thiocyanate), generally, red-colored $[Fe(SCN)_6]^{3-}$ complex is formed which increases the intensity and stability of the color. The acidic medium suppress the hydrolysis of Fe^{3+} ions

$Fe^{3+} + 3H_2O \rightarrow Fe(OH)_3 + 3H^+$

A series of the standard solution of Fe(III) is treated with potassium thiocyanate in a nitric acid medium to get an intense red-colored complex of ferric thiocyanate and is diluted to definite volume. Based on Beer Lambert's Law The absorbance of each of the arithmetically increasing Fe(III) volume concentration and test solution is measured at 480nm against reagent blank. The wavelength of 480nm is selected as a proper filter so that the metal ion absorbs maximum light at this wavelength.

When a calibration curve is plotted, the graph is drawn between optical density (absorbance) against concentration will be a straight line for that solution that obeys Beer-Lambert's Law. Using a standard solution of different Concentrations, a calibration curve for the suitable solute in solution indrawn.

By Experimentation for colorimeter Estimation of Iron that is Fe(III) in various Ferric Ammonium sulphate solution (stock solution) reacts with reagent Potassium thiocyanate KCNS in acidic medium of Nitric acid HNO_3 generates a complex of red-colored $[Fe(SCN)_6]^{3-}$ which obeys the beer lambert law and generates a linear line in calibration curve plot of Optical Density (absorbance) against Concentration of solution. Thus we can infer we can model this problem with a linear regression algorithm that could be used to mimic colorimeter values to obtain Optical density against the concentration of Sample.

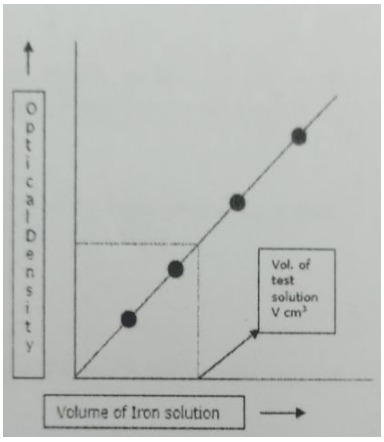


Fig. 3 Calibration curve for the colorimetric estimation of Iron

CALCULATIONS:

1000 cm^3 of stock solution contains 'a' g of ammonium ferric ammonium sulphate = ...g

Molecular mass ferric ammonium sulphate → 2 atoms of Fe

964.36g of ammonium ferric sulphate → 111.7g of Fe

"a"g of ammonium ferric sulphate = $\frac{111.7 \times 8.64}{964.36}$ g of Fe

'a' g of Ferric alum= (55.84/482.19) x a = ...g of Fe in 1000cm^3.

1cm3 of Ferric alum = ...mg of Fe (say b)

'c' cm^3 of the test solution = b x V mg of Fe= .....mg of Fe.

Fig. 4 Calculation of weight of Fe(mg) in solution

III.INTRODUCTION TO LINEAR REGRESSION

Regression is a form of predictive modeling technique that investigates the relationship between a dependent and an independent variable. Regression Modeling involves graphing a line over a set of data points that most closely fit the overall set of data. It depicts the changes in the dependent variable on the x-axis to the Independent variable on the y-axis.

Linear Regression is part of Regression which is a statistical model that represents the relationship between two variables with a linear equation. It is widely used for problems involving determining the strength of the predictors, dealing with continuous variables or finding a correlation between two variables. Linear regression models are well-documented algorithm as it offers low computational complexity and is highly comprehensive and transparent compared to other ML algorithms.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 8 Issue II Feb 2020- Available at www.ijraset.com

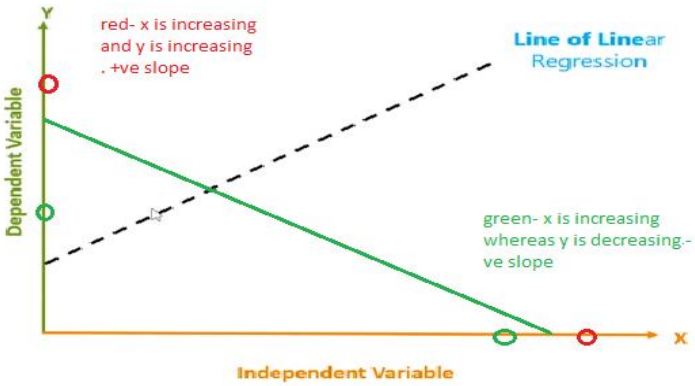


Fig. 5 Line of Linear Regression

In the above Fig 4, we have the independent variable(feature) on the x-axis and independent variable(output) on the y-axis. Consider the red points representing an increase in the x-axis and there is a corresponding increase on the y-axis then we would have a +ve line of linear regression (dotted line) as it has a +ve slope. Similarly consider green points that increase on the x-axis but decrease on the corresponding y-axis then we would have a –ve line of linear regression (green line) as it has –ve slope.
Mathematically the Line of Linear regression is given by
Equation: **y=mx+c**, where m is the slope and c is the y-intercept

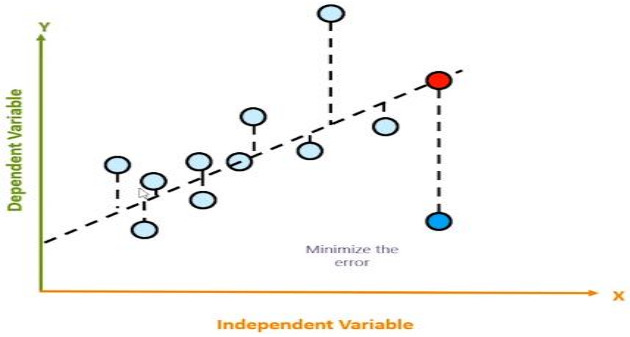


Fig. 6 Minimizing the Error

Suppose we have plotted an initial regression line for the set of data points as Fig 5, the regression line is capable of predicting the output based on the value of independent variable denoted by red point in comparison to the actual value denoted by dark blue point. The main objective is to reduce the distance between the Estimated value to the predicted value called as Error, the best fit line to the data point will have the least error or the distance between the actual value to the predicted values for all set of data points denoted by dotted vertical lines.

