Vocational Course Recommender System using Machine Learning

Srinivas Adapa
Dept. of Computer Science & Engineering
Coastal Institute of Technology & Management
(Affiliated to JNTUK)
Vizianagaram, India
srinivas2804@gmail.com

Netaji Gandi
Dept. of Information Technology
Vignan Institute of Technology for Womens
(Affiliated to JNTUK)
Visakhapatnam, India
netaji.gandi@gmail.com

Alekya Vechalapu
Dept. of Computer Science & Engineering
Coastal Institute of Technology & Management
(Affiliated to JNTUK)
Vizianagaram, India
alekya257@gmail.com

Hema Pilla
Dept. of Computer Science & Engineering
Coastal Institute of Technology & Management
(Affiliated to JNTUK)
Vizianagaram, India
hemapilla@gmail.com

Abstract- This paper measures and assess an individual interest in vocational areas/occupational courses using Thurstone Interest Schedule and Machine Learning model. These days parents are forcing their children to do certain education to reach particular professions to which the child is having no interest and ultimately he/she is going to end up jobless because of the huge gap between their Passion and Profession. This Vocational Course Recommender System is going to resolve this problem using Machine Learning.

Keywords—Vocational, Occupational, Thurstone Interest Schedule, Machine Learning, Education, Passion, Profession, Recommender System, Model

I. INTRODUCTION

This paper aims in measuring an individual interest in vocational areas/occupational using Thurstone Interest Schedule and Machine Learning model. These days parents are forcing their children to do certain education to reach particular professions to which the child is having no interest and ultimately he/she is going to end up jobless because of the huge gap between their Passion and Profession. This model is beneficial for students who complete their 10th class - SSC/CBSE and entering into 10+2 for their decision to choose their passion stream after their upper primary schooling. The Thurstone Interest Schedule is a checklist by which a student can systematically clarify his understanding of his vocational interests.[1] This Thurstone Interest Schedule is part of psychological test. The basic function of psychological test is to measure differences between individuals or between the reactions of the same individual on different occasions.[5]

In this paper we are going to discuss on how we give inputs to the system and receive the label outputs based on Machine Learning algorithms. Under machine learning algorithms, Supervised learning algorithm is used to make

predictions based on a set of examples.[2] With supervised learning, we have an input variable that consists of labeled training data and a desired output variable. We use an algorithm to analyze the training data to learn the function that maps the input to the output. Under Supervised learning we will be using Classification technique. When the data are being used to predict a categorical variable, supervised learning is also called classification. [2]

I. THURSTONE INTEREST SCHEDULE

The schedule consists of around eighty occupations and the subject is asked to mark his/her preference (likes or dislikes). This questionnaire form available online at https://docs.google.com/forms/d/1m561PwwyAXwUUbXNWnY3jEXXBQfRk8rULDv_Ox86bpo/edit?usp=drive_web [6] consists of around 200 objective preferences for the subject to select. The accumulated scores in the profile represent interest in ten vocational fields as shown in below table [1]:

Table-1: Thurstone Interest Schedule Vocational Courses

S.No.	Vocational Code	Vocational Area	Professions
1.	PS	Physical Science	Engineering, Laboratory Technician, Systems / Software Engineer, Mathematician, Research Associate, ...etc.
2.	BS	Biological Science	Biotechnologist, Medicinal Chemist, Pharmacologist, Zoologist, Dentist, General practice doctor, Veterinary nurse, Zookeeper, Research scientist (life sciences & medical)...etc
3.	C	Capitational	Banking, Accounting, Auditor, Economic Support Specialist,...etc.
4.	B	Business	Businessman, Entrepreneurship,
5.	E	Executive	Managerial Professions
6.	P	Persuasive	Marketing Executive, Sales executive, Retail manager,...etc.
7.	L	Linguistic	Journalism, Idealism, Lexicographer, Teaching assistant, Speech and language

			therapist,...etc.
8.	H	Humanitarian	Affairs officer, Social Services, Education advisor,...etc.
9.	A	Artistic	Art editor, Architect, Animator, Advertising art director, Photographer...etc.
10.	M	Musical	Instrumentalist, Singer, Music Director, Radio DJ, ...etc.

