Assignment 01 - Recent Trends in Data Science

| Roll No. | 58 |
| --- | --- |
| Name | Ninad Rao |
| Class | D15-A |
| Subject | DS using Python Lab |
| LO Mapped | LO1: Understand the concept of Data science process and associated terminologies to solve real-world problems |
|  |  |

**Date of Submission**: 20th Feb 2022

**Explain the term Data Science**

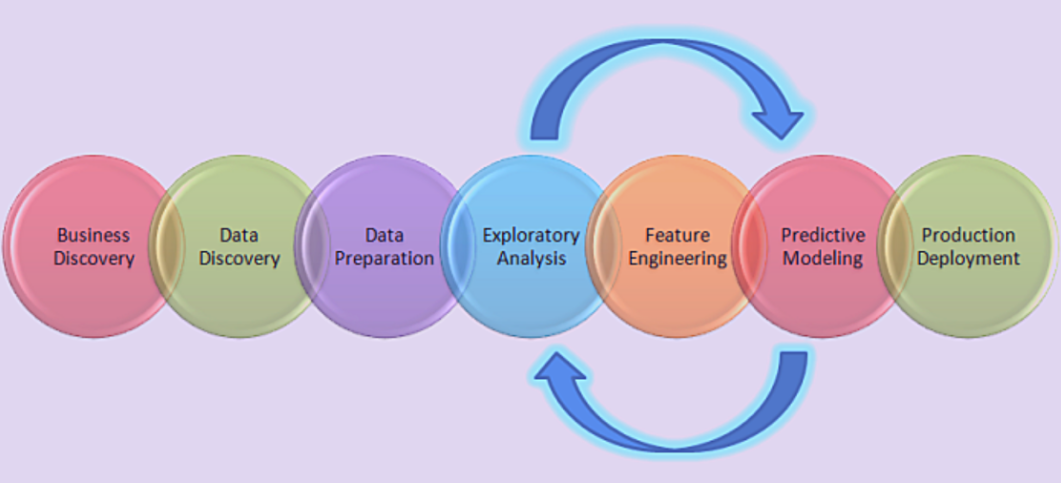
Data science is the field of applying advanced analytics techniques and scientific principles to extract valuable information from data for business decision-making, strategic planning and other uses. Data science helps organizations increase operational efficiency, identify new business opportunities and improve marketing and sales programs, among other benefits. Ultimately, they can lead to competitive advantages over business rivals.

Data science incorporates various disciplines like data engineering, data preparation, data mining, predictive analytics, machine learning and data visualization, as well as statistics, mathematics and software programming. It's primarily done by skilled data scientists, although lower-level data analysts may also be involved. In addition, many organizations now rely partly on citizen data scientists, a group that can include business intelligence (BI) professionals, business analysts, data-savvy business users, data engineers and other workers who don't have a formal data science background.

This comprehensive guide to data science further explains what it is, why it's important to organizations, how it works, the business benefits it provides and the challenges it poses. You'll also find an overview of data science applications, tools and techniques, plus information on what data scientists do and the skills they need.

**Elaborate the tasks involved in Data Science and explain its lifecycle**

A data science life cycle is nothing but a repetitive set of steps that you need to take to complete and deliver a project/product to your client. Although the data science projects and the teams involved in deploying and developing the model will be different, every data science life cycle will be slightly different in every other company. However, most of the data science projects happen to follow a somewhat similar process.



The tasks involved in Data Science are:

1. **Understanding the Business Problem**:

In order to build a successful business model, it's very important to first understand the business problem that the client is facing. It is important to take consultation from domain experts and finally understand the underlying problems that are present in the system.

A Business Analyst is generally responsible for gathering the required details from the client and forwarding the data to the data scientist team for further speculation. Even a minute error in defining the problem and understanding the requirement may be very crucial for the project hence it is to be done with maximum precision.

2. **Data Collection**:

The data science project starts with the identification of various data sources, which may include web server logs, social media posts, data from digital libraries, data accessed through sources on the internet via APIs, web scraping, or information that is already present in an excel spreadsheet. Data collection entails obtaining information from both known internal and external sources that can assist in addressing the business issue.