A. Understanding Linear Regression Algorithm with an Example

Suppose we have data points grey colored as Fig 6 left part, the algorithm first find the slope using the equation
$m = \sum(x-\bar{x})(y-\bar{y}) / (x-\bar{x})^2$, the computation is given in Fig 6 for each coordinate, thus mean is computed as m=0.4 based on above Equation. Also, the mean coordinates of (x,y) are as (3,3.6) denoted by ∑x and ∑y. which is given as purple dot in the plot, thus the line of regression has to pass through coordinate (3,3.6) Based on the algorithm.

x	y	x-x̄	y-ȳ	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4	2.8
∑ x= 3	∑ y= 3.6			∑ = 10	∑ = 4

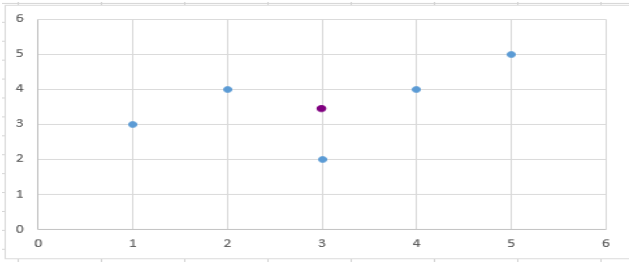


Fig. 7 Data points coordinates on the left and plotted points with mean coordinate (3,3.6) on the right

Substituting the values of coordinates of mean (3,3.6) and m=0.4in line equation y=mx+c, we get c=2.4
Thus predicted values for x={1,2,3,4,5} are as follows,
y1=0.4*1+2.4=2.8
y2=0.4*2+2.4=3.2
y3=0.4*3+2.4=3.6
y4=0.4*4+2.4=4.0
y5=0.4*5+2.4=4.4, the plotted coordinates (1,2.8),(2,3.2),(3,3.6),(4,4.6) and (5,4.4) form the regression line, Error is denoted as dotted black line distance between predicted value to the actual value, the objective of linear regression algorithm is to minimize the error to get the best-fitted regression line.

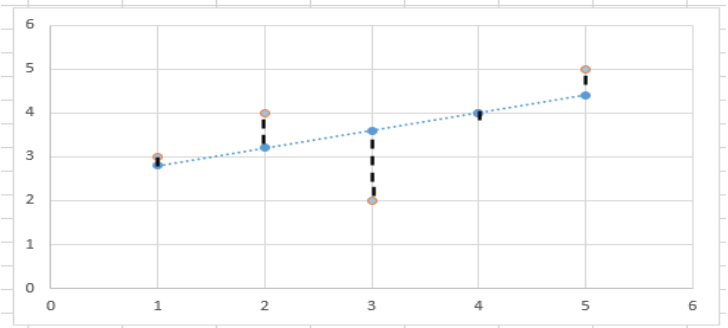


Fig. 8 Error computed for determining the best fit line.

1) *Loss Function:* The loss is the error in our predicted value of **m** and **c**. Our goal is to minimize this error to obtain the most accurate value of **m** and **c**.

We will use the Mean Squared Error function to calculate the loss. There are three steps in this function:

a) Find the difference between the actual y and predicted y value(y = mx + c), for a given x.
b) Square this difference.
c) Find the mean of the squares for every value in X.

$$E = \frac{1}{n}\sum_{i=0}^{n}(y_i - \bar{y}_i)^2$$

Mean Squared Error Equation

Here y_i is the actual value and $\bar{y}_i$ is the predicted value. Let's substitute the value of $\bar{y}_i$:

$$E = \frac{1}{n}\sum_{i=0}^{n}(y_i - (mx_i + c))^2$$

Substituting the value of $\bar{y}_i$

So we square the error and find the mean. hence the name Mean Squared Error.

2) *Measure for Goodness of Fit:* R – squared value is a statistical measure of how close the data are to the fitted regression line. Given as the difference between Distance of actual –mean vs Distance of predicted – mean. Mathematically given as

$$R^2 = \frac{\sum (y_p - \bar{y})^2}{\sum (y - \bar{y})^2}$$

Computation for R-squared value is given below for the same as the above example

x	y	$y-\bar{y}$	$(y-\bar{y})^2$	y_p	$y_p - y$	$(y_p - y)^2$
1	3	-0.6	0.36	2.8	-0.8	0.64
2	4	0.4	0.16	3.2	-0.4	0.16
3	2	-1.6	2.56	3.6	0	0
4	4	0.4	0.16	4	0.4	0.16
5	5	1.4	1.96	4	0.8	0.64
	$\bar{y}$ =3.6		∑ = 5.2			∑ =1.6

Fig. 9 Calculation for R squared value

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 8 Issue II Feb 2020- Available at www.ijraset.com

Substituting values in Equation of R square we get around 0.3 value, which signifies the regression line is not a best-fitted line, as Best fitted line should be close to 1 indicating perfectly fitted line.
Thus we need an Iterative approach to minimize the Log function to obtain high R squared value to obtain the best-fitted line this is accomplished by the gradient descent algorithm.

