CHRIST
(DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

DATA GEEK

STATISTICS IN DATA SCIENCE

VOLUME 4 ISSUE 1 OCTOBER 2022

SCHOOL OF BUSINESS AND MANAGEMENT

BUSINESS ANALYTICS SPECIALIZATION

EDITOR'S DESK

"The goal is to turn data into information, and information into insight."

- Carly Fiorina

Statistics is at the core of Data Science. The processing of information is the most important aspect of any Data Science approach. When we talk about developing insights from data, we are essentially digging for possibilities. Statistical Analysis is the name given to these possibilities in Data Science. When we have data in numerical format, we have an infinite number of ways to understand it. The power of statistics is not limited to understanding data, it also provides methods to measure the success of insights, obtaining different approaches to the same problem, and determining the best mathematical approach for the data in hand. Hence statistics and data science go hand in hand and understanding the key concepts of statistics is imperative to work on any dataset to derive insights from it.

With this, we present an enthralling Volume 3 Issue 1 of DataGeek newsletter focused on a pivotal theme of Statistics in Data science. It includes numerous interesting articles, infographics, insightful interview and fun crossword to keep you engaged and leave you with immense learnings. The team would like to extend sincere thanks and gratitude to our esteemed alumnus Mr.Arun Suresh, Data Science Manager at AB InBev for his industry insights. I would like to extend gratitude to our Dean, Dr. Jain Mathew, Associate Deans Dr. Georgy Kurien and Dr. Jeevananda S, Head of Specialization - BA, Dr. Lakshmi Shankar Iyer for their guidance in making this issue a success. Also, a special appreciation to the newsletter team for the effort, time and inputs without which this issue would not have been possible. A big thanks to all the students who have provided their valuable inputs. Once again congratulations to the entire team.

Please reach out to us for any queries or suggestions at datageek@mba.christuniversity.in

WITH REGARDS,
DR. TRIPTI MAHARA

TABLE OF CONTENTS

STATISTICS AND DATA SCIENCE IN POLICY MAKING

Source: https://emeritus.org/in/learn/why-become-a-data-scientist/

Data Science is an interdisciplinary field focusing on obtaining value from heterogeneous data collections, moving from data collection to data analysis and visualizations. Data Science acts on two pillars: First, the emphasis is given to Big Data, and second is the elaboration process, consisting of the cycle of phases required to transition from raw data and generate knowledge.

The study of data collection, analysis, classification, and interpretation is the fundamental definition of statistics. The Central Statistical Organization (CSO) is the statistical organization in charge of collecting data on Indian economic, social, and general situations and activities. The CSO facilitates the ability for users to use statistics for well-informed decisions in all environments effectively. The Indian National Data Warehouse on Official Statistics offers remote access options to end users through a network, allowing all data consumers to quickly access the published and unpublished validated data from a single source. It makes historical data sets and statistical tools more easily accessible to assist in future planning. Effective policy-making requires reliable statistics that are transparent, which is a necessary part of the enabling environment for improving development outcomes.

Statistics is a mathematically-based field that seeks to collect and interpret quantitative data. In contrast, data science is a multidisciplinary field that uses scientific methods, processes, and systems to extract knowledge from data in various forms. Both Statistics and Data Science are essential for effective policy-making. Data science is an interdisciplinary field, and Statistics is an integral part and an absolute requirement for effective policy-making.

Government can benefit significantly from data science and statistics. A growing understanding of how data "can decrease vulnerability about the best game plan" in policy design has emerged in recent decades. For instance, it can inform better policy-making processes and result in progressively adequate, increasingly successful national policies. With predictive causal analytics, prescriptive analytics, and machine learning, data science can aid in extracting pertinent information and knowledge from vast amounts of data to enhance government decision-making or provide the insights needed to make data-driven decisions. Through enhanced predictive analytics, the government can use the data to address implementation gaps, identify overlaps, target the appropriate beneficiaries, and contribute to smart policy-making. The NSSO (National Sample Survey Office) gathers data on the subject. After a year of data gathering and research, the government receives the document to formulate policy.

Predictive analytics uses data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on historical data. There are many potential benefits of predictive analytics in government, including more effective, efficient, and less biased decision-making. On the other hand, some risks and challenges must be adequately considered. Risks exist in implementing any predictive analytics program, but in the government sphere, the risks increase as the prediction outcomes potentially significantly affect a citizen. The risks of false positives and false negatives are heightened. One of the machine learning techniques used in policy formulation is clustering. Clustering is the task of dividing the population or data points into several groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In public policy, subgroups of a more significant population can be vital. Subgroups are complicated, and nothing is unidimensional. It is essential to consider the number of factors that might be associated to characterize the most significant contrasts in any population being examined. The clustering technique considers the revelation of these fundamental groups across numerous factors that might otherwise remain hidden.