For each vocational area, we have several job occupations available and this paper may cover as per the mind maps for each vocational area is shown below:

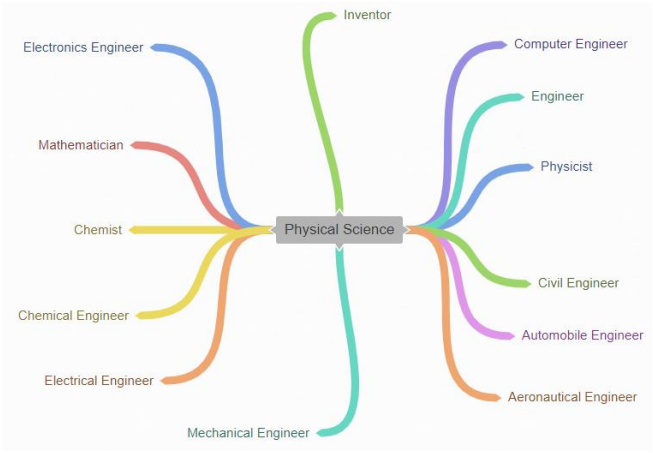

Fig-1: Physical Science Vocational Area Occupations Mind Map

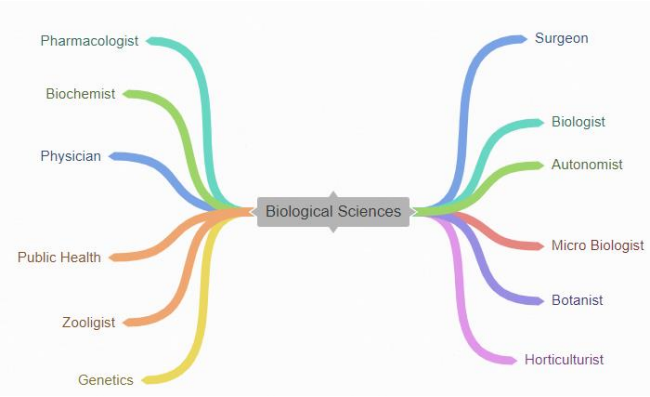

Fig-2 :Biological Science Vocational Area Occupations Mind Map

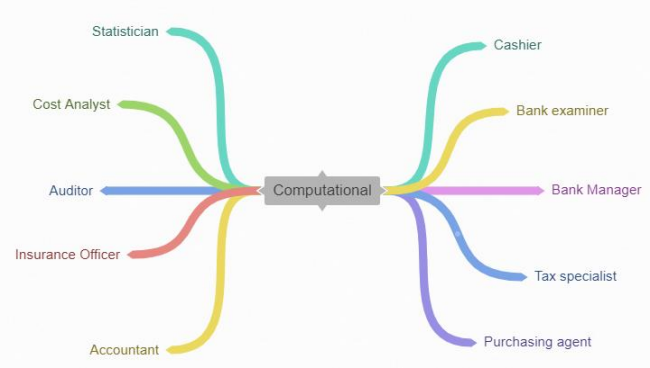

Fig-3 : Computational Vocational Area Occupations Mind Map

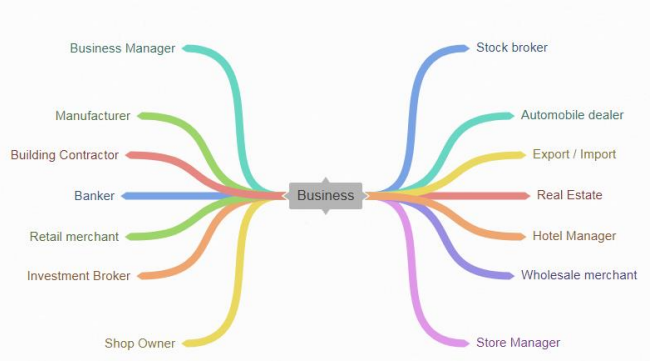

Fig-4 : Business Vocational Area Occupations Mind Map

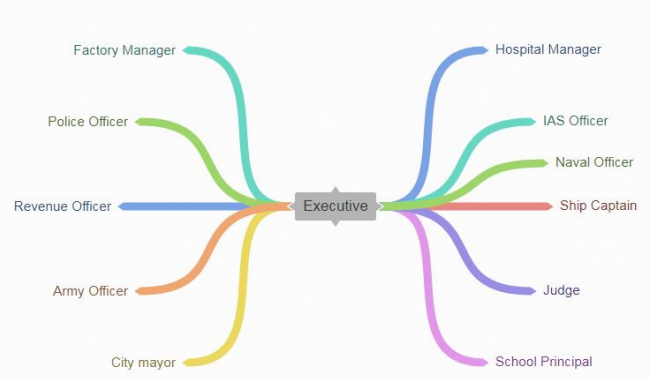

Fig-5 : Executive Vocational Area Occupations Mind Map

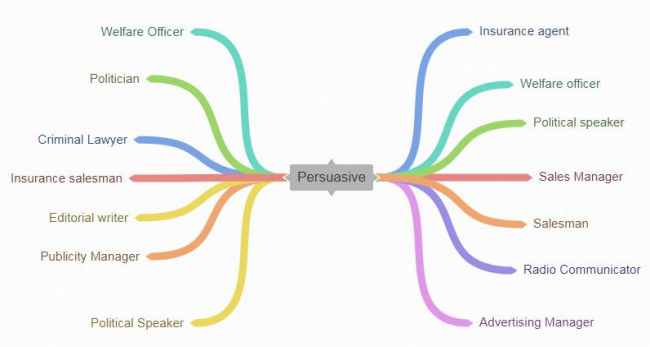

Fig-6 : Persuasive Vocational Area Occupations Mind Map

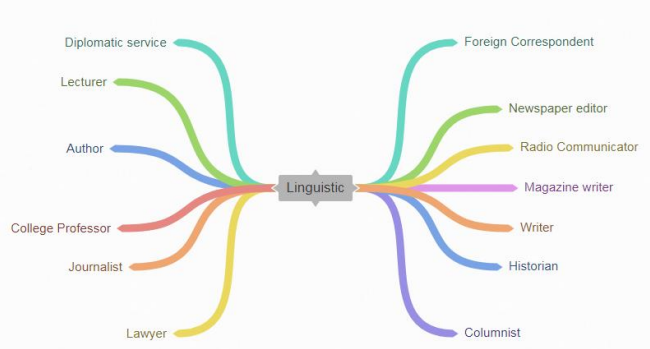

Fig-7 : Linguistic Vocational Area Occupations Mind Map

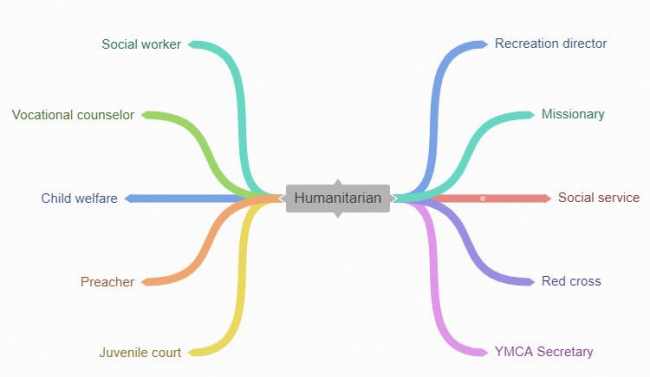

Fig-8 : Humanitarian Vocational Area Occupations Mind Map

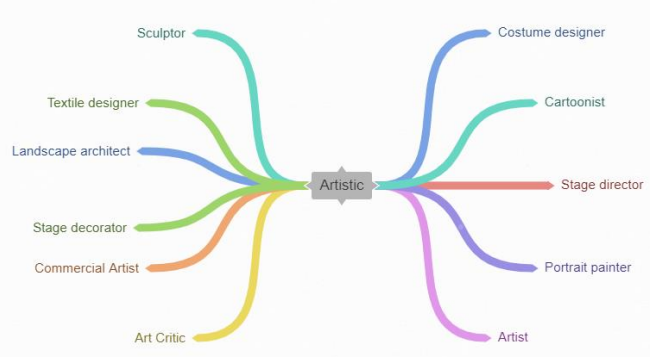

Fig-9 : Artistic Vocational Area Occupations Mind Map

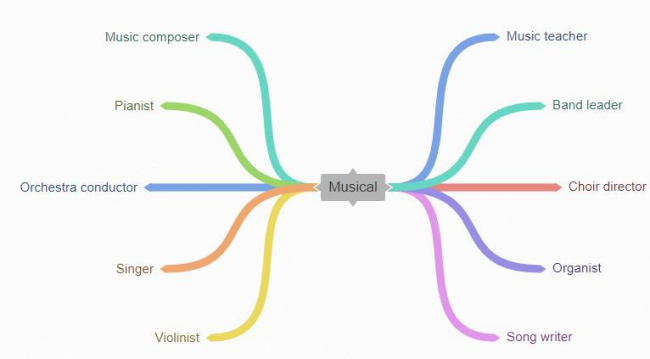

Fig-10 : Musical Vocational Area Occupations Mind Map

Based on all these mind maps shown in above figures the questionnaire is created to gather the interest of the subject in deciding the Vocational Areas which he/she is suitable.

II. CHALLENGES FACED IN ANCIENT THURSTONE INTEREST SCHEDULE