B. Gradient Descent Algorithm

Gradient descent is an optimization algorithm that iteratively finds the values of learnable parameters of a function (f) to minimize the cost function (or error rate).

1) *Analogy:* Imagine a valley and a person who is blindfolded and wants to get to the bottom of the valley. He goes down the slope and takes large steps when the slope is steep and small steps when the slope is less steep. He decides his next position based on his current position and stops when he gets to the bottom of the valley which was his goal.

The Algorithm explained in simplistic terms to applying gradient descent to m and c step by step:

a) Initially let m = 0 and c = 0. Let L be our learning rate. This controls how much the value of m changes with each step. L could be a small value like 0.0001 for good accuracy.

b) Calculate the partial derivative of the loss function with respect to m, and plug in the current values of x, y, m and c it to obtain the derivative value D.

$$D_m = \frac{1}{n}\sum_{i=0}^{n} 2(y_i - (mx_i + c))(-x_i)$$

$$D_m = \frac{-2}{n}\sum_{i=0}^{n} x_i(y_i - \bar{y}_i)$$

D_m is the value of the partial derivative with respect to m. Similarly, let's find the partial derivative with respect to c,

$$D_c = \frac{-2}{n}\sum_{i=0}^{n}(y_i - \bar{y}_i)$$

c) Now we update the current value of m and c using the following equation also know as hyperparameters (θ_1 and θ_2)**:**

$$m = m - L \times D_m$$

$$c = c - L \times D_c$$

d) We repeat this process until our loss function is a very small value or ideally 0 (which means 0 error or 100% accuracy). The value of m and c that we are left with now will be the optimum values.

So for the analogy m can be considered the current position of the person. D is equivalent to the steepness of the slope and L can be the speed with which he moves. Now the new value of m that we calculate using the above equation will be his next position, and L×D will be the size of the steps he will take. When the slope is steeper (D is more) he takes longer steps and when it is less steep (D is less), he takes smaller steps. Finally, he arrives at the bottom of the valley which corresponds to our loss = 0.

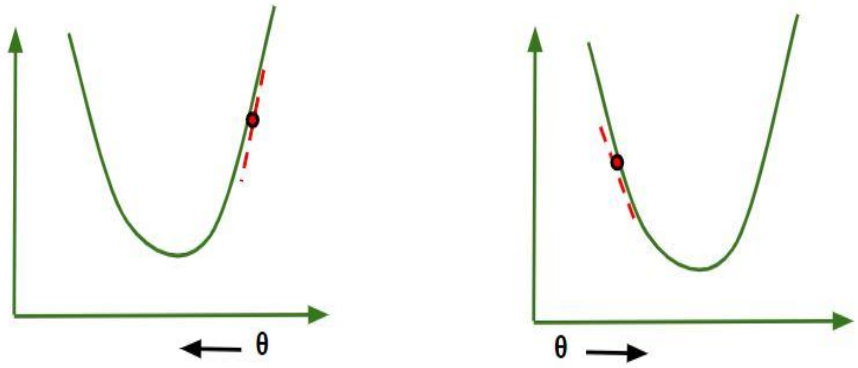


Fig. 10 Regardless of the slope is +ve or –ve, updating parameters in step 3 will ensure reducing loss function and reaching local minima

Fig. 11 Learning rate L value is optimized as if too large (left side) gradient Descent can overshoot the minimum and if L is too small Gradient Descent will take small steps to reach local minima and will take a longer time to reach minima.

IV. THE PROPOSED METHODOLOGY FOR ESTIMATING IRON

A. Procedure

Transfer the given ammonium ferric sulphate solution (stock solution) to a burette and draw out 0.5, 1.0, 1.5 till 7.5 comprising 15 variations of FAS solution concentration into series of 25cm^3 volumetric flasks. Add 4 cm^3 KCNS and 1.5 cm^3 HNO_3 (4M) to each of them and dilute up to the mark with distilled water in 25ml Standard flask, Stopper the flask and mix solution well. After 5 min, transfer a part of the solution to cuvette to measure the absorbance of the solution against the reagent blank at 480nm using colorimeter.

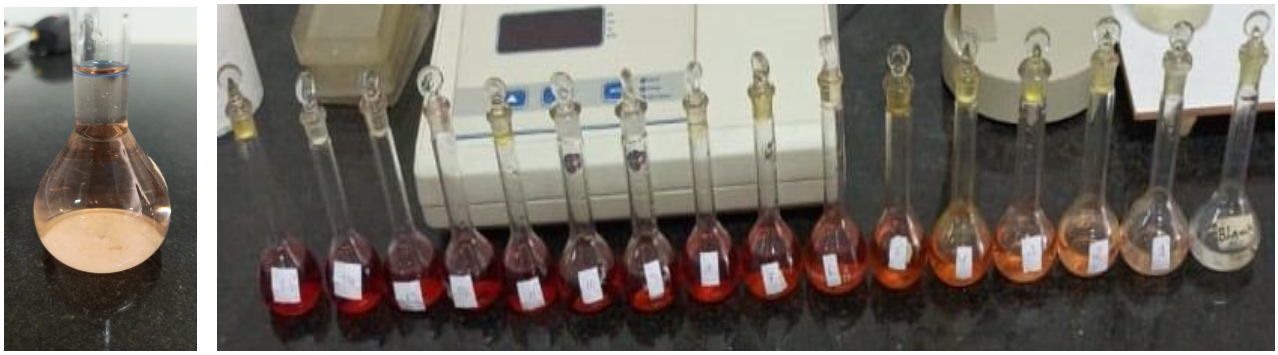


Fig. 12 Standard Flask with varied FAS concentration,4cm3 KCNS, and 1.5 HNO3 cm3, filled with distilled water up to flask mark (left side) and different concentration of Ferric solution in Standard flask