CONTRIBUTORS

Survesh G (2127412)

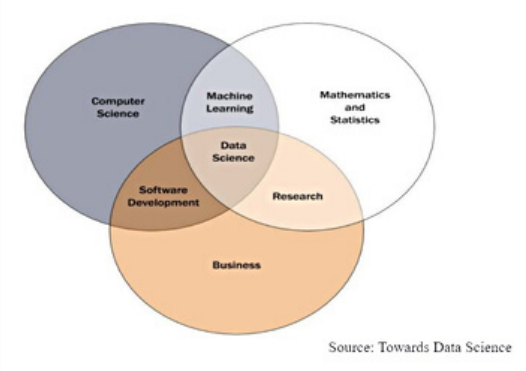
Sree Lakshmi Deevi (2127437)

Ajay A Kashyup (2127004)

HOW MUCH DO DATA SCIENTISTS NEED TO KNOW ABOUT STATISTICS?

"A Data Scientist is a person who is better at statistics than any programmer and better at programming than any statistician." - Josh Wills.

Data Science is a way of deriving meaningful insights from data that helps improve decision-making. This can be achieved by first understanding the data, which can be done using statistical measures. Combining computer science, mathematics, statistics, and business knowledge enable Data Scientists to develop suitable solutions and design a system to bring about improvements.



Source: Towards Data Science

Data scientists must comprehensively understand the statistical techniques given below, as any knowledge gap will lead to data manipulation or false conclusions.

- **Descriptive Statistics:** Descriptive statistics helps data scientists present data in a meaningful way by identifying basic features of the data, providing summaries and descriptions of the data, and visualizing data.

- **Probability Distribution:** Probability distribution refers to the likelihood of occurrence of an event, and several properties, such as expected value, variance, skewness, and kurtosis, can be measured, which will help understand the spread of the data.

- **Hypothesis Testing:** Hypothesis Testing evaluates two statements about a population and supports the certainty of the findings of an experiment.

- **Dimension Reduction:** By using feature selection and feature extraction, data scientists lower the number of random factors that need to be considered. This reduces the process of putting data into algorithms and simplifies data models.

- **Bayesian Statistics:** Frequency statistics analyze the likelihood of an event based on previously collected data. However, Bayesian statistics expands on this idea by considering variables that we expect to be true in the future.

Statistics help Data Scientists select relevant features and find the existing relationships between the dependent and independent features. Hypothesis testing is widely used in data science to determine whether a claim or hypothesis regarding a real-world occurrence can be taken as true or false based on the available data.

Customer churn is one of the major problems companies face, as the cost of acquiring new customers is much more than retaining existing customers. In this case, hypothesis testing can be used to check if a company's marketing efforts significantly impact the churn of customers. If the company concludes that marketing efforts do not significantly impact churn, it might focus on other factors that will increase the churn rate. Therefore, it is essential for data scientists to equip themselves with the necessary statistical techniques to arrive at the right decision.
Not everyone is familiar with machine learning algorithms' performance measurements, such as f-score, recall, precision, accuracy, root mean squared error, etc. Instead, suitable statistical techniques will help data scientists explain the data better and make the right decisions, which can be done only with statistics.

CONTRIBUTORS

Shivani Gaonkar (2128050)

Midhuna M (2128364)

Taniya Jose (2128354)

CHALLENGES AND FUTURE OF STATISTICS IN DATA SCIENCE

"Data Science" appeared in the title of a statistical conference in 1996, and was introduced by IFCS (International Federation of Classification Societies). Even though the term "Data Science" was coined by statisticians, the public perception of "data science" places a much greater emphasis on the significance of computer science and business applications, particularly in the era of "Big Data".

Processing information is the most crucial component of any Data Science approach. Creating insights from data is essentially the same as mining for possibilities. These options are referred to as statistical analysis in data science. Data science tools that primarily tackle problems through statistical analysis can process a variety of enormous data volumes. The focus of data science is on analyzing the massive amounts of data created in actual time by applying statistical techniques that use mathematics, machine learning algorithms, and other tools. Although statistics may not be portrayed as being as vital as the other fields, it is more crucial for the areas of data collection, enhancement, and sophisticated modeling required for prediction.

Any real-world problem in this field of study can be solved using a balanced approach that combines statistical methods with computational algorithms. The number of data sets is expanding exponentially, and this trend does not appear to decrease. In order to perform thousands or millions of simultaneous tests on trillions of data points, we require procedures that are both effective and efficient. Currently, inefficient methods are applied to the data, which is not a practical solution.

Following are the ways statistics will progress within data science:

- Widespread use of statistical methods by non-statisticians.
- New sampling schemes and experiment designs are inspired by the internet and big data applications.

Better methods for inference with dependent observations.

Statistics are needed more certainly for advanced modeling, data acquisition, and enrichment that is required for prediction. Using statistical techniques and computing algorithms together to reach a balanced method to find the answer to any actual problem in the area of research. It can also be done to find a solution to any problem in this area of research. Additionally, it can further highlight that it helps to resolve major mathematical approaches in which data science needs a lack of projecting huge crucial information from data sets in the actual time frame.

The fundamental ideas of statistics will not change in the future because it is a component of Data Science. As Data Science develops, the number of raw statistics used would decrease. Currently, data science is in its early phases and would require a high application of statistics. Statistics still has the issue of detecting systematic errors in datasets, especially in domains where the random errors are large. This can shift the study in another direction and dealing with such errors will be hard for the person conducting the study. Updation of the probability when there is a new addition of data into the dataset also becomes a problem in statistics when used in data science.

