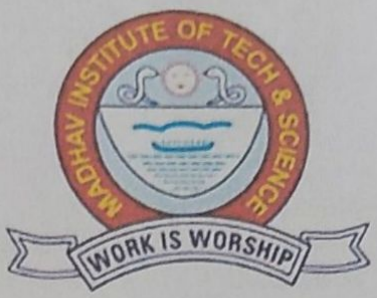
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)



Internship Report

on

Emerging Business Opportunities

A project report submitted in partial fulfilment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

Submitted by:

Abi Soni

0901IT181005

Internship - 160801

Faculty Mentor:

Dr. Tej Singh

Assistance Professor

Information Technology

Submitted to:

DEPARTMENT OF INFORMATION TECHNOLOGY
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE
GWALIOR - 474005

MAY-JUNE 2022

1

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

CERTIFICATE

This is certified that **Abi Soni** (0901IT181005) has submitted the internship report titled Emerging Business Opportunities under the mentorship of prof. Tej Singh, in partial fulfilment of the requirement for the award of degree of Bachelor of Technology in **Information Technology** from Madhav Institute of Technology and Science, Gwalior.

Dr. Tej Singh 03/06/2022
Assistant Professor
Information Technology

Dr. Akhilesh Tiwari 3/6/22
Professor and Head,
Department of IT

2

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR
(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

DECLARATION

I hereby declare that the work being presented in this project report, for the partial fulfilment of requirement for the award of the degree of Bachelor of Technology in Information Technology at Madhav Institute of Technology & Science, Gwalior is an authenticated and original record of my work under the mentorship of Dr. Tej Singh, Assistance Professor, Information Technology.

I declare that I have not submitted the matter embodied in this report for the award of any degree or diploma anywhere else.

Date: 03/06/2022
Place: Gwalior

Abi Soni
0901IT181005
IV Year,
Information Technology

3

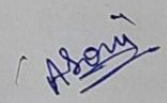
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR
(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary/interdisciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Department of Information Technology,** for allowing me to explore this project. I humbly thank **Dr. Akhilesh Tiwari**, Professor and Head, Department of Information Technology, for his continued support during the course of this engagement, which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance of Dr. Tej Singh, Assistance Professor, Information Technology, for his continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department.



Abi Soni
0901IT181005
IV Year,
Information Technology

4

ABSTRACT

Data science is a growing field in industries now days, as the large amount of data is being produced, companies have started utilizing that data for their benefits by applying data science techniques on that data.

Data Science is study of data with the help of modern tools and technologies to find insights and based on those insights make business decision. Data science uses machine learning to build predictive models.

Organization Information:

Tiger Analytics is an example of what analytics can do in order to give solution of some of the toughest problems faced by organizations around the world. They develop solutions using data science technology for several (Fortune 500) companies. They have offices in multiple cities across the United Kingdom, United States, Singapore, and India and a substantial remote global workforce.

Background:

Manufacturer A is a leading Food & Beverage manufacturer. he wants to understand the growth pattern preferences of consumers and evaluate their brands positioning across different themes. he wants to know the sales drivers of their products.

Objective:

Client has given us his sales and related data along with data of its competitors in that field. We must study and apply the data science skills to understand its growth pattern in different themes and its positioning across different themes.

To know the drivers to its sale we must build a machine learning model.

5

TABLE OF CONTENTS

6

Introduction

Data science is helpful for modern day business to get the beneficial information insights from the data which is useful in their business. This helps them know the market trends, helps in price management and find new opportunities in the market. Data is not only useful for business analysis but also in other fields. Voice assistants for the recommendation also requires knowledge science. Thus it is also needed in self-driving automobile, a computer program that gives required output for client service. These all are the real life application for data science.

Data science is used in mining of large amount of knowledge of information, from structured as well as unstructured dataset to find patterns, extract useful insights from them. Data science uses concepts like statistics, applied science, analytics, machine learning and other new technologies to achieve insights from huge knowledge.

In the next few upcoming years there will be large amount of data produced every day and the amount of knowledge that exists will grows exponentially. At any instance of time 90% of the amount of data present is generated in last 2 years in accordance with the source like IBM.