Normally, the data analyst team is responsible for gathering the data. They need to figure out proper ways to source data and collect the same to get the desired results. There are two ways to source the data:

* Through **web scraping** with Python
* Extracting Data with the use of **third party APIs**

3. **Data Preparation**:

This stage is referred to as Data Cleaning or Data Wrangling. It entails steps such as selecting relevant data, combining it by mixing data sets, cleaning it, dealing with missing values by either removing them or imputing them with relevant data, dealing with incorrect data by removing it, and also checking for and dealing with outliers. By using feature engineering, you can create new data and extract new features from existing ones. Format the data according to the desired structure and delete any unnecessary columns or functions. Data preparation is the most time-consuming process, accounting for up to 90% of the total project duration, and this is the most crucial step throughout the entire life cycle.

**Exploratory Data Analysis** (EDA) is critical at this point because summarizing clean data enables the identification of the data’s structure, outliers, anomalies, and trends. These insights can aid in identifying the optimal set of features, an algorithm to use for model creation, and model construction.

4. **Data Modeling**:

Throughout most cases of data analysis, Data Modeling is regarded as the core process. We first tend to select the appropriate type of model that would be implemented to acquire results, whether the problem is a regression problem or classification, or a clustering-based problem. Depending on the type of data received we happen to choose the appropriate machine learning algorithm that is best suited for the model. Once this is done, we ought to tune the hyperparameters of the chosen models to get a favorable outcome.

Finally, we tend to evaluate the model by testing the accuracy and relevance. In addition to this project, we need to make sure there is a correct balance between specificity and generalizability, which is that the created model must be unbiased.

5. **Model Deployment**:

Before the model is deployed, we need to ensure that we have picked the right solution after a rigorous evaluation. Later on, it is then deployed in the desired channel and format. This is naturally the last step in the life cycle of data science projects. Please take extra caution before executing each step in the life cycle to avoid unwanted errors.

If you choose the wrong machine learning algorithm for data modeling then you will not achieve the desired accuracy and it will be difficult in getting approval for the project from the stakeholders. If your data is not cleaned properly, you will have to handle missing values or the noise present in the dataset later on. Hence, in order to make sure that the model is deployed properly and accepted in the real world as an optimal use case, you will have to do rigorous testing in every step.

**Highlight the similarities and differences between Data Science, Business Analytics, Data Analytics and Big Data**

**Data Science**

1. Meaning - Data Science is a field that refers to the collective processes, theories, concepts, tools and technologies that enable the review, analysis and extraction of valuable knowledge and information from raw data.
2. Skills Required - Combines data, Capturing, Statistics, Mathematics, Programming and Problem solving required.
3. Application Areas - Digital advertisements, Internet research, Recommender system and Image/Speech recognition.
4. Tools and Languages - Python, SAS and SQL.
5. Annual Salary - $130,323

**Business Analytics**

1. Meaning - Business Analytics is the process by which businesses use statistical methods and technologies for analyzing historical data in order to gain new insight and improve strategic decision-making.
2. Skills Required - Statistics and probability, Data retrieval, Statistical tools, Statistical programming and Visualization required.
3. Application Areas - Finance, Marketing, HR Professionals and Manufacturing.
4. Tools and Languages - SAS, Excel and Tableau Public.
5. Annual Salary - $74,458

**Data Analytics**

1. Meaning - Data Analytics is the process of examining the data sets in order to draw the conclusions about the information they contain, increasingly with the aid of specialized systems and software.
2. Skills Required - Analytical, Math and Statistical programming skills required.
3. Application Areas - Gaming, Travel, Energy management and Healthcare.
4. Tools and Languages - R, Tableau Public and Apache Spark.
5. Annual Salary - $69,845

**Big Data**

1. Meaning - Big Data refers to the voluminous amounts of structured or unstructured data that organizations can potentially mine and analyze for business gain.
2. Skills Required - Business skills, Creativity, Analytics, Math and Statistics programming.
3. Application Areas - Communication, Retail, Financial services and Education.
4. Tools and Languages - Hadoop, NoSQL and Hive.
5. Annual Salary - $62,006

The similarities between them are:

1. **Analyze and understand data**

In data science, your understanding of data will determine how successful your model is. For instance, you are building a model to forecast the sale. If you know what is the most important or combination of factors that affect sales, then you can input those factors into your model. Therefore, your model will be able to give reasonable predictions.

In Business intelligence, this is also very important. Every day, you will be flooded with business requests. No matter whether you are trying to answer business questions or provide insights, you need to have a concrete understanding of data. For example, you can’t give insight into which kind of customers convert more without knowing the customer journey.