Also, the validity of futuristic prediction based on the techniques available in data science can be held doubtful because most of it is based on fundamental statistical concepts which have not been updated over the years. New techniques have not been discovered which could help in obtaining better prediction accuracies. The addition of other factors that we usually do not take into consideration while analyzing data includes the other factors in a domain which we usually tend to ignore. These factors can creep in during the actual event that is happening and will alter the predictions made by models. These issues will also affect the quality of predictions as statistics has no techniques to consider these negligent factors.

CONTRIBUTORS

Allen Denny (2127708)

Akshara Rajendran (2127534)

Neola Cleo Rego (2127649)

Alen Alosious (2127504)

Vaishakh Mohandas (2127628)

CORPORATE INTERVIEW

WITH

Mr. Arun Suresh

Data Science Manager at AB InBev

Christ (Deemed To Be University)

MBA Marketing, BATCH (2010 - 2012)

Mr. Arun Suresh is a passionate problem solver who enjoys answering complex questions using quantitative analysis or exploring mysteries of unstructured data. Finding actionable insights from large data sets and learning something new excites him. He has an extensive experience 11 years in Analytics, Data Mining, exploratory data analysis and data visualization. With this, he helps businesses gain greater value and maximize their return on investment. ML, Optimization (Linear/Non-Linear) Data Engineering, Data analysis, Data wrangling, and data visualization are a few of the areas of his strength.

1. What do you think is the importance of Statistics in Data Science?

Data science is a methodology- an approach, and stats is the engine that drives it. So if you do not comprehend data science statistics, it is very difficult to have the right approach. It is a two-way stream, I have seen statistics guys who could not connect the dots downstream and software engineers who could not connect the dots upstream, so it's a double-edged sword, which makes stats one of the most important things out there.

2. You must have met a lot of data scientists, students, and job seekers along your way, what, according to you, is the most misunderstood statistical concept and why?

Basically, statistics usually come as KPI, we use it to evaluate models. A lot of time, it can be used as advantageous, and you can completely miss the boat. One of the most important things that I have seen is MAE(Mean Absolute Error), and by definition, it should be only one, you cannot have multiple MAE. So all these factors like R square, RMSE, MAE, etc., anything with an "M" is mean, so I have seen new folks who are without complete understanding, trying to compute it in different ways. You will get an answer, but that might not be the right answer. One another important thing that people get wrong is the hypothesis. Null and alternate -Which one to reject and which one to accept? When to do what? These are the things that I have seen people kind of get confused about. I would say that is because of a lack of practice or exposure, if you have not done that enough, then you might go wrong.

3. You must have dealt with several datasets so Far. By far what has been the most challenging part of dealing with the dataset?

So, data engineering is one of the most important aspects of anybody who wants to be a data scientist. Nobody is going to get you the data; they are just going to point at the data, and the data can be jumbled up, multiple sources, and agnostic and completely centralized. So usually, the most challenging part is getting the data data-ready for data science. You cannot have a closed mind, usually - it is more exploratory, you would want to know how the features are going to play and the kind of encoding that you might need to do.

The most important thing is missing data- how do you treat them, are you going to omit the rows or impute them and cover them up with mean, mode, or median and replace them is kind of tricky because what you are going to do with data is going to have a long-term effect in terms of how your output is going to be.

4. What are your favorite statistics and/or data science websites? What publications, websites, blogs, conferences, and/or books do you read/attend that are helpful to your work?

I do have a few of my favorites- StackOverflow, StackExchange, MLMastery, and GeeksforGeeks. I would recommend you check which website's UX syncs with you and your requirements, everybody needs to find what works out the best for them. Once you google, you will find a lot of websites, it sometimes depends on the topic's popularity as well. If the topic is very popular, then every other website will have a view of it, and the more niche the topic, the limited your options are.

5. What would be the key statistics concept that you would recommend every aspiring data scientist to be thorough with?

There are a few of them, but data science, basically in terms of the majority of aspects, is regression and classification, this is where you have a lot of stats in play. Especially in regression, many stats are involved, so I would recommend all the pivotal metrics there- R square, RMSE, MAE, MSE, etc. You need to understand each of them to a level to relate to what it means, can R square be negative? Why should it be closer to 0? What is the difference between R square and RMSE? Why does RMSE change after each and every iteration while R square remains the same? What are MAE, MSE, and MAPE? The ideal part of statistics, and why I recommend these things, is understanding how you understand and present your models. On the other end, I would recommend understanding Coefficients and Elasticity, explainable AI is a big thing now in the industry. On the classification side of things, you have a false positive, precision, recall, and F1 score - what does this all mean, in terms of KPIs, these are all important. Apart from that, I would suggest understanding the Euclidean distance and the correlation- understanding the types and their applicability in various situations. Data scientists explain a problem with the said KPIs to the business partners, and the business partners might not essentially be aware of these KPIs, so they might be intrigued to know better about them. Hence it is recommended to have a good understanding of the same.

6. Do you see a gap between what is taught in academics and what is being used in the industry? If so, how can it be bridged?

