INTERNSHIP REPORT

ONE MONTH INTERNSHIP
(JAN 2021 – FEB 2021)

Data Science & Business Analytics

At THE SPARKS FOUNDATION

Prepared for

Department of Computer Science,
University of Mumbai
Ranade Bhavan, B-Wing, Ground Floor,
Vidyanagari Campus, Kalina,
Santacruz (East), Mumbai-400098.

Master in Computer Science
Specialization in Data Science

Prepared by

Manjiri Harishchandra Sawant
UDCSDS 312

May 7, 2021

Table of Contents

GRIP - The Sparks Foundation

...inspiring, innovating, integrating

The Sparks Foundation (TSF) is a non-profit organization registered in India and Singapore. We envision a world of enabled and connected little minds, building the future. We aim to inspire students, help them innovate, and let them integrate to build the next generation of humankind. We help the students to integrate and help each other, learn from each other, and do well together.

The Graduate Rotational Internship Program (GRIP) is the flagship program of TSF in which students, recent graduates, and professionals focus on technical skills development as well as professional profile improvement on LinkedIn. The program gives young aspiring minds a learning platform to connect with students and professionals from varied diversity, background, skills, and countries. During the Covid-19 pandemic, the format of GRIP is a 1-month, unpaid, and virtual internship.

Our Vision Statement

A world of enabled and connected little minds, building future.

Our Mission Statement

To inspire students, help them innovate and let them integrate to build the next generation humankind.

Inspire

To inspire, motivate and encourage students to learn, create and help build a better society.

Innovate

To teach new ways of thinking, to innovate and solve the problems on their own.

Integrate

To let the students integrate, and help each other, learn from each other and do well together.

Our Values

Resilience Commitment Integrity

Respect People Training Excellence

Quality Professionalism

Guiding Principles Based On Our Goals, Mission And Core Values

We remain aware of the mission

We aim to collaborate, and help students through our programmes to inspire and integrate them, using innovative ways.

We care for children and people

We believe children are gifts of god and any service to them is service to god. We are committed to providing everyone and opportunity to learn, grow and help others grow.

We treat everyone with respect

We treat all with respect, dignity and honesty. We show respect by listening to and seeking to understand different points of views, being sensitive to cultural and religious diversity, and by working collaboratively with others to achieve common goals and a harmonious work culture.

We collaborate to build partnerships

We actively seek to partner with organizations with similar goals, individuals, and corporates to enhance the effectiveness of our programs and advance common interests.

We honour the responsibilities entrusted to us

We constantly seek opportunities and innovative ways to create value through the work that we do to further our mission.

Advisors

Ms. Marie Ang - Founder and Managing Director of Solomon's Guild and the Asia Pacific President of Academic Management Group, LLC

Prof. Jasjit Singh - INSEAD professor since 2004

Dr. Allen Lai - Senior Vice President of ACT Genomics based in Singapore

Dr. Alan Prem Kumar - Principal Associate at Cancer Science Institute of Singapore, National University of Singapore and an Assistant Professor at Medical Science Cluster and Department of Pharmacology of Yong Loo Lin School of Medicine, NUS

Ms. Marta Vlatchkova - Director (FP&A) at Essilor, Singapore with Asia, Middle East, Russia and Africa.

Mr. Nedir Boudriga - Director of Market Risk & Product Control at Julius Baer, Singapore.

Prof. Roger Lehman - Professor at INSEAD Business School, Singapore, specializes in entrepreneurship, leadership, life-long learning, leading change and innovation, Director of Executive Masters in Consulting and Coaching for Change program at INSEAD.

Prof. Vishal Narayan - Associate Professor, Marketing at the NUS Business School.

Founding Team

Priyesh Kumar
Founder

Priyesh is a technology enthusiast at his core. He is known for architecting, designing and developing large scale applications, and ensures security, scalability, and maintainability. A graduate of Indian Institute of Technology Delhi, and National University of Singapore, he works with VISA, NUS, and the likes with some of the large-scale projects of impact and his expertise is in Application Innovation, Machine Learning and Artificial Intelligence. He brings in more than 12 years of versatile experience to build The Sparks Foundation into an innovative social enterprise with impact at scale.

