ABSTRACT

In my 2 months training period I have learnt about basics of Python 3.0 and machine learning like data cleaning, data preprocessing, how to apply various models and algorithms on training data, how to make predictions out of current and previous data and many more. I have solved various code challenges given on the daily basis which polishes the concept which I have studied so far. I have also made various projects like Classified Advertisements, Movies Reviews and many more during my 2 months training period.

CHAPTER 1

INTRODUCTION

1.1 Significance of Training

The meaning of training is "to learn a skill". Companies always look for candidates who are technically sharp, creative, with analytical bent of mind. They are in search of people who don't need training and can immediately start working. That is the reason that the budding B.Tech Engineers all say that fresher's face a number of problems when they apply for jobs after finishing their degree. The ultimate solution which came in this scenario was to give an industrial exposure to these students so that both employer and employee, both are benefitted to some extent.

- Summer Training is an effective period which helps in gaining depth technical knowledge of the field.
- Industrial exposure helps in enhancing technical skills of the person that makes clearly differentiates him from the crowd at the time of interview.
- Practical oriented learning and working with a team helps a student in knowing more about the industrial environment.
- The certificate which one gets is globally accepted.
- Most of the industries follow demonstrations and workshops of projects which helps in learning more.
- Students get a chance to interact with software developer trainers, corporate trainers and software engineer trainers. Here they can solve all their queries regarding the tasks.
- The certification provided by the reputed companies is the highlight of your resume and reflects your practical industrial knowledge too.
- For hiring fresh B.Tech candidates most of the alleged IT and software visit the "Industrial Training Organizations". Because they are not at all interested in giving training to candidates and they just want the maximum output from day one.
- The project on which students work during the training helps the employers in judging the capabilities of the candidate.

Marks are the most important eligibility criteria in most of the cases, but the technical and aptitude skills which one learns during the training also play a great role in acquiring the dream job.

Besides this, it is very clear from the above discussion, that if you miss the summer training then you are ignoring a number of factors which can lead to a successful career. So, take it serious.

You get to work in a real environment

College students that complete an internship have a great advantage over those that don't. They learn how to apply the knowledge that they have gained in the classroom into the workplace. Internships give you the chance to test-run the knowledge that you have gained while learning new skills in the process. Moreover, internships give you the chance to learn more about the industry you're interested in which is impossible to learn within a classroom.

1.2 Background of Company

Forsk Technologies is a Startup located in M5 Startup Oasis, Industrial Area, Sitapura Jaipur. Forsk is working for a paradigm shift in industry and academia partnership for skill enhancement and improved industry engagement through studio and flip classroom concept using technology and data.

Forsk Technologies work as industry partner providing end to end solution for lab sessions from content development, execution of lab assignments in the form of sub modules/projects and evaluation of candidates.

- Forsk was founded by Dr. Sylvester and Yogendra Singh in mid 2015. Sylvester and Yogendra spent over 12 years in industry.
- It basically focuses on project based learning. Project-based learning is a dynamic classroom approach in which students actively explore real-world problems and challenges and acquire a deeper knowledge.
- The mission of Forsks Technologies is to bring industry approach of software development and product engineering during lab sessions as part of academic curriculum in engineering universities and colleges.
- Forsks prepare students industry ready in the domain of Data Science, IOT, Android, Cloud, Full stack Web Development and making a mark.

It basically focuses on project based learning. Project-based learning is a dynamic classroom approach in which students actively explore real-world problems and challenges and acquire a deeper knowledge. Forsk Provides it's trainees with handsome practice on code challenges as per the topic covered on the daily basis. Forsk Technologies has following specialities:- Programming Boot Camps, IoT Training, Industry Academia Partnerships, Data Science Training, Placement Readiness Program, Android App Development Training, and Full stack Web Development Training.

Forsk Labs is connected with Industries and Startups who require emerging skills and can provide a meaningful Internship experience. These internships are mostly aimed at hiring the candidate for a long term. Students learn the real application by working on Live projects which are deployed on Global Scale. These internships also provide opportunities to work for companies in the US, Europe, and other nations.

1.3 Nature of Business

Forsk Technologies Private Limited is a Private incorporated on 29 July 2015. It is classified as Non-govt company and is registered at Registrar of Companies, Jaipur. Its authorized share capital is Rs. 100,000 and its paid up capital is Rs. 100,000. It is involved in Other computer related activities [for example maintenance of websites of other firms/ creation of multimedia presentations for other firms etc.]. Forsk is working for a paradigm shift in industry and academia partnership for skill enhancement and improved industry engagement through studio and flip classroom concept using technology and data.

1.4 Products and Services

a) Forsk Technologies is based out of Jaipur and works with universities to prepare students for skill based hiring using data and technologies.
b) Currently focused on IoT, Machine Learning, Deep Learning, Cloud, Big Data, Full stack and Android.
c) Forsk's ESTP (Engineering Specialist Training Program) helps students fill industry gap by improving skills in emerging technologies, better connect with industry and secure quality career in the industry.
d) Forsk Technologies has following specialities:- Programming BootCamps, IoT Training, Industry Academia Partnerships, Data Science Training, Placement Readiness Program,

1.5 Conclusion

Forsk Technology is a startup which is working forsk a paradigm shift in industry and academia partnership for skill enhancement and improved industry engagement through studio and flip classroom concept using technology and data. It works with universities to prepare students for skill based hiring using data and technologies. It basically focuses on project based learning. Project-based learning is a dynamic classroom approach in which students actively explore real-world problems and challenges and acquire a deeper knowledge.

CHAPTER 2

COMPANY INFRASTRUCTURE

2.1 Introduction

Forsk is working for a paradigm shift in industry and academia partnership for skill enhancement and improved industry engagement through studio and flip classroom concept using technology and data.

It basically focuses on project based learning. Project-based learning is a dynamic classroom approach in which students actively explore real-world problems and challenges and acquire a deeper knowledge.

Forsk Technologies work as industry partner providing end to end solution for lab sessions from content development, execution of lab assignments in the form of sub modules/projects and evaluation of candidates.

2.2 Network Structure

There is flexibility in the work environment of the company. The training can easily contact the mentor for any issues. Also in case of any problem or complaint anyone can easily approach the top level authorities and resolve it through their guidance and support.

2.3 Hardware and Software

The company not only deals with software projects but also hardware products dealing with Embedded & VLSI System like Advanced Wireless Communication Systems. Forsk also give training in IOT as well which is completely hardware based.