Tabulated the reading is as below

TABLE I
OD RECORDED FOR FAS SAMPLES

Sno.	Fe^{3+} ($cm^{3)}$	KCNS (cm^3) Constant	HNO_3 (cm^3) Constant	Absorbance (Optical Density)
1.	0.5	4	1.5	0.08
2.	1	4	1.5	0.13
3.	1.5	4	1.5	0.22
4.	2	4	1.5	0.31
5.	2.5	4	1.5	0.38
6.	3	4	1.5	0.47
7.	3.5	4	1.5	0.55
8.	4	4	1.5	0.60
9.	4.5	4	1.5	0.70
10.	5	4	1.5	0.75
11.	5.5	4	1.5	0.84
12.	6	4	1.5	0.92
13.	6.5	4	1.5	0.98
14.	7	4	1.5	1.03
15.	7.5	4	1.5	1.11

B. Data Collection

The variation of Fe^{3+} (FAS) solution with $4cm^3$ of KCNS as a reagent and $1.5cm^3$ of HNO_3 for acidic medium and filled with distilled water till standard Flask mark, is transferred into the dry cuvette which is used to record the Colorimeter values at 480nm filter with pre-adjusted with respect to blank solution. These Cuvette images acquired using a digital camera with at least five megapixels of resolution. It is advisable to take sample images with a background of consistent background to record red generated colored $[Fe(SCN)_6]^{3-}$ complex accurately.

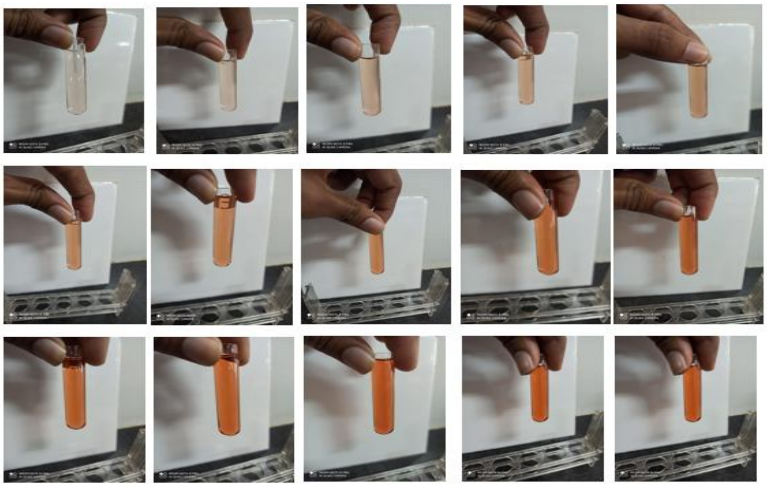


Fig. 13 Images of the solution in cuvette used in colorimeter as Table 1

The images acquired as Fig 15, saved in hard disk is latter cropped to extract only the cuvette solution patch and not the background of the cuvette for accurate Feature extraction.

Fig. 14 cropped images of a cuvette for feature Extraction

C. Feature Extraction

From the cropped Image as in Fig 15, we resize the input cropped image to the standard size of 150x150 pixels using the misc.imresize function provided the scipy package. Once the input image is resized to standard 150x150, we extract the mean pixel values for various components of multiple colorspaces. 11 Colorspaces were considered in Feature Extraction taken as Mean non zero array values rounded to 4 decimal places are RGB, HSV, XYZ, RGB cie , Gray, LAB, YUV, YCbCr, yib, ybpr and ydbdr scale (y can be excluded from yib,ybpr, and ydbdr as it has same value as Y in YUV). Since we have Image Features extracted for various colorspaces the model is far more accurate and reliable than eye observation and far more convenient and faster than traditional laboratory methods for estimating iron in the solution sample. Each Component of the above mentioned 11 colorspaces gives us 28 features and we have considered two independent variables Fe^{3+} (cm^3) sample concentration and other is the optical density value of colorimeter.computing mean values of colorspace features were repeated iteratively for all 16 variations of samples including blank solution and data values were stored in CSV format using the Pandas package in python for further assessments.

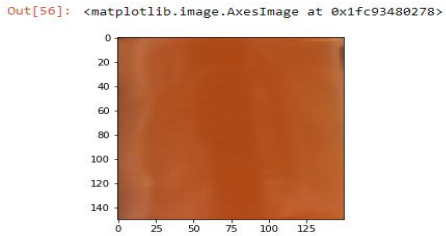


Fig. 15 Resizing the cropped Input image for feature extraction

The data saved in the XLSX file is read and processed using the Pandas package in python.

D. Data Exploration for Feature Selection

In this stage, we will determine the best features to be considered for modeling Multiple linear regression, for more comprehensibility we will model based on two use cases. The first use case considers $Fe^{3+}(cm^3)$ FAS solution concentration as independent/target variable against extracted mean color space features and the second use case considers Optical Density values recorded from colorimeter against corresponding extracted mean colorspace features. Both use cases are important for estimating Iron concentration in solution as Use case 1 gives more clarity on the $Fe^{3+}(cm^3)$ in the solution based on mean feature values extracted from cuvette image, whereas Use case 2 will elaborate on the expected colorimeter value for the sample at 480nm pre-calibrated to blank solution.