The answer is yes, a big-time yes! Education talks more about proof of concepts and best data sets. Hence, how many of the data scientists have done data engineering is very limited. The industry is very diverse. In my transition from being a MBA in marketing guy to doing my MS, I can vouch for the fact that nothing that I learned there could have been used or would have prepared me for the industry. It is something that would need a lot of self-learning. While my experience in UpGrade as well, my teaching style was focused on problem solving. You might know Python, R,sklearn, Tensorflow and what not but a lot of the companies have amazing in-house resources, some of them use Azure, AWS, GCP so you might have to unlearn and relearn. So the only thing that is constant is problem-solving.

When provided with a dataset, your thought process should be - which of it is the target variable and what are the features that I can use, and you need not have an algorithm background to understand that. You need to understand and identify the key right because if that goes wrong, the entire project goes for a toss and I would say that is the biggest gap. During my MBA this was my key takeaway wherein we were taught how to think from the problem-solving aspect. It is the ability to solve the problem faster and the thinking style that is going to make the difference and take you places.

7. What are some of the suggestions that you would give to someone who is aspiring to be a business analyst?

You need to be a problem solver, you need to understand the domain. Start by understanding from introspection - Do you see a problem that could be solved? What could be the approaches that you can use? It can be very minor as well, let's say someone has been working on 10 different excels and they have been copy-pasting it, can you automate it? Data science is nothing but finding a better way to do stuffs and it starts with data engineering.

HOW IS STATISTICS AND DATA SCIENCE RELATED?

"It is the mark of a truly intelligent person to be moved by statistics." - George Bernard Shaw

In the month of April this year, analytics jobs witnessed a 30.1% increase in open jobs compared to the same period last year. What does this say? It states the difference between the number of open jobs this year and the last year for the month of April. Thus, statistics make acquiring inferences from massive amounts of data more efficiently. So what is the relationship between Statistics and Data Science?

Statistics has a significant role to play in data science. It follows a process for solving a problem using data science, like

- Defining the problem
- Identifying the required data
- Data Pre-processing
- Modeling the data
- Training and testing
- Verifying and deploying

Statistics in data science seek structure and relationships among various unflustered data. Data structuring aids in revealing various valuable insights hidden within the collected data. As businesses progress, there are some stressful situations that businesses encounter wherein they have to choose. How can informed decisions be made? Well, statistics come into play here. It helps in reducing the uncertainty in decision-making that affects the business. A simple t-test can be rather satisfying, especially when it comes to decision-making. Sometimes, a more rigorous analysis of statistical data can provide useful information.

Statistics is a problem-solving process that seeks answers to questions through data. Statistics present the information in organized structures through plots and graphs like bar plots, histograms, pie charts, and others. At the base of any machine learning algorithm, statistics is needed.

It helps to structure and find relationships between various clusters of data for deriving valuable insights from it. Statistics are used by data scientists to apply quantifiable numerical models to proper factors. Statistics play a crucial role in the field of Data Science, as it is all about analysis, storage, practical application of data, and mobility. It helps in structuring the raw data and quantifying the uncertainty in it.

An intermediate level of statistical understanding is required for every algorithm, big data analysis, and targeted market research. Statistics stands out as the tool to understand, interpret and draw conclusions from data. Statistics thus contributes significantly to Data Science's advancement to the levels it has reached today.

CONTRIBUTORS

Nikitha Reddy
(2128240)

Rakshantha A
(2127552)

Tania Chakraborty
(2127557)

APPLICATION OF STATISTICS AND DATA SCIENCE IN MODERN WARFARE AND CYBER SECURITY

“Statistics and Data Science are buzzwords and are composites of many concepts,” stated the US Standards Technology Institute in the 2015 framework. Statistics is the discipline that concentrates on the collection, organizing, analyzing, interpreting, and introduction of data. For using statistics to address issues in industry, society, or science, it is standard to start by studying a statistical population or statistical model. While data is used in many scientific investigations, statistics is concerned with how data is used when there is ambiguity and how to make decisions when there is doubt.

Cybersecurity studies safeguard information's availability, confidentiality, and integrity while preventing unwanted access or improper use of data, devices, and networks. Data science is the method of extracting valuable information from vast quantities of unstructured data, also referred to as "Big Data," by fusing subject-matter expertise with programming skills, understanding of mathematics, and expertise in statistics. Data scientists often use algorithms, procedures, scientific methods, tools, techniques, and systems for their data processing activities. They then apply the insights gained across a variety of disciplines. Data science and cyber security are fundamentally related since the latter needs the safeguards and protection the former offers. Data scientists need clean, uncompromised data to obtain their findings and guarantee that the information they analyze is secure. Now the application of statistics and data science are applied in modern warfare for efficient and effective execution of projects and for the better utilization of resources. The article covers the various applications of data science and statistics in modern warfare and cyber security.

Source: Towards data science

Collecting effective and required information for operation and functions is the first initial step to deriving a fruitful outcome. And in the online world, Data is generated every frequent second; it may be in the Military, education, health, sports, retail, etc. but utilizing that Data leads to a difference in society. As per the 2012 military report of Afghanistan, Most of the requested information was already within the system, but still, the requester was unable to locate or access it. Moreover, Indian militaries are currently lacking such dedicated positions for data scientists, which results in a loss of edge over adversaries when formulating military plans. But maybe on a slow level, things are developing; many agencies and forces have started using data science and statistics in their operation, such as BSF (Border Security Forces) using cattle movement data to predict the chance of cross-border intrusion. Data science is restricted not only to surveillance, border management, and operation planning but also to financial analysis, logistics, and disaster management.