Management Team

Tanwi Kaushik
Co – Founder and Executive Manager

Tanwi is a social entrepreneur and researcher. She is a double masters in Life Sciences from Cornell University (USA) and NUS (Singapore) with 6 years of experience in research and development. Having worked in India, USA and Singapore has given her a tremendous exposure and a chance to understand the academia and the industry internationally. After a vast understanding of academics in various countries from her own and others' experiences, she comprehended the gaps that still persist in the education system and hinder students from progressing to their full potential.

Graduate Rotational Internship Program

Why

Workshops/guidance for personality insights, LinkedIn profile building, interviews skills, further education, etc. Paid Job Offer after 12 months.

A unique global, multicultural & rare diverse working environment available for interns. We are located across the world and ensure successful projects.

We are students and alumni from top colleges (IIT, IIM, NUS, Stanford, etc.) and experts from industry.

How

An intern will start with a function he/she is most comfortable with. The function will remain unchanged for at least three months.

At the end of three months, the intern can choose to move to any other function of his/her interest.

Functions Available

Website & Apps Expert Mentors Relations Content & Marketing

Legal & Strategy Student Engagements Finance & Risk

Corporate Partners

Xaltius

AINE AI

Code for India

Description about Activities

How Many Tasks to Do

***Minimum Requirement** to be eligible to get an internship completion certificate:*

- LinkedIn Profile Improvement - Improve your professional profile on LinkedIn. It is **MANDATORY** FOR ALL.
- Technology (only Tech interns) - Complete **AT LEAST ONE TASK** from the list of tasks given under internship function. After that, intern can do as many tasks as he/she wants for learning & LoR (Letter of Recommendation).
- Non-Tech (only non-tech interns) - Complete **AT LEAST ONE TASK** from the list of tasks given under internship function. Intern can do as many tasks as he/she wants for learning & LoR.
- Peer-evaluation (mandatory for all): **Watch and comment on the at least 5 task videos on LinkedIn posted by fellow interns**. Refer to FAQs for the steps of peer evaluation: https://lnkd.in/gnGiBbb
- Additional tasks for LoR (optional) - This will be shared via email. Intern can also refer to FAQs for this: https://lnkd.in/gnGiBbb

Instructions for all Tasks

- Due to similar nature of skills, some tasks are combined. Example, tasks of Web Development and Mobile App Development are merged into one category - Web & Mobile Development.
- Intern can do as many tasks he/she can do from his/her domain category for learning & skills development.
- In case of any **query related to tasks**, Intern can ask in TSF Network. Please refer to section 'Internship Period & Tasks' in the FAQs document: https://lnkd.in/gnGiBbb
- Intern's posts in TSF Network needs approval. We may not approve a query which has recently been answered. Interns are suggested to scroll down or use 'Search' to make sure your query is unique in last few days.
- For peer-evaluation (mandatory), please refer to FAQs document.

Data Science & Business Analytics Tasks

Task 1

Prediction using Supervised ML (Level – Beginner)

- Predict the percentage of a student based on the no. of study hours. This is a simple linear regression task as it involves just 2 variables.
- Intern can use R, Python, SAS Enterprise Miner or any other tool. Data can be found at http://bit.ly/w-data
- What will be predicted score if a student studies for 9.25 hrs/ day?
  Sample Solution: https://bit.ly/2HxiGGJ

Task 2

Prediction using Unsupervised ML (Level – Beginner)

- From the given iris dataset, predict the optimum number of clusters and represent it visually.
- Use R or Python or perform this task.
- Dataset : https://bit.ly/3kXTdox
- Sample Solution : https://bit.ly/3cGyP8j

Task Submission:

1. Host the code on GitHub Repository (public). Record the code and output in a video. Post the video on YouTube
2. Share links of code (GitHub) and video (YouTube) as a post on **Intern's LinkedIn profile**, not TSF network.
3. Submit the LinkedIn link in Task Submission Form when shared.
4. Please read FAQs on how to submit the tasks.