2.4 Available Policy/Plan

1. To find a client for the website.
2. To have suggestion for expert.
3. To collect more information in the field of architecture growth.
4. To constantly update changes in the website with the latest information.
5. To come up with new and brand new concepts for the website.

2.5 Conclusion

Forsk Technologies is startup which basically focuses on training of students. It gives freshers the knowledge about new technologies like Data Science, IOT, Android, Blockchain and many more. Different areas of Forsk, had originated at different times, many of which came out as a result of identification of opportunities. It is a place for perfect exposure to the IT industry for freshers.

CHAPTER 3

TRAINING ATTENDED

3.1 Introduction

Mr. Sylvester Fernandes is HR and Director of Forsks Technology and Mr. Yogendra Singh is also the Director of Forsks Technology who also used to train us on the technology Machine Learning and Python. There are 3 mentors under them who used to help us when we got stuck anywhere. My training is totally based on Data Science for which firstly I have to learn Python. Data Science is concept used to tackle big data and includes data cleaning, preparation and Analysis. Machine Learning is branch of AI based on Idea that Machines can learn from data, identify patterns and make decisions with minimal human intervention. We used to code on Spyder IDE (in Anaconda).

3.2 Introduction to Machine Learning

Not a single discussion about the technologies and the future of technology is complete without the mention of machine learning. Slowly and gradually machine learning has become probably the most important technological breakthrough of recent time. The machine learning impacts have the potential to deliver will be huge, and it's going to be a game changer for the future.

Several companies are aware of the machine learning and its importance, but not every company has the kind of required expertise to make full use of this technology. Lack of expert resources is the biggest challenge companies are facing.

Machine Learning can be used in detecting anomalies, recommending new products, enhancing customer services and a lot of things. In fact, with the amount of data that is available along with the exceptional computing power machine learning can prove to be extremely useful for future. More than 80% of companies, around the world, are already planning to used to enhance their customer experience. There is a lot of innovation happening in the field of machine learning. There is no doubt about the importance of machine learning but the biggest challenge

is the lack of expertise. There is a huge potential but the lack of resources is slowing things down a bit.

Over the years, data science has become an integral part of many industry like agriculture, marketing optimization, risk management, fraud detection, marketing analytics and public policy among others.

By using data preparation, statistics, predictive modeling and machine learning, data science tries to resolve many issues within individual sectors and the economy at large.

Data science emphasizes the use of general methods without changing its application, irrespective of the domain. This approach is different from traditional statistics tend to focus on providing solutions that are specific to particular sectors or domains.

The traditional methods depend on providing sectors with solutions that tailored to each problem rather than applying the standard solution.

Today, data science has far reaching implications in many fields, both academic and applied research domains like machine translation, speech recognition, digital economy on one hand and fields like healthcare, social science, medical informatics, on the other hand.

It effects the growth and development of brand by providing a lot of intelligence about consumers and campaigns, through techniques like data mining and data analysis.

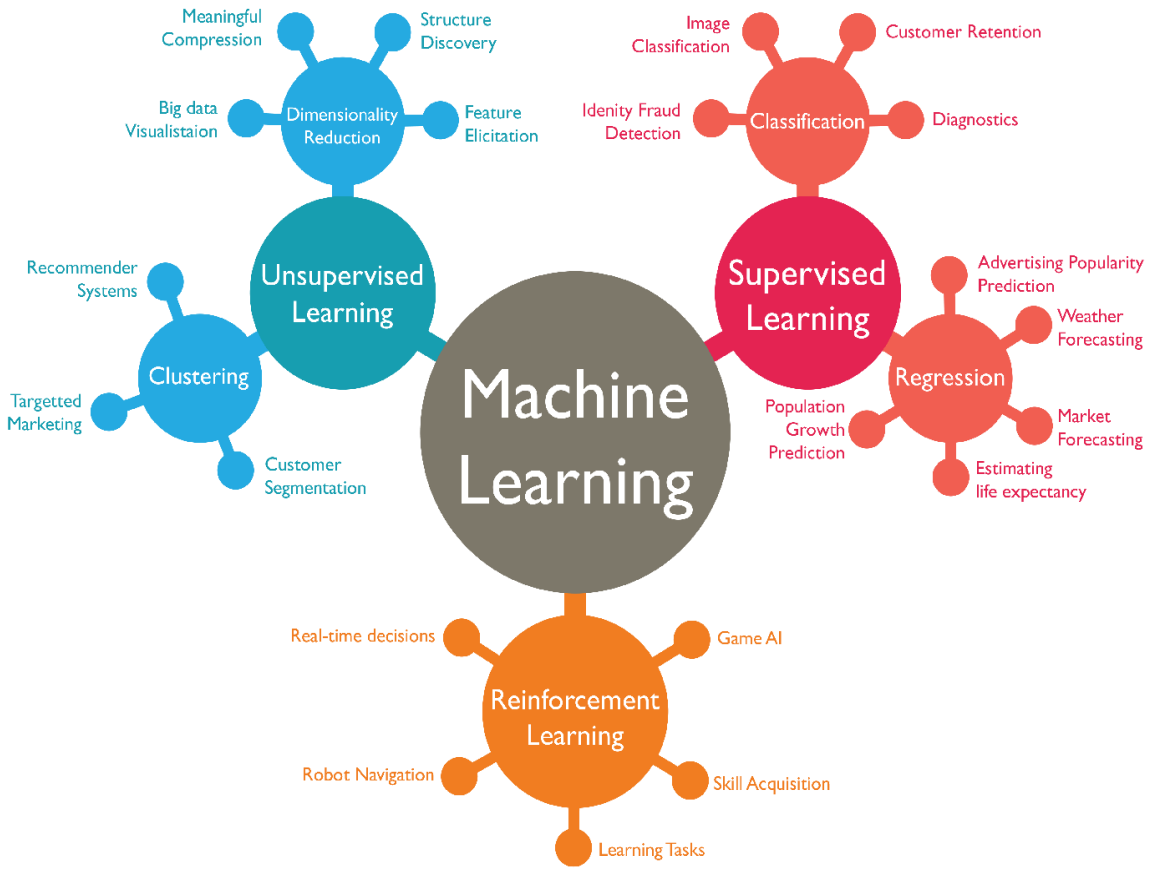
The growing importance of data science has in turn led to the growth and importance of data scientists. These data scientists professionals are now an integral parts of brands, businesses, public agencies and non-profit organizations.

