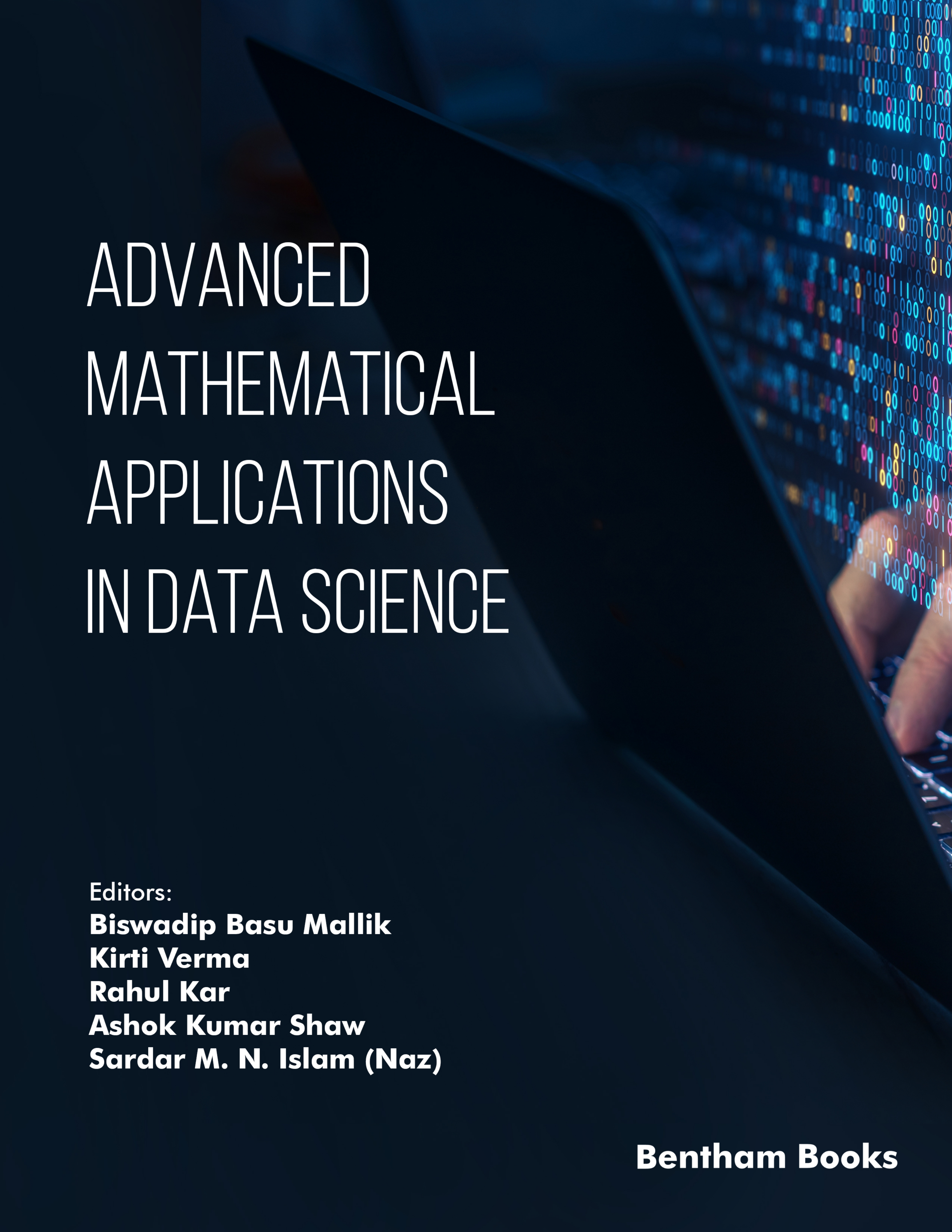

ADVANCED MATHEMATICAL APPLICATIONS IN DATA SCIENCE

Editors:
Biswadip Basu Mallik
Kirti Verma
Rahul Kar
Ashok Kumar Shaw
Sardar M. N. Islam (Naz)

Bentham Books

Advanced Mathematical Applications in Data Science

Edited by

Biswadip Basu Mallik
Department of Basic Science and Humanities
Institute of Engineering & Management, Kolkata
West Bengal, India

Kirti Verma
Department of Engineering Mathematics
Lakshmi Narain College of Technology, Jabalpur
Madhya Pradesh, India

Rahul Kar
Department of Mathematics
Kalyani Mahavidyalaya, Kalyani
West Bengal, India

Ashok Kumar Shaw
Department of Basic Sciences and Humanities
Budge Budge Institute of Technology
Budge Budge, Kolkata
West Bengal, India

&

Sardar M. N. Islam (Naz)
ISILC, Victoria University
Melbourne, Australia

Advanced Mathematical Applications in Data Science

Editors: Biswadip Basu Mallik, Kirti Verma, Rahul Kar, Ashok Kumar Shaw
& Sardar M. N. Islam (Naz)

ISBN (Online): 978-981-5124-84-2

ISBN (Print):978-981-5124-85-9

ISBN (Paperback): 978-981-5124-86-6

Published by Bentham Science Publishers Pte. Ltd. Singapore. All Rights Reserved.

First published in 2023.

Disclaimer:

Bentham Science Publishers does not guarantee that the information in the Work is error-free, or warrant that it will meet your requirements or that access to the Work will be uninterrupted or error-free. The Work is provided "as is" without warranty of any kind, either express or implied or statutory, including, without limitation, implied warranties of merchantability and fitness for a particular purpose. The entire risk as to the results and performance of the Work is assumed by you. No responsibility is assumed by Bentham Science Publishers, its staff, editors and/or authors for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products instruction, advertisements or ideas contained in the Work.

Limitation of Liability:

In no event will Bentham Science Publishers, its staff, editors and/or authors, be liable for any damages, including, without limitation, special, incidental and/or consequential damages and/or damages for lost data and/or profits arising out of (whether directly or indirectly) the use or inability to use the Work. The entire liability of Bentham Science Publishers shall be limited to the amount actually paid by you for the Work.