Major part of this data is being generated by the web users. By the year 2025 there will be generation of one hundred and sixty five zettabytes per annum. All it means is that there is a lot of labor in data science of which much is yet uncover.

Simple knowledge analysis will analyze data from one supply or a limited amount of data. However, Data science And its tools are useful in understanding large data and data from multiple sources in a very fluent way. Data science had frequent application in fields like healthcare, marketing, banking and finance.

7

1.1 Problem statement:

We are given the data of a Food and Beverages company. We are expected to study the data and do the necessary analysis using data science tools to extract useful business insights from that data.

1.2 Data

It will provide the following data for the project:

- Sales_Data – At UPC level for both Client and Competitors
- Social_Media_Data – Mentions of theme across all social media Platforms
- Google_Search_Data – Search volume of the Theme
- Theme Product List – Product to theme Mapping
- Product Manufacturer List – Product to Manufacturer Mapping
- Theme List – Theme Names

Manufacturer will also provide description of the dataset including all the features present. It will make available knowledgeable personnel to provide any necessary background on data and business context. These personnel will also help Tiger in working with the end consumers of our models and identifying which features could be actionable. They would also help in corroborating findings from the models and can help provide an insider's business perspective

1.3 Deliverables:

1.4.1 Data Preparation from social media, google search and sales data

Demonstrates:

- Provide the list of themes available across all data sources
- Understands consumer preference(themes) available in each data source
- Provide a report for data sufficiency, sparsity and anomalies in each data source
- Recommend the time granularity (Daily/Weekly/Monthly/Quarterly/Yearly) for the analysis

8

1.4.2 Data exploration and Hypothesis Validation

Demonstrates:

- Merge the required data sources
- Understand the overall market share of our client
- Find the potential competitors for our client in each theme
- What are the themes which are emerging in social media, Google Search & Sales?
- Validate the hypothesis: Trend flows from Social -> Search -> Sales
    - What is the latency observed?
    - Is the latency significantly different across themes?
    - Pictorially represent transition between sources

1.4.3 Build the sales model and identify the driver of sales

Demonstrates:

- Perform appropriate data transformation/aggregation
- Create a dependent variable by aggregating sales of our client to corresponding theme level
- Identify the right model technique and select the suitable variables
- Estimate the impact on sales due to social trends, search trends, own price and competitor effects
- Justify the estimated impacts are accurate
    - Model Performance
    - Hold out validation

1.4.4 Recommend levers for business growth

Demonstrates:

- Use EDA and Sales models
- What are the controllable factors which client could leverage to increase sales across themes?
- How to achieve a 5% increase in sales overall?

9

2.1 System Requirements

Python

Python is a high-level programming language. It is an object-oriented and an interpreted language which means its code is interpreted and run directly line by line unlike other languages
Where the whole code is first compiled then that compiled file is executed to get the output. Python also have high-level built-in data structure which makes it very useful in fields like data science and data analysis. Python's simple syntax is also easy to learn and read. Python support many packages and module for different functionalities and use which makes it a versatile programming language. Python is also freely distributed.

Libraries and packages:

NumPy

NumPy provides N-Dimensional arrays which are very useful. Other then that NumPy offers Mathematical functions, random number generator, linear algebra routine etc. NumPy is written using C language and is open sourced.

Pandas

Pandas is a python library. It is mainly used for data analysis and other manipulation of tabular data in DataFrames. Almost all type of file formats can be imported in pandas easily and can be converted into dataframes.

Some other modules are: matplotlib. pyplot, seaborn, sklearn, missingno.

Jupyter Notebook

It is a web-based interactive platform. It provides simple and interactive data science environment for many programming languages. It is not limited to IDE, but also a good presentation or educational tool.

Hardware requirements:

Processor - At least 2 GHz; Recommended 2.4 GHz or more.
Hard Drive - At least 100 GB; Recommended 500 GB or more.
Memory (RAM) - At least 4 GB.