Task 3

Exploratory Data Analysis – Retail (Level – Beginner)

- Perform 'Exploratory Data Analysis' on dataset 'SampleSuperstore'.
- As a business manager, try to find out the weak areas where you can work to make more profit.
- What all business problems you can derive by exploring the data? Intern can choose any of the tools of his/her choice.
- (Python/R/Tableau/PowerBI/Excel/SAP/SAS)
- Dataset: https://bit.ly/3i4rbWl
- Beginner Level - Create dashboards. Screen-record along with intern's audio explaining the charts and interpretations.

Task 4
Exploratory Data Analysis – Terrorism (Level – Intermediate)

- Perform 'Exploratory Data Analysis' on dataset 'Global Terrorism'.
- As a security/defense analyst, try to find out the hot zone of terrorism.
- What all security issues and insights you can derive by EDA?
- Intern can choose any of the tool of your choice
  (Python/R/Tableau/PowerBI/Excel/SAP/SAS)
  Dataset: https://bit.ly/34SRn3b
  Intermediate Level - Create storyboards. Screen-record along with your audio explaining the charts and interpretations. Use annotations, animation and images.

Task 5
Exploratory Data Analysis – Sports (Level – Advanced)

- Perform 'Exploratory Data Analysis' on dataset 'Indian Premier League'.
- As a sports analyst, find out the most successful teams, players and factors contributing win or loss of a team.
- Suggest teams or players a company should endorse for its products.
- Intern can choose any of the tool of your choice
- (Python/R/Tableau/Powerbase/Excel/SAP/SAS)
- Dataset: https://bit.ly/34SRn3b
- Advanced Level - Create storyboards. Screen-record along with your audio explaining the charts and interpretations. Use annotations, animation and images.

Task submission:
1. Create the dashboards and/or storyboard and record it
2. Upload the recording on Youtube, share the link on LinkedIn
3. Submit LinkedIn post link in Task Submission Form when shared
4. Please read FAQs on how to submit the tasks.

Task 6
Prediction Using Decision Tree Algorithm (Level – Intermediate)

- Create the Decision Tree classifier and visualize it graphically.
- The purpose is if we feed any new data to this classifier, it would be able to predict the right class accordingly.
- Dataset : https://bit.ly/3kXTdox
- Sample Solution : https://bit.ly/2G6sYx9

Task submission:
1. Host the code on GitHub Repository (public). Record the code and output in a video. Post the video on YouTube.
2. Share links of code (GitHub) and video (YouTube) as a post on Intern's LinkedIn profile.

3. Submit the LinkedIn link in Task Submission Form when shared.
4. Please read FAQs on how to submit the tasks.

Task 7

Stock Market Prediction using Numerical & Textual Analysis (Level – Advanced)

- Objective: Create a hybrid model for stock price/performance prediction using numerical analysis of historical stock prices, and sentimental analysis of news headlines.
- Stock to analyze and predict - SENSEX (S&P BSE SENSEX)
- Download historical stock prices from finance.yahoo.com
- Download textual (news) data from https://bit.ly/36fFPI6
- Use either R or Python, or both for separate analysis and then combine the findings to create a hybrid model
- Intern is free to select a different stock to analyze and news dataset as well while not changing the objective of the task.

Task 8

Timeline Analysis: Covid -19 (Level – Advanced)

- Create a storyboard showing spread of Covid-19 cases in your country or any region (Asia, Europe, BRICS etc) using Tableau, Power BI or SAP
- Use animation, timeline and annotations to create attractive and interactive dashboards and story
- Identify interesting patterns and possible reasons helping Covid-19 spread with basic as well as advanced charts
- Screen-record the completed storyboard along with your audio explaining the charts and giving recommendations.
- Dataset: Daily updated .csv file on https://bit.ly/30d2gdi

Task submission:

1. Create the dashboards and/or storyboard and record it
2. Upload the recording on YouTube, share the link on LinkedIn
3. Submit LinkedIn post link in Task Submission Form when shared
4. Please read FAQs on how to submit the tasks.