1) Use Case 1: Independent/Target variable being $Fe^{3+}(cm^3)$ FAS solution concentration

a) Correlation Matrix Compilation: From below computed correlation matrix, we find that features that have high correlation values against independent variable $Fe^{3+}(cm^3)$ concentration values are S, A, B.1 (b component of LAB), V.1 (v component from YUV), i,q, br and cr as underlined below. Other features have low correlation value and most of them have –ve values that are not suitable for linear regression modeling. We will explore it graphically in the next section.

TABLE III
CORRELATION RECORDED FOR FAS SAMPLES

Sno.	Features	Correlation value	Sno.	Features	Correlation value
1.	R	0.094091	16.	B.1	0.975631
2.	G	-0.978167	17.	Y1	-0.967900
3.	B	-0.975865	18.	U	-0.963801
4.	H	-0.711237	19.	V.1	0.986290
5.	S	0.976791	20.	i	0.981910
6.	V	0.072974	21.	q	0.958245
7.	X	-0.923823	22.	pb	-0.963724
8.	Y	-0.952868	23.	br	0.986314
9.	Z	-0.932677	24.	cb	-0.967892
10.	Rcie	-0.120653	25.	y2	-0.963750
11.	Gcie	-0.976300	26.	cr	0.986290
12.	Bcie	-0.976349	27.	db	-0.963774
13.	Grey	-0.971686	28.	dr	-0.986290
14.	L	-0.961918	29.	Fe3+(cm3)	1.000000
15.	A	0.993645			

b) Scatter Matrix: Graphically we can understand S, A, B.1 (b component of LAB), V.1 (v component from YUV), i,q, br and cr have high correlation values because scatter plots show that these features +vely correlate with the independent variable Fe3+(cm3), also the data points in the yellow highlighted plots are linear we make them best features for selection. In the case of other features, we see other patterns such as –very linear correlation, polynomial and beautiful quadratic curves. Since our focus is on Modeling Mulitple linear regression model we will not utilize these association patterns.

```
import seaborn as sns

pp = sns.pairplot(data=dataset,
                  y_vars=['Fe3+(cm3)'],
                  x_vars=features[0:7])
pp = sns.pairplot(data=dataset,
                  y_vars=['Fe3+(cm3)'],
                  x_vars=features[7:14])
pp = sns.pairplot(data=dataset,
                  y_vars=['Fe3+(cm3)'],
                  x_vars=features[14:21])
pp = sns.pairplot(data=dataset,
                  y_vars=['Fe3+(cm3)'],
                  x_vars=features[21:])
```

Fig. 16 scatter plot using Seaborn

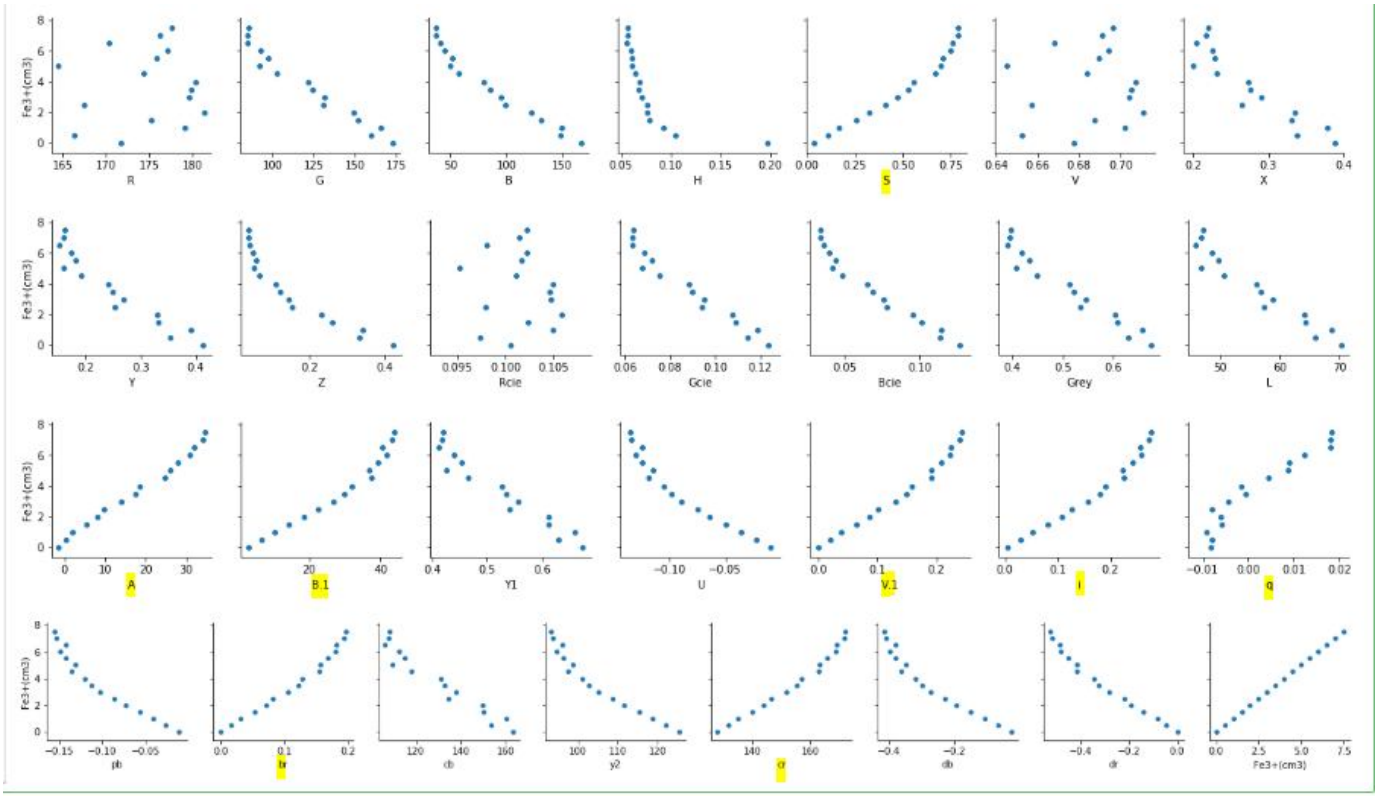


Fig. 17 Scatter plot of Feature against Fe3+(cm3) concentration, Highlighted features are selected for modeling

c) *Heat Map:* Observation recorded from the correlation matrix can be visually assessed by the Heat map provided by seaborn, as expected the selected features S, A, B.1, V.1), i,q, br and cr have the darkest shades of blue at last column in comparison to the target variable.