General:

1. Any dispute or claim arising out of or in connection with this License Agreement or the Work (including non-contractual disputes or claims) will be governed by and construed in accordance with the laws of Singapore. Each party agrees that the courts of the state of Singapore shall have exclusive jurisdiction to settle any dispute or claim arising out of or in connection with this License Agreement or the Work (including non-contractual disputes or claims).
2. Your rights under this License Agreement will automatically terminate without notice and without the

need for a court order if at any point you breach any terms of this License Agreement. In no event will any delay or failure by Bentham Science Publishers in enforcing your compliance with this License Agreement constitute a waiver of any of its rights.

3. You acknowledge that you have read this License Agreement, and agree to be bound by its terms and conditions. To the extent that any other terms and conditions presented on any website of Bentham Science Publishers conflict with, or are inconsistent with, the terms and conditions set out in this License Agreement, you acknowledge that the terms and conditions set out in this License Agreement shall prevail.

Bentham Science Publishers Pte. Ltd.
80 Robinson Road #02-00
Singapore 068898
Singapore
Email: subscriptions@benthamscience.net

CONTENTS

FOREWORD ... i

PREFACE ... ii

LIST OF CONTRIBUTORS ... iii

CHAPTER 1 THE ROLE OF MATHEMATICS IN DATA SCIENCE: METHODS, ALGORITHMS, AND COMPUTER PROGRAMS ... 1

Rashmi Singh, Neha Bhardwaj and *Sardar M. N. Islam (Naz)*

INTRODUCTION ... 1
DATA SCIENCE ... 2
MAIN MATHEMATICAL PRINCIPLES AND METHODS IMPORTANT FOR DATA SCIENCE ... 3
Linear Algebra ... 3
Matrices ... 3
System of Linear Equation ... 4
The Number of Solutions ... 5
Vectors ... 5
Loss Function ... 6
Regularization ... 7
Support Vector Machine Classification ... 7
Statistics ... 8
Probability Theory ... 8
Normal Distribution ... 8
Z Scores ... 10
The Central Limit Theorem ... 11
Some Other Statistical Methods ... 11
Skewness ... 11
Kurtosis ... 12
Applications of Statistics in Data Science through Machine Learning Algorithms ... 13
Regression ... 13
Machine Learning Using Principal Component Analysis to Reduce Dimensionality ... 14
Mathematical Basis of PCA ... 15
Classification ... 16
K-Nearest Neighbor ... 16
Naive Bayes ... 17
Calculus ... 17
Optimization or Operational Research Methods ... 18
Dynamic Optimization Model ... 19
Stochastic Optimization Methods ... 19
Some Other Methods ... 21
Computer Programs ... 21
CONCLUDING REMARKS ... 21
REFERENCES ... 21

CHAPTER 2 KALMAN FILTER: DATA MODELLING AND PREDICTION ... 24

Arnob Sarkar and *Meetu Luthra*

INTRODUCTION ... 24
Why Kalman Filter? ... 24
UNDERSTANDING THE KALMAN FILTER ... 25
What is Kalman Filter? ... 25

State Space Approach 26
Mean Squared Error 26
KALMAN FILTER EQUATIONS 27
GENERAL APPLICATIONS OF KALMAN FILTER 33
KALMAN FILTER EQUATIONS IN ONE DIMENSION 33
EXAMPLE 1: FINDING THE TRUE VALUE OF TEMPERATURE 35
First Iteration 36
Second Iteration 36
EXAMPLE 2: FINDING THE TRUE VALUE OF ACCELERATION DUE TO GRAVITY 37
EXAMPLE 3: VERIFYING HUBBLE'S LAW 41
LIMITATIONS OF KALMAN FILTER 45
OTHER FILTERS 46
FUTURE PROSPECTS 46
CONCLUDING REMARKS- KALMAN FILTER IN A NUTSHELL 47
APPENDIX – BASIC CONCEPTS 47
A.1. LINEAR DYNAMIC SYSTEMS 47
A.2. ERROR COVARIANCE MATRIX 48
A.3. TULLY FISHER RELATION 49
A.4. RED SHIFTS AND RECESSIONAL VELOCITY 49
REFERENCES 50

CHAPTER 3 THE ROLE OF MATHEMATICS AND STATISTICS IN THE FIELD OF DATA SCIENCE AND ITS APPLICATION 51
Sathiyapriya Murali and *Priya Panneer*
INTRODUCTION 51
Data Science 51
DATA SCIENCE IN MATHEMATICS 52
MATH AND DATA SCIENCE IN EDUCATION 52
TYPES OF DATA SCIENCE IN MATH 53
Linear Algebra 53
APPLICATION OF LINEAR ALGEBRA IN DATA SCIENCE 53
Loss Function 53
Mean Squared Error 53
MEAN ABSOLUTE ERROR 54
COMPUTER VISION 54
CALCULUS 54
CALCULUS IN MACHINE LEARNING 54
APPLICATIONS IN MEDICAL SCIENCE 55
APPLICATION IN ENGINEERING 55
APPLICATIONS IN RESEARCH ANALYSIS 55
APPLICATIONS IN PHYSICS 55
STATISTICS 55
Types of Statistics in Data Science 56
Descriptive Statistics 56
Inferential Statistics 56
Application of Statistics in the Field of Study 57
VITAL STATISTICS IDEAS OBTAINING STARTED 57
DISTRIBUTION OF DATA POINT 57
APPLIED MATH EXPERIMENTS AND SIGNIFICANCE TESTING 57
NONPARAMETRIC STATISTICAL METHODS 58