Description about Work Performed

*There are **total 8 tasks** has been assigned out of which **4 tasks has been accomplished** during internship.*

Task 1

Prediction using Supervised ML (Level – Beginner)

Predict the Percentage of Students Based on the No. of Study Hours

Supervised learning is the types of machine learning in which machines are trained using well **"labelled" training data**, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

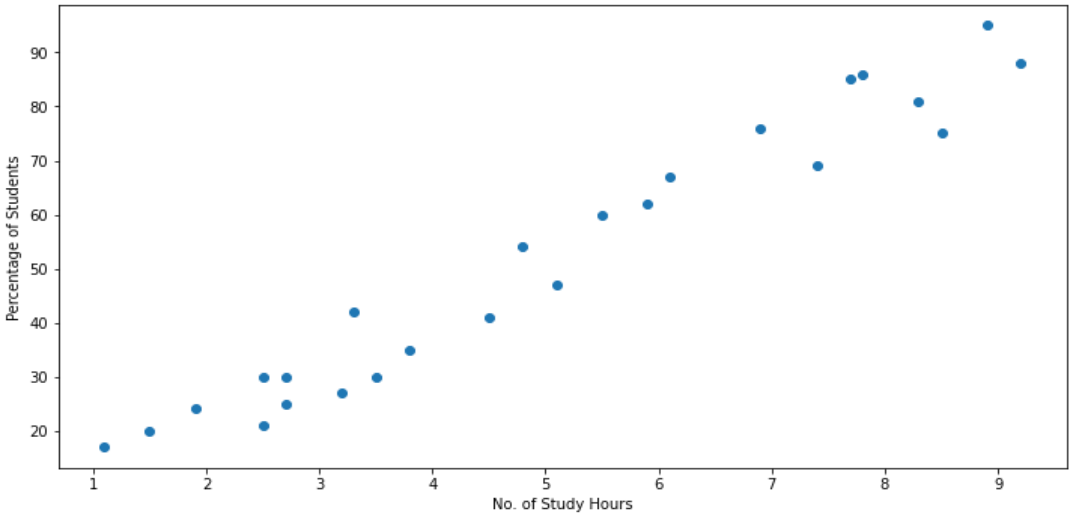
Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function **to map the input variable(x) with the output variable(y).**

Examples of Supervised Machine Learning

- Application of Photo Tagging
- Loan Approval
- Targeted Online Advertisement

This is **simple Linear Regression Task** as it involves just two variables i.e. 'Hours' & 'Scores'. *No. of Study Hours* is an independent variable and *Percentage of Students* is a dependent variable. As the No. of study Hours increases the Percentage of students would also increase.

Tool Used: Python Jupyter Notebook



Problem Statement: What will be predicted score if a student studies for 9.25hrs/day?

Here is the Solution:

Code: https://github.com/ManjiriSDS/The-Spark-Foundation/blob/main/Task%201%20The%20Sparks%20Foundation.ipynb

Recorded Output in YouTube Video:
https://www.youtube.com/watch?v=lkTiUUGPrU4&t=1s

Task 2
Prediction using Unsupervised ML (Level – Beginner)

From the given 'Iris' dataset, predict the optimum number of clusters and represent it visually.

Unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things.