These data scientists work tirelessly to make sense of large amount of data and discover relevant patterns and designs in them, so that they can be effectively utilized to realize future goals and objectives.

This means that data scientists are gaining prime importance and understanding data in a proper manner is reflected in their rising salaries as well.

According to a recent study by McKinsey Global Institute, there is a shortage of analytical and managerial talent, especially as they are need to make sense of the large amount of data available in the world.

This is one of the most pressing challenges in the current times. Further, this report estimates that by 2020, there will be a requirement of four to five million data analysts.



3.3 Description About Machine Learning topics.

Firstly trainers taught us about python in which we have covered various topics like python datatypes, functions, lists, tuples, dictionaries, Regular Expressions, Libraries like Numpy, Pandas, Matplotlib, scikitlearn. In Machine Learning I have learnt data preprocessing which includes missing values handling, data splitting, handling of categorical values and feature scaling. Then we have learnt about various algorithms and models to be applied for predictions. We have broadly learnt supervised and unsupervised Machine learning algorithm which is further classified in regression, classification and unsupervised in clustering and association. Then we have learnt all types of regression and classifications algorithms like Linear Regression, Decision Tree Regressor and Classifier and many more.

3.4 Tools Used For Python and Machine Learning

A. Jupyter Notebook

Jupyter Notebook (formerly IPython Notebooks) is a web-based interactive computational environment for creating Jupyter notebooks documents. The "notebook" term can colloquially make reference to many different entities, mainly the Jupyter web application, Jupyter Python web server, or Jupyter document format depending on context. A Jupyter Notebook document is a JSON document, following a versioned schema, and containing an ordered list of input/output cells which can contain code, text (using Markdown), mathematics, plots and rich media, usually ending with the ".ipynb" extension.

Jupyter notebooks document can be converted to a number of open standard output formats (HTML, presentation slides, LaTeX, PDF, ReStructuredText, Markdown, Python) through 'Download As' in the web interface, via the nbconvert library or 'jupyternbconvert' command line interface in a shell.

To simplify visualisation of Jupyter notebook documents on the web, the nbconvert library is provided as a service through NbViewer which can take a URL to any publicly available notebook document, convert it to HTML on the file and display it to the user.

Fig 2.2: Jupyter Notebook Home

Jupyter Notebook provides a browser-based REPL built upon a number of popular open-source libraries:

- IPython
- ØMQ
- Tornado (web server)

- jQuery
- Bootstrap (front-end framework)
- MathJax

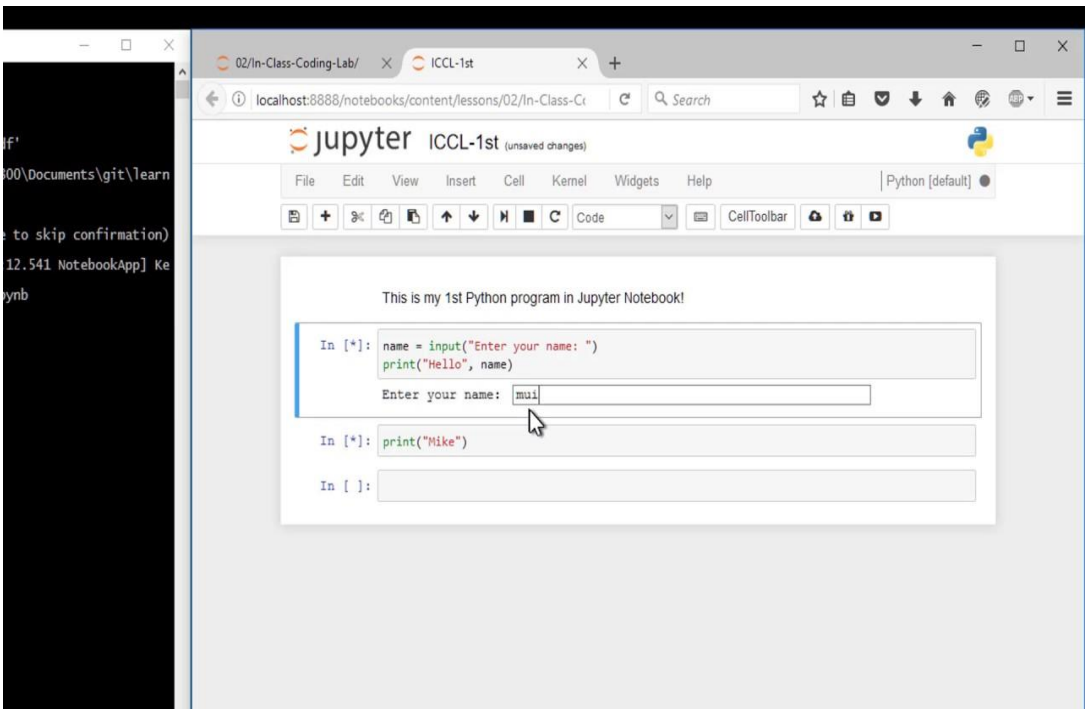


Fig 2.1: Jupyter Notebook Workspace

Jupyter Notebook can connect to many kernels, (by default Jupyter Notebook ships with the IPython kernel) to allow programming in many languages. As of the 2.3 release (October 2014), there are currently 49 Jupyter-compatible kernels for as many programming languages, including Python, R, Julia and Haskell.

The Notebook interface was added to IPython in the 0.12 release (December 2011), renamed to Jupyter notebook in 2015 (IPython 4.0 – Jupyter 1.0). Jupyter Notebook is similar to the notebook interface of other programs such as Maple, Mathematica, and SageMath, a computational interface style that originated with Mathematica in the 1980s. According to The Atlantic, Jupyter interest overtook the popularity of the Mathematica notebook interface in early 20

Jupyter kernels

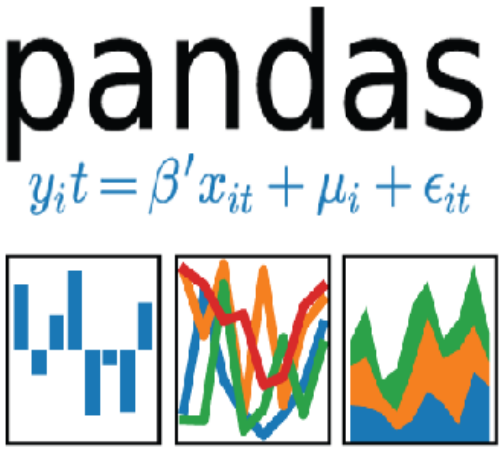
A Jupyter kernel is a program responsible for handling various types of request (code execution, code completions, inspection), and providing a reply. Kernels talks to the other components of Jupyter using ZeroMQ over the network, and thus can be on the same or remote machines. Unlike many other Notebook-like interface, in Jupyter, kernels are not aware they are attached to a specific document, and can be connected

to from many client at once. Usually Kernels are implemented and allow execution of a single language with a couple of exceptions.