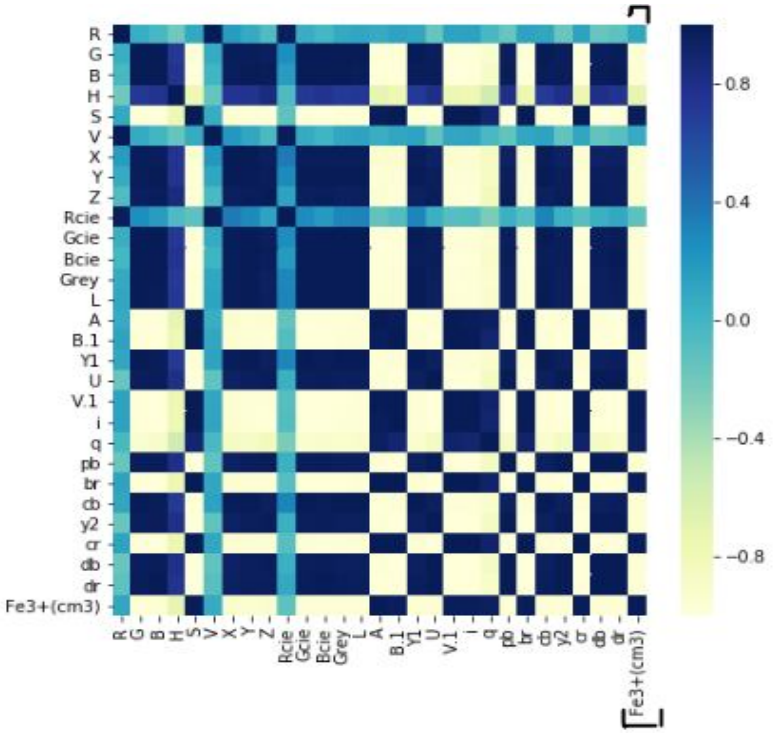


Fig. 18 HeatMap for Fe3+(cm3) target variable

2) *Use Case 2:* Independent/Target variable being Optical Density for the colorimeter

a) *Correlation Matrix Compilation:* From below computed correlation matrix, we find that the features that have high correlation values against independent variable Optical density values are S, A, B.1 (b component of LAB), V.1 (v component from YUV), i,q, br and cr are also having high correlation against recorded colorimeter value.

TABLE IIIV
CORRELATION RECORDED FOR OPTICAL DENSITY VALUES

Sno.	Features	Correlation value	Sno.	Features	Correlation value
1.	R	0.096814	16.	B.1	0.981443
2.	G	-0.981834	17.	Y1	-0.971733
3.	B	-0.981205	18.	U	-0.970659
4.	H	-0.716189	19.	V.1	0.990722
5.	S	0.982201	20.	i	0.986961
6.	V	0.075546	21.	q	0.953990
7.	X	-0.930628	22.	pb	-0.970599
8.	Y	-0.959308	23.	br	0.990738
9.	Z	-0.941541	24.	cb	-0.971726
10.	Rcie	-0.117667	25.	y2	-0.970617
11.	Gcie	-0.980145	26.	cr	0.990721
12.	Bcie	-0.981544	27.	db	-0.970637
13.	Grey	-0.975440	28.	dr	-0.990721
14.	L	-0.966463	29.	OD	1.000000
15.	A	0.996002			

b) *Scatter Matrix:* Graphically we can understand same S, A, B.1 (b component of LAB), V.1 (v component from YUV), i,q, br and cr have high correlation values because scatter plots show that these features +vely correlate with the independent variable OD values, In case of other features we see other patterns such as –vely linear correlation, polynomial and beautiful quadratic curves. Since our focus is on Modeling Mulitple linear regression model we will not utilize these association patterns.