2. **Using data to solve business problems**

For instance, the business intelligence department is required to forecast the sales to allocate a budget for each sales team to spend. By performing simple regression on historical data in the excel sheet, and also manually tweaking some variables to comprehend future events, it can just be done using some formulas in an excel sheet.

If this problem is required to be tackled by the data science team, based on what I observed, the tool used will be very different. For instance, they will try out with a sophisticated machine learning model or a deep learning model that cannot be achieved by just using excel.

Although the tool is different, both of them are just trying to forecast the sales as accurately as possible. A combination of both teams will actually lead to a more efficient result. This is because the tasks of the business intelligence team are answering questions or providing insights. They will understand the data more and therefore if both teams work together, the data science team can save time to find out those important variables to be used to train the model.

3. **Storytelling and Visualization**

Imagine you are able to find out some valuable insights, but you end up plotting a graph that is not able to clearly show your insights or maybe you plot the right graph, but you are not able to communicate with your customers or clients for them to understand your insights clearly.

After all the hard work of cleaning and analyzing the data, you need to also make sure that you can present your results in a clear manner. However, the tool to plot the graph could be different. For instance, the business intelligence team uses Tableau, Excel or inhouse built’s business intelligence dashboard to plot the graph.

On the other hand, the data science team will utilize programming languages such as Python or R to visualize by coding. No matter what tools both teams are using, the end result will be to do PowerPoint slides to show their work.

4. **Tools and Insights**

Both demonstrate how the world of BI has changed, one in a very literal sense. Big data, just as a term, personifies the paradigm shift the field underwent. There’s just more data to work with. And because there’s more available, there’s more that can be done with it.

While the two are distinct terms, there is a significant overlap between them. Both are trying to glean insights from data analysis. Big data analytics tools can perform business analytics and have led to an extreme shift in how it is done and what results it can produce. But there are some differences, as well.