10

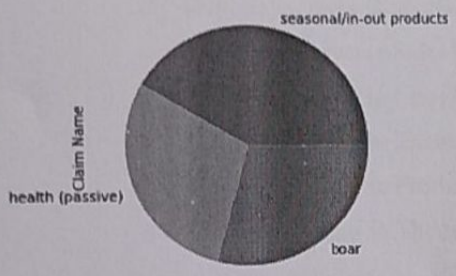
3.1 Data Preparation:

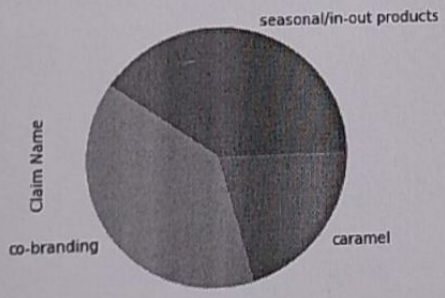
3.1.1 List of themes across Data sources:

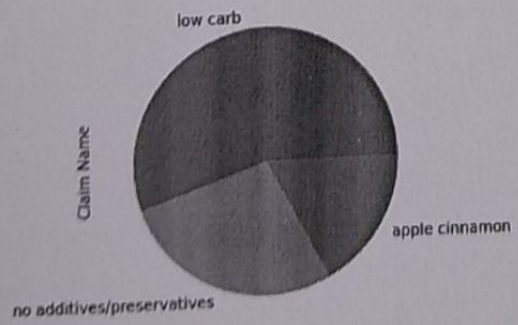
- Google Search: 160
- Social Media: 192
- Sales Data: 49
- Total Unique Themes Available Across all Data sets: 208

3.1.2 Consumer preferences across data sources:

Selecting top 3 themes across data sources based on total post, search volume and sales dollars value.







11

3.1.3 Sparsity of each data sources:

- **We define the sparsity of a Data Frame as the ratio of count of zero elements to the total number of elements.**
    - Sparsity of Social media data source 0.054
    - Sparsity of Sales Data Source 0.003
    - Sparsity of Google search Data Source 0.0
    - Sparsity of Theme product list 0.159
    - Sparsity of Product Manufacturer 0.0
    - Sparsity of Theme 0.002

3.1.4 Anomalies in each data sources:

- **Anomalies basically means duplicate data, and we need to reduce the redundancy in each data source.**
    - Duplicated rows in social media: 16487
    - Duplicated rows in Sales Data: 0
    - Duplicated rows in Google Search Data: 40
    - Duplicated rows in Theme Product list Data: 0
    - Duplicated rows in Product Manufacturer Data: 0
    - Duplicated rows in Themes Data: 0

12

4.1 Data Exploration

4.1.1 Market share of our client

- VENDOR — SALES Value
  - A — 3.093203e+10
  - B — 1.538073e+10
  - D — 1.196682e+10
  - E — 1.939263e+09
  - F — 4.995875e+09
  - G — 1.951155e+09
  - H — 2.277656e+09
  - Others — 1.854397e+10
  - Private Label — 9.753021e+09

4.1.2 Emerging Themes in each data sources

- Social Media

	Theme Name	%Growth
0	peanut butter	4166.000000
1	venison	310.000000
2	butternut squash	300.000000
3	cranberry	235.211268
4	peanut	191.056911
5	single shot	87.240356
6	pepper	72.727273
7	dill	67.132867
8	red raspberry	60.707269
9	microwaveable	50.962196

13

- Google

	Claim Name	%Growth
0	convenience - ready prepared	3000.000000
1	poultry	712.495057
2	fruit	550.000000
3	peanut butter	279.657795
4	coconut	279.098361
5	oral health	195.191298
6	natural	113.461582
7	carob	100.000000
8	seeds	98.214286
9	parsley	91.498741

- Sales

	Claim Name	%Growth
0	halal	415.925206
1	ethical - not specific	120.332640
2	poultry	119.480444
3	vegetarian	41.409127
4	beef hamburger	23.665851
5	low sugar	21.912724
6	beans	11.312246
7	herbs	10.116459
8	convenience - easy-to-prepare	-3.554983
9	high/source of protein	-4.203324
10	nuts	-4.325802

