The Key Questions in Data Sciences and Machine Learning – A Literature Review

Aman Kumar, Nidhi Upadhyay, Ankita Singh, Ankit Raj

***Abstract*— Amongst one of the most emerging fields of computation and sciences, data sciences have proved their metal in the world today. With the blending and convergence of multiple technological sectors and human life, the amount of data generated have increased exponentially today, making data the new oil and amongst one of the most exquisite resources available. With problems ranging from medical sciences to addressing problems of business intelligence, the application of data sciences in various domains is accepted as a major factor for decision making which is widely accepted now. This multi-disciplinary field has concepts overlaying with data-driven technologies like Big Data, Machine Learning, Statistical Inferences, Cloud computing and mathematics. Undoubtedly just not being a tool-oriented field, the multi-disciplinary field of data sciences, today has become a practical and practice-oriented field. It has been these features, which makes data science amongst the most demanded skills of the 21st century, with very high demands of skilled professionals. With analysis of the use of data science and its trends, this research paper highlights the various general and common terminologies, analogous to this field and throws light on some key areas where data-driven decision-making methods are used widely. With understanding the basic methodology of the decision-making policy, the paper also studies the availability of open-source tools that could be utilized effectively in this field. Since all of the areas of science where data analysis has been the priority of research and data is the centre of focus are evolving rapidly, it has moulded data science into a must to have skill. This not only is improving the decision-making capabilities but helping individuals and stakeholders to experience and answer the most challenging questions with the best data-based answers.**

Keywords*— *Data Sciences, Machine Learning, Big Data, Statistics, Medical Sciences, Business Intelligence

I. INTRODUCTION

Availability of data has become one of the most critical resources of economic currency in the world today. IBM predicts that by 2020, 28% of the digital jobs would be using concepts of data science and machine learning. The Quant Crunch report of IBM states that – "Machine learning, big data, and data science skills are the most challenging to recruit for, and can potentially create the greatest disruption if not filled.". Today as high as 45% of the vacancies available in the field of data sciences go vacant [1]. Concepts of Data Sciences and Machine Learning, in a very short period, are becoming part of the education curriculum for students and professionals across the world. With the research done, methods have also been suggested that could be used to design an effective curriculum for Data Sciences for academic teaching [2,3]. Methods like this could help to bridge the gap between unskilled and trained workforce [4]. With the data generated today at such a rapid rate from varied sources, the usage for the same is increasing at a rapid pace. Sectors as in healthcare and using methodologies as seen in biostatistics, epidemiological modelling is done today for an analysis of COVID-19 patients. For businesses, problems of facility location could be addressed using methods of data science using tools of spatial data analysis. Forecasting of sales, weather prediction patterns, sentiment analysis and natural language processing, all make use of various tools used in machine learning and data sciences [5]. This knowledge discovered by analysis of data is key to solving problems, hence the surge in the demand of professional data scientists now is justifiable.

Aman Kumar, Amity School of Engineering & Technology, Amity University, Patna, India.

Nidhi Upadhyay, Amity School of Engineering & Technology, Amity University, Patna, India.

Ankita Singh, Amity School of Engineering & Technology, Amity University, Patna, India.

Ankit Raj, Gaya College, Magadh University, Gaya, India.

II. BACKGROUND

Implementation of data sciences and collection of data for problem analysis hasn't been a new field. Analysis of real-time data with the help of data science methodologies has been observed earlier in fields like - Epidemiological and Clinical Studies [6,7] and Traffic Conditions and stock markets [8,9]. The methods and tools used here process data in a pipeline method to address the underlying constraints. Later with the help of tools available in statistics, the hypothesis is verified to draw a conclusive response for the problem. Methodologies as such, have increasingly been used today with the advancements in edge and cloud computing which had enabled parallel processing of data, earlier that wasn't achievable.

A. Data collection and availability

This research makes use of text references and data for trend and pattern analysis which is fetched from multiple open-sourced services. The datasets fetched are further wrangled for building graphs for visual analysis. The following are the set of datasets used in this research study:

- Google Trends Search Result Patterns [10]
- Kaggle 2020 Data Science Survey Dataset [11]

The research involves the description of the various fields and sub-fields in data sciences with analysis of tools used in the domain of data sciences and machine learning. Inferences are drawn from the open-sourced Kaggle Surveys on data science. Textual references of about and methodologies involved in this field are also mentioned under subsequent subtopics in this paper.