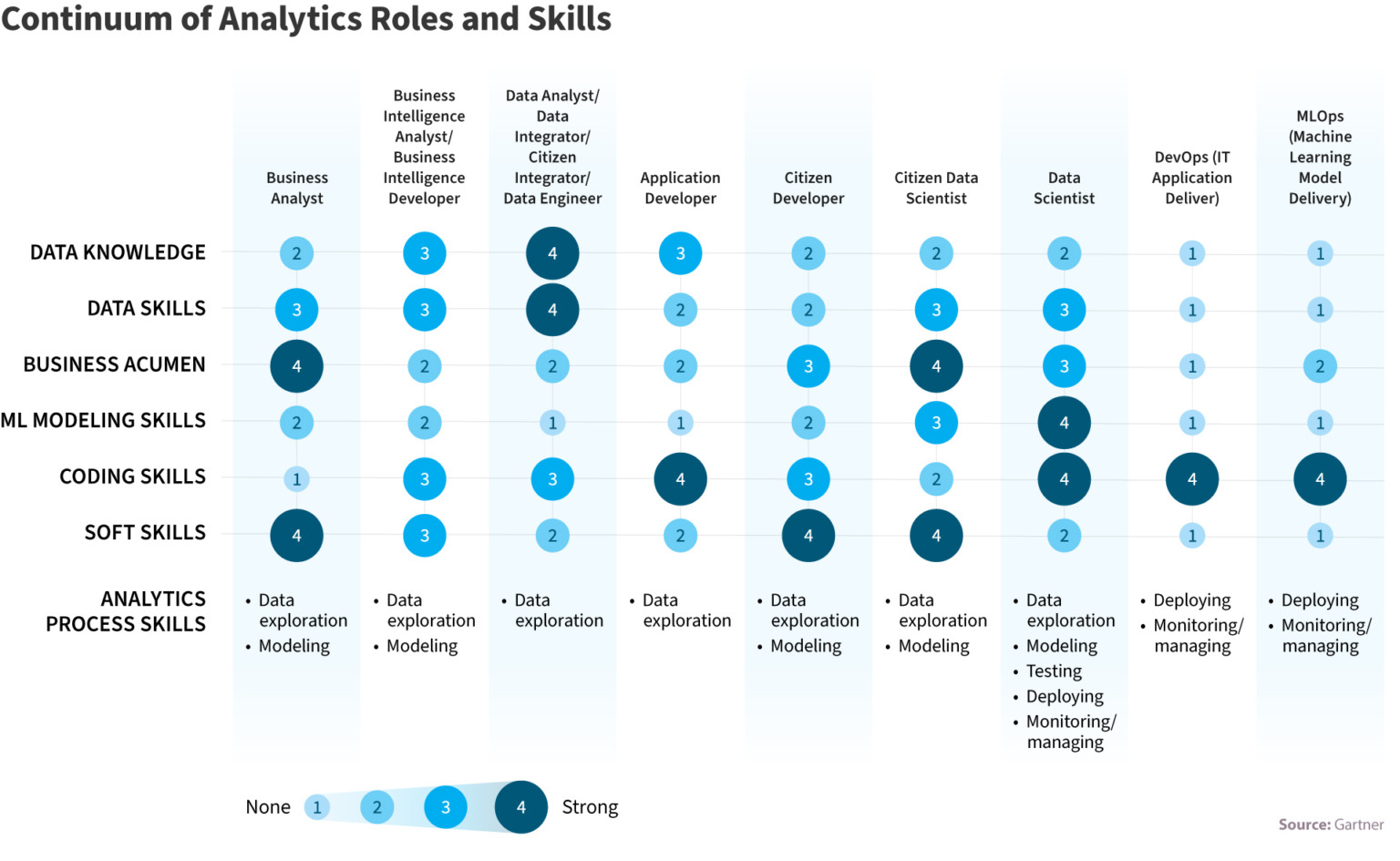
5. **Data Gathering**

Both these roles are in fact, similar in a lot of ways, since both involve data gathering, inference accumulation and data modeling. The scope of data science and business analytics often overlap and the skill sets are not mutually exclusive

**Explain various roles in a Data Science project**

When thinking about data science team roles, there are two things to consider. There are two types of data scientists. Type A data scientists look after analysis. These are the data scientists that work with data and look after data cleaning, modeling, and forecasting.

Type B data scientists are strong software programmers with good engineering skills. Type B scientists are responsible for building and as such, they build recommendation systems as well as use cases. Within any organization focused on data science, you can expect to have the following data science roles in place.



1. **Chief Data Officer/Chief Analytics Officer**

When building a data science team, this role is a critical one. The Chief data officer (CDO) looks after lots of different data-related functions. This includes areas of focus like data quality and data management as well as the creation of the overall data strategy. The Chief data office and chief analytics officer (CAO) are both unique roles, however, based on the organization, they could be filled by the same individual.

2. **Business Analyst**

A business analyst has the same role as a Chief analytics officer but their focus is more tactical versus strategic. They use data to determine project requirements and deliver recommendations and reports to stakeholders.

3. **Data Architects and Data Engineers**

Data architects and engineers work together to build a solution. The architect visualizes the requirements for the framework, while the engineer builds the digital framework.

4. **Data Analyst**

The data analyst looks at data that has been collected and makes sure that it is useful and comprehensive. The analyst is responsible for interpreting the data so many businesses look for an analyst with strong visualization skills.

5. **Data Scientist**

Data scientists fulfill a dual role. They have the skills needed to solve complex technical issues, but they also have a natural curiosity so they know what questions need to be asked. Data scientists can develop ML models, but they require access to copious amounts of data. Having access to data helps the data scientist detect patterns and relationships helping them build theories.

6. **Machine Learning Engineer**

Machine learning engineers are distinct from data scientists. Machine learning engineers combine software engineering skills with machine modeling abilities. They determine which model to use and what data is required for the model.

7. **Business Intelligence Engineer**

This role is somewhat unique as it’s not a requirement for all data science teams. However, with specialized data science models, the role of the data visualization engineer is crucial. Successful data visualization engineers need to have a solid foundation of UI skills to help create unique data visualization elements for stakeholders. He also defines which metrics and charts would be the most beneficial for business.