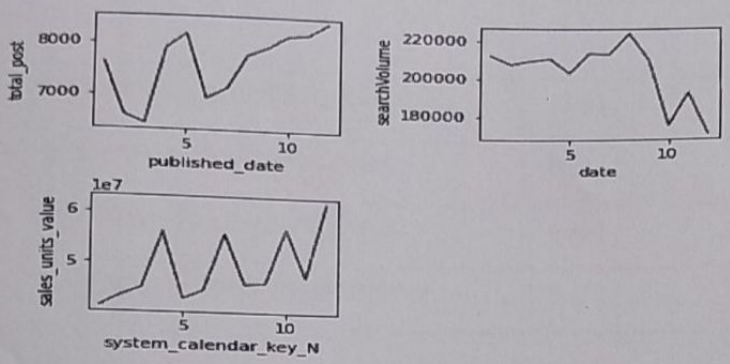
14

5.1 Hypothesis Validation

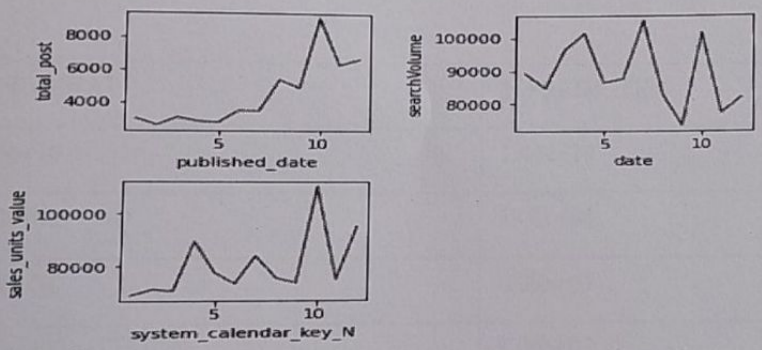
We must validate the hypothesis that trend flow from social -> google -> sales.

This hypothesis has been proved using time graph for a particular theme for a particular year on monthly basis in each data sources.

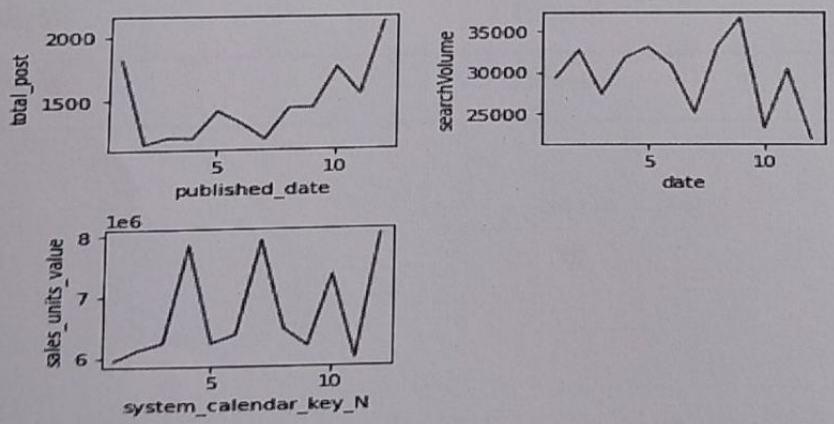
Theme 1:



Theme 2:



Theme 3:



Trend flow can be observed from above graphs and the time latency is also different for different themes.

15

6.1 sales model and Drivers of sales:

6.1.1 Model selection

Different models are built and then the correct model is selected based on R2 value.