By default Jupyter ships with IPython as a default kernel and reference implementation via the ipykernel wrapper

B. Pandas

Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, **real world** data analysis in Python. Additionally, it has the broader goal of becoming **the most powerful and flexible open source data analysis / manipulation tool available in any language**. It is already well on its way toward this goal.



	BandName	WavelengthMax	WavelengthMin
0	CoastalAerosol	450	430
1	Blue	510	450
2	Green	590	530
3	Red	670	640
4	NearInfrared	880	850
5	ShortWaveInfrared_1	1650	1570
6	ShortWaveInfrared_2	2290	2110
7	Cirrus	1380	1360

Fig 2.2: Pandas

Pandas is well suited for many different kinds of data:

- Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet
- Ordered and unordered (not necessarily fixed-frequency) time series data.

- Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels
- Any other form of observational / statistical data sets. The data actually need not be labeled at all to be placed into a pandas data structure

The two primary data structures of pandas, Series (1-dimensional) and DataFrame (2-dimensional), handle the vast majority of typical use cases in finance, statistics, social science, and many areas of engineering. For R users, DataFrame provides everything that R's data.frame provides and much more. pandas is built on top of NumPy and is intended to integrate well within a scientific computing environment with many other 3rd party libraries.

Here are just a few of the things that pandas does well:

- Easy handling of **missing data** (represented as NaN) in floating point as well as non-floating point data
- Size mutability: columns can be **inserted and deleted** from DataFrame and higher dimensional objects
- Automatic and explicit **data alignment**: objects can be explicitly aligned to a set of labels, or the user can simply ignore the labels and let *Series*, *DataFrame*, etc. automatically align the data for you in computations
- Powerful, flexible **group by** functionality to perform split-apply-combine operations on data sets, for both aggregating and transforming data
- Make it **easy to convert** ragged, differently-indexed data in other Python and NumPy data structures into DataFrame objects
- Intelligent label-based **slicing**, **fancy indexing**, and **subsetting** of large data sets
- Intuitive **merging** and **joining** data sets
- Flexible **reshaping** and pivoting of data sets

- **Hierarchical** labeling of axes (possible to have multiple labels per tick)
- Robust IO tools for loading data from **flat files** (CSV and delimited), Excel files, databases, and saving / loading data from the ultrafast **HDF5 format**

Time series-specific functionality: date range generation and frequency conversion, moving window statistics, moving window linear regressions, date shifting and lagging, etc.

3.5 Classification of Machine Learning Algorithms

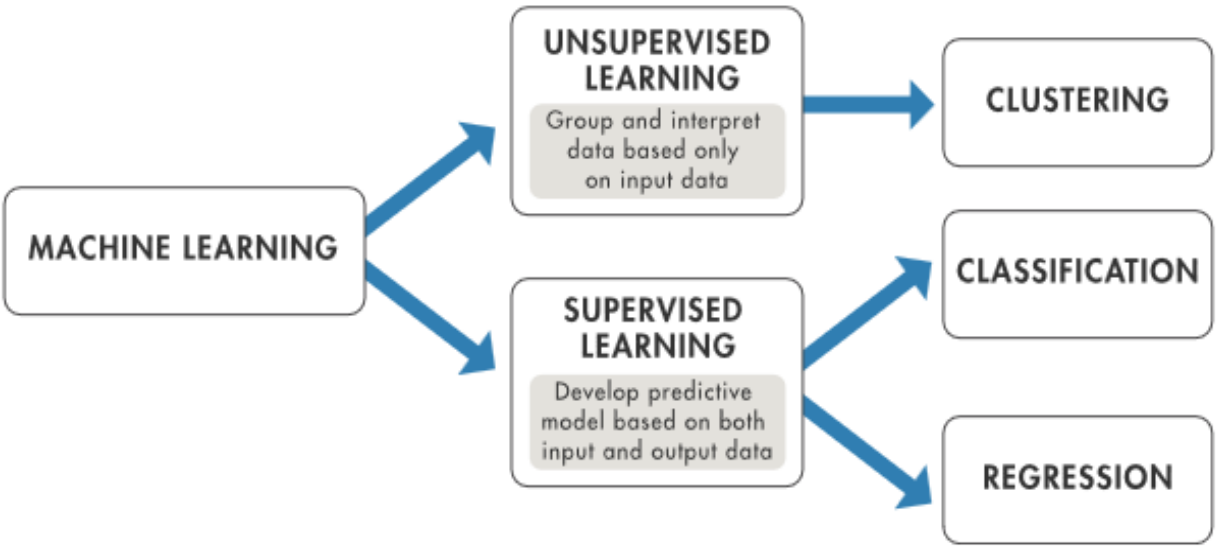
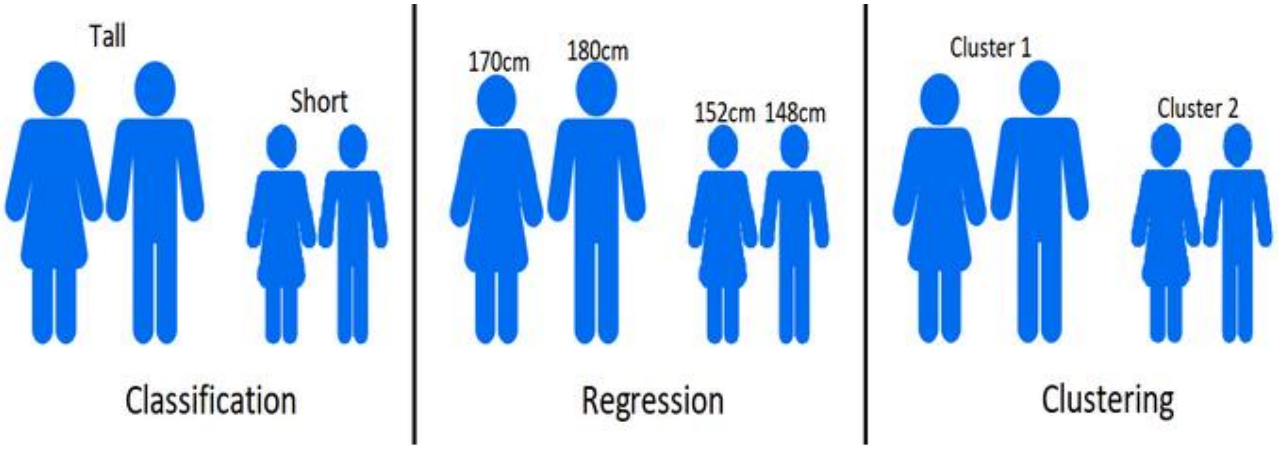


Figure 1. Machine learning techniques include both unsupervised and supervised learning.