APPLICATION OF STATISTICS IN DATA SCIENCE ANALYZING AND CATEGORIZING DATA 58
NUMERIC DATA & CATEGORICAL DATA 58
EXPLORATORY KNOWLEDGE ANALYSIS 58
SIGNIFICANCE TESTS 58
Null Hypotheses 59
Alternative Hypotheses 59
CHI-SQUARED CHECK 59
STUDENT’S T-TEST 59
ANALYSIS OF VARIANCE CHECK (ANOVA) 59
Unidirectional 60
Two-ways 60
RESERVATION AND PREDICTION 60
Linear Regression 60
Logistic Regression 60
CLASSIFICATION OF KNOWLEDGE SCIENCE IN STATISTICS 60
Naive Mathematician 61
K-nearest Neighbors 61
PROBABILITY 61
FREQUENCY TABLES 61
HISTOGRAM 62
CONTINUOUS RANDOM VARIABLES 63
SKEWNESS DISTRIBUTION 63
RIGHT SKEW DISTRIBUTION 63
LEFT SKEW DISTRIBUTION 63
NORMAL DISTRIBUTION 63
EXPONENTIAL DISTRIBUTION 64
UNIFORM DISTRIBUTION 64
POISSON DISTRIBUTION 64
IMPORTANT OF INFORMATION SCIENCE 64
DATA WHILE NOT KNOWLEDGE SCIENCE 65
DATA CAN PRODUCE HIGHER CLIENT EXPERTISE 65
DATA USED ACROSS VERTICALS 65
POWER OF INFORMATION SCIENCE 65
FUTURE OF INFORMATION SCIENCE 65
DATA SCIENCE IN TRADE 66
BENEFITS OF KNOWLEDGE SCIENCE 66
STATISTICAL INFORMATION 66
DATA SCIENCE IS VERY IMPORTANT IN THE MODERN WORLD 66
DATA INDIVIDUAL 66
DATA SCIENCE WORKS 67
CONCLUDING REMARKS 67
REFERENCES 67

CHAPTER 4 BAG OF VISUAL WORDS MODEL - A MATHEMATICAL APPROACH 68
Maheswari
INTRODUCTION 68
HISTOGRAM REWEIGHTING – TF – IDF APPROACH 69
COST MATRIX GENERATION 69
EUCLIDEAN DISTANCE AND COSINE DISTANCE 70
MODEL DESCRIPTION 71

Histogram Generation for Image 71
Computation of Cost Matrix 71
Reweighting of Histogram using TF – IDF 72
Visualization of Original Euclidean, Reweighted Euclidean 72
Normalization of Original Histogram 74
Checking for Similarity of the Normalized Histogram 74
Visual Comparison of Histograms 76
CONCLUSION 78
REFERENCES 78

CHAPTER 5 A GLANCE REVIEW ON DATA SCIENCE AND ITS TEACHING: CHALLENGES AND SOLUTIONS 80
Srinivasa Rao Gundu, Charanarur Panem and *J. Vijaylaxmi*
INTRODUCTION 80
THE IMPACT OF DATA SCIENCE ON THE SOCIETY 83
EDUCATIONAL GOALS OF DATA SCIENCE 83
DATA SCIENCE IN PRACTICE AS A PROBLEM SOLVING 84
LITERATURE REVIEW 85
DEMANDS OF THE DATA SCIENCE INDUSTRY AND THE DATA SCIENCE CURRICULUM 86
INHERENT PROBLEMS IN DATA SCIENCE CURRICULA DEVELOPMENT 88
TEACHING DATA SCIENCE 89
CONCLUDING REMARKS 90
REFERENCES 90

CHAPTER 6 OPTIMIZATION OF VARIOUS COSTS IN INVENTORY MANAGEMENT USING NEURAL NETWORKS 92
Prerna Sharma and *Bhim Singh*
INTRODUCTION 92
RELATED WORK 93
ASSUMPTION AND NOTATIONS 94
MATHEMATICAL FORMULATION OF MODEL AND ANALYSIS 96
MULTILAYER-FEED FORWARD NEURAL NETWORKS 99
WORKING ON PROPOSED SYSTEM 102
EXPERIMENTAL RESULTS AND ANALYSIS 102
CONCLUDING REMARKS 102
REFERENCES 102

CHAPTER 7 CYBER SECURITY IN DATA SCIENCE AND ITS APPLICATIONS 105
M. Varalakshmi and *I. P. Thulasi*
INTRODUCTION 105
DATA SCIENCE TODAY 105
MOTIVE AND SIGNIFICANCE OF DATA SCIENCE 106
IMPORTANCE OF DATA 107
IMPORTANCE OF DATA SCIENCE 107
MOTIVATION OF DATA IMPORTANT INDUSTRIES 107
DATA SCIENCE FOR PREFERABLE TRADE 107
DATA ANALYTICS FOR CLIENT ACQUISITION 108
DATA ANALYTICS FOR REVOLUTION 108
DATA SCIENCE FOR ENHANCESURVIVAL 108
PART OF DATA SCIENCE IN CYBER SECURITY 108
CONNECTION ALLYING SUBSTANTIAL DATA AND CYBER SECURITY 109

DATA SCIENCE USED IN CYBER SECURITY 109
Negative Hoping on “Lab-based” Order 110
Utilize Entrance to Sufficient Data 110
Specialize in this Irregularity 110
Utilize Data Science in a Logical Approach 111
UPCOMING CHALLENGES IN CYBER SECURITY DATA SCIENCE 111
OPERATE CLASSIFICATION ISSUES IN CYBERSECURITY DATAFILE 111
RELIABILITY SCHEME RULE 111
AMBIENCE PERCEPTON IN CYBER SECURITY 112
ATTRIBUTE ENGINEERING IN CYBER SECURITY 112
PROMINENT SECURITY ACTIVE CREATION AND ARRAY 112
DISCUSSION 113
CONCLUDING REMARKS 114
REFERENCES 115

CHAPTER 8 ARTIFICIAL NEURAL NETWORKS FOR DATA PROCESSING: A CASE STUDY OF IMAGE CLASSIFICATION 116
Jayaraj Ramasamy, R. N. Ravikumar and *S. Shitharth*
INTRODUCTION 116
ARCHITECTURE OF ANN 117
Input Layer 118
Hidden Layer 118
Output Layer 118
BENEFITS OF ARTIFICIAL NEURAL NETWORK (ANN) 119
Ability for Processing 119
Network-based Data Storage 119
Capacity to Function Despite a Lack of Knowledge 119
Transmission of Memory 119
Acceptance for Faults 119
DISADVANTAGES 120
Ensure that the Network Structure is Correct 120
Network Activity that has Gone Unnoticed 120
Network's Life Expectancy is Unknown 120
WORKING OF ANN 120
TYPES OF ANN 121
Feedback ANN 121
Feed-Forward 121
SIMPLE NEURAL NETWORK 121
LITERATURE REVIEW 122
PROPOSED SYSTEM 123
RESULTS AND DISCUSSION 124
CONCLUSION 125
REFERENCES 125