REGRESSION MODEL	PERFORMANCE (R2 VALUE)
• MULTIPLE LINEAR REGRESSION	0.781
• POLYNOMIAL REGRESSION	0.816
• DECISION TREE REGRESSION	0.891
• RANDOM FOREST REGRESSION	0.417

6.1.2 Model Training (Decision Tree)

Hold out Validation

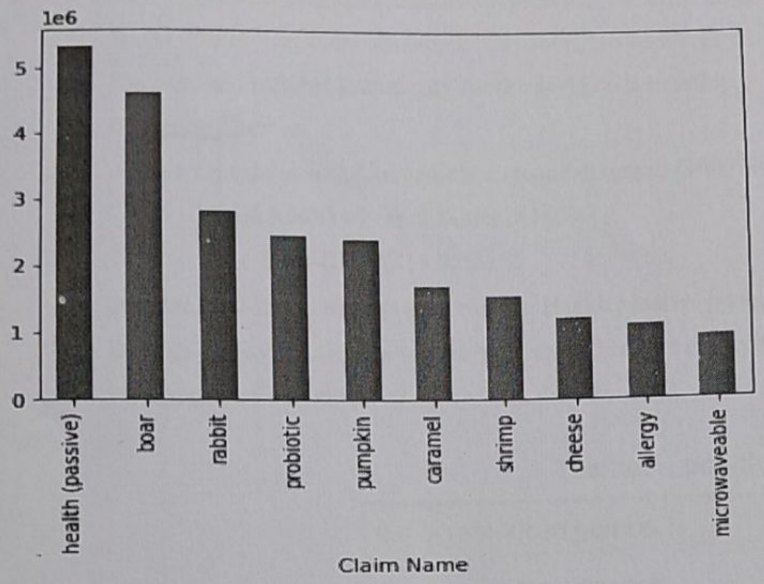
Y-ORIGINAL	Y-PREDICTED
1.86e+10	1.44e+10
7.55e+07	1.04e+08
1.28e+08	4.09e+07
5.29e+06	1.69e+07
2.40e+09	2.38e+09
2.17e+05	5.67e+05

16

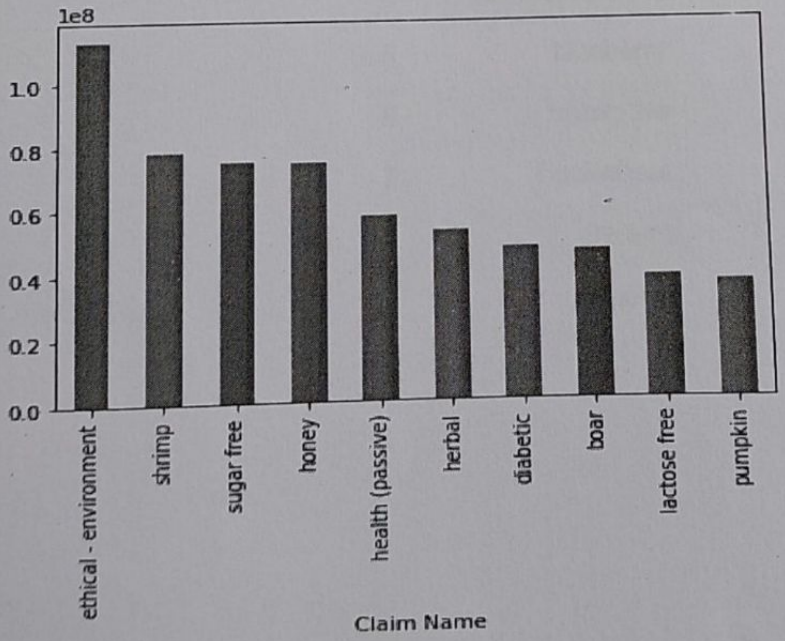
6.2.1 Drivers of sales

Themes with high business opportunities.

Social media:



Google Searches:



17

7.1 Conclusion

- The overall market share of our client A is 32%.
- The consumer preference in each data source is:
    - Google Search- seasonal/in-out products, health (passive) , boar
    - Social media- seasonal/in-out products, caramel, co-branding
    - Sales data-low carb, apple cinnamon, No additives/preservatives
- The best and suitable granularity for the analysis is monthly.
- Emerging Themes:
    - Google search-Convenience –ready prepared (3000%)
    - Social Media –Peanut Butter (4166%)
    - Sales Data-Halal (415.92%)
- Themes with high business opportunity: Health passive, pumpkin
- Drivers Of Sales: sales lbs values, total post (Features that can increase the sales)
- Positioning of our client:

	Themes	positioning
0	american gumbo	4
1	american southwest	2
2	apple cinnamon	3
3	beans	3
4	beef hamburger	2
5	blueberry	2
6	brown ale	2
7	buckwheat	3
8	chicken	2
9	cookie	2

8.1 List of References:

- Udemy
- W3 school
- Realpython.