There are more than 100 occupational jobs in vocational areas and it is very difficult for the person to calculate and finalize the result manually based on subject interest. Previously it was used to calculate based on the questionnaire form offline and calculated based on the Thurstone Interest Schedule Manual.

In this paper, we take the inputs from the subject using the online questionnaire form available at https://goo.gl/forms/ucXKg8OxBJJaxDQT2 [6]. This online questionnaire form is created using google forms and the inputs are collected using excel sheet after which we process this data using machine learning to get instant accurate results compared to ancient manual methodology.

This questionnaire form consists of 200 questions out of which some questions are being repeated intentionally to know the psychological response of the subject on the same question in several sections. This questionnaire is developed based on the Psychology Test called Thurstone Interest Schedule [1] and this questionnaire also translated in regional Telugu language because the subjects who belong to rural areas may not to be able to understand english completely.

III. OVERCOMING CHALLENGES USING MACHINE LEARNING

A computer program is said to learn from **experience E** with respect to some **class of tasks T** and **performance measure P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**. [**Mitchell**]

There are several types of machine learning[9]:

- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning
- Reinforcement Learning

In this paper we will implement Supervised Learning.

WIKI Supervised Learning Definition:

Supervised learning is the Data mining task of inferring a function from **labeled training data**. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal).

A **supervised learning algorithm** analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels of **unseen instances**. This requires the learning algorithm to generate from the training data to unseen situations in a "reasonable' way. [7]

Given:

- a set of input features x1,x2,...xn
- A target feature Y
- a set of training examples where

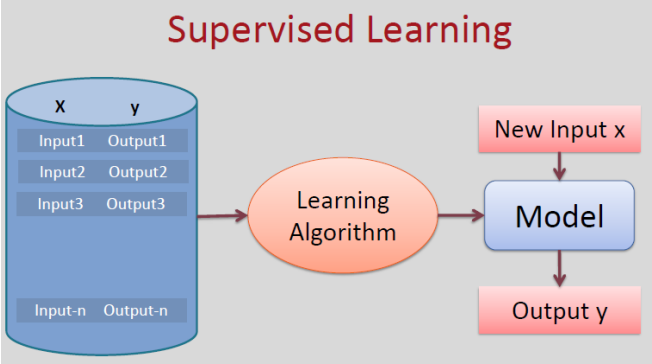

Fig-11: Block Diagram of Supervised Learning [8]

From the above Fig-11, we can train our Supervised Learning model to determine the subject's Vocational Course Interest.

- x, y (pre-classified training example)
- given an observation x, what is the best label for y?

So x will be the inputs (preference of trades) from the subject and y will be the target output (vocational course interest) i.e., the desired vocational course that particular subject is interested. Before we actually finalize the Supervised-Classification Learning model, we need to train and test with several data to build accuracy in the model.