CHAPTER 9 CARBON EMISSION ASSESSMENT BY APPLYING CLUSTERING TECHNIQUE TO WORLD’S EMISSION DATASETS 128
Nitin Jaglal Untwal
INTRODUCTION 128
Research Methodology 130
Limitations of the Study 130
Feature Extraction and Engineering 131
Data Extraction 131

Standardizing and Scaling 132
Identification of Clusters by Elbow Method 132
Cluster Formation 134
RESULTS AND ANALYSIS 134
Cluster One – High Rainfall 134
Cluster Two 141
Cluster Three 141
Cluster Four 142
Cluster Five 142
Cluster Six 142
CONCLUSION 143
REFERENCES 143

CHAPTER 10 A MACHINE LEARNING APPLICATION TO PREDICT CUSTOMER CHURN: A CASE IN INDONESIAN TELECOMMUNICATION COMPANY 144
Agus Tri Wibowo, Andi Chaerunisa Utami Putri, Muhammad Reza Tribosnia, Revalda Putawara and *M. Mujiya Ulkhaq*
INTRODUCTION 144
LITERATURE REVIEW AND CONTRIBUTION 146
RESEARCH DESIGN 147
Dataset 147
Data Preparation 148
Exploratory Data Analysis 148
Features Selection 150
MACHINE LEARNING APPLICATION 151
Ridge Classifier 151
Gradient Booster 151
Adaptive Boosting 151
Bagging Classifier 152
k-Nearest Neighbor 152
Decision Tree 153
Logistic Regression 153
Random Forest 153
MODEL PERFORMANCE AND EVALUATION 154
RESULT 155
CONCLUDING REMARKS 159
REFERENCES 160

CHAPTER 11 A STATE-WISE ASSESSMENT OF GREENHOUSE GASES EMISSION IN INDIA BY APPLYING K-MEAN CLUSTERING TECHNIQUE 162
Nitin Jaglal Untwal
INTRODUCTION 163
Introduction to Cluster Analysis 163
Research Methodology 165
Data Source 165
Period of Study 165
Software used for Data Analysis 165
Model Applied 165
Limitations of the Study 165
Future Scope 165
Research is Carried Out in Five Steps 165
Feature Extraction and Engineering 165

Data Extraction 166
Standardizing and Scaling 166
Identification of Clusters by Elbow Method 166
Cluster formation 169
RESULTS AND ANALYSIS 169
Cluster One 171
Cluster Two 173
Cluster Three 174
CONCLUSION 175
REFERENCES 175

CHAPTER 12 DATA MINING TECHNIQUES: NEW AVENUES FOR HEART DISEASE PREDICTION 177
Soma Das
INTRODUCTION 177
ADVERSE IMPACT OF CARDIOVASCULAR DISEASES IN INDIA 179
Smoking 179
Hyperglycaemia 179
Hypertension 179
Obesity 180
Dyslipidaemia 180
Dietary Habits and Exercise 180
Genetic Risk Factors 180
Treatment Gaps 181
The Multilayer Perceptron (MLP) 181
Coactive Neuro-Fuzzy Inference System (CANFIS) 182
Aptamer Biochip-based CDSS –ensemble (Apta CDSS-E) 182
Intelligent Heart Disease Prediction System (IHDPS) 182
Intelligent and Effective Heart Attack Prediction System (IEHPS) 183
Decision Tree Fuzzy System (DTFS) 183
CONCLUDING REMARKS 183
REFERENCES 184

CHAPTER 13 DATA SCIENCE AND HEALTHCARE 186
Armel Djangone
INTRODUCTION 186
So, What is Data Science? 186
Data Science Techniques vs. Data Mining 187
Now, Why is Data Essential? 187
What is an Ideal Data Scientist? 188
Technical and Soft Skills for Healthcare Data Scientists 188
Technical Skills 189
Soft Skills 189
Why is Data Science so Crucial for Organizations? 190
HEALTHCARE DATA: CHALLENGES AND OPPORTUNITIES 191
Opportunities 191
Defining Big Data 193
Challenges 194
Data Science Opportunities for Healthcare 195
HEALTHCARE LEADERSHIP 197
Transactional leader 198
Transformational leadership 198

CONCLUDING REMARKS 200
REFERENCES 200

SUBJECT INDEX 201

FOREWORD

There is a need to provide a new, up-to-date, comprehensive, and innovative review of the developments to show, integrate, synthesize and provide future research directions in the applications of advanced mathematics in data science. Therefore, this book has made a valuable contribution to the literature by providing systematic reviews on the interrelationships between mathematics, statistics, and computer science.

Data Science is one of the most significant advances of this century. It deals with the collection, preparation, analysis, visualization, management, and preservation of this data – both structured and unstructured. Data science incorporates several technologies and academic disciplines to discover, extract, compile, process, analyze, interpret, and visualize data. It includes mathematics, statistics, computer science and programming, statistical modeling, database technologies, signal processing, data modeling, artificial intelligence, machine learning, natural language processing, visualization, and predictive analytics.

Mathematics is very important in the field of data science as concepts within mathematics aid in identifying patterns and assist in creating algorithms. Understanding various statistics and probability theory notions is key to implementing such algorithms in data science.

This book provides a comprehensive account of the areas of the applications of advanced mathematics in data science. It has covered many significant issues, methods, and applications of data science and mathematics in some crucial areas, such as The Role of Mathematics in Data Science, Mathematical Modeling in Data Science, Mathematical Algorithms for Artificial Intelligence and Big Data, Soft Computing in Data Science, Data Analytics: Architecture, Opportunities, And Open Research Challenges, Linear Regression, Logistic Regression, Neural Networks, and a Review on Data Science Technologies.

The book has implications for data science modeling and many real-life applications. Many readers, including undergraduate university students, evening learners, and learners participating in online data science courses, will be benefitted from this book.

I recommend this book to all interested in data science technologies, mathematical modeling, and applications.