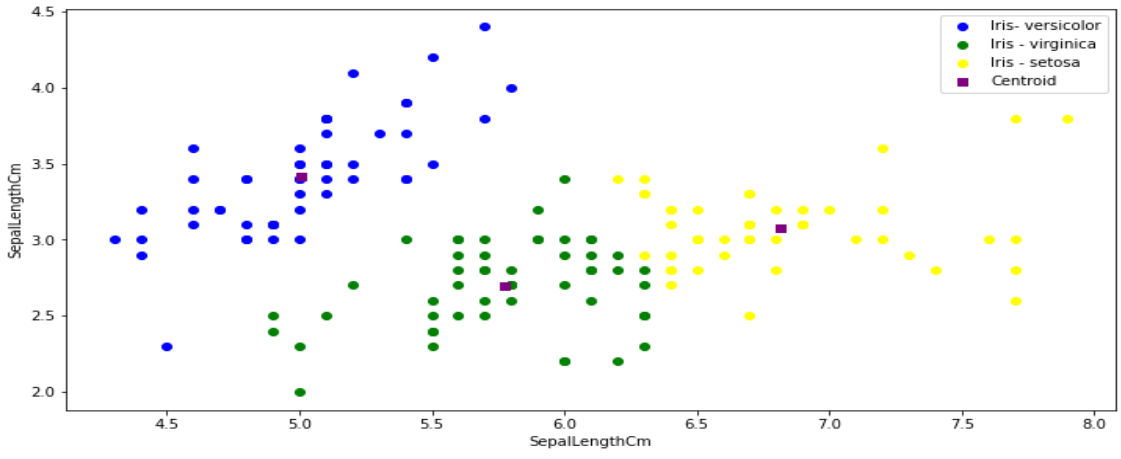
Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we **have the input data but no corresponding output data.** The goal of unsupervised learning is **to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format**.

Examples of Unsupervised Machine Learning

- Pattern Recognition
- Customer segmentation
- Clustering DNA patterns to analyze evolutionary biology.

Clustering With K-Means
Clustering is often referred to as Unsupervised Machine Learning technique. Clustering is the process of dividing the datasets into groups, consisting of similar data points.
Correct no of Clusters (K) has been determined using Elbow Method.

Tool Used: Python Jupyter Notebook



Here is the solution:

Code: https://github.com/ManjiriSDS/The-Spark-Foundation/blob/main/Task%202%20The%20Sparks%20Foundation.ipynb

Recorded Output in YouTube Video:
https://www.youtube.com/watch?v=Tk7fWt_TNcs&t=2s

Task 5
Exploratory Data Analysis – Sports (Level – Advanced)

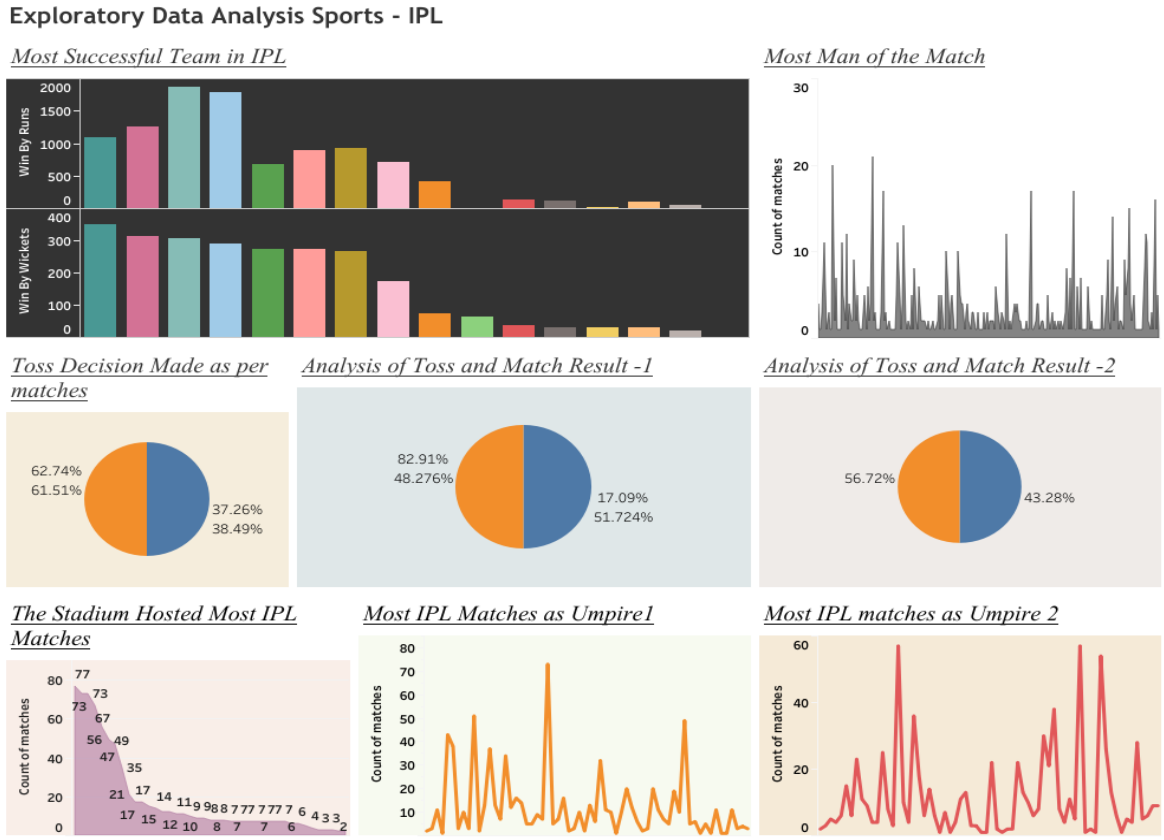
Perform 'Exploratory Data Analysis' on dataset 'Indian Premier League'.