19

TIGER ANALYTICS

Abi Soni
Jan 18, 2022

Dear Abi,

We are pleased to extend to you an offer of internship with **Tiger Analytics India Consulting Private Limited** (the **Company**).

This contract is valid only for the period of internship, and you will be required to sign a separate contract should you take up a full-time role with the Company.

Your internship is subject to the following terms and conditions:

1. Date of Commencement
   The internship is for a period of 4 Months - Jan 24, 2022 to May 31, 2022

2. Place of Work
   Your internship will be administered remotely.

3. Stipend
   You will be paid a stipend of INR 30000 (pre-tax) per month during your internship. This will be deposited into your bank account.

4. Benefits
   Benefits available to full-time employees such as Provident Fund and Medical Insurance are not applicable to Interns.

5. Leave Entitlements
   During your internship period, you are entitled to leave as approved by your manager. Leave cannot be encashed.

6. Safety
   The Company is committed to providing a safe working environment for all employees and therefore required to abide by all safety rules and procedures operating within the Company.

7. Conduct
   You will be expected to dress appropriately for a business setting. Business casual attire as outlined below is considered appropriate:
   a. A collared shirt, pants, and shoes for men
   b. Equivalent Indian or Western business casuals for women

   Behaviour unbecoming of a business environment (as perceived the Company) will not be tolerated and might lead to termination of employment.

FPR 1

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	ABI SONI		Department	Data science	
Industry/Organization	Tiger Analytics		Date/Duration	15 days	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					
Learning capacity/Knowledge up gradation					
Performance/Quality of work					
Behavior/Discipline/Team work					
Sincerity/Hard work					
Comment on nature of work done/Area/Topic	Abi soni has started with campus training Program; she is going through the basics of data science, Stats concepts. She is able to complete all the respective assignments on time.				
OVERALL GRADE (Any one)	**VERY GOOD**				
Name of Industry Mentor	Padmajothi				
Signature of Industry Mentor	Padmajothi				

Receiving Date	**15/02/22**	**Name of Faculty Mentor**	**Dr. Tej Singh**	**Sign**	**xxx**

21

FPR 2

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	ABI SONI		Department	Data science	
Industry/Organization	Tiger Analytics		Date/Duration	15 days	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					
Learning capacity/Knowledge up gradation					
Performance/Quality of work					
Behavior/Discipline/Team work					
Sincerity/Hard work					
Comment on nature of work done/Area/Topic	Abi soni is continuing with campus training Program;				
OVERALL GRADE (Any one)	**VERY GOOD**				
Name of Industry Mentor	Padmajothi				
Signature of Industry Mentor	Padmajothi				

Receiving Date	**01/03/22**	**Name of Faculty Mentor**	**Dr. Tej Singh**	**Sign**	

22

FPR 6

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	ABI SONI		Department	Data science	
Industry/Organization	Tiger Analytics		Date/Duration	15 days	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					
Learning capacity/Knowledge up gradation					
Performance/Quality of work					
Behavior/Discipline/Team work					
Sincerity/Hard work					
Comment on nature of work done/Area/Topic	Abi soni is going through the regression concepts this month.				
OVERALL GRADE (Any one)	**VERY GOOD**				
Name of Industry Mentor	Padmajothi				
Signature of Industry Mentor	Padmajothi				

Receiving Date	**30/04/22**	**Name of Faculty Mentor**	**Dr. Tej Singh**	**Sign**	**xxx**

FPR 4

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	ABI SONI		Department	Data science	
Industry/Organization	Tiger Analytics		Date/Duration	15 days	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					
Learning capacity/Knowledge up gradation					
Performance/Quality of work					
Behavior/Discipline/Team work					
Sincerity/Hard work					
Comment on nature of work done/Area/Topic	Abi soni moved ahead in the campus training Program;.				
OVERALL GRADE (Any one)	**VERY GOOD**				
Name of Industry Mentor	Padmajothi				
Signature of Industry Mentor	Padmajothi				