S.B. Goyal
Faculty of Information Technology
City University
Petaling Jaya, 46100, Malaysia

PREFACE

The title of our book is Advanced Mathematical Applications in Data Science. The book is dealing specially Data Analysis – Mining and analysis of Big Data, Mathematical modelling in Data science, Mathematical Algorithms for Artificial Intelligence and Big Data, using MATLAB with Big Data from sensors and IOT devices, the relationship between Big data and Mathematical modelling, Big IOT Data analytics, Architecture, opportunities and open research challenges, the role of Mathematics in Data science, linear regression, logistic regression, Neural networks, Decision tree, applications of linear algebra in Data science, Big Data and Big Data analytics, concepts, types and techniques, foundation of Data science, fifty year of Data sciences, Health Bank – a world health for Data science applications in Healthcare, Radio frequency identification, a new opportunity for Data science, towards a system building agenda for data, semantic representation of Data science properly, a review on Data science techniques, Big Data: the next era of Information and Data science in medical imaging, Data science and healthcare, soft computing in Data science, foundation for private, fair and robust Data science, Data science fundamental principles, practical Data sciences for Actuarial task *etc*.

The scope of this book is not only limited to above highlighted areas but much more than that. Today as all of us are aware that most of the decision making and marketing strategies are data driven. So the research in this field is very much important and useful for any kind of day to day decision making and for marketing strategies *etc*. Finally we would thank the Bentham Science publishing house for giving us an opportunity to explore this field.

Biswadip Basu Mallik
Department of Basic Science & Humanities
Institute of Engineering & Management
Kolkata, West Bengal
India

Kirti Verma
Department of Engineering Mathematics
Lakshmi Narain College of Technology
Jabalpur Madhya Pradesh
India

Rahul Kar
Department of Mathematics
Kalyani Mahavidyalaya, Kalyani
West Bengal, India

Ashok Kumar Shaw
Department of Basic Sciences and Humanities
Budge Budge Institute of Technology, Budge Budge
Kolkata, West Bengal
India

&

Sardar M. N. Islam (Naz)
ISILC, Victoria University
Melbourne, Australia

List of Contributors

Armel Djangone	Dakota State University, Business Analytics and Decision Support, Washington Ave N, Madison, United States
Arnob Sarkar	National Atmospheric Research Laboratory, Department of Space, Andhra Pradesh, Government of India
Agus Tri Wibowo	Department of Consumer Service, PT Telekomunikasi Indonesia, Jakarta, Indonesia
Andi Chaerunisa Utami Putri	Department of Consumer Service, PT Telekomunikasi Indonesia, Jakarta, Indonesia
Bhim Singh	Department of Basic Science, Sardar Vallabh Bhai Patel University of Agriculture and Technology, Meerut (U.P.), India
Charanarur Panem	Department of Cyber Security and Digital Forensics, National Forensic Sciences University Tripura Campus, Tripura, India
J. Vijaylaxmi	PVKK Degree & PG College, Anantapur, Andhra Pradesh, India
Jayaraj Ramasamy	Department of IT, Botho University, Gaborone, Botswana
M. Varalakshmi	Marudhar Kesari Jain College for Women, Vaniyambadi, Tirupattur(dt), Tamilnadu, India
Meetu Luthra	Department of Physics, Bhaskaracharya College of Applied Sciences, University of Delhi, Delhi, India
Maheswari	Department of Computer Applications, Fatima College, Madurai, India
Muhammad Reza Tribosnia	Department of Consumer Service, PT Telekomunikasi Indonesia, Jakarta, Indonesia
M. Mujiya Ulkhaq	Department of Industrial Engineering, Diponegoro University, Kota Semarang, Indonesia Department of Economics and Management, University of Brescia, Brescia BS, Italy
Neha Bhardwaj	Department of Mathematics, School of Basic Sciences and Research, Sharda University, Noida, Uttar Pradesh, India
Nitin Jaglal Untwal	Maharashtra Institute of Technology, Aurangabad, India
I. P. Thulasi	Marudhar Kesari Jain College for Women, Vaniyambadi, Tirupattur(dt), Tamilnadu, India
Priya Panneer	Department of Mathematics, Mathematics Marudhar Kesari Jain College for Women, Vaniyambadi, Tirupattur, Tamilnadu, India
Prerna Sharma	Department of Basic Science, Sardar Vallabh Bhai Patel University of Agriculture and Technology, Meerut (U.P.), India
Rashmi Singh	Amity Institute of Applied Sciences, Amity University, Noida, Uttar Pradesh, India
R. N. Ravikumar	Department of Computer Engineering, Marwadi University, Gujarat, India
Revalda Putawara	Department of Consumer Service, PT Telekomunikasi Indonesia, Jakarta, Indonesia

Sardar M. N. Islam (Naz)	ISILC, Victoria University, Melbourne, Australia
Sathiyapriya Murali	Department of Mathematics, Mathematics Marudhar Kesari Jain College for Women, Vaniyambadi, Tirupattur, Tamilnadu, India
Srinivasa Rao Gundu	Department of Digital Forensics, Malla Reddy University, Dhulapally, Hyderabad, Telangana, India
S. Shitharth	Department of Computer Science, Kebri Dehar University, Kebri Dehar, Ethiopia
Soma Das	Life Science, B.Ed. Department, Syamaprasad Institute of Education and Training, Kolkata, India. Honorary Guest Faculty, Sports Science Department, University of Calcutta, Kolkata, India

CHAPTER 1

The Role of Mathematics in Data Science: Methods, Algorithms, and Computer Programs

Rashmi Singh[1,*], **Neha Bhardwaj**[2] and **Sardar M. N. Islam (Naz)**[3]

[1] *Amity Institute of Applied Sciences, Amity University, Noida, Uttar Pradesh, India*

[2] *Department of Mathematics, School of Basic Sciences and Research, Sharda University, Noida, Uttar Pradesh, India*

[3] *ISILC, Victoria University, Melbourne, Australia*

Abstract: The field of data science relies heavily on mathematical analysis. A solid foundation in certain branches of mathematics is essential for every data scientist already working in the field or planning to enter it in the future. In whatever area we focus on, data science, machine learning engineering, business intelligence development, data architecture, or another area of expertise, it is important to examine the several kinds of mathematical prerequisites and insights and how they're applied in the field of data science. Machine learning algorithms, data analysis and analyzing require mathematics. Mathematics is not the only qualification for a data science education and profession but is often the most significant. Identifying and translating business difficulties into mathematical ones are a crucial phase in a data scientist's workflow. In this study, we describe the different areas of mathematics utilized in data science to understand mathematics and data science together.