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Typical graphical techniques used in EDA are:

- Box plot
- Histogram
- Multi-vari chart
- Run chart
- Pareto chart
- Scatter plot

Tool Used: Tableau Software



Problem Statement: As a sports analysts, find out the most successful teams, players and factors contributing win or loss of a team. Suggest teams or players a company should endorse for its products.

Here is the Solution:
Dashboard:
https://public.tableau.com/profile/sonu75769#!/vizhome/IPLSportEDA/Dashboard1
Exploratory Data Analysis IPL - Manjiri Sawant | Tableau Public

Recorded Output in YouTube Video:
https://www.youtube.com/watch?v=Gu-IBUyP-CU&t=2s

Task 6

Prediction Using Decision Tree Algorithm (Level – Intermediate)

Create the Decision Tree classifier and visualize it graphically.

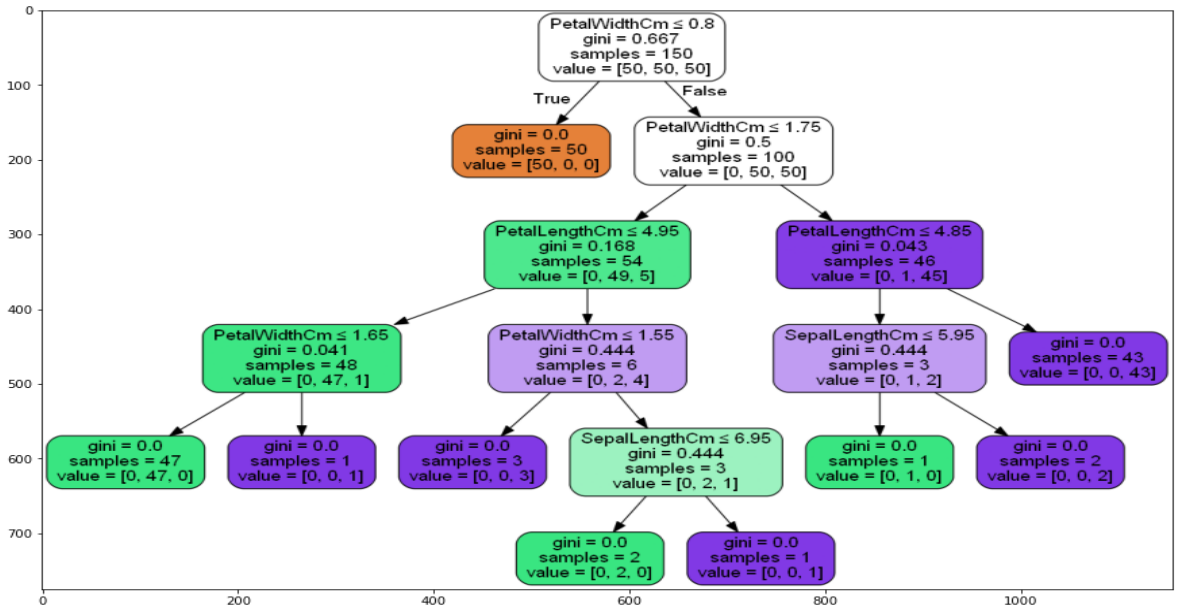
Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.**

In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node**. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

It is a graphical representation for getting all the possible solutions to problem/decision based on given conditions.