But having more data leads to the invitation of fraudulent activities in the system, such as stealing of data, but Data science and statistics themselves bring a solution to this as the Anti-Fraud Command Center (AFCC) has identified and terminated 500,000 such attacks in its last eight years. By analyzing the trend of all the parameters, machine learning will immediately notify the data scientist of any suspicious and unusual activities to maintain the level of cyber security.

A better method of detecting and predicting intrusions: There are numerous ways for hackers to break into various systems. Additionally, they have a reputation for frequently altering their tools, appearances, and techniques to evade capture. Early intrusion detection is essential because of this. By utilizing such data science, businesses can supply machine learning algorithms with current and historical data on such invasions. To better manage their systems and foresee upcoming attacks, businesses might use this type of analysis to uncover patterns that help detect intrusions. You can find security gaps in your environment before hackers use machine learning techniques.

In addition to intelligence, surveillance, border, marine, and space management, operational planning, logistics, financial management, disaster management, future technologies, cognitive analysis, and analysis of past information are all areas where data analytics are used in the military. But eventually, soldiers from all across the world will realize how important data scientists are. No matter how huge the nation's economic database on national security may be, it will always take enormous teams of highly qualified people to analyze it and bring it to the table in a legible state.

By reducing the chance of failure, better understanding the anti-national components, getting novel insights, assisting in effective expansion, and enhancing event predictions, data scientists in the military, intelligence, and law enforcement serve the organization. These advantages will further lower operations casualties, collateral damage, crime rates, terror attack rates, and cross-border intrusion rates. Data science will soon play a crucial part in a country's security, and that day is not far off.

The main instrument for data science work in the cybersecurity industry is a source dataset. The vast majority of currently available datasets are obsolete and might not be sufficient to comprehend the most recent behavioral patterns of different cyber-attacks. Despite the fact that the data can be usefully understood after many processing steps, there is still a dearth of information detailing the traits of current attacks and their recurrence patterns. The desired decisions may therefore be made with a low accuracy rate by using additional processing or machine learning approaches.

CONTRIBUTORS

Nirmal Joseph (2127823)

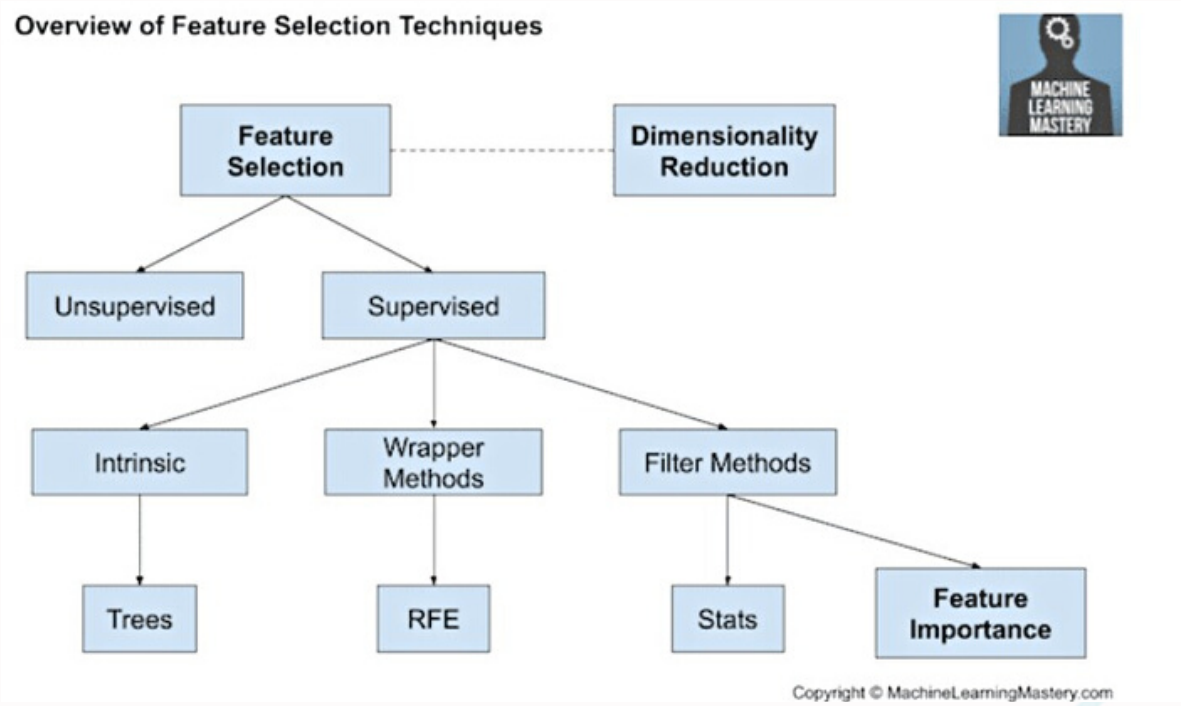
Shubham Sharma (2128327)

Tejaswi Harsh (2127831)

FEATURE SELECTION

Feature Selection is a process in data analysis that helps an analyst determine which variables to include in the model. In the world of Big Data, no one can argue with the need for this process. One of many cases is facial recognition software improvement by selecting and removing irrelevant data from images before recognizing key attributes like mouth, nose, or eye. Other fields can use the same approach where there is a lot of noise and irrelevant information (e.g., clicking on an advertisement).