Keywords: Baye's theorem, Classification, Computer programs, Data science, Linear algebra, Machine learning, Matrices, Normal distribution, Optimization, Regression, System of linear equations, Vectors.

INTRODUCTION

To analyze data for the sake of decision making, "Data Science" combines different subfields of work in mathematics/statistics and computation in order to accomplish this. The use of the word "science" suggests that the discipline in question follows methodical procedures to arrive at findings that can be verified.

[*] **Corresponding author Rashmi Singh:** Amity Institute of Applied Sciences, Amity University, Noida, Uttar Pradesh, India; E-mail: rsingh7@amity.edu

Biswadip Basu Mallik, Kirti Verma, Rahul Kar, Ashok Kumar Shaw & Sardar M. N. Islam (Naz) (Eds.)

The discipline makes use of ideas that are derived from the fields of mathematics and computer science since the solutions to the following problems can be found in the findings that are achieved *via* kinds of columns given below. such processes: making a Netflix movie suggestion, financial projections for the company, a home's price can be estimated by comparing it to other properties of a similar size and quality in terms of factors like the number of rooms and square footage, a song suggestion for Spotify playlist as discussed [1, 2, 3, 4]. How, therefore, does mathematics come into play here? In this chapter, we give evidence for the claim that mathematics and statistics are crucial because they provide the means to discover patterns in data. Furthermore, newcomers to data science from other fields can benefit greatly from familiarity with mathematics.

DATA SCIENCE

Data science uses the tools and methods already available to discover patterns, generate meaningful information, and make decisions for businesses. Data science builds prediction models with machine learning.

As discussed [5], data can be found in a variety of formats, but it is useful to think of it as the result of an unpredictable experiment whose outcomes are up to interpretation. In many cases, a table or spreadsheet is used to record the results of a random experiment. To facilitate data analysis, variables (also known as features) are typically represented as columns and the items themselves (or units) are represented as rows. To further understand the utility of such a spreadsheet, it is helpful to consider three distinct kinds of columns given below:

- In most tables, the first column serves as an identifier or index, where a specific label or number is assigned to each row.

- Second, the experimental design can be reflected in the columns' (features') content by identifying which experimental group a given unit falls under. It is not uncommon for the data in these columns to be deterministic, meaning they would remain constant even if the experiment was repeated.

- The experiment's observed data is shown in the other columns. Typically, such measurements are not stable; rerunning the experiment would result in different results [6].

Many data sets can be found online and in various software programs.

Data science study may be divided as follows:

1. Acquire, enter, receive, and extract information from signals and data using these key phrases related to data capture. At this juncture, we are collecting both structured and unstructured data in their raw forms.

2. Data Architecture, Data Processing, Data Staging, Data Cleansing, and Data Warehousing all need regular upkeep. At this point, the raw data will be taken and transformed into a format that the next stage can utilize.

3. Data processing consists of data mining, data summarization, clustering and classification, data wrangling, data modeling, *etc.* Once the data has been prepared, data scientists evaluate its potential for predictive analysis by looking for patterns, ranges and biases.

4. Some analytics/analysis methods are exploratory, confirmatory, predictive, text mining, and qualitative. At this point, the data will be analyzed in several ways.

5. Communication is required in a number of different areas, including the reporting of data, the display of data, business intelligence, and decision-making. The final step in the process involves analysts producing the findings in formats that are simple to grasp, such as charts, graphs, and reports.

Applying such algorithms in data science requires familiarity with numerous topics, from mathematics, probability theory, and statistics. However, almost every single topic of today's data science methods, including machine learning, is rooted in rigorous mathematics.

MAIN MATHEMATICAL PRINCIPLES AND METHODS IMPORTANT FOR DATA SCIENCE

Linear Algebra

The fields of data science and machine learning can benefit tremendously from using linear algebra, a branch of mathematics. Learning linear algebra is the most important mathematical ability for anyone interested in machine learning. The vast majority of machine learning models may be written down as matrices. A dataset is frequently represented as a matrix in its own right. Linear algebra is employed in data pre-processing, data transformation, and model evaluation (see [4, 5, 7, 8]).

Matrices

The building elements of data science are matrices. They appear in a variety of linguistic personas, from Python's NumPy arrays to R's data frames to MATLAB's matrices.

CHAPTER 2

Kalman Filter: Data Modelling and Prediction

Arnob Sarkar[1] and **Meetu Luthra**[2,*]

[1] *National Atmospheric Research Laboratory, Department of Space, Andhra Pradesh, Government of India*

[2] *Department of Physics, Bhaskaracharya College of Applied Sciences, University of Delhi, Delhi, India*

Abstract: We provide here an analysis of Kalman filter, which has wide applications in the experimental and observational fields. Kalman filter is a data fusion algorithm or a mathematical tool which is based on the estimation theory. It basically is a set of mathematical equations which provide a computational mechanism for evaluating the state of discrete processes with noisy data. In fact, observations and data analysis is a very key aspect of all theories. In any set of data, to make it useful, one has to minimize the error/noise by taking into consideration various aspects like the estimated values (the theoretical values), the measurement values, experimental errors and the estimated errors. We have shown here how this can be done using Kalman Filtering technique. Kalman Filter is a tool which can take the observational data and improvise it to identify the best possible value of the parameters involved. Kalman filter and its variants such as the extended Kalman filter have wide applications mainly in the field of communication *e.g.*, in GPS receivers (global positioning system receivers), radio equipment used for filtering and removing noise from the output of laptop trackpads, image processing, face recognition and many more.

Keywords: Acceleration, Big data, Data science, Extended kalman filter, GPS, Kalman filter, Mathematical modelling, Noise, Signals, Speed, Uncertainty.

INTRODUCTION

Why Kalman Filter?