Tool Used: Python Jupyter Notebook



Applications of Decision Tree

- Fraud detection
- Credit Risk
- Loan Default
- Predicting Sales of retail Outlet
- Customer Churn

Problem Statement: Using Classifier Predict Right Class Accordingly.

Here the solution:

Code: https://github.com/ManjiriSDS/The-Spark-Foundation/blob/main/Task%206%20The%20Sparks%20Foundation.ipynb

Recorded Output in YouTube Video:
https://www.youtube.com/watch?v=ubZ2saIFQj0&t=5s

Submitted ALL Tasks link:

Enclosed herewith:
https://drive.google.com/file/d/1xgIZy0lkOvj_NauecRBZab6bfx52jpn4/view?usp=sharing

Self-Assessment

Lesson Learned

An internship is a learning experience of its own kind. The importance it has got over the years in building one's career is not exaggerated; given what it has to give back.

Neither is it spoon-fed school learning, nor pressure filled workload. And in between, we not only learn the basics of work life but also the soft skills required for a brighter professional career.

1. *Problem Solving Skills*
   An internship introduces us to real-life work problems and hence develops problem-solving skills.
2. *Work Ethics*
   We won't really learn about work ethics until we are in a tangible work environment. It is only when we are introduced to the actual environment that we learn work ethics.
3. *Adaptability Skills*
   Being adaptive to your surroundings easily is one of the most useful soft skills not only desirable to employers but also important to our self-growth.
4. *Responsibility*
   Internship experience makes us more responsible and accountable for what decisions we make and how we execute what's been allocated to us.
5. *Time Management*
   Internship helps us learn to manage our time better by maintaining a balance between our work and personal life, without harming any of them.

Relation of the Internship to Career Goals

This Data Science and Business Analytics virtual Internship contributes to learn about

- ✓ Data analyst job roles and responsibilities.
- ✓ What are the skills required to become a data analyst
- ✓ Gain real world data science experience
- ✓ Build and tune machine learning model using R, Python and scikit learn
- ✓ Use BI tools such as Tableau to analyse data, find important patterns and design, visualization dashboards on the given dataset
- ✓ Build Strong Public Profile through LinkedIn Profile Improvement, present given task and submissions; improve skills through various activities as a part of the internship

THE SPARKS FOUNDATION

THIS IS PRESENTED TO

MANJIRI HARISHCHANDRA SAWANT

for successful selection as an intern at The Sparks Foundation for function Data Science & Business Analytics.

PRANAV DUBEY
DIRECTOR

28-DEC-2020
DATE

CODE : CCL9PMZPCW
Verify at:
https://truecertificates.com/verification

THE SPARKS FOUNDATION
INSPIRE, INNOVATE, INTEGRATE

CERTIFICATE OF COMPLETION

This Certificate is presented to

Manjiri Harishchandra Sawant

for an outstanding contribution during the session (Jan 2021 - Feb 2021) of Graduate Rotational Internship Program at The Sparks Foundation on 07-Feb-2021.

Certificate Number: P7ZG2EXKGU

Verification at:
https://truecertificates.com/verification/

PRANAV DUBEY
MANAGING DIRECTOR

Peer Evaluation on LinkedIn

Enclosed Herewith:

https://www.linkedin.com/posts/manjiri-sawant-3893b757_task1-linear-regression-activity-6757317864069369856-mgkJ

https://www.linkedin.com/posts/manjiri-sawant-3893b757_task2-prediction-using-unsupervised-ml-thesparksfoundation-activity-6756945761516208129-l-GS

https://www.linkedin.com/posts/manjiri-sawant-3893b757_task5exploratory-data-analysis-sports-activity-6757743373530017792-r_09

https://www.linkedin.com/posts/manjiri-sawant-3893b757_task6-activity-6757308681441853440-P7hk