**Describe following Data Science Case Study/Application for the following domain:**

**Customer Sentiment Analysis (Retail)**

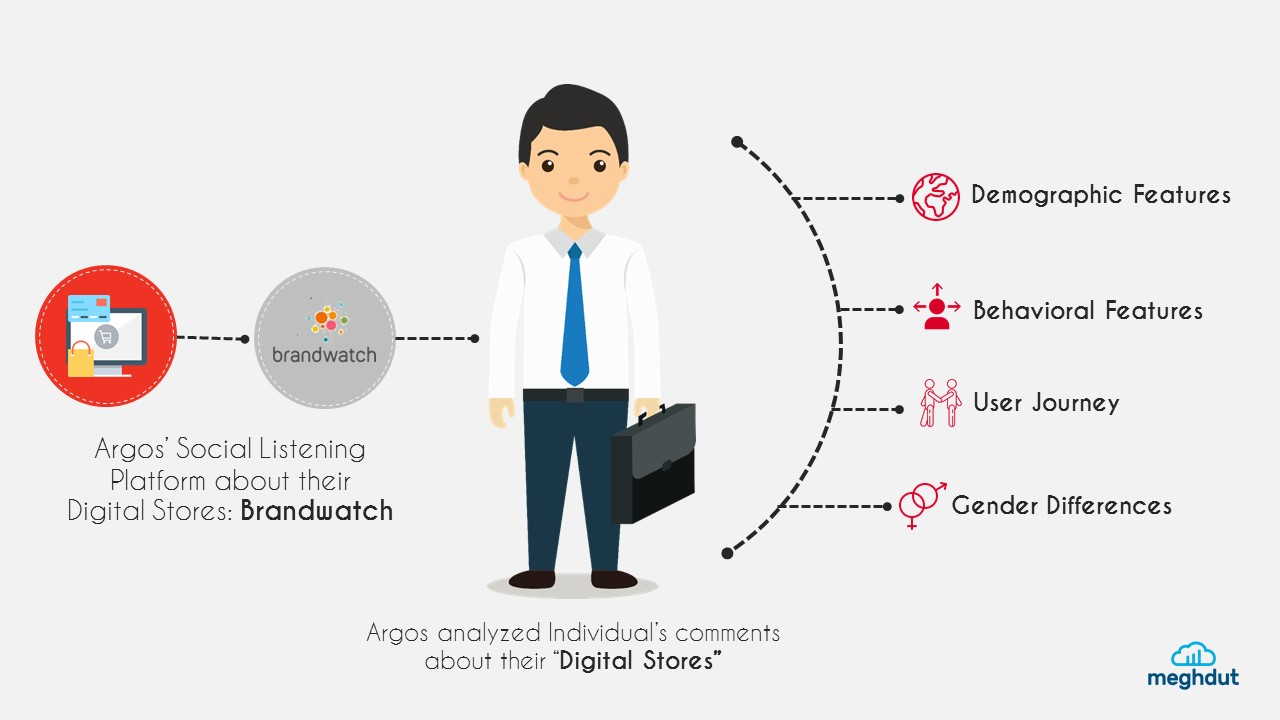
Customer experience is at the heart of any business. But for retailers, it cannot get any truer than this. In this digital era, information has become readily accessible for customers. With a simple click, customers can learn about a host of information about retail stores, ranging from product availability, discount offers to customer feedback, payment details and so on. Today’s shoppers are socially connected individuals, they feel empowered to voice their opinion about anything in social media. Retail industry players should learn how to convert this wealth of social data to achieve business objectives.

The public perception of certain aspects of their business can be tracked. Based on the findings, retailers should make modifications to delight the customers. For example, if a retail chain makes any change in store design or loyalty offers, they can track and analyze social data to understand if this change is working favorably for its business. Enhancing customer experience should be a continuous process and it has to be driven by data insights.

**Customer Sentiment analysis** is not just limited to the Retail Industry but also plays a very significant role in many other industries. The implementation of Data Science in retail has made it much simpler.

Customer sentiment analysis is performed by using several complex Machine Learning algorithms. The algorithm works on the customer data that is collected through various social media platforms and online service feedback. So that they can understand the customer’s attitude towards the product. This analysis makes use of language processing for identifying the words that show either a positive or a negative attitude of the customer. All the detected customer sentiments are analyzed and the output represents the overall sentiments of the text.

**Argos** is the United Kingdom’s one of the leading retailers for toys, home furnishing and personal care offering more than 43,000 products. In 2013, the company wanted to embrace a **digital-first** approach, so they went for a colossal project of opening 53 new digital stores across the UK.