Difference between Classification, Regression and Clustering



3.6 Conclusion

In my 2 months Summer Training period I have good exposure to new technologies like Machine Learning. It helps me to solve various real world problems and I was successfully able to make my projects out of this which is Classified Advertisements and Movies Reviews. It is quiet beneficial to get insight about how industry works what are the new updates available recently in market and many more.

CHAPTER 4

PROJECT INFORMATION

4.1 Introduction

The title of my project is "**CLASSIFIED ADVERTISEMENTS**".

Classified Advertisement is form of Small Advertisement which is common in newspapers, periodicals etc. They are cheaper ads usually column wide, grouped under particular sections.

In this we were given task to predict the category under which upcoming advertisement will be posted given features as section, cities and heading and we have to also find top 5 categories which has highest no. of posts.

This Project in Python 3.0 and Machine Learning technology.

4.2 Project Description

In this we have used "Ad_category" Dataset which is taken from "Analytics Vidhya". This is the data for local classified advertisements. It has 9 prominent **sections**: jobs, resumes, gigs, personals, housing, community, services, for-sale and discussion forums. Each of these sections is divided into subsections called *categories*. For example, the services section has the following categories under it: beauty, automotive, computer, household, etc.

For a set of sixteen different cities (such as New York, Mumbai, etc.), we have provided data from **four sections**

- for-sale
- housing
- community
- services

Also we have selected a total of 16 **categories** from the above sections.

- activities
- appliances
- artists
- automotive
- cell-phones
- childcare
- general
- household-services
- housing
- photography
- real-estate
- shared
- temporary
- therapeutic
- video-games
- wanted-housing

Each category belongs to only 1 section.

About Data:

- city (string) : The city for which this Craigslist post was made.
- section (string) : for-sale/housing/etc.
- heading (string) : The heading of the post.

Each of the fields have no more than 1000 characters. The input for the program has all the fields but *category* which you have to predict as the answer.

A total of approximately 20,000 records have been provided to you, proportionally represented across these sections, categories and cities. The format of training data is the same as input format but with an additional field "category", the category in which the post was made.

Task:

- Given the city, section and heading of an advertisement, we predict the category under which it was posted
- Also we have to show top 5 categories which has highest number of posts

Process:

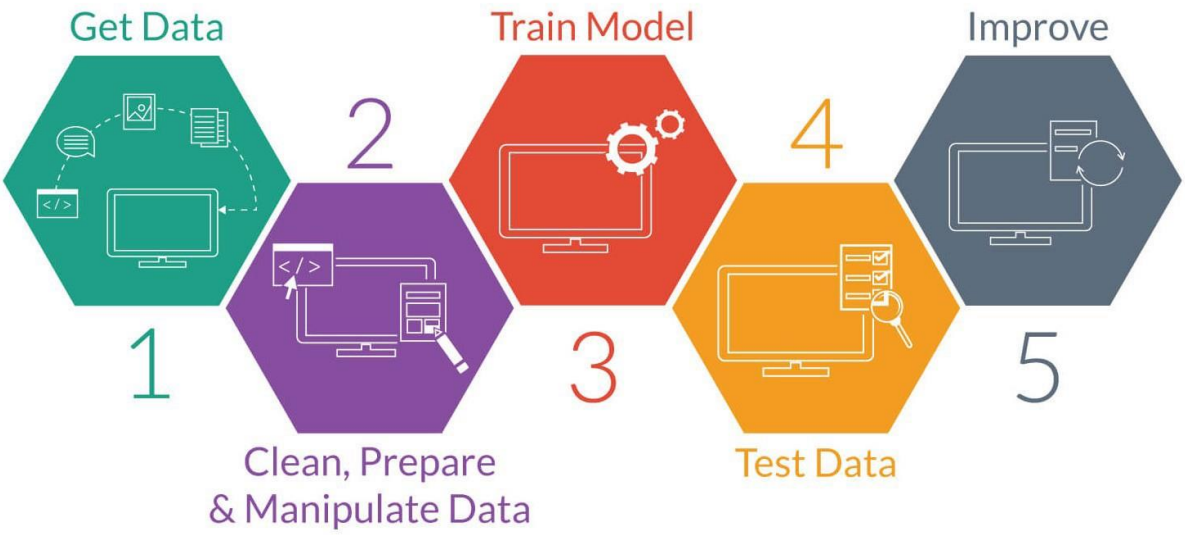


Fig 2.1: Machine Learning process

Procedure:

a) Understanding the problem

b) Gathering data

c) Preparing that data

d) Choosing a model

e) Training

f) Evaluation

g) Hyper parameter tuning

h) Prediction

1. Understanding the problem:

Our first step is to understand the problem. In this step, we deeply analyze the client requirements. Client is asked about his/her budget, tools they have and what product they want.

2. Gathering Data:

Once you know exactly what you want and the equipments are in hand, it takes you to next step of machine learning- Gathering Data. This step is very crucial as the quality and quantity of data gathered will directly determine how good the predictive model will turn out to be. The data collected is then tabulated and called as Training Data.

During gathering step, our client gave us a database of around 300 tables and then we also web scrapped and downloaded data from various web sources.

3. Data Preparation:

After the training data is gathered, you move on to the next step of machine learning: Data preparation, where the data is loaded into a suitable place and then prepared for use in machine learning training. Here, the data is first put all together and then the order is randomized as the order of data should not affect what is learned.

This is also a good enough time to do any visualizations of the data, as that will help you see if there are any relevant relationships between the different variables, how you can take their advantage and as well as show you if there are any data imbalances present. Also, the data now has to be split into two parts. The first part that is used in training our model, will be the majority of the dataset and the second will be used for the evaluation of the trained model's performance. The other forms of adjusting and manipulation like normalization, error correction, and more take place at this step.