B. Data Selection

The following data was fetched from the source datasets and were used for this research:

- Google's Search Trends dataset. Fetched from API calls over Python to know the search trends of "data science" and "machine learning" keywords over time, on Google Search Engine.
- Kaggle Data Science Survey 2020. This dataset contains the list of common questions on tools, topics, implementation methods, knowledge and background of professionals who are working in the data science domain

III. ANALYSIS AND AREAS IN DATA SCIENCES

A. Evolution of Data Sciences in the last 5 years.

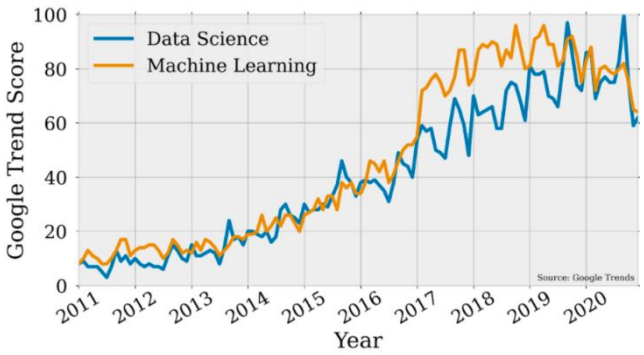

Fig.1 – Google Search Trends for keywords – "Machine Learning" and "Data Science" between 2011-2020.

Using Google Trends, the data on search trends for data science and machine learning across the decade was fetched. The data could easily be loaded by an API call over the Google Trends report. From analysis from Fig.1, in 2011-2012 the Google Trend Score for the searched keywords – "Data Science" and "Machine Learning", had a trend score of 15.

With the latter half of the decade, the search trends for the keywords have increased rapidly as the sectors of data science and machine learning boomed. The emergence of data science as a must-have skill is associated with the perks it brings with it. Analytical and informative decision making is the need of the hour, which is achieved best when knowledge of data science is brought into action.

B. The terminology – Data Science

The necessity of Data scientists in the 21st century is at an all-time high, resulting in a flourishing career in this field. Data Science produces numerous opportunities for freshers as well as professionals. It leads to more comprehensive analysis in Business development, enhanced operational efficiency, prediction and minimized risk vulnerability through accurate forecasting models. For being a Data Scientist, we need a lot of skills and practice involved in this domain. Data Science, as a career opportunity, is expected to grow exponentially for the next three decades and will generate millions of jobs in coming years.

Simply said, the science of using data to answers the questions of interest is called Data Sciences. It has emerged as the domain of application of knowledge where large and big volumes of data are fetched and analyzed with the modern tools in place to see the underlying patterns in it. These underlying patterns could be used later to build predictive models and derive meaningful answers to the problem which we try to analyze. In technical terms, data science is said to be an interdisciplinary branch of science which uses algorithms and process to derive insights from structured and unstructured datasets. This field is closely related to data mining, big data and machine learning [12]. The field has seen an unparalleled interest in the second half of the decade (Fig.1), which has contributed to the demand for skilled people. Machine learning, mathematical modelling, statistics, programming, databases and the act of data storytelling combined with knowledge of data ethics are the prerequisites for anyone who is thinking or is making a career in the field of data sciences.

C. The educational background of Data Scientists

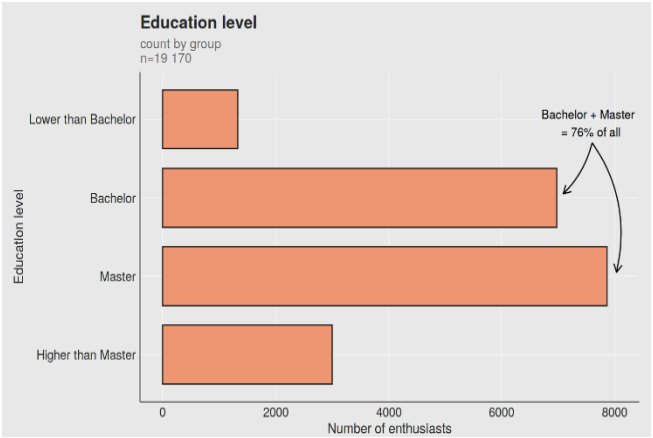

Fig.2 – Education Level vs the number of people interested in learning data sciences and machine learning.