Feature selection can be divided into two parts: variable selection and model selection. Variable selection is selecting the most critical, relevant variables from a dataset, while model selection is selecting an appropriate statistical model to produce an accurate prediction.



In this study, we will discuss the case of feature selection through variable selection.

PCA is an exploratory technique for uncovering patterns of data sets by reducing the dimensionality. That is, it reduces the set of observations into a new set of values called principal components (PCs). The first principal component (PC1) is called the Pareto principle because it accounts for 80% of the total variance. Thus, PCA can reduce data set dimensions while retaining as many important variables as possible in a model. These restrictions must be considered when concluding using PCA.

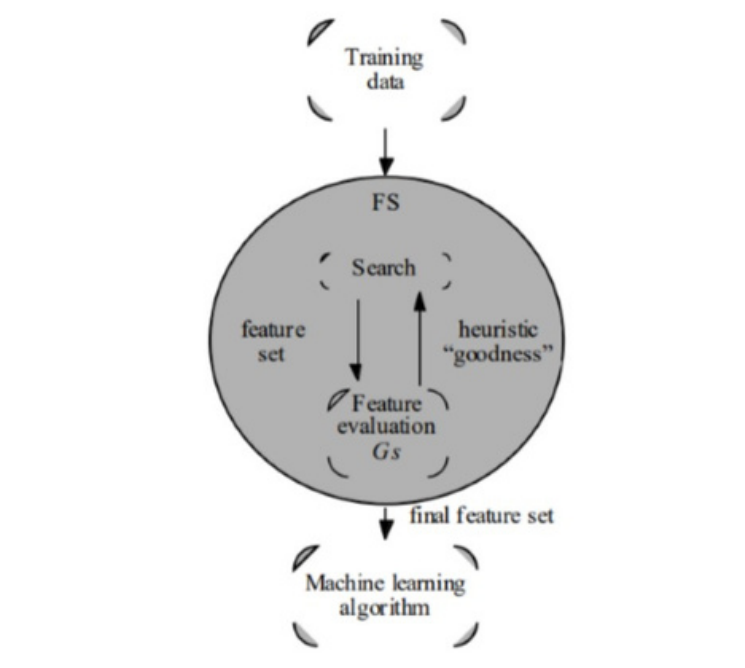
PCA can be exploited, just like any other statistical technique. It is crucial to avoid scaling variables to fit prior knowledge of the data because scaling can lead to differing PCA results. Using PCA, you may easily explore the data to understand its significant variables and detect outliers. PCA is a tool for identifying the principal axes of variance within a data set. It is one of the most effective techniques in the data analysis toolbox when appropriately used. Apart from PCA, there is an innovative approach to feature evaluation for machine learning that makes use of a correlation-based heuristic.

Applying the feature selector to data as a pre-processing step for three popular machine learning algorithms allows for the evaluation of the feature selector's efficacy. Using PCA, you may easily explore the data to understand its significant variables and detect outliers. PCA is a tool for identifying the principal axes of variance within a data set. It is one of the most effective techniques in the data analysis toolbox when appropriately used. Apart from PCA, there is an innovative approach to feature evaluation for machine learning that makes use of a correlation-based heuristic. Applying the feature selector to data as a pre-processing step for three popular machine learning algorithms allows for the evaluation of the feature selector's efficacy.

LDA, or linear discriminant analysis, is a multi-class classification predictive modeling approach. Dimensionality reduction is the process wherein lowering the number of variables used as inputs into a predictive model takes place. LDA can also be used to reduce the number of variables in a dataset by projecting a training dataset in a way that optimally divides the instances into their respective classes. A simpler predictive model with fewer input variables may perform better when making predictions based on fresh data.

Correlation-based Feature Selection employs a search algorithm and a function to assess the value of feature subsets, similar to most feature selection methods. The heuristic used by CFS to assess the "quality" of feature subsets considers both the degree of intercorrelation between the features and their effectiveness for predicting the class label. The following is a statement of the heuristic's underlying hypothesis:

Features that are significantly correlated with (predictive of) the class but uncorrelated with (not predictive of) each other make up the best feature subsets.



Source: researchgate.net

The performance and intercorrelations of the features are used by the algorithm (CFS) to direct its search for a suitable subset of characteristics. The preliminary findings are promising and support the potential of CFS as a useful feature selector for popular machine learning techniques. By increasing the learning algorithms' accuracy and simplifying their output, the correlation-based evaluation heuristic used by CFS seems to select feature subsets that are helpful to the algorithms' learning capabilities. Future research will try to comprehend why CFS performs better in some areas than others. The domains where CFS has not fared as well may benefit from addressing the difficulties stated above (measure bias and feature interactions). Future research will examine the correlation between CFS's evaluation heuristic and the actual performance of machine learning algorithms for randomly selected subsets of attributes.