4. Choosing a model:

The next step that follows in the workflow is choosing a model among the many that researchers and data scientists have created over the years. Make the choice of the right one that should get the job done.

5. Training:

After the before steps are completed, you then move onto what is often considered the bulk of machine learning called training where the data is used to incrementally improve the model's ability to predict.

The training process involves initializing some random values for say A and B of our model, predict the output with those values, then compare it with the model's prediction and then adjust the values so that they match the predictions that were made previously.

This process then repeats and each cycle of updating is called one training step.

6. Evaluation:

Once training is complete, you now check if it is good enough using this step. This is where that dataset you set aside earlier comes into play. Evaluation allows the testing of the model against data that has never been seen and used for training and is meant to be representative of how the model might perform when in the real world.

7. Parameter Tuning:

Once the evaluation is over, any further improvement in your training can be possible by tuning the parameters. There were a few parameters that were implicitly assumed when the training was done. Another parameter included is the learning rate that defines how far the line is shifted during each step, based on the information from the

previous training step. These values all play a role in the accuracy of the training model, and how long the training will take.

For models that are more complex, initial conditions play a significant role in the determination of the outcome of training. Differences can be seen depending on whether a model starts off training with values initialized to zeroes versus some distribution of values, which then leads to the question of which distribution is to be used. Since there are many considerations at this phase of training, it's important that you define what makes a model good. These parameters are referred to as Hyper parameters. The adjustment or tuning of these parameters depends on the dataset, model, and the training process. Once you are done with these parameters and are satisfied you can move on to the last step.

8. Prediction:

Machine learning is basically using data to answer questions. So this is the final step where you get to answer few questions. This is the point where the value of machine learning is realized. Here you can Finally use your model to predict the outcome of what you want.

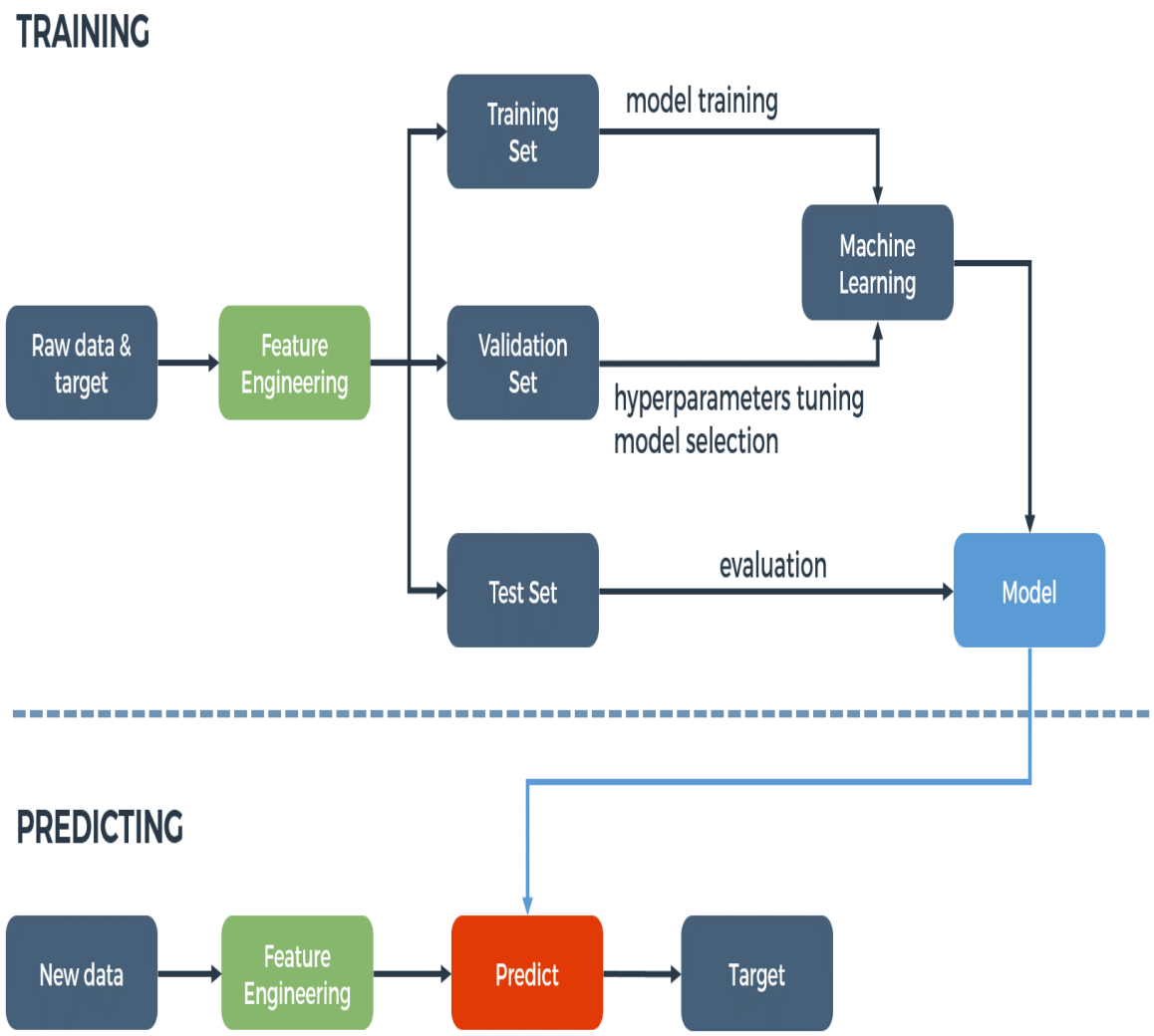


Fig 2.2: Machine Learning training & prediction

Methodology:

- **Reading .csv file** - To read .csv file first we need to Convert text file into .csv format.

- **Data Preprocessing**- Before perform the operation on the data, we have to clean the data to remove anomalies and redundancy. Also we have to handle the missing values.

- **Splitting the file** - Use sklearn.model_selection to split the data into training and testing.

- **Classification -** In machine learning and statistics classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. Classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available. An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm, that maps input data to a category Clustering - k-means clustering is a Machine Learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The kmeans algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics, and related fields.

- **Data Visualization -** Libraries ‘matplotlib’ are used to visualise the different graphs

Result Analysis:

As the result, I performed prediction on the text file which was the original data consisting of approximately 20,000 records, proportionally represented across these sections, categories and cities. On each line, based on the sections and cities we obtain the category of the advertisement.