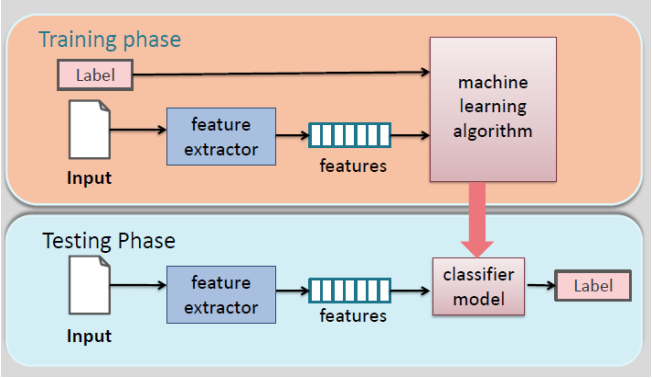

Fig-12: Training and Testing Phase in Machine Learning [8]

Often, the individual observations are analyzed into a set of quantifiable properties which are called features. May be

- categorical (e.g. "A", "B", "AB", for blood type)
- ordinal (e.g. "large", "medium" or "small")
- integer-valued (e.g. the number of words in a text)
- real-valued (e.g. height)

In this paper, our feature is categorical where we provide output/target any of the vocational areas of the subject as displayed in Table-1. We will gather around 500 samples of data out of which we use 400 samples are used for training data and 100 samples are used for test data.

IV. CLASSIFICATION LEARNING PROBLEM

Under Supervised Machine learning there are two types of Learning problems available such as:

- Classification Learning Problem
- Regression Learning Problem

If the prediction for target feature is discrete then Classification Learning will be used and if the target feature is continuous then we will apply Regression Learning.
Below Fig-13 illustrates with an example for Classification learning.

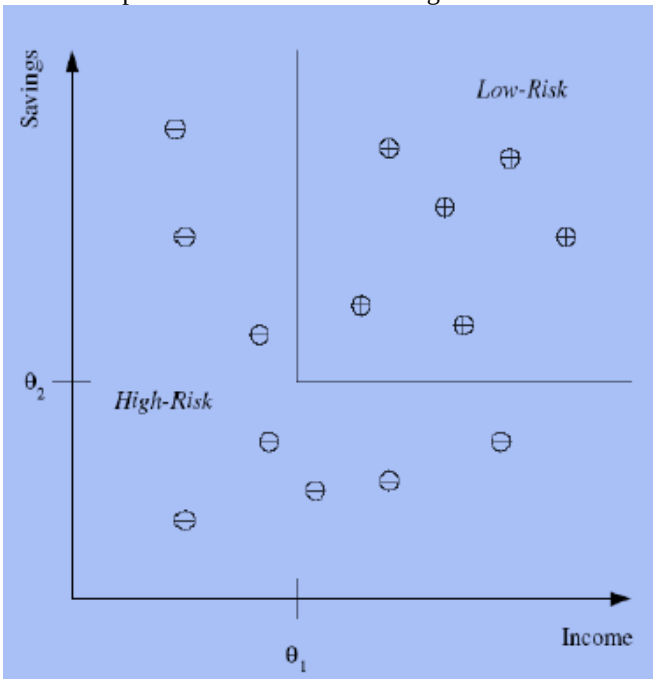

Fig-13: Classification Learning example for Credit Scoring[8]

This figure above differentiates between **low-risk** and **high-risk** customers from their income and savings.

In this paper our aim is to determine the Vocational Course recommendations to the subject based on his/her Thurstone Interest Schedule inputs through online questionnaire form [6]. So our target feature here is discrete as we are going to determine the Vocational Course Area so we use Classification Learning Problem.

V. MACHINE LEARNING TOOLS & ALGORITHMS

Supervised learning algorithms try to model relationships and dependencies between the target prediction output and the input features such that we can predict the output values for new data based on those relationships which it learned from the previous data sets [9].
List of popular Supervised Learning Algorithms[9]:

- Nearest Neighbor
- Naive Bayes
- Decision Trees
- Linear Regression
- Support Vector Machines (SVM)
- Neural Networks

In this paper we are not going to discuss on these algorithms as different models can use different algorithms based on the purpose to get different results & accuracy.