Some Statistical Methods for Feature Selection: -

1. Chi-Square Test (For Categorical Feature and Categorical Response):

The Chi-square test basically examines the validity of the premise that there is no relationship between the input and outcome. We anticipate a low p-value if the feature really does influence the response. As a result, features with low p-values are chosen when picking features.

2. ANOVA (For Categorical Feature and Continuous Response):

Analysis of Variance (ANOVA) evaluates the statistical significance of the intergroup difference by comparing the means of several groups (mean of responses for each categorical characteristic data). We anticipate observable disparities between the means of several groups if a trait is relevant.

3. Correlation (For Continuous Feature and Continuous Response):

Pearson's coefficient, which examines the linear relationship between the characters and the response, is the most widely used correlation coefficient. When a relationship is non-linear, Pearson's correlation yields low values.

The accuracy and stability of our models can be considerably improved with good features. To increase the chance of finding a useful feature, it is sensible to include as many features as possible. Since not all the features would contribute positively towards achieving the end goal, many data practitioners favor feature selection as using many features can be detrimental to training efficiency, accuracy, and interpretability while model building.

CONTRIBUTORS

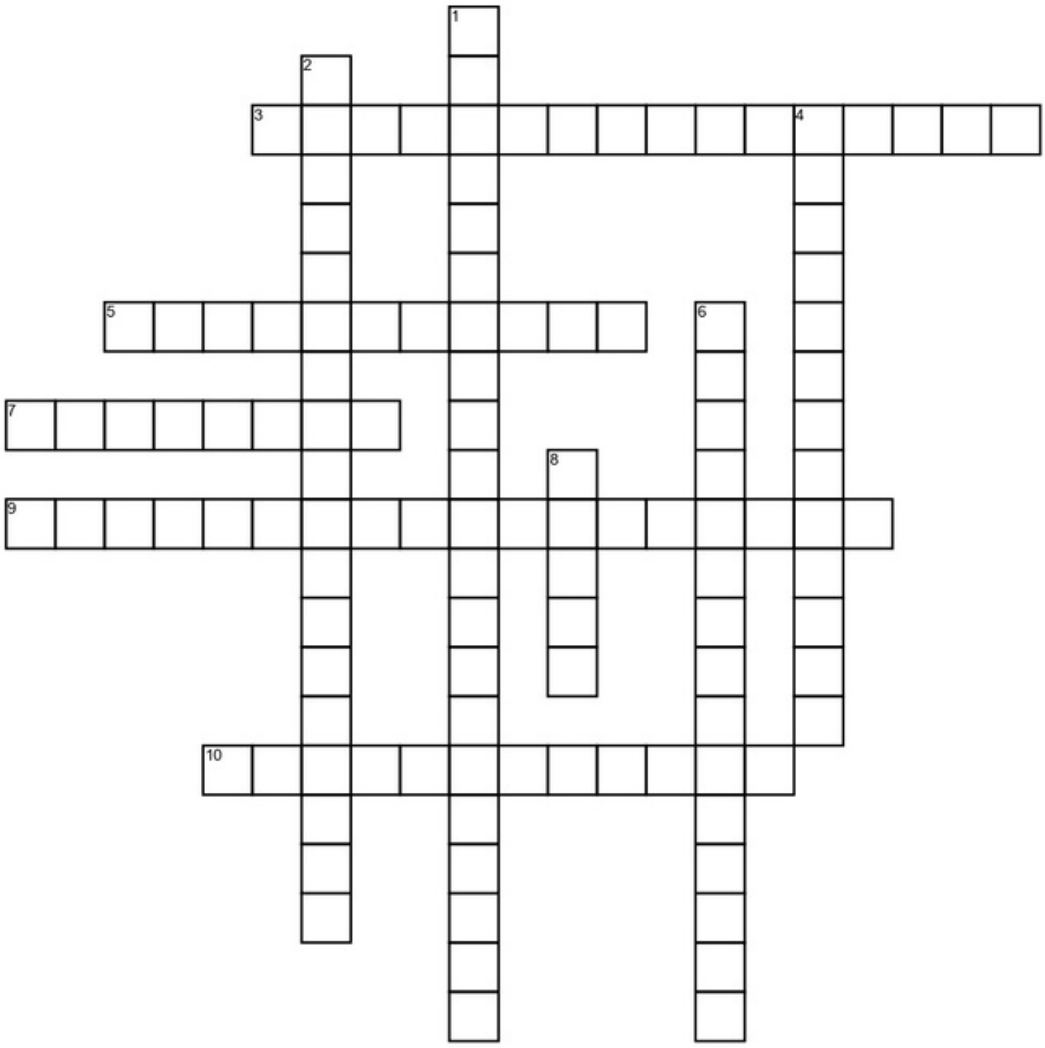
Ganavi CA (2127135)

Chandana Sai S (2127044)

Poorvi Prakash (2127252)

CROSSWORD

Statistics In Data Science



Across

3. Instance where the dependent variable's variance is constant across all data.
5. A graph of bivariate where each data pair is represented by a point on the graph.
7. A numerical value that describes the deviation of each value in the dataset from the mean value and illustrates how widely distributed the individual figures in a set of data are around the mean.
9. A range of values that explains the uncertainty surrounding an estimate.
10. A statistical technique used with an objective to assess the adequacy of a classification, given the group memberships.