When we have a large set of data, efficient parameter estimation remains one of the most important tasks to perform [1]. There are a number of methods where data estimation or prediction algorithms are implemented to get a proper result. Situations arise when there exists noise in the signal. These noises are the unwanted signals which are responsible for incorrect and undesired output [2].

[*] **Corresponding author Meetu Luthra:**Department of Physics, Bhaskaracharya College of Applied Sciences, University of Delhi, Delhi, India; E-mail: meetu.luthra@bcas.du.ac.in

Biswadip Basu Mallik, Kirti Verma, Rahul Kar, Ashok Kumar Shaw & Sardar M. N. Islam (Naz) (Eds.)

There are various methods to remove the noise and get appropriate results. These include: Linear curve fitting, Quadratic curve fitting and other degrees of curve fitting.

Linear curve fitting, which solely depends on least square curve fitting, is optimal for small number of data points. But, with continuous noisy data inputs and undefined total number of data points, a suitable algorithm is required where input at every consecutive time interval is taken which includes noise, and a suitable path is estimated or predicted. This task can be accomplished by Optimal Filters.

There are various optimal filters to achieve this and get appropriate predicted results. One such filter is Kalman Filter. Using Kalman Filter, it is possible to counter unwanted signals or data inputs with an unknown total number of data points and make predictions of the desired variables. Kalman Filter is not the only filter to do so. Other filters like *Extended Kalman Filter*, *Unscented Kalman Filter* and *Particle Filters* also perform the same task, but with greater efficiency and are more practical. Although these filters are better, the computation costs are generally higher than the computation costs of Kalman Filter.

Kalman Filter was originally developed by Rudolf E. Kalman in his paper in 1960, where he intended to use the state space approach to an earlier known filter, the Wein's Filter [3].

UNDERSTANDING THE KALMAN FILTER

What is Kalman Filter?

Kalman Filtering is a mechanism by which the true value of the quantity (*e.g.*, velocity, position, *etc.* of an object) can be estimated by using a set of equations and consecutive data values through an *iterative process.* The values being measured contain some or the other kind of errors that are not predictable or are *random* and in addition, there exist inherent uncertainties or variations in the data too.

Kalman Filter is a *two*-step process consisting of a filter and a smoother. It begins with the first step and makes a prediction for the required variables for the next step. *State equations* and *error matrices* are updated. This process is called filtering. After the filtration process, the estimation does not have any information about previously estimated values. Once, a variable of interest has been filtered, it is then smoothed.

State Space Approach

There are systems that depend on time and may change their behaviour over time. Such systems are known as *Dynamic Systems*[1]. One such type of Dynamic System is the Linear Dynamic System where the system varies linearly with time.

It is important to predict how the system behaves after a given time. This is achieved in numerous ways, one of which has been discussed below:

For a Linear Dynamic System[2], a State-Space approach is used. For this State-Space approach, there should be a finite dimensional representation of a particular problem [4].

Let x be a state vector with N x1 dimension. The variation of this state vector x equals a constant matrix φ multiplied by the state vector (x) itself. The variation can take two forms:

a) **Flow**: System varies continuously with time (t)

$$\frac{dx(t)}{dt} = \phi x(t) \quad \textbf{(1)}$$

b) **Discrete**: The system changes its state at discrete intervals (..., $m-1$, m, $m+1$,).

$$x_{m+1} = \phi . x_m \quad \textbf{(2)}$$

Here, m is an integer representing the time step t_m.

Mean Squared Error

The following sections build up the basis for the Kalman Filter equation.[3]

A general way of describing a linear dynamic system is:

$$y_k = a_k + n_k \quad \textbf{(3)}$$

y_k is the observed variable which is time-dependent,

a_k is the gain term

x_k is a variable which bears the information

CHAPTER 3

The Role of Mathematics and Statistics in the Field of Data Science and its Application

Sathiyapriya Murali[1,*] and **Priya Panneer**[1]

[1] *Department of Mathematics, Mathematics Marudhar Kesari Jain College for Women, Vaniyambadi, Tirupattur, Tamilnadu, India*

Abstract: Mathematics is the rock-solid foundation of everything that happens when science is present, and it is also extremely important in the field of data science since mathematical ideas assist discover models and facilitate the development of algorithms. But, the concepts they present and the tools they enable are the only reasons statistics and arithmetic are so crucial to data science. There is a particular type of mathematical reasoning that is necessary to grasp data, beyond the fundamentals of calculus, discrete mathematics, and linear algebra. For the implementation of such algorithms in data science, a thorough understanding of the various principles of probability and statistics is essential. Machine learning is one of the many modern data science techniques that has a strong mathematical base. The evidence presented in this chapter backs up our earlier claim that math and statistics are the fields that offer the greatest tools and approaches for extracting structure from data. For newcomers coming from other professions to data science, math proficiency is crucial.

Keywords: Applications in medical science, Bayes' theorem, Binomial, Bernoulli, Computer vision, Calculus, Calculus in machine learning, Gaussian normal, Linear algebra, Loss function, Mean squared error, Mean absolute error, Nonparametric statistical methods, Regression.

INTRODUCTION

Data Science

"Data Science" combines many statistical disciplines with computer technology to clarify the significance of data as a factor in decision-making. The term "science" implies that it is a field that relies on systematic process to achieve results that can be tested. Because machine learning is a process that requires arithmetic to complete analyses and during the hunt for data insights, the field of data science requires a working knowledge of mathematics. Math is typically one of the most

[*] **Corresponding author Sathiyapriya Murali:** Department of Mathematics, Marudhar Kesari Jain College for Women, Vaniyambadi, Tirupattur, Tamilnadu, India; E-mail:sathiyamathvlr@gmail.com

Biswadip Basu Mallik, Kirti Verma, Rahul Kar, Ashok Kumar Shaw & Sardar M. N. Islam (Naz) (Eds.)

important subjects, even though it may not be the one that is most desired for your academic and professional path in data science. One of the most important paces in a data scientist's communication is to establish and comprehend work daring and condition others toward mathematical ones. Whether a data scientist, engineer who specializes in machine learning, developer who understands business intelligence, data architect, or another industry expert; it's possible that we still don't fully understand how our careers in data science will be organized. Nevertheless, various mathematical techniques and everything that data science uses them for should be considered. We may finally better follow the option of teaching mathematics since we would have a superior understanding of knowledge and attention.