Using the Kaggle Data Science Survey 2020 dataset, the graph for 19170 respondents is plotted. The outcome for the same is present in Fig.2 which shows the education level of respondents vs. the no of people who were keen to learn data sciences and machine learning. 76% of the data science enthusiast respondents have either a bachelors or master's degree. However, interest in learning data sciences has also been observed from people having no undergraduate degree.

D. Is there an age to learn Data Sciences and Machine Learning?

One of the key questions while starting learning data sciences and machine learning today is the age. People often have confusions on topics like – "Is it too soon to begin?" or "Am I too late?". With the Kaggle Data Science Survey 2020, 19637 respondents age were recorded who was keen on learning data sciences and machine learning, with the dataset and with the use of libraries plots a graph that records to count of respondents, their age and education level.

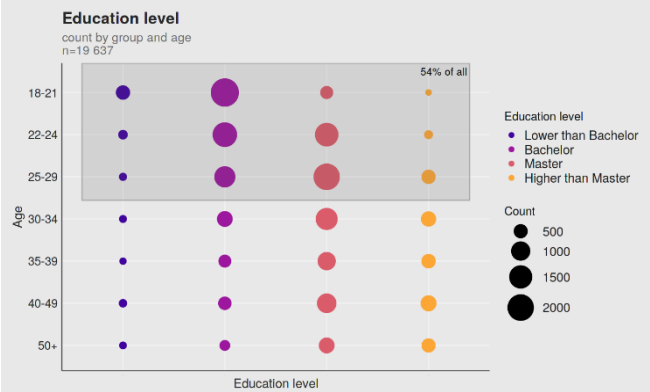

Fig.3 – Education Level vs Age of Data Science Enthusiasts.

From Fig.3, the age bracket of 18-29 years has the highest count. People across this age group shows a keen interest in learning data science and machine learning. This highlights that there isn't a specifically defined age to begin acquiring data science skills.

E. Prerequisites in learning Data Sciences & Machine Learning

Fig.1 in this research paper is generated using the codes written in Python Programming Language, while Fig.2 and Fig.3 are generated with the code written in R Language. The choice of the best programming language with the confusions on what are the essential prerequisites in learning data sciences is amongst other questions one faces while getting started into this field. Below are some of the essential prerequisites one must know before learning data sciences.

- **Machine learning:** Considered as the skeleton of data sciences, the concepts of the mathematics behind machine learning algorithms (regression, classification and clustering) is a must knowledge to have.

- **Linear Modelling:** A data scientist must know how to frame equations from the problems provided. Knowledge of mathematical modelling, linear algebra, vectors and matrices amongst the best methods to learn this concept.

- **Statistics:** Concepts of probability, measures of central tendency, association and dispersion is considered as the core of data sciences. Knowledge on these topics is a good add-on helping to achieve accurate results.

- **Databases:** Since it's the database where the data is stored, a familiarity with Query Languages, relational and non-relational databases should be present.

A Data Scientist is expected to have knowledge and exposure to programming languages such as Python, R, Scala, along with extensive know-how of the database systems. With the help of programming language, the working logic of algorithms is designed for a model. Such algorithms fetch and modify data from its source and use it for machine learning purposes. Python and R is the most preferred language to write a code and analyze on data. The choice of programming language to learn data science is subjective and often it drills down to personal preferences. With Python and R being the most popular languages to implement data science and machine learning algorithms, any person who has the technical know-how of either of them is good, to begin with.

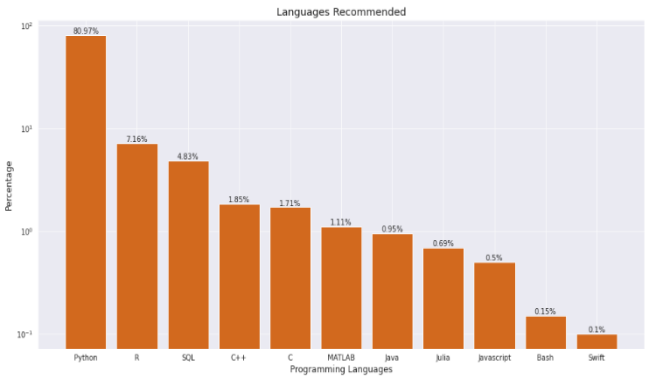

Fig.4 – 1st Programming Language preference for data science workflows of people for Kaggle Data Science Survey 2020

From Fig.4, 80.97% of respondents have Python Programming Language as their 1st preference for data science workflows. The easy learning curve of python and the availability of open-source libraries have in majority contributed to this trend.

F. What coding experience is required to get into Data Science?

Using the Kaggle 2020 Data Science Survey Dataset, for 8742 respondents the data on coding experience was noted. Later with the help of Python, a heatmap was plotted (Fig.5). Fig.5 shows the number of respondents. For X-axis the age of the respondents was highlighted, whereas for the Y-axis coding experience (in years) were recorded. The respective count of the respondents was present as cells in the heatmap.

- For the age bracket 25-29 years, 3-5 years of coding experience was noticed. 17.57% of the respondents of the survey had 1-5 years of coding experience for the age bracket 25-29 years.

- With age, the coding experience tends to increase.

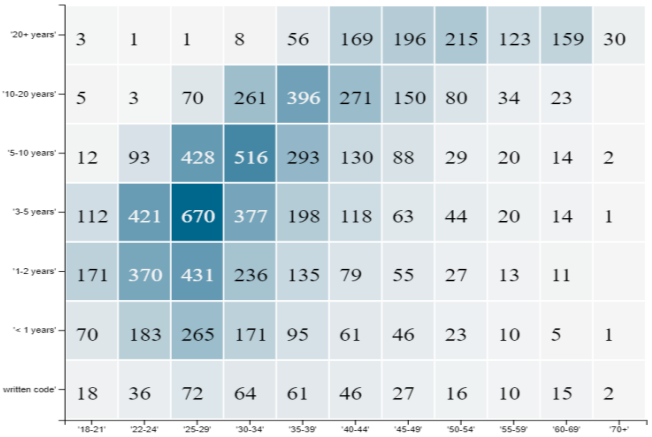

Fig.5 – Heatmap for the count of no of respondents for years of coding experience vs. age of respondents.

The majority of the students and new learners of Data Science fell under the age bracket of 22-29 years. 28.63% of the respondents in this age bracket had coding experience

between 1-5 years. With deeper analysis on the Dataset, we get to know the following granular details for the heatmap:

- Database Engineers and Research Scientists working in the domain of Data Sciences seems to be having a high coding experience than others.

- 44% of the respondents who were Database Engineers and Research Scientists reported 5-10 years of coding experience. Also, 65% of the same respondents reported 3-5 years of coding experience.

Coding experience is one of the most sought-after factors in Data Science. With 3-5 years of coding experience being common among students, it is generally observed that coding experience increases with time in this domain. Hence, even with a few years of coding experience in languages as Python, R, C or C++ is a good to-go-to factor while getting into Data Science.

G.How does the daily workflow of data scientist look like?

The main job of a data scientist is to derive a data-driven conclusion by extracting key insights and patterns from the data. For a provided problem, a data scientist typically:

- ✓ Develop a set of questions to be answered.
- ✓ Recognize what data would be important to collect.
- ✓ Collect data across the sources to solve the problem.
- ✓ Process the raw data with mathematical methods.
- ✓ Develop a machine learning algorithm
- ✓ Feed clean data to the machine learning algorithm.
- ✓ Check if the model is ethical and explainable.
- ✓ Share the results with stakeholders.

To achieve this workflow in a streamlined process, knowledge of multiple tools proves often to be beneficial:

- **Spreadsheets:** Tools like Google Sheets and Microsoft Excel could help in basic data wrangling and getting it cleaned for analysis.

- **Visualizations:** Knowledge of Tableau, PowerBI could help to make charts and dashboarding.

Data dashboarding with the help of data storytelling is one of the key communications and presentation skills to have when one is building a career in data sciences. These soft skills help to communicate findings of analytic processes to stakeholders. Problem definition, data investigation, model development and deployment with iterative enhancements and regular updates is how a typical data science life cycle seems.

H.The career prospective and areas where Data Science & Machine Learning is used:

Glassdoor reports the average salary of a data scientist in the United States as $113,000/annum and in India 9,07,000 Rupees/annum. Also, the U.S. Bureau of Labor Statistics predicts Data Science would alone create as much as 11.5 million jobs by 2026. The areas where Data Science and Machine Learning has highlighted unprecedented growth are:

- **Healthcare** – Using the concepts of biostatistics and modelling, the pandemic spread could be calculated. Computer vision tools could be implemented for disease identification at very early stages.

- **Business Intelligence** – Problems as facility location, logistics, profit maximization, stock market predictions are carried best with data science tools.

- **Recommender Systems** – These systems help to suggest appropriate content to the end-users to maximize sales or viewership. Online video streaming platforms to e-shopping websites, the fundamental method to maximize sales by these companies is done using recommender systems.

- **Education** – Predictive analytics methodologies helping in student evaluation without bias. With advancements in AI content delivery is much better today.

- **Defense and Border Surveillance -** reduce the risk of failure of current systems with prior prediction, helpful in gaining a better understanding of the anti-national elements, receiving unique insights, aiding in efficient expansion, and improving event predictions.

Not just restricted to these domains, with the proper blend of data sciences and machine learning almost the majority of the sectors where data-driven decision making is an essential process in business, concepts of data science, machine learning and business intelligence could prove its metal.

I. What are the few challenges in getting a Data Science job?

Some of the key challenges in getting a data science job today involves the following aspects:

- Lack of expertise knowledge
- Non-accuracy of datasets
- Collaboration with data Engineers
- Understanding of the Business problem
- Misconceptions about the role
- Working with worst algorithms

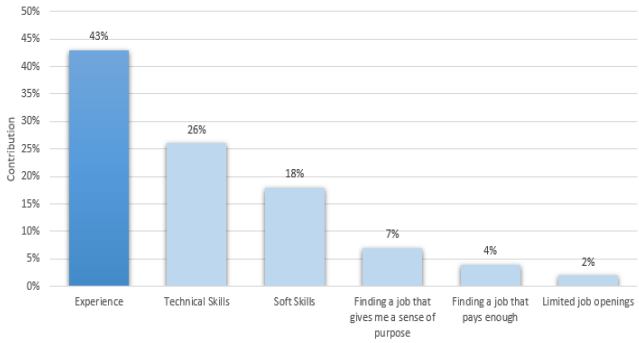

Fig.6 – Biggest obstacles in landing a Data Science Job

Along with the topics mentioned, the Anaconda Data Science report says experience amongst the biggest obstacles in landing a job.

Landing a Data Science job and understanding the job market in Data Science is equally essential to excel in this domain. Often, there are multiple factors preventing novice learners to land a job and secure an ideal position of responsibility in this domain. Anaconda's State of Data Science Report 2020 shed light on the factors and obstacles in the way for students and learners to get a job in Data Sciences. The reported responses are highlighted in Fig.6. The following are the important patterns captured from this data:

- With ample job opportunities available, the number of positions open for this domain isn't a factor of concern for candidates seeking a job.
- 43% of company respondents reported lack of experience as the major concern in this field. This is particularly an important metric as Data Science job opportunities demand experience.

Lack of experience hence becomes the biggest factor, because of which multiple Data Science offerings go vacant. Particularly for students, this becomes a more mundane problem. With no straight solution for this - research work, internships and side-projects involvement are amongst the probable solutions for this problem. These areas both increases the skills, technical know-how as well as increases experience among students.

IV. CONCLUSIONS

The upsurge of interest in the field of data sciences and machine learning has not only to be observed academically but also socially. Search trends of Google search for these keywords have observed a rise lately in after 2015's, with more people keen to learn and contribute with their skills in this field. With data science being a multidisciplinary field attracting people from multiple domains, survey results have highlighted that people in graduate and postgraduate years of study are amongst the most enthusiastic people, keen to learn and practice data sciences.

People under the age group 18-29 years were amongst the keen learners of data science, proving the fact that there isn't any defined age to start learning data sciences. With mathematics, statistics as one of the key prerequisites in learning data sciences and machine learning, knowledge about linear modelling and databases are amongst the must-have skills. For programming languages, Python and R have a great inclination to process the data science workflows. With the set of sequences of steps for typical data science and machine learning project workflow and tools used, any individual could leverage the best of this world to build a prospective career in this sector.

V. DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationship that could have appeared to influence the work reported in this paper.

VI. ACKNOWLEDGEMENTS

We would like to thank Kaggle Platform for access to relevant 'Data Science – Survey' datasets and Google Trends (2020) for data on Google keywords search patterns. The research would not have been possible without this support.

VII. REFERENCES

[1] IBM Quant Cruch Report, IBM (2020). Retrieved from: https://www.ibm.com/downloads/cas/3RL3VXGA

[2] Y. Demchenko et al., "EDISON Data Science Framework: A Foundation for Building Data Science Profession for Research and Industry," 2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom), 2016, pp. 620-626, doi: 10.1109/CloudCom.2016.0107.

[3] Z. Qiang, F. Dai, H. Lin and Y. Dong, "Research on the Course System of Data Science and Engineering Major," 2019 IEEE International Conference on Computer Science and Educational Informatization (CSEI), 2019, pp. 90-93, doi: 10.1109/CSEI47661.2019.8938944.

[4] E. W. Bethel, "Towards a Data-Centric Research and Development Roadmap for Large-Scale Science User Facilities," 2017 IEEE 13th International Conference on e-Science (e-Science), 2017, pp. 462-464, doi: 10.1109/eScience.2017.72.

[5] C. K. Leung, Y. Chen, S. Shang and D. Deng, "Big Data Science on COVID-19 Data," 2020 IEEE 14th International Conference on Big Data Science and Engineering (BigDataSE), 2020, pp. 14-21, doi: 10.1109/BigDataSE50710.2020.00010.

[6] C.K. Leung et al., "Data science for healthcare predictive analytics", IDEAS, pp. 8:1-8:10, 2020.

[7] J. Souza et al., "An innovative big data predictive analytics framework over hybrid big data sources with an application for disease analytics", AINA, pp. 669-680, 2020.

[8] Y. Huang et al., "Diffusion convolutional recurrent neural network with rank influence learning for traffic forecasting", IEEE TrustCom/BigDataSE, pp. 678-685, 2019.

[9] K.J. Morris et al., "Token-based adaptive time-series prediction by ensembling linear and non-linear estimators: a machine learning approach for predictive analytics on big stock data", IEEE ICMLA, pp. 1486-1491, 2018.

[10] Google Trends – Explore what the world is searching, Google. [Online]. Accessed: May, 2021. Retrieved from: https://trends.google.com/trends/?geo=IN

[11] Kaggle Survey 2020 dataset, Kaggle [Online]. Accessed: May, 2020. Retrieved from https://www.kaggle.com/c/kaggle-survey-2020/data

[12] Hayashi, Chikio et al. Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Japan. pp. 40–51. doi:10.1007/978-4-431-65950-1_3. ISBN 9784431702085.

Mr. Aman Kumar is currently a student at Amity School of Engineering and Technology, Amity University, Patna (India). He is also associated with Indian Institute of Technology, Madras (India) as a student at Computer Science Department. His research interests lie in the field of Data Sciences, Big Data, Machine Learning and Cloud Computing. Aman expertise in Data Visualization and Data Analytics, he is an IBM Certified Data Science Professional. He is certified on Oracle Cloud Infrastructure as an Architect | Developer at Professional and Associate levels and is having a keen interest in Microsoft Azure and Amazon Web Services.

Ms. Nidhi Upadhyay is a web developer enthusiast with a penchant for exploring more web tools and keen to learn new technologies. Currently pursuing B. Tech in Computer Science from Amity University, Patna, Nidhi has obtained a wealth of knowledge in python libraries, bootstrap framework, CSS, HTML, and JS. She loves and is known for applying responsive design principles in web pages and want to keep her creative flair intact. She is currently working over some datasets for machine learning-based projects and grabbing some programming knowledge with concepts, and frameworks to brush up her skills.

Ms. Ankita Singh is currently a final year student at Amity School of Engineering and Technology, Amity University, Patna (India), pursuing her Bachelors in Technology. A keen problem solver, Ankita if found of contributing her skills to solve societal problems. With interest in Data Science & computer networks, her research interests lie in the field of Cloud Computing, Data Sciences, Machine Learning and Computer Networks.

Ankit Raj is a companionable and instinctual person who loves to interact with technology to find out the hands-on solution of obstacles occurred in the real life of the community. Ankit Raj has a keen interest in computer programming, Data Science and Machine Learning. Currently, he is in 2nd year, pursuing a Bachelor of Computer Application (BCA) in Gaya College, Gaya (A constituent of Magadh University at Bodh-Gaya). At present he is deeply going through key concept of Data Sciences and Machine Learning as well as finding solution of various other real-life obstacles which can be solved through AI technologies.