

**SUMMER TRAINING REPORT**

on

**DATA SCIENCE & MACHINE LEARNING**

(June – July 2022)

Submitted by

**NIVASH G**

**Registration Number: 12010380**

**[B. Tech (Computer Science Engineering) Hons] – LPU**

**Specialization: Data Science**

Under the Guidance of

**PRANAY SHARMA**

School of Computer Science & Engineering

Lovely Professional University, Phagwara

**DECLARATION**

I hereby declare that I have completed our six weeks of summer training at  **All Soft Solutions(IBM), Punjab** through online classesfrom **June 6** to **July 8** under the guidance of **Pranay Sharma.** I declare that I have worked with full dedication during this training and our learning outcomes fulfill the requirements of training for the award of the degree of **B.Tech (Hons)**, Lovely Professional University, Phagwara.

**NIVASH G**

Date:15/07/2022

Expert from IBM(Teacher): **Pranay Sharma**

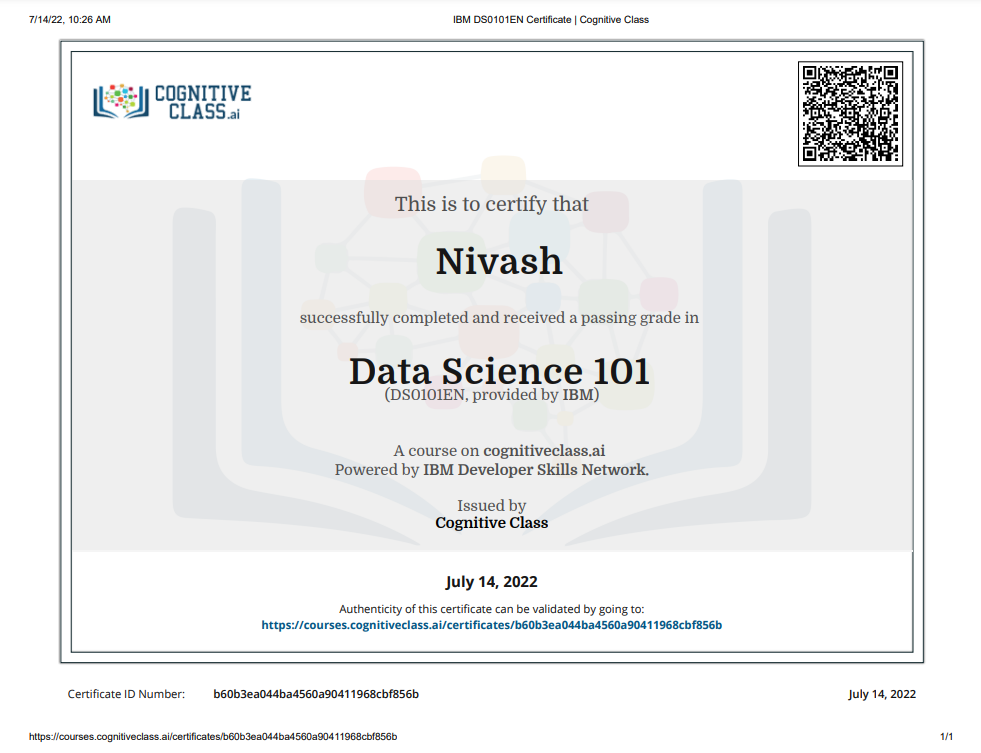
**ACKNOWLEDGEMENT**

The success and final outcome of this project required a lot of guidance and assistance from many people and I are extremely fortunate to have got this along the completion of our project work. Whatever I have done is only due to such guidance and assistance and I would not forget to thank them.

I would like to express our respect and gratitude to the teachers of Lovely Professional University and the coordinator for providing this opportunity to do this training program and project regarding **data science & machine learning.**

I would also like to extend our gratitude to **Pranay Sharma**(Teacher from IBM/All Soft Solutions) who helped us by giving guidance and support all through the program and also in completing the project.

**SUMMER TRAINING CERTIFICATE**





**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **S.NO** | **TITLE** |
| 1 | Cover Page |
| 2 | Declaration of the Student |
| 3 | Acknowledgment |
| 4 | Summer Training Certificate |
| 5 | Table of Contents |
| 6 | Introduction |
| 7 | Technology Learnt |
| 8 | Reason for choosing Data Science and Machine Learning |
| 9 | Learning Outcome |

**INTRODUCTION**

Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization’s data. It has incalculable benefits in business, research, and our everyday lives. Your route to work, your most recent search engine query for the nearest coffee shop, your Instagram post about what you ate, and even the health data from your fitness tracker are all important to different data scientists in different ways. Sifting through massive data lakes, and looking for connections and patterns, data science is responsible for bringing us new products, delivering breakthroughs, insights and making our lives more convenient. These insights can be used to guide decision-making and strategic planning.