But a huge challenge also came up with this project. Argos wanted to know their customers' reaction to this heavy change in store’s look and feel. They decided to track and monitor what customers were saying about them in social media channels. **Brandwatch Analytics** became their provider of social listening platforms. The platform offered sentiment analysis features which allowed Argos to understand people’s general sentiment about the change. Other significant insights that Argos received with this platform are:

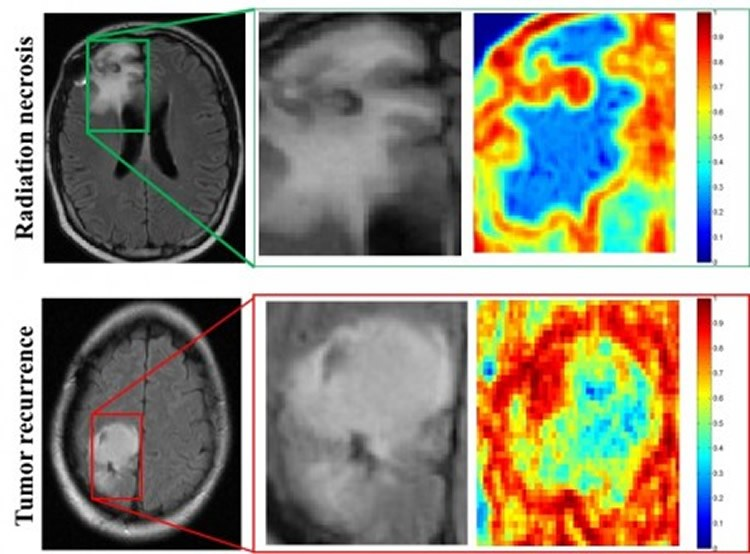
* Demographic and behavioral features of the individuals who commented about Argos
* Difference of sentiments according to gender, region and user journey

The social monitoring data provide Argos some valuable inputs. For example, the male population welcomed the digital change more cordially than their female counterpart. Londoners were more favorable to the new concept of stores. Overall, this data continuously helps Argos to mold their operations to increase customer satisfaction level and boost profitability.

**Medical Image Analysis (Healthcare)**

Medicine and healthcare together form a promising field for utilizing technological advancements. The healthcare sector is acquiring new heights due to the advancements in Data Science. It is helping in various aspects, and one of them is the analysis of medical images. It is one of the most interesting areas of study in image recognition technology. IBM estimates that the medical images contain around 90% of the overall medical data. Doctors use the medical imaging technique to effectively visualize the interior parts of the body.

Data Science helps in the recognition of scanned images to figure out the defects in a human body for helping doctors make an effective treatment strategy. These medical image tests include **X-ray**, **sonography**, **MRI** (Magnetic Resonance Imaging), **CT scan**, and many more. Proper analysis of the images of these tests helps gain valuable insights for the doctors to provide the patients with better treatment.



Numerous methods are used to tackle the difference in modality, resolution, and dimension of these images. Many more are being developed to improve the image quality, extract data from images more efficiently, and provide the most accurate interpretation.

These are the general imaging techniques. But, the involvement of Data Science has made these imaging techniques further revolutionize the healthcare industry. There are various methods in Data Science that find the differences between the states of image and resolution and check the orthogonality. Data Scientists are working on creating more advanced techniques to improve the quality of the image analysis so that the patient’s data from an image is extracted efficiently.

There is a recent study published by **Google AI** on diagnosing skin diseases using Deep Learning. The Deep Learning model is created in such a way that it can diagnose **26** diseases related to skin with an accuracy of **97** per cent. The diagnosis is performed using deep neural networks, Machine Learning, and Data Science. Now, let us look at the three common algorithms used in medical image analysis:

1. **Anomaly detection algorithm** - This algorithm helps in identifying conditions such as bone fracture and displacements.
2. **Image processing algorithm** - The image processing algorithm helps in analyzing images and enhancing and denoising them.
3. **Descriptive image recognition algorithm** - It visualizes and extracts data from images, interprets it, and makes use of it to form a bigger picture (for example, merging the images of the brain scan and designating them accordingly).

Several supervised and unsupervised Machine Learning algorithms can be used for this purpose. Using some Deep Learning based algorithms provides the result with more accuracy. The deep-learning based algorithms increase the diagnostic accuracy by learning from the previous examples and then suggest better treatment solutions.

The most popular image-processing techniques focus on enhancement, segmentation, and denoising that allows deep analysis of organ anatomy, and detection of diverse disease conditions. Some of the advanced applications are cancer detection, tumor detection, etc. One of the most popular analytical frameworks, Hadoop contributes to image analysis in several tasks by using Machine Learning algorithms like SVM, Wavelet analysis, etc.