4.3 Roles and Responsibilities

It was the project assigned to the team of 2. So I have managed the part of data cleaning and Natural Language Processing of heading column in this project.

4.4 Scope of the System

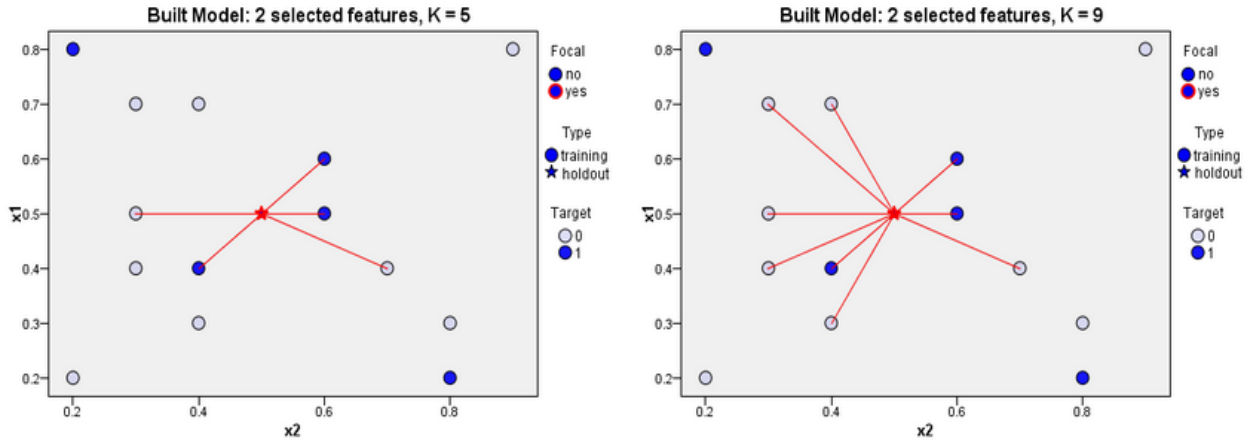
Classified Advertisement is basically more cheaper than any other business advertisement which comes on television etc.

This basically helps to segregate our advertisements on the basis of respective category of advertisement like Samsung Galaxy S8 can come under category of cell phones which is present in section of appliances. It also helps to find top 5 categories under which most of the advertisements are posted.

4.5 System Development/ Implementation

For the development of project we have to make predictions in order to predict category of new advertisement. So for that on training dataset we have applied KNN (K Nearest Neighbors) Algorithm with the value of k=7 and p=2.

KNN is basically a supervised Machine Learning Algorithm which is used to classify new query instance or new object on the basis of previous or already trained training samples and predictions are made on testing dataset for that we take the no. of neighbors(k) as given and Euclidean distance(p) between new query instance objects and training sample should be minimum.



K Nearest Neighbours

4.6 Snapshots of Project

Training dataset:-

df_train - DataFrame

Index	category	city	heading	section
383	cell-phones	newyork	IPHONE Repair charging por…	for-sale
384	cell-phones	newyork	WE REPAIR COMPUTERS	for-sale
385	cell-phones	newyork	Speaker repair for i…	for-sale
386	cell-phones	newyork	iphone 5 LCD Repair FOR O…	for-sale
387	cell-phones	newyork	power button for iPhone 4…	for-sale
388	cell-phones	newyork	Fix your iPhone 5 Scr…	for-sale
389	cell-phones	newyork	map	for-sale
390	cell-phones	newyork	iPhone Factory Unlo…	for-sale
391	cell-phones	newyork	Iphone 4 32gb Factory Unlo…	for-sale
392	cell-phones	newyork	Bold Blackberry U…	for-sale
393	cell-phones	newyork	Blackberry Bold 9000 - …	for-sale
394	cell-phones	newyork	Blackberry Storm 9530 -…	for-sale
395	appliances	newyork	Large Air conditioner	for-sale
396	appliances	newyork	Freestanding Cooling and …	for-sale
397	appliances	newyork	FAN 18" Floor unit 3 speed…	for-sale
398	appliances	newyork	Vacume Cleaner,Biss…	for-sale
399	appliances	newyork	VITA MIX 3600 BLENDER MOTO…	for-sale
400	appliances	newyork	Welbilt bread machine	for-sale
401	appliances	newyork	map	for-sale
402	appliances	newyork	__*Portable A/C , Washer…	for-sale
403	appliances	newyork	SUPER AFFORDABLE: …	for-sale

Format | Resize | Background color | Column min/max

Testing Dataset:-

df_test - DataFrame

Index	city	heading	section
0	chicago	Madden NFL 25 XBOX 360. Br…	for-sale
1	paris.en	looking for room to rent.	housing
2	newyork	two DS game	for-sale
3	seattle	map	housing
4	singapore	Good Looking Asian Sensat…	services
5	newyork	map	for-sale
6	singapore	Panasonic , 2doors fridg…	for-sale
7	dubai.en	SWEET FRIENDLY CO…	services
8	london	✯ Indulge your…	services
9	chicago	Enough	community
10	london	Brand New Xperia T	for-sale
11	london	Shaadi Single Muslims	community
12	seattle	ER Nurse wants a resp…	housing
13	seattle	490 $ ROOM for rent ava…	housing
14	london	Lovely two bedroom Los …	housing
15	seattle	PERFECT CONDITION MO…	for-sale
16	chicago	CONPRAMOS CARROS VIEJO…	services
17	singapore	For Sale - Samsung Gala…	for-sale
18	newyork	Full Time Nanny Needed…	community
19	seattle	206-375-3387 opening for …	services
20	newyork	Babysitter Looking For	community

Status of corpus after Applying NLP on heading (stopwords removal and stemming):-

corpus - List (20217 elements)

Index	Type	Size	Value
0	str	1	new batteri c s2 blackberri 7100 7130 8700 curv pearl
1	str	1	brand new origin samsung galaxi note 2 batteri
2	str	1	samsung galaxi siii 999 marbl white mobil smartphon brand new
3	str	1	ipad mini 64gb 4g sim unlock
4	str	1	htc evo 4g lte trade
5	str	1	iphon mobil unlock phone new
6	str	1	map
7	str	1	amp iphon 5 32gb black good condit scratch 550
8	str	1	iphon 5 white tmobil lifeproof case
9	str	1	phone 4
10	str	1	phone 4
11	str	1	lg spectrum 2 shell holster case
12	str	1	otterbox defend case iphon 4s
13	str	1	factori unlock iphon 5 16gb
14	str	1	samsung galaxi s3 16g
15	str	1	iphon 5 white 16 gb amp factori unlock tmobil simpl 16gb
16	str	1	unlock blackberri bold 9900
17	str	1	iphon 5 tmobil 32gb black 10 10
18	str	1	samsung phone unlock amp imei fix servic
19	str	1	samsung phone unlock amp imei fix servic