Receiving Date	**31/03/22**	**Name of Faculty Mentor**	**Dr. Tej Singh**	**Sign**	

FPR 3

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	ABI SONI		Department	Data science	
Industry/Organization	Tiger Analytics		Date/Duration	15 days	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					
Learning capacity/Knowledge up gradation					
Performance/Quality of work					
Behavior/Discipline/Team work					
Sincerity/Hard work					
Comment on nature of work done/Area/Topic	Abi soni is continuing his campus training.				
OVERALL GRADE (Any one)	**VERY GOOD**				
Name of Industry Mentor	Padmajothi				
Signature of Industry Mentor	Padmajothi				

Receiving Date	**16/03/22**	**Name of Faculty Mentor**	**Dr. Tej Singh**	**Sign**	

FPR 5

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	ABI SONI		Department	Data science	
Industry/Organization	Tiger Analytics		Date/Duration	15 days	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					
Learning capacity/Knowledge up gradation					
Performance/Quality of work					
Behavior/Discipline/Team work					
Sincerity/Hard work					
Comment on nature of work done/Area/Topic	Abi soni is going through the git, git-hub, ubuntu, and other handy tools of data science.				
OVERALL GRADE (Any one)	**VERY GOOD**				
Name of Industry Mentor	Padmajothi				
Signature of Industry Mentor	Padmajothi				

Receiving Date	**15/04/22**	**Name of Faculty Mentor**	**Dr. Tej Singh**	**Sign**	

FPR 7

FORTNIGHTLY PROGRESS REPORT (FPR) FROM INDUSTRY MENTOR

Name of student	ABI SONI		Department	Data science	
Industry/Organization	Tiger Analytics		Date/Duration	15 days	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					
Learning capacity/Knowledge up gradation					
Performance/Quality of work					
Behavior/Discipline/Team work					
Sincerity/Hard work					
Comment on nature of work done/Area/Topic	Abi Soni along with basic training doing a case study given to him.				
OVERALL GRADE (Any one)	**VERY GOOD**				
Name of Industry Mentor	Padmajothi				
Signature of Industry Mentor	Padmajothi				

Receiving Date	**31/05/22**	**Name of Faculty Mentor**	**Dr. Tej Singh**	**Sign**	**xxx**

27

Appendix - Role of Mentors

I am very grateful to my Institute and Industry mentors who helped me throughout my internship period during my last semester. Whenever I was stuck, they were always there to help me through my problems and guide me through. They were in continuous contact with me and helped me to successfully complete my internship.

Dr. Tej Singh
Assistant Professor
Information technology

Abi Soni
0901IT181005

28

turnitin

Similarity Report ID: oid:28506:17955715

PAPER NAME

Original_IT_Project_Report (5).pdf

WORD COUNT	CHARACTER COUNT
1874 Words	11521 Characters
PAGE COUNT	FILE SIZE
20 Pages	941.7KB
SUBMISSION DATE	REPORT DATE
May 31, 2022 2:31 PM GMT+5:30	May 31, 2022 2:32 PM GMT+5:30

15% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 2% Internet database
- Crossref database
- 13% Submitted Works database
- 0% Publications database
- Crossref Posted Content database

Excluded from Similarity Report

- Bibliographic material

Summary

turnitin

Similarity Report ID: oid:28506:17955715

● 15% Overall Similarity

Top sources found in the following databases:

- 2% Internet database
- Crossref database
- 13% Submitted Works database
- 0% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Korea National University of Transportation on 2020-07-06 Submitted works	10%
2	Kaplan College on 2021-08-09 Submitted works	1%
3	techtwinklers.com Internet	<1%
4	grin.com Internet	<1%
5	University of Portsmouth on 2020-09-24 Submitted works	<1%
6	bspace.buid.ac.ae Internet	<1%
7	Trine University on 2022-03-05 Submitted works	<1%
8	Djidenou Hans Amos Montcho. "Bayesian variable selection using Data... Crossref posted content	<1%

Sources overview