Tools are a big part of machine learning and choosing the right tool can be as important as working with the best algorithms. Machine learning tools are not just implementations of machine learning algorithms.

There are many tools & frameworks available online which includes open source as well as proprietary softwares. We can use some of the tools & frameworks to run these machine learning algorithms like [10]:

- WEKA (GUI)
- Apache Singa (API)
- KNIME (GUI)
- RapidMiner (GUI)
- Orange (GUI)
- Waffles (CUI)
- WEKA Machine Learning Workbench (CUI)
- Pylearn2 for Python (API)
- Deeplearning4j for Java (API)
- LIBSVM for C (API)

GUI - Graphical User Interface
CUI - Command Line User Interface
API - Application Programming Interface

Remote Tools:
Some tools are hosted on a server remotely and called from our local environment. These tools are often referred to as Machine Learning as a Service (MLaaS).
Examples of remote tools:

- Google Prediction API
- AWS Machine Learning
- Microsoft Azure Machine Learning

VI. TRAINING & MODEL EVALUATION

In this paper we are going to show the glimpse of the code as this paper implementation is used as our students final semester project.
Training a model is the process of iteratively improving our prediction equation by looping through the dataset multiple times, each time updating the weight and bias values in the direction indicated by the slope of the cost function (gradient).
There are two parameters (coefficients) in our cost function we can control: weight m and bias b. Since we need to consider the impact each one has on the final prediction, we use partial derivatives.

$$f(m,b) = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

We can calculate the gradient of this cost function as:

$$f'(m,b) = \begin{bmatrix} \frac{df}{dm} \\ \frac{df}{db} \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum -2x_i(y_i - (mx_i + b)) \\ \frac{1}{N}\sum -2(y_i - (mx_i + b)) \end{bmatrix}$$

Training is complete when we reach an acceptable error threshold, or when subsequent training iterations fail to reduce our cost as shown in Fig-14 below.

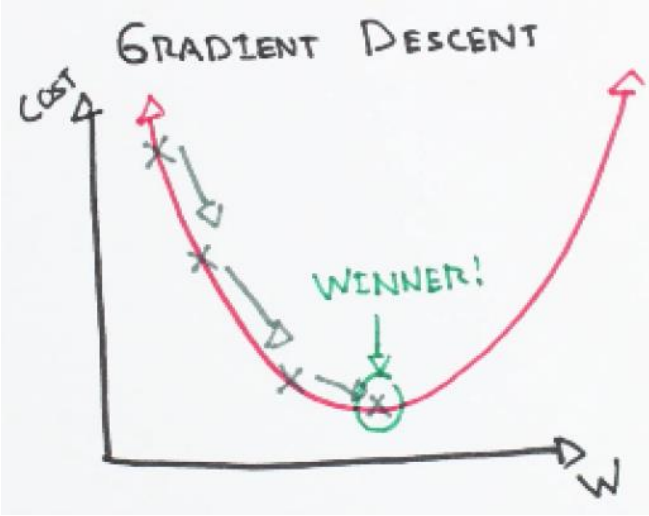

Fig-14: Gradient Descent - Decreasing Cost Function [11]

Before training we need to initialize our weights, set our hyper parameters (learning rate and number of iterations), and prepare to log our progress over each iteration.

Code [11]:

```
def train(vocational_areas, interest, weight, bias, learning_rate, iters):
cost_history = []
for i in range(iters):
weight,bias = update_weights(vocational_areas, interest, weight, bias, learning_rate)
#Calculate cost for auditing purposes
cost = cost_function(features, targets, weights)
cost_history.append(cost)
# Log Progress
if i % 10 == 0:
print "iter: "+str(i) + " cost: "+str(cost)
return weight, bias, cost_history
```

Growing complexity & Normalization [11]:

As the number of features grows, calculating gradient takes longer to compute because in our case there are around 10 features (so called vocational areas). We can speed this up by "normalizing" our input data to ensure all values are within the same range.