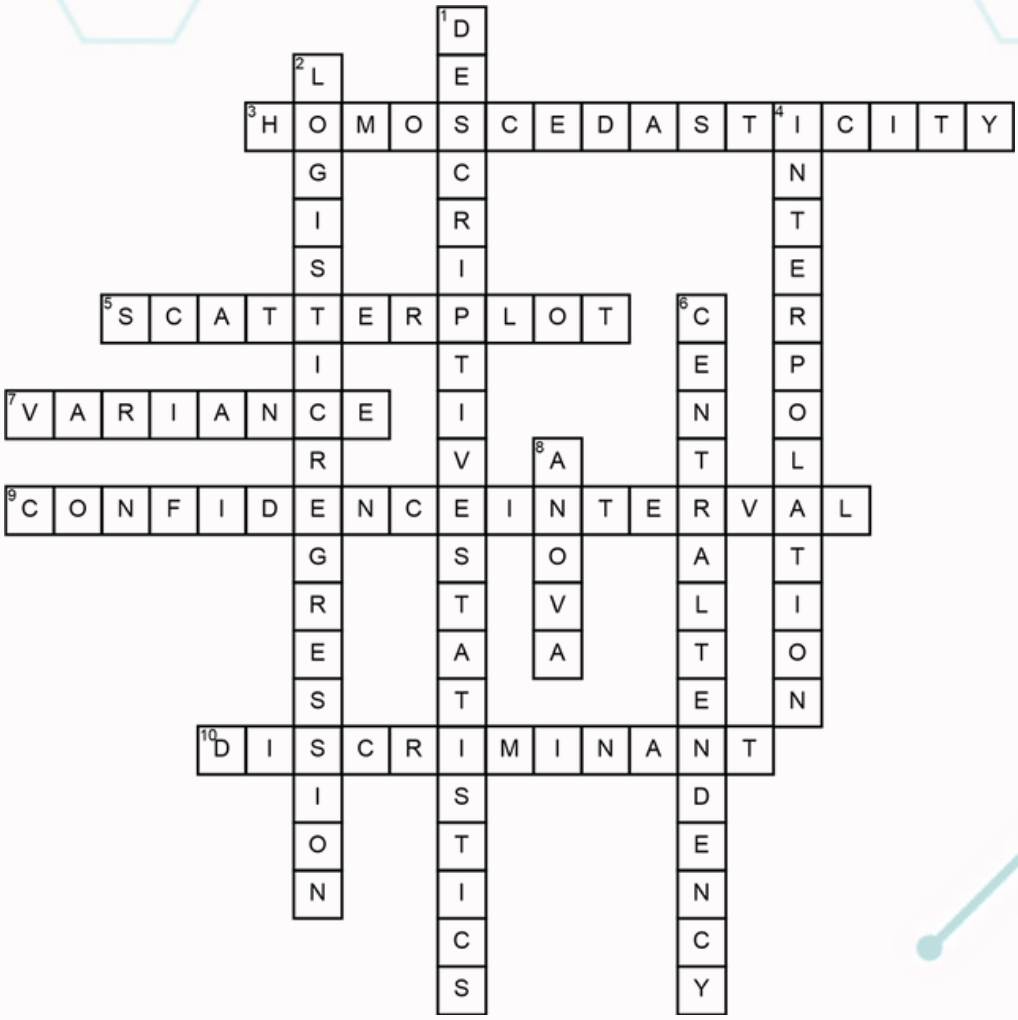
Down

1. Way of analysing and identifying the basic features of a data set.
2. A statistical model that uses a sigmoid function to model a binary dependent variable.
4. Calculating a function's value based on the value of other datapoints in a given sequence.
6. The statistical measure that identifies a single value as representative of an entire distribution.
8. Used to test the hypothesis that differences in the means of two or more independent groups are the same.

CONTRIBUTOR

Aishwarya Shenoy KP
(2127635)

Solution



Across

3. Instance where the dependent variable's variance is constant across all data.

5. A graph of bivariate where each data pair is represented by a point on the graph.

7. A numerical value that describes the deviation of each value in the dataset from the mean value and illustrates how widely distributed the individual figures in a set of data are around the mean.

9.A range of values that explains the uncertainty surrounding an estimate.

10. A statistical technique used with an objective to assess the adequacy of a classification, given the group memberships.

Down

1. Way of analysing and identifying the basic features of a data set.

2. A statistical model that uses a sigmoid function to model a binary dependent variable

4. Calculating a function's value based on the value of other datapoints in a given sequence.

6. The statistical measure that identifies a single value as representative of an entire distribution.

8. Used to test the hypothesis that differences in the means of two or more independent groups are the same.

NEWSLETTER TEAM

Aishwarya Shenoy KP
(2127635)

Tania Chakraborty
(2127557)

Joanna Maria Vinukanth
(2127742)

Kingsly N
(2127724)

Rakshantha A
(2127552)

Renuka Prasad A
(2127528)

Dheeraj S
(2128212)

Nikhil
(2127220)