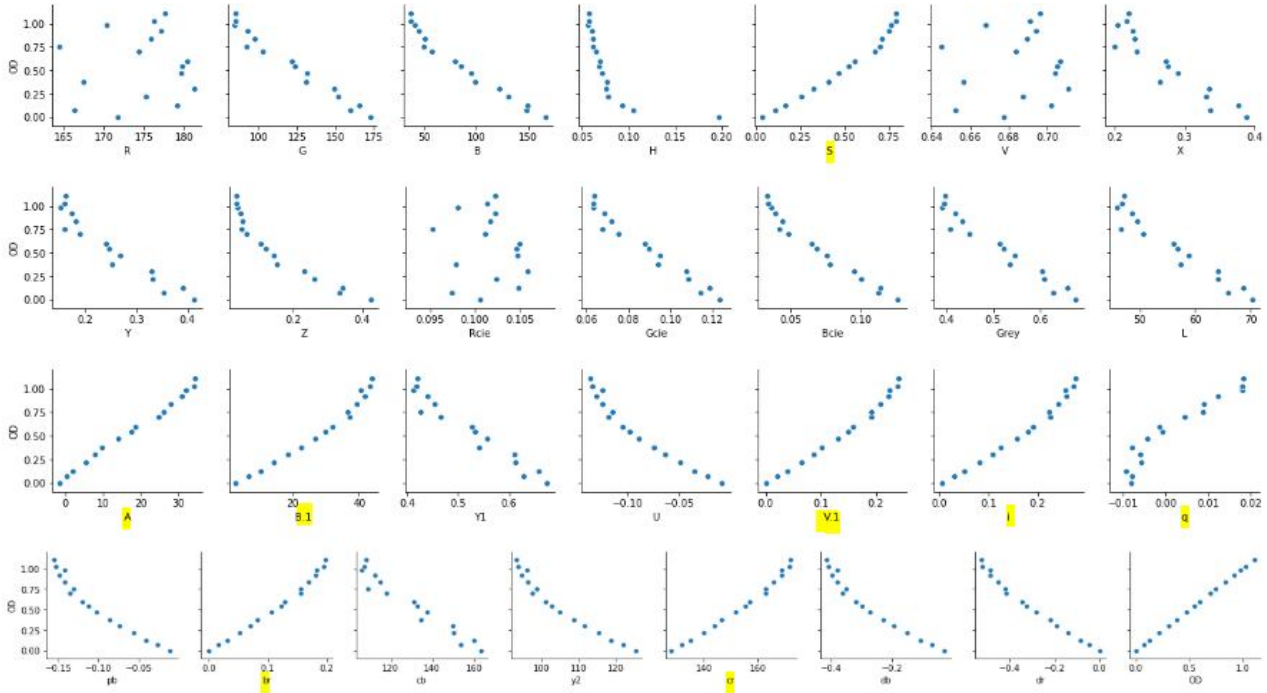


Fig. 19 Scatter plot of Feature against OD values, Highlighted features are selected for modeling

c) *Heat Map:* Heat map provided by seaborn, as expected has the selected features S, A, B.1, V.1), i,q, br and cr with the darkest shades of blue at last column in comparison to target variable OD.

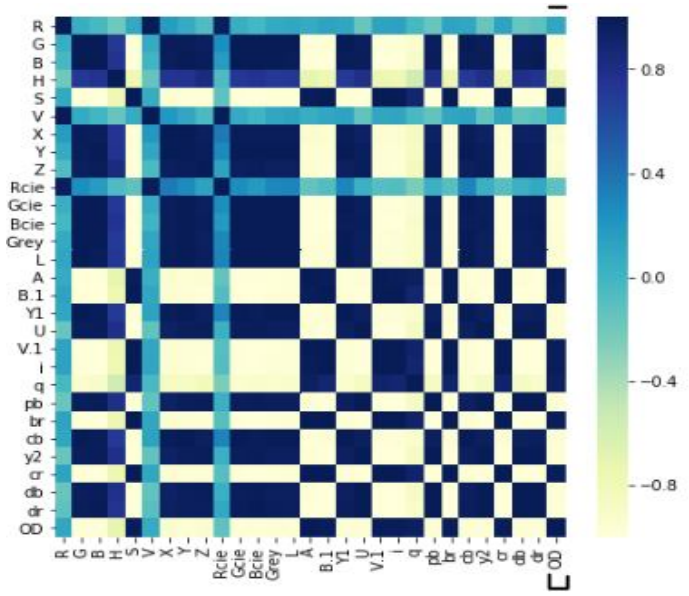


Fig. 20 HeatMap for OD target variable

E. Modeling Multiple Linear Regression Model and Evaluation:

We first filter out the data frame as X having the selected features and Y having the target feature using pandas loc function. Then Sklearn package is used for train_test_split for splitting data into train and test to 70% as train data and 30% as test data for evaluation. We then use the train set to build the Multiple Linear model using the Sklearn package, code is given below.

```
X = pd.DataFrame(dataset.loc[:,['S','A','B.1','V.1','i','q','br','cr']]) #feature selected independent varaiables
Y = pd.DataFrame(dataset.iloc[:,-2]) #independent variable Fe3+(cm3)
# In case of OD ; Y = pd.DataFrame(dataset.iloc[:,-1])

from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test = train_test_split(X,Y,test_size=0.3,random_state=5)

from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train,Y_train)

y_pred =regressor.predict(X_test)
y_pred =pd.DataFrame(y_pred,columns=['Predicted'])
```

Fig. 21 Training Multiple Linear Regression Model

1) *Evaluation Metrics For Test Set Were As Below:* We can see we have a much better metrics profile for target variable OD use case 1 compared to Use case 2 of target variable Fe3+(cm3), but both have good r2value indicating both models perform well for predicting values based on extracted features from images.

TABLE V. Evaluation Metrics

Use Case 1 (Fe3+(cm3))		Use Case 1 (OD)
Metrics	Value	Value
Mean Absolute Error:	0.4714575388902085	0.07598513538212773
Mean Squared Error:	0.4880024566666279	0.012550034967724142
Root Mean Squared Error:	0.5938096245950424	0.11202693858052241
R square value :	0.7040649753087853	0.8028460008840621

The models trained for use cases 1 and 2, can be stored in Pickle file for future use.

```
import pickle
from sklearn.externals import joblib
# save the model to disk
filename = 'finalized_model.sav'
pickle.dump(regressor, open(filename, 'wb'))
# load the model from disk
loaded_model = pickle.load(open(filename, 'rb')
```