The accelerating volume of data sources, and subsequently data, has made data science one of the fastest growing fields across every industry. As a result, it is no surprise that the role of the data scientist was dubbed the “sexiest job of the 21st century” by Harvard Business Review. Organizations are increasingly reliant on them to interpret data and provide actionable recommendations to improve business outcomes.

**What is Data?**

Data is a collection of information. One purpose of Data Science is to structure data, making it interpretable and easy to work with. Data can be categorized into two groups:

* Structured data - Structured data is organized and easier to work with.
* Unstructured data – Unstructured data is information that either does not have a pre-defined data model or is not organized in a pre-defined manner.
* Python is a programming language widely used by Data Scientists.
* Python has in-built mathematical libraries and functions, making it easier to calculate mathematical problems and to perform data analysis.
* Python has libraries with large collections of mathematical functions and analytical tools.

**TECHNOLOGY LEARNT**

**Python Variables**

Variables are containers for storing data values. A variable is created the moment we first assign a value to it. A Python variable is a name given to a memory location. It is the basic unit of storage in a program.

**Python Data Types**

Data Types represents the kind of value that tells what operations can be performed on a particular data.

**Python Numbers**

Numeric data type represent the data which has numeric value. Numeric value can be integer, floating number or even complex numbers. These values are defined as int, float and complex class in Python.

**Python Casting**

Casting is the method to convert the variable data type into a certain data type in order to the operation required to be performed by users.

**Python Strings**

String is the collection of the characters surrounded by single quotes, double quotes, or triple quotes. The computer does not understand the characters; internally, it stores manipulated character as the combination of the 0's and 1's.

**Python Booleans**

Booleans type is one of the built-in data types provided by Python, which represents one of the two values i.e. True or False. Generally, it is used to represent the truth values of the expressions.

**Python Operators**

Operators in general are used to perform operations on values and variables. These are standard symbols used for the purpose of logical and arithmetic operations.

**Python Lists**

Lists are used to store multiple items in a single variable. list is a collection of things, enclosed in [ ] and separated by commas.

**Python Tuples**

Python Tuple is a collection of objects separated by commas. In some ways, a tuple is similar to a list in terms of indexing, nested objects, and repetition but a tuple is immutable, unlike lists which are mutable.

**Python Sets**

Set is the collection of the unordered items. Each element in the set must be unique, immutable, and the sets remove the duplicate elements. Sets are mutable which means we can modify it after its creation.

**Python Dictionaries**

Dictionary is used to store the data in a key-value pair format. The dictionary is the data type in Python, which can simulate the real-life data arrangement where some specific value exists for some particular key. The dictionary is defined into element Keys and values.

**Python If...Else**

If. Else provides the block of the code for the false case of the condition to be checked. If the condition provided in the if statement is false, then the else statement will be executed.

**Python While Loops**

Python while loop repeats a statement or group of statements while a given condition is TRUE. It tests the condition before executing the loop body.

**Python For Loops**

For loop executes a code block multiple times and abbreviates the code that manages the loop variable.

**Python Functions**

Python functions are simple to define and essential to intermediate-level programming. The exact criteria hold to function names as they do to variable names. The goal is to group up certain often performed actions and define a function. It is also used to define an unnamed function.

**Python Lambda**

Lambda Functions in Python are anonymous functions, implying they don't have a name. The def keyword is needed to create a typical function in Python,

**Python Arrays**

An array is defined as a collection of items that are stored at contiguous memory locations. It is a container which can hold a fixed number of items, and these items should be of the same type. It is an idea of storing multiple items of the same type together and it makes easier to calculate the position of each element by simply adding an offset to the base value.

**Python Classes/Objects**

Class creates a user-defined data structure, which holds its own data members and member functions, which can be accessed and used by creating an instance of that class. A class is like a blueprint for an object.

**Python Modules**

A module is a file containing Python code, definitions of functions, statements, or classes. An example\_module.py file is a module we will create and whose name is example\_module.

**Python Dates**

Datetime module supplies classes to work with date and time. These classes provide a number of functions to deal with dates, times and time intervals. Date and datetime are an object in Python, so when you manipulate them, you are actually manipulating objects and not string or timestamps.