Score=89.5% after applying KNN:-

labels_train1	DataFrame	(20217, 1)	Column names: category
score	float64	1	0.895187218677350 74
strs	str	1	[{"city":"chicago","se

Prediction of category column:-

labels_pred - NumPy array

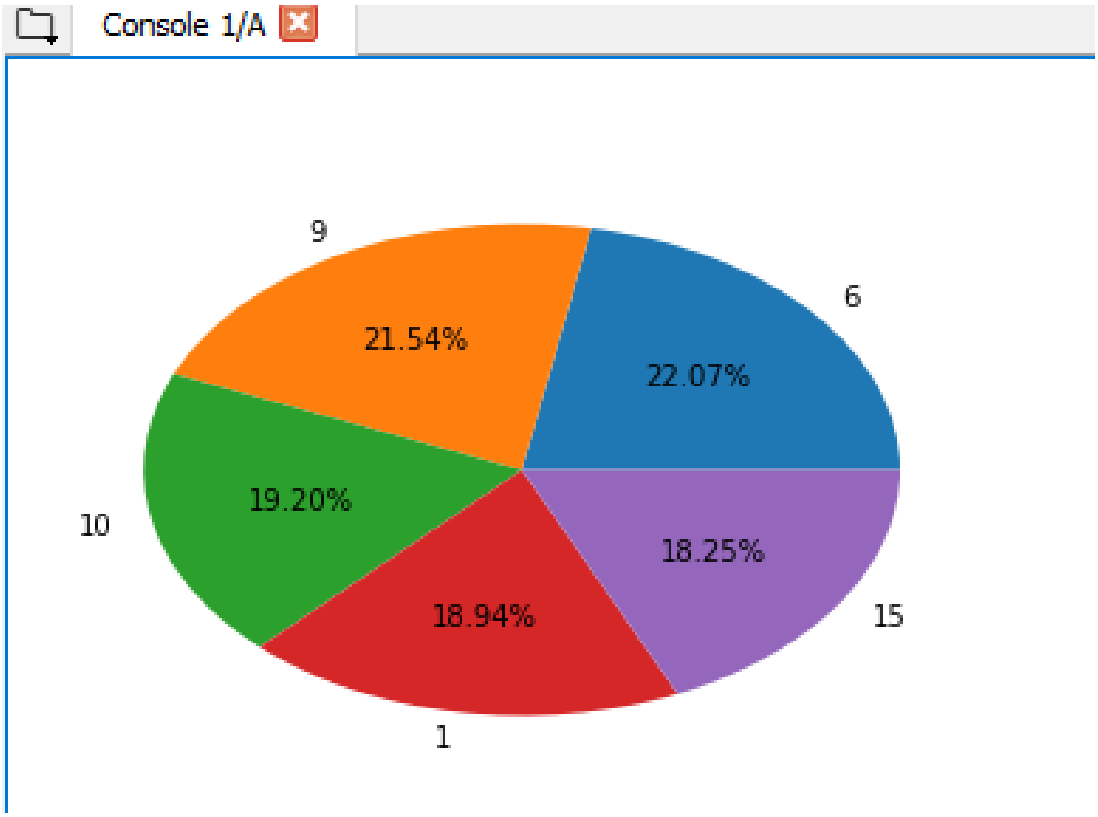
	0
0	14
1	11
2	1
3	15
4	13
5	9
6	1
7	13
8	13
9	6
10	9
11	2
12	15
13	15
14	11
15	1
16	7
17	1
18	2

Top 5 categories:-

category_most - Series

Index	0
6	2083
9	2033
10	1812
1	1788
15	1722

Pie Chart Of top 5 categories:-



4.7 Conclusion

This project is basically more cheaper than any other business advertisement which comes on television etc. It helps to segregate our advertisements on the basis of respective category of advertisement like Samsung Galaxy S8 can come under category of cell phones which is present in section of appliances.

It also helps to find top 5 categories under which most of the advertisements are posted and we have drawn the pie chart for the same.

CHAPTER 5

CONCLUSION

5.1 Introduction

Internships are a unique opportunity that one will never again encounter throughout his/her career. Summer training is an effective period which helps in gaining depth technical knowledge of the field. The summer training is very essential because during these training programs student gets opportunities to cover all those important aspects which they miss during their four years of engineering curriculum and every crucial like great technical background on their respective technologies. Summer training in a short time covers all basic requirements for great career in future. Summer training in other words is an opportunity where students can improve their technical and personality aspects for a secure future in the technical field and it is only way out towards a great career and must be done with best possible efforts.

5.2 Lessons Learned & Skills Developed

- Ability to solve real world problems, various code challenges related to real world can be solved.
- Capability to communicate effectively.
- Ability to learn a lot about new technology by the handsome amount of hands on practice on problems and code challenges.
- Ability to identify, formulate and model problems and field engineering solutions based on a system approach.
- Understanding of the importance of sustainability and cost- effectiveness in design and development of engineering solution.
- Ability to be a multi- skilled engineer with goods technical knowledge, management, leadership and entrepreneurship skills.
- Awareness of the social, culture, global and environmental responsibility as an engineer.
- Capability and enthusiasm for self-improvement through continuous professional development and life-long learning.

5.3 Knowledge Gained

The training at Forsk Technologies provided me an insight of Machine Learning which is the flourishing technology in market today and its recent trends in industry. Interaction with employees & great support from training coordinator enabled my successful completion of training. It provided industrial exposure and acted as a stepping stone for my career ahead.

5.4 Key Learning

- Exposure to new and emerging Technology which is Machine Learning.
- Learnt how to manage projects and how to complete them before deadlines.
- Team work.
- Real time Code Challenge Solving Ability.

REFERENCES

- Forsks Techologies edx platform.
- http://www.openedx.forsk.in/courses/Forsk_Labs/ST101/2018_19/info
- Forsks Technologies website:- www.forskslabs.com
- www.analyticsvidhya.com