Our input is a 500 X 10 matrix containing all the 10 vocational areas such as Physical Sciences, Biological Sciences, Computation, Business, Executive, Persuasive, Linguistic, Humanitarian, Artistic and Musical.

Our output is a normalized matrix of the same shape with all values between -1 and 1.

Code [11]:

```
def normalize(features):
**
features - (500, 10)
features.T - (10, 500)
We transpose the input matrix, swapping
cols and rows to make vector math easier
**
for feature in features.T:
fmean = np.mean(feature)
frange = np.amax(feature) - np.amin(feature)
#Vector Subtraction
feature -= fmean
```

```
#Vector Division
feature /= frange
return features
```

After training our model through 1000 iterations with a learning rate of 0.0005, we finally arrive at a set of weights we can use to make predictions. Fig-15 below shows the Error Rate after several iterations with a learning rate of 0.0005.

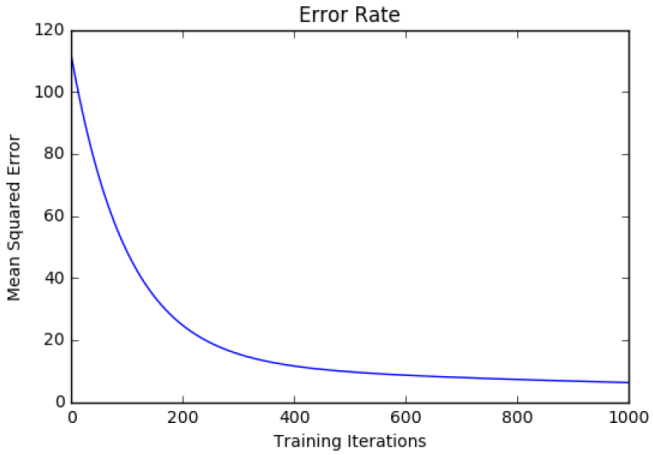

Fig-15: Training Iterations Vs Mean Squared Error [11]

CONCLUSION

As every concept have pro's and con's, same for this paper. In this paper we try to cover maximum 10 Vocational areas and 100+ occupations however in real life there are many more occupations available which also increases the complexity of the system. This paper covers all the theoretical concepts to implement Vocational Course Recommender system using Machine Learning.

ACKNOWLEDGMENT

I am thankful to my family members for supporting me in writing this paper after office hours at home. I also thank my co-authors who contributed in moral support. Last but not the least i would like to thank my student project team members (Ms.G.Rupa; Ms. DVSS Alekhya; Mr. M.N Prasad; Mr. A Vinodh Kumar and Mr. K.V.Appala Raju) without which i couldn't get primary source of information (questionnaire data filled) from the subjects to train, validate and test the Machine Learning System.

REFERENCES

[1] Thurstone, L. L., *Thurstone Interest Schedule Manual*. New York: Psychological Corporation, 1947.

[2] Hui Li, "*Which machine learning algorithm should I use?*" April 12, 2017. https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/

[3] Jared P. Lander, *R for Everyone - Advanced Analytics and Graphics* (2015)

[4] B. Yegnanarayana, *Artificial Neural Networks.* **417** (2010)

[5] *Experimental Psychology and Psychological - Assessment Practicals*, School of Distance Education, Andhra University, Visakhapatnam 2017.

[6] Thurstone Interest Schedule Questionnaire online Form, https://docs.google.com/forms/d/1m561PwwyAXwUUbXNWnY3jEXXBQfRk8rULDv_Ox86bpo/edit?usp=drive_web

[7] Eileen Mcnulty, What's the Difference between Supervised and Unsupervised Learning, January 8, 2015. Source: https://dataconomy.com/2015/01/whats-the-difference-between-supervised-and-unsupervised-learning/

[8] Sudeshna Sarkar, IIT Kharagpur, Introduction to Machine Learning, Module-1, Part-B - Introduction.

[9] David Fumo, Types of Machine Learning Algorithms You Should Know, Jun 15, 2017. Source: https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861

[10] Jason Brownlee, Machine Learning Tools, December 28, 2015.
Source: https://machinelearningmastery.com/machine-learning-tools/

[11] ML Cheatsheet Documentation, September 12, 2018, Github.