**Python Math**

Python has a set of built-in math functions, including an extensive math module, that allows you to perform mathematical tasks on numbers.

**Python JSON**

JSON is a widely used data format for data interchange on the web. JSON is the ideal format for organizing data between a client and a server. The main objective of JSON is to transmit the data between the client and the web server. It is easy to learn and the most effective way to interchange the data.

**Python RegEx**

A RegEx, or Regular Expression, is a sequence of characters that forms a search pattern. It can be used to check if a string contains the specified search pattern.

**Python Try...Except**

Try and Except statement is used to handle these errors within our code in Python. The try block is used to check some code for errors i.e the code inside the try block will execute when there is no error in the program. Whereas the code inside the except block will execute whenever the program encounters some error in the preceding try block.

**NumPy Tutorial**

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It can also be used as an efficient multi-dimensional container of generic data.

**Pandas Tutorial**

Pandas is an open-source library that is built on top of NumPy library. It is a Python package that offers various data structures and operations for manipulating numerical data and time series. It is mainly popular for importing and analyzing data much easier.

**SciPy Tutorial**

SciPy is a python library that is useful in solving many mathematical equations and algorithms. It helps make ML models is that it also makes a strong programming language available for use in developing less complex programs and applications.

**Python Matplotlib**

Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

**SYSTEM IMPLEMENTATION**

**Python:** It is an interpreted, high-level general-purpose programming language. Created by Guido Van Rossum and first released in 1991. Its language constructs and objects-oriented approach aim to help programmer with clear, logical code for small and large-scale tools. Python is used for web development (server-side), software development, mathematics, it can be used alongside software to create workflows, it can connect to database systems, it can also read and modify files, it can be used to handle big data and perform complex mathematics and can be used for rapid prototyping, or for production-ready software development.

**Pandas:** This is an open-source Python libraries which is mainly used in Data Science and machine learning subjects. This library provides analysis tool for data manipulation, using its data structures this are used for analyzing data for manipulating time series analysis and numerical data.

**NumPy:** NumPy can be name come from Numeric Python, it is a data analysis library for Python that contains various numerical functions and methods for numerical analysis and also having multi-dimensional array objects and to process these arrays contains collection of routines.

**Matplotlib**: Matplotlib is easy to use and an amazing visualizing library in Python. It is built on NumPy arrays, and it work with the broader SciPy stack and consists of several plots like pie, line, bar, graph, scatter, histogram, etc. In this project, Matplotlib is used for various visualizations for analysis of whatsapp chats. Visualizations like bar charts, line charts, pie charts are used.

**Seaborn**: Seaborn is a library mostly used for statistical plotting in Python. To make statistical plots more attractive it provides beautiful color palettes and default styles. In this project, Seaborn is used for heatmap visualization for showing 24 hours with 7 day with different scale of color for getting hour with max to min messages.

**NLP:** In this project, Features of NLP are used like Parsing Text, Eliminating stop words and Analyzing Text. Parsing text is used for splitting messages into words for analysis like total words and mostly used words. A file is used that contains all stop words which is given to the python program to show meaningful words only by eliminating all stop words. Text analysis is used to identify how many media are shared; how many links are shared.

**SURVEY**

As a demo Survey analysis on the usage and Impact of WhatsApp Messenger, Various Studies and analysis has been done on the usage and impact of WhatsApp. Some of these studies are for finding the **impact of WhatsApp** on the students and some are based on for the public in a local region.

In a study of southern part of India was conducted on the age group of between 18 to 23 years to investigate the importance of WhatsApp among youth. Though this study, it was found that students spent 8 hours per day on using WhatsApp and remain online almost 16 hours a day. All the respondents agreed that they are using WhatsApp for communicating with their friends. They also exchange images, audio and video files with their friends using WhatsApp. It was also proved that the only application that the youth uses when they are spending time on their smart phone is WhatsApp. Methods used in this survey is to analyze the **intensity of WhatsApp usage** and its popular services and to identify the degree of positive or negative impacts of using WhatsApp.

**ADVANTAGES OF WHATSAPP CHAT ANALYSIS PROJECT**

• Works on all devices.

• Shows based on whatsapp chat file.

• Shows different visualizations.

• Total Messages.

• Total words.

• Media shared.

• Link shared.

• Monthly timeline.

• Most busy day.

**TECHNOLOGY LEARNT**

**Module-1 Python in Data Science**