Fig. 22 Save Model using Pickle for Future Usage

V. APPLICATIONS OF METHODOLOGY

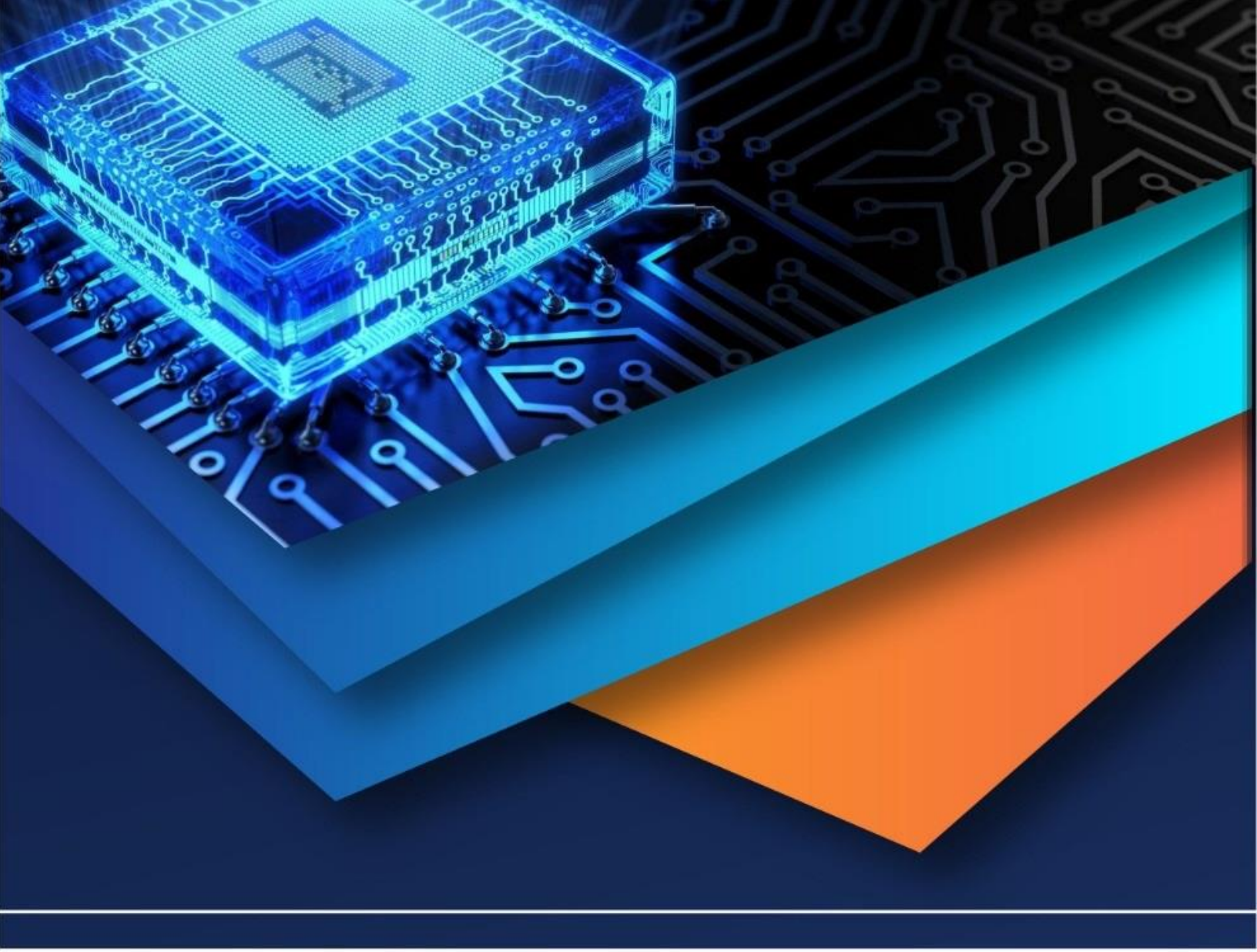
The proposed system is capable of mimicking colorimeter reading for determining any chemical element which has a suitable reagent that generated unique color, the Model trained on the gathered data from images and corresponding colorimeter value can replace the use of colorimeter providing faster, robust and efficient way to get corresponding stimulated colorimeter value without the worry of pre adjustments, storage of developed colored samples to preprocessing and much more approachable for researchers, Laboratory Technician, scientist and farmers in case of simpler regents to determine the amount of element in Sample. This methodology is a boost to determine NPK elements which is the most crucial for crop yield and plant health. The same approach can be used for paper [5] making researching insights and analytics easier and faster without reliance on colorimeter, gather large historical data and Feasible alternative

VI. ACKNOWLEDGMENT

we would like to thank Ms. Archana S Puthran Faculty at the Department of Basic Sciences Jain University Bangalore for assisting in facilitating this experiment and guiding us throughout for performing experiments and data collection. Your guidance has helped us immensely to understand the chemistry concepts behind this research.

REFERENCES

[1] Rout and Sahoo, “Role of iron in plant growth and metabolism”, Reviews in Agricultural Science, 3:1-24, 2015. DOI: 10.7831/ras.3.1
[2] Nazanin Abbaspour, Richard Hurrell, Roya Kelishadi “Review on iron and its importance for human health” .J Res Med Sci. 2014 Feb; 19(2): 164–174. PMCID: PMC3999603
[3] Adarsh Menon, "Linear Regression using Gradient Descent” https://towardsdatascience.com/linear-regression-using-gradient-descent-97a6c8700931
[4] Azam sayeed, "Regression Models” https://medium.com/@azamsayeed123/supervised-learning-4e082468f282
[5] Robert Tatina “A Colorimeter for Measuring Phosphorous in Solution", The American Biology Teacher 63(Mar 2001):190-193 DOI: 10.1662/0002-7685(2001)063[0190:ACFMPI]2.0.CO;2



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY