Journal of Modern Technology and Engineering
Vol.8, No.3, 2023, pp.235-243

Jomard Publishing

COMPARATIVE ANALYSIS OF DATA DRIVEN TECHNOLOGIES OF ARTIFICIAL INTELLIGENCE

Resmiye Nasiboglu

Dokuz Eylul University, Faculty of Science, Department of Computer Science, Izmir, Turkiye

Abstract. Artificial Intelligence (AI) technology is indisputably the one of the most popular and effective technologies of recent years. AI technologies can be broadly divided into two large groups: logic-based or rule-based approaches and data-driven machine learning approaches. In particular, data-based methods have been developing more rapidly in recent years. The theoretical basis of these methods is generally statistical analysis. Other fields in this group are data mining and machine learning. In this study, the characteristics of statistical modelling, data mining and machine learning, which are data-driven fields of AI, are examined and their comparative analyzes are made. For clarity, the analysis results are presented in tables. It is clear that this study will be especially useful for young computer scientists working in the field of Artificial Intelligence.

Keywords: Artificial intelligence, Data-driven approaches, Statistical learning, Machine learning, Data mining.
Corresponding author: Resmiye Nasiboglu, Department of Computer Science, Faculty of Science, Dokuz Eylul University, Izmir, Turkiye, e-mail: *resmiye.nasiboglu@deu.edu.tr*
Received: 8 August 2023; Revised: 12 november 2023; Accepted: 3 December 2023;
Published: 30 December 2023.

1 Introduction

The use of modern computers has accelerated large-scale statistical computations and also made possible new methods for analyses that were impractical to perform manually. Today, in all areas involving decision making, artificial intelligence (AI) technologies based on computer science and statistical methodologies are applied to make accurate inferences from aggregated data and to make decisions in the face of uncertainty.

The science of statistics actually emerged as a generalized concept that includes the application of applied statistics, descriptive statistics and inferential statistics. Statistical theory is used to understand and realize the logic underlying statistical inferences. Mathematical statistics is used to improve various aspects of computational statistics and experimental design, as well as the manipulation of probability distributions necessary to obtain results relevant to prediction and inference methods (Nikoletseas, 2014; Anderson et al., 1994; Breiman, 2001).

Fields of AI such as data science, data mining and machine learning, which are popular today, can be described as a new wave based on statistics and computer science. But of course, not only statistical science but also rapidly developing computer technologies have made important contributions to the emergence of these fields. Statistics, data mining and machine learning are also closely related to fuzzy logic and fuzzy set theory, which is another field of science that focuses on decision making under uncertainty. In recent years, there have been many theoretical and applied studies in this common area (Zadeh, 2011; Nasibov, 2011; Nasiboglu, 2022; Nasibov and Nasiboglu, 2022, 2023).

In this paper, the basic fields of AI such as data science, data mining and machine learning are discussed from various perspectives, and their comparative analyses are given.

2 Rule based versus data driven approaches

We can divide the techniques developed to solve Artificial Intelligence problems into two main headings: Approaches developed based on data, i.e. data driven approaches and approaches developed based on rules or logical inferences.

The statistician's approach to problem solving can generally be summarized as a model-based approach. In order to collect the data required for the solution of a particular problem, it is first decided which data and how much data should be collected. In other words, with a statistical approach, it is first decided which model to use and data is collected accordingly. Classical statistical solutions are generally considered for situations where the number of data is not very large. Many classical statistical analysis methods have been developed depending on whether the number of observations is less than or more than 30. The process of model building in statistics generally aims at the creation of a specific mathematical function. The inputs and outputs of the observations are used to optimally adjust the parameters of the function. In short, in the statistician approach, the function patterns to be used in the model are certain, and observation data is used to fine-tune them. Various functions (models) are tried when necessary. This can be called a mathematical or statistical-based, i.e. model-based approach.

Approaches such as data science, data mining and machine learning can normally be described as data driven approaches. There is much more data and it is more important to reduce the number of observations and data size to be used. In the creation of models, algorithmic approaches come to the fore, not mathematical functions. The model is no longer created as a mathematical function, but as an algorithm. It is possible to find algorithms (models) that give more accurate results by trying various combinations of different features with different algorithms.

3 Statistics and tasks of statistical studies

Sciences such as data science, statistics, data mining, machine learning are interesting disciplines based on data analysis that help businesses and administrators make optimal decisions and positively impact the growth of the business. General definitions of these disciplines are given below and their main features are discussed.

3.1 Statistics

Nearly all data mining and machine learning algorithms are theoretically based on statistics. Statistics is the science of collecting, investigating and analyzing data and making inferences and predictions about the future. Statisticians are interested in designing surveys and experiments to obtain quality data that can be used to make predictions about the whole population. The tasks of statistics can be formally listed as follows:

1. Survey and experiment design;
2. Summarizing and understanding data;
3. Predicting population behavior;
4. Ability to make predictions or predictions.

Statistics is often used as a technique of summarizing numbers to find descriptive statistics such as mean, median, mode, standard deviation, variance, percentiles and hypothesis tests. The basic aim of a statistical research project is to investigate causality between dependent and independent variables. Particularly, it draws a conclusion regarding the effect of changes in the values of predictors on the dependent attributes.

3.2 Tasks of statistical studies

Since the science of statistics is completely data-based, how data is collected significantly affects the direction of statistical studies and the tools to be used. For example, when complete census data cannot be collected, statisticians collect sample data by developing specific experimental designs and survey samples. Statistics itself also provides tools for forecasting and prediction through statistical models. One of the most effective tools used in statistics is sampling theory. Sampling theory is part of probability theory. Mathematical statistics and probability theory are used to study the sampling distributions of statistics and, more general statistical procedures. Population parameters are estimated from the statistics obtained from the sample using statistical inference techniques.

Statistical studies can be divided into two groups as experimental studies and observational studies. The basic steps in the experimental work are as follows:

1. Planning the research, including finding the number of necessary experiments of the study;

2. Random assignment of actions to design experiments that use grouping to reduce the influence of interactive variables and to provide unbiased estimates of experimental error;

3. To perform the experiment and analyze the data by following the experiment protocol.

In the observational study, on the other hand, the analysis is carried out on the data collected impromptu without designing the experiment. As an example of an observational study, a study investigating the relationship between smoking and lung cancer can be taken. This type of study typically uses a questionnaire to collect observations related to the area of interest and then performs statistical analysis.

4 Major data driven approaches of AI

One of the main reasons why statistical modeling, data mining and machine learning models, which are data-oriented approaches, have come to the fore in recent years is the increased ease of collecting data. In particular, the development of data-based technologies has accelerated considerably as a result of the 4th industrial revolution (Industry 4.0), the widespread use of the digitalization process, the Internet of Things (IoT), etc. It is obvious that this revolution will be a very comprehensive change, unlike any revolution humanity has ever experienced before. In this regard, everyone from the public and private sectors to academia and civil society will need to keep up with this revolution. (Schwabd, 2015) emphasized that the 4th Industrial Revolution is a technological revolution that will fundamentally change the way we live, work and establish relationships between people. Among the Industry 4.0 components, Artificial Intelligence technology has a special importance. In the study (Yablonsky, 2019), the potential relationship of artificial intelligence, big data (BD) and advanced analytics (AA) technologies with digital business platforms was investigated. In this context, a multidimensional BD-focused AI innovation classification framework has been developed, including the AA-BD-AA innovation value chain, relevant BD levels, and analytical maturity improvement. The study mainly focused on data-based human-machine relationships.

As can be seen from the above literature analysis, the usage areas of artificial intelligence are very diverse. But despite this, the data-driven approaches used in this context can be divided into main approaches such as data science, data mining and machine learning. The first two of these approaches is the approach of researchers with a statistical background, and the third is the approach of researchers with a computer scientist background. Below is information about these approaches.

4.1 Data science

Data science is an interdisciplinary scientific field that uses statistics, scientific methods, algorithms and systems to extract information or make predictions using structured or unstructured data that may contain noise. Data science integrates data analysis methods as well as domain knowledge gained from the core application area (e.g. medicine, natural sciences, social sciences etc.). Data science can be defined as a profession, as a discipline, as a research paradigm, as a research method, and as a science (Donoho, 2017; Dhar, 2013; Danyluk and Leidig, 2021; Mike and Hazzan, 2023; Hayashi, 1998; Cao, 2018).

Data science combines informatics, data analysis, statistics, and other related methods to understand and analyze data and real problem. Theories and techniques from many areas relating the mathematics, statistics, computer science, information science and domain knowledge are used by statisticians (Figure 1).

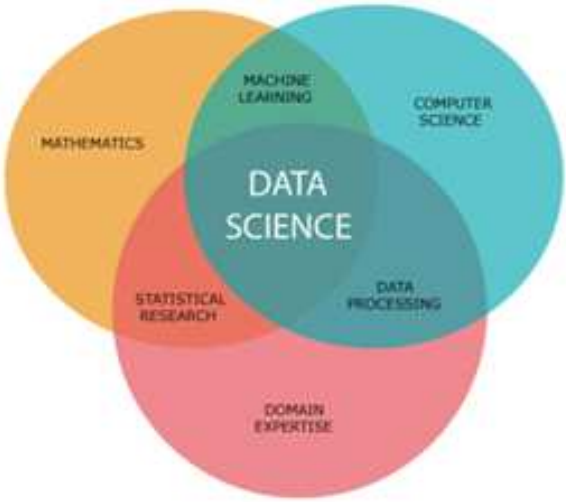

Figure 1: The place of data science among other related fields

A data scientist is someone who creates programming code and combines it with statistical knowledge to generate insights from data. But data science is different from computer science and information science. Statistics and data science are largely overlapping fields. Although they have many common methods, there are differences in the way they approach the problem. In terms of proximity, statistics is a branch of science separated from mathematics. Data science, on the other hand, is a more comprehensive science inspired by computer science and is closely related to many disciplines, including statistics (Bell et al., 2009; Davenport and Patil, 2012).

4.2 Data mining

Another concept that is closely intertwined with data science is data mining. In fact, sometimes these two concepts are used as the same thing. Like data science, the concept of data mining is closer to the statistician perspective. The model creation process in data mining is an important process, but the importance of the data collection and preparation process becomes more prominent. The created model is generally intended to be used to discover certain patterns in the data set.

Statistics generally deals primarily with numerical data. However, data sets considered within the scope of data mining may contain a mixture of text, images, audio, video, and geographic data. Finding interesting patterns hidden in the data is the main aim of data mining. Data mining is the first step of the data science study. Data mining is an area where we try to identify patterns in data and reveal the first insights. Data mining uses statistics, machine learning and database techniques to mine large databases and find patterns. In particular, it

mostly consists of techniques such as cluster analysis, anomaly detection, and association rule mining. From a data usage perspective, although data is collected for a specific purpose in statistics, one aim is to find models or patterns in data mining according to the data already available.

4.3 Machine learning

Since the machine learning approach is more of a computer scientist approach, the data collection and preparation process in this approach lags behind the model creation process. The machine learning approach focuses more on algorithms for training models from data. The created model is generally used for the automation of prediction, classification, etc. purposes in the future.

Machine learning is a field of science that focuses on ensuring that computers can learn on their own from sample data sets and predicting new incoming data (Kohavi and Provost, 1998; Lindsay, 1964; Aggarwal, 2018; Haykin, 2009; Cortes, 1995). Machine learning, as part of data science, uses the power of statistics and learns from the training dataset. Machine learning models are statistical and probabilistic models that capture patterns in data using computational algorithms. Machine learning algorithms build models using sample data to make predictions or decisions without being explicitly programmed. The model can be developed automatically through trial and error and the use of data. For example, classification, regression etc. models are created using training datasets, and the accuracy of these models is validated using test datasets.

There are general approaches used in machine learning such as Supervised learning, Unsupervised learning, Semi-supervised learning and Reinforced learning.

Examples of commonly used models in machine learning and data mining include Artificial neural networks, Decision trees, Support Vector Machines, Regression analysis, Bayesian networks, K-nearest neighbor algorithm etc.

5 Comparative Analysis of the different data-driven fields

The need for Data Science emerges in most modern scientific disciplines, including engineering, natural sciences, computer and information sciences, economics, business, commerce, environment, healthcare, and life sciences. In the study of (Tsihrintzis et al., 2019), it was stated that data analytics emerged as a need and some scientific and technological fields in which it could play an important role were investigated. The general characteristics of research conducted in these areas are that they take a data-based approach. Provided that they are data-centric, these researches can be approached from the perspectives of data science, data mining and machine learning. Below, the differences between these perspectives are explained in some detail.

5.1 Statistical Modeling versus Machine Learning

Formulating the relationships between variables in the form of mathematical equations is one of the main purposes of statistical modeling. In general, statistical modeling is more related to mathematics. The foundations of statistical modeling date back more than a hundred years. But machine learning is a very new technology. Towards the end of the 20th century, as computer technologies replaced manual analyzes with less data, statistical analysis methods evolved and formed a new field, data mining. The unmanageable volume and complexity of big data that the world is currently awash in has increased the potential and need for machine learning.

Machine Learning is a process to construct relationship between variables using input-output pairs of data. No logical connections or rule-based programming are used to determine this relationship. It is a subfield of computer science and artificial intelligence that deals with constructing systems driven from data rather than clearly programmed instructions.

As the name suggests, machine Learning requires minimal human effort (Figure 2). Machine learning consists of algorithms in which the computer tries to find patterns hidden in data. The predictive power is often very strong, as the machine does this work on extensive data and is free of all assumptions. The statistical model is math intensive and based on coefficient estimation. The modeler must understand the relationship between the variable and other variables before inserting it into the model.

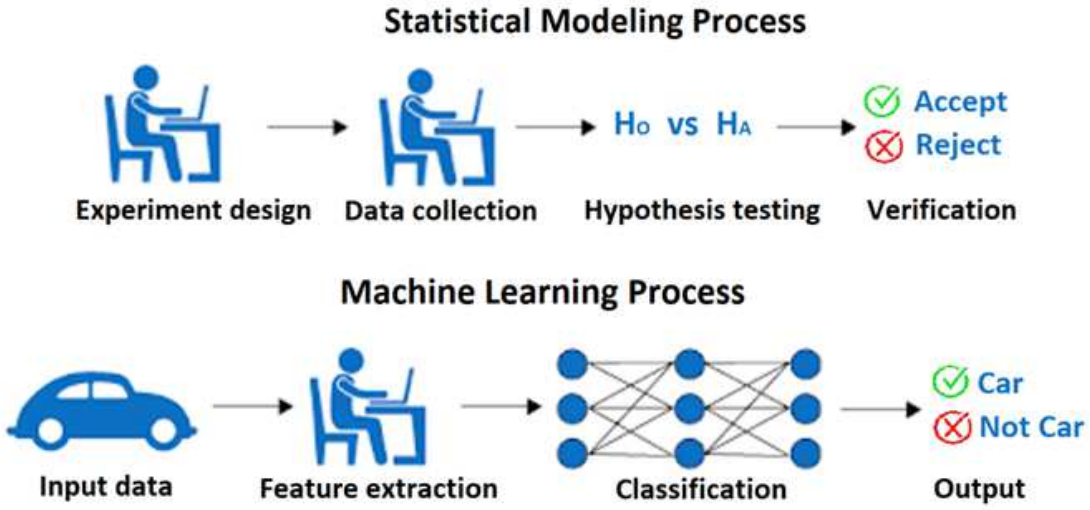

Figure 2: Statistical modeling vs Machine learning

The common goal behind using any of the machine learning and statistical modeling tools is to learn from data. Both approaches aim to gain insight into the underlying phenomenon using the data generated in the process. But there are also differences in the main goals of machine learning and statistics in general: statistics infer populations from a sample, while machine learning finds generalizable prediction patterns (Bzdok et al., 2018). The similarities and differences between machine learning and statistics can be summarized as in Table 1.

Table 1: Statistical modeling vs machine learning

Statistical modeling	**Machine learning**
Statistics quantifies data from the sample and predicts population behavior.	Machine learning constructs models from data and predicts outputs for new inputs.
Usually it is about sampling, population and hypothesis.	Usually predictions are related to supervised and unsupervised learning.
It's about transforming data into aggregated numbers, statistics to understand the structure in the data.	It's about creating algorithms that help machines mimic human learning.
While both are concerned with learning from data, statistics is more concerned with inference from the model.	The focus of machine learning is more concerning with performance and optimization.
Statistical learning involves formalization a hypothesis (making assumptions), and model (test) is validated only for a given hypothesis.	Machine learning algorithms deal with data by running on the various situations, rather than directing the data in accordance with the initial hypothesis.

5.2 Data mining vs Statistical analysis

The purpose of both Data Mining and Statistics is to analyze data, but there are some differences. Data mining process involves modeling, estimating and optimizing up to dataset while statistics

analyses how efficient a dataset is. The similarities and differences of data mining and statistics are given in Table 2 below (ProjectPro, 2023).

Table 2: Data mining vs statistical analysis

Data mining	**Statistical analysis**
It is exploratory – first examines data, discovers new patterns, and then builds theories.	It is confirmatory - First builds theory and then tests it using various statistical tools.
Data cleaning is one of the important stages of data mining.	Applies statistical methods on clean data.
It generally deals with large datasets.	It generally deals with small datasets.
It uses intuitive thinking.	It does not involve intuitive thinking.
It is an inductive process.	It is a deductive process. It does not involve making any guesses.
It uses numeric and non-numeric data together.	It generally uses numeric data.
Not much attention is paid to data collection.	It is more related to the issue of data collection.
Some of the popular data mining methods are: Prediction, Classification, Neural Networks, Clustering, Association and Visualization.	Some of the popular statistical methods: Descriptive Statistics, Inferential statistics.

5.3 Data mining vs Machine learning

Data mining is generally used in the preliminary stage of machine learning. Data collection, data preparation, and accumulation of clean and complete data in a data warehouse are the main tasks of data mining. Machine learning, on the other hand, deals with creating prediction models by considering data prepared generally using data mining tools. The similarities and differences between data mining and machine learning are given in Table 3 (ProjectPro, 2023).

Table 3: Data mining vs machine learning

Data mining	**Machine learning**
Data mining is used to extract hidden patterns from large datasets.	Machine learning is focused on generating models from data and, essentially, then using that model for prediction.
In data mining, rules are obtained from existing data.	Algorithms used in machine learning teach the computer to learn rules. Essentially, in the future, these rules will be used for predictive purposes.
Data mining requires more human intervention and serves the purpose of making data easier to process and make sense of by humans.	Machine learning is less dependent on human influence. For this purpose, emphasis is placed on self-learning of the machine. In this case, human intervention is mostly limited to tuning the optimal parameters of the algorithms.
In data mining, the model does not aim to constantly optimize itself.	In machine learning models, the model attempts to optimize itself when new data is added to the system.
Data mining is used to make predictions for the business given large structured on non-structured historical datasets.	Machine learning algorithms generally uses structured data to construct prediction models.

6 Conclusion

In this study, we discussed and examined data driven approaches, which are the most important component of Artificial Intelligence (AI) technology, which is an increasingly important technology in today's world. These approaches are Statistical modeling, Data mining and Machine learning technologies. Because of their close proximity to each other, sometimes even because they are so intertwined, these technologies can lead to confusion. In order to clarify these questions, in this study, the basic purposes of use of the mentioned technologies, the conveniences they provide and their main differences from each other were analyzed. To be more effective, the comparison results are presented in tables.

We think that this study, which includes a comparative analysis of these technologies that are most used among computer scientists today, will be more useful, especially for young researchers.

References

Aggarwal, C.C. (2018). *Neural Networks and Deep Learning*. A Textbook. Springer, 497 p.

Anderson, D.R., Sweeney, D.J., & Williams, T.A. (1994). *Introduction to Statistics: Concepts and Applications*. West Group.

Bell, G., Hey, T., & Szalay, A. (2009). Computer Science: Beyond the Data Deluge. *Science, 323*(5919), 1297–1298.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science, 16*(3), 199–231.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. Nature Methods, 15, 233–234.

Cao, L. (2018). Data Science: A Comprehensive Overview. *ACM Computing Surveys 50*(3), 1–42.

Cortes, C., Vapnik, V.N. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Danyluk, A., Leidig, P. (2021). Computing Competencies for Undergraduate Data Science Curricula. ACM Data Science Task Force Final Report.

Davenport, T.H., Patil, D.J. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review, 90*(10), 70–76.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM, 56*(12), 64–73.

Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics, 26*(4), 745–766.

Hayashi, C. (1998). What is Data Science? Fundamental Concepts and a Heuristic Example. In Hayashi C et al. (eds). *Data Science, Classification, and Related Methods*. Springer, Japan. pp. 40–51.

Haykin, S. (2009). *Neural Networks and Learning Machines*. Pearson Education, 3rd edition.

Kirchner, F. (2020). A Survey of Challenges and Potentials for AI Technologies. In: Kirchner, F., Straube, S., Kühn, D., Hoyer, N. (eds) *AI Technology for Underwater Robots. Intelligent Systems, Control and Automation: Science and Engineering*, vol 96. Springer, Cham.

Koby, M., Orit, H. (2023). Why is it Hard to Define Data Science? cacm.acm.org. Access date: 03.02.2023.

Kohavi, R, Provost, F. (1998). Glossary of terms. *Machine Learning 30*(2–3), 271–274.

Lindsay, R.P. (1964). The Impact of Automation On Public Administration. *Western Political Quarterly, 17*(3), 78–81.

Nasiboglu, R., Nasibov, E. (2022). FyzzyGBR—A gradient boosting regression software with fuzzy target values. *Software Impacts*, 14, Art. no 100430.

Nasiboglu, R., Nasibov, E. (2023). WABL method as a universal defuzzifier in the fuzzy gradient boosting regression model. *Expert Systems with Applications*, 212: Art. no 118771.

Nasiboglu, R. (2022). Analysis of different approaches to regression problem with fuzzy information. *Journal of Modern Technology and Engineering, 7*(3), 187-198.

Nasibov, E. (2011), Fuzzy Logic in Statistical Data Analysis. In Lovric M (ed.), *International Encyclopedia of Statistical Science*, Springer-Verlag Berlin Heidelberg, pp.558-563.

Nikoletseas, M.M. (2014). Statistics: Concepts and Examples.

ProjectPro (2023) Data Mining vs. Statistics vs. Machine Learning, https://www.projectpro.io/article/data-mining-vs-statistics-vs-machine-learning/349, Access date: 29.01.2023.

Schwabd, K. (2015). The fourth industrial revolution—what it means and how to respond. *Foreign Aff. December* 12, 2015.

Tsihrintzis, G.A., Sotiropoulos, D.N., & Jain, L.C. (2019). Machine Learning Paradigms: Advances in Data Analytics. In: Tsihrintzis, G., Sotiropoulos, D., Jain, L. (eds) *Machine Learning Paradigms. Intelligent Systems Reference Library*, vol 149 . Springer, Cham.

Yablonsky, S.A. (2019). Multidimensional Data-Driven Artificial Intelligence Innovation. *Technology Innovation Management Review, 9*(12), 16-28.

Zadeh, L. (2011). Fuzzy Set Theory and Probability Theory: What is the Relationship? In Lovric M (ed.), *International Encyclopedia of Statistical Science*, Springer-Verlag Berlin Heidelberg, pp.563-566.