DATA SCIENCE IN MATHEMATICS

Learning the theoretical underpinnings of data science or machine learning can be demoralizing because they include a wide range of mathematical disciplines and extensive web resources. In this essay, the goal is to present resources to raise the mathematical foundation needed for data science practical/research activity.

Data science relies heavily on mathematics since mathematical concepts are essential for understanding design and creating algorithms. Understanding various concepts from probability theory and statistics is essential for using these algorithms in data science. Regression, maximum likelihood estimation, knowledge of distributions (Binomial, Bernoulli, Gaussian (Normal)), and the Bayes' Theorem are all included in this course [1].

MATH AND DATA SCIENCE IN EDUCATION

For data scientists, regardless matter how far in the future their business careers take them, math is a crucial educational strength. It ensures that you can help a firm solve issues and grow more quickly, improve model displays, and successfully integrate complicated data into the management of business risks [2].

It should be ensured to develop the necessary mathematics knowledge and skill sets using a major online workshop provider like Easy to learn. They provide Data Science Certification Courses that guide students through all they need to know in order to pursue a career in data science, including math-related courses and applications.

TYPES OF DATA SCIENCE IN MATH

Linear Algebra

Ability to use linear algebra in data science processes and applications that are different and diverse. We have categorized these applications into several areas where linear algebra will be used to become a strong data scientist. dimension-holding property of basic machine learning. Computer vision, natural language processing, and reduction [2].

First and foremost, the fact is aware of linear algebra and deep learning techniques. Two pertinent data science applications of linear algebra are shown, together with the most valuable product of basic linear algebra's intended use. Simply put, we may refer to linear algebra as the “math of vectors” and “mathematics” of matrices. Principal Component Analysis is one example of how linear algebra is used in data science and machine learning to reduce the amount of data that can be processed. Deep Learning, neural networks, natural language processing, and other applications also use linear algebra.

According to Encyclopedia Britannica, Linear Algebra is a mathematical control that assigns with Vectors and Matrices and more typically Vector Spaces and Linear Transformation. A two-dimensional (or rectangular) array of numbers is one way to define a matrix. Linear algebra would play the role of Robin if Data Science were Batman. It's common to overlook this real sidekick. Nonetheless, it actually drives important areas of data science, such as computer vision and natural language processing.

APPLICATION OF LINEAR ALGEBRA IN DATA SCIENCE

Loss Function

The most effective machine learning algorithm collects data, analyzes it, and then builds a copy using various ways (linear regression, logistic regression, decision tree, random forest, *etc.*). They could then estimate a prospective data enquiry after discovering the solutions [2].

Mean Squared Error

Mean Squared Errors (MSE), which are straightforward to comprehend and typically determined total ably in the majority of regression issues, are almost definitely the most frequently used dropping error speak to. Data research was aided by the Mass Python Athenaeum, Scikit, Tens or Flow each featuring a built-

CHAPTER 4

Bag of Visual Words Model - A Mathematical Approach

Maheswari[1,*]

[1] *Department of Computer Applications, Fatima College, Madurai, India*

Abstract: Information extraction from images is now incredibly valuable for many new inventions. Even though there are several simple methods for extracting information from the images, feasibility and accuracy are critical. One of the simplest and most significant processes is feature extraction from the images. Many scientific approaches are derived by the experts based on the extracted features for a better conclusion of their work. Mathematical procedures, like Scientific methods, play an important role in image analysis. The Bag of Visual Words (BoVW) [1, 2, 3] is one of them, and it is helpful to figure out how similar a group of images is. A set of visual words characterises the images in the Bag of Visual Words model, which are subsequently aggregated in a histogram per image [4]. The histogram difference depicts the similarities among the images. The reweighting methodology known as Term Frequency – Inverse Document Frequency (TF-IDF) [5] refines this procedure. The overall weighting [6] for all words in each histogram is calculated before reweighting. As per the traditional way, the images are transformed into the matrix called as Cost matrix. It is constructed through two mathematical: Euclidean distances and Cosine distances. The main purpose of finding these distances is to detect similarity between the histograms. Further the histograms are normalized and both distances are calculated. The visual representation is also generated. The two mathematical methods are compared to see which one is appropriate for checking resemblance. The strategy identified as the optimum solution based on the findings aids in fraud detection in digital signature, Image Processing, and classification of images.

Keywords: Bag of visual words, Cost matrix, Cosine distance, Euclidean distance.

INTRODUCTION

Recent advances in information retrieval from images, in particular with mathematical methods, are helpful for research [7, 8]. The dissimilarities among a

[*] **Corresponding author Maheswari:** Department of Computer Applications, Fatima College, Madurai, India; E-mail: kpmshri123@gmail.com

Biswadip Basu Mallik, Kirti Verma, Rahul Kar, Ashok Kumar Shaw & Sardar M. N. Islam (Naz) (Eds.)

set of images can be identified by extracting the features from the images. The analysis of images using Bag of Words model is one of the efficient ways.

It describes the images as 'Visual words' rather than pixel values. Visual word is a generalised feature descriptor [9], most frequently, it is a mean value of the cluster of images. The aggregate occurrence of the words is represented as a histogram. It is impossible to distinguish the images from the histogram since it is not very expressive. In anology with this case, the images, mathematical computation and other processes are done through Python Programming.

HISTOGRAM REWEIGHTING – TF – IDF APPROACH

To get the better similarity from the histogram, an approach called reweighting of the histogram can be used. This is implemented through TF – IDF (Term Frequency – Inverse Document Frequency) weighting. The term "frequency" essentially converts histograms into units of length. Inverse Individual dimensions (words) are given weights based on how frequently they appear in all the images. Every bin in a histogram is reweighted, and the "uninformative" terms (*i.e.,* characteristics that appear frequently in images/everywhere) are downweighted and enhance the importance of rare words.

The reweighted words in the histogram are computed using TF-IDF formula:

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i}$$

where

n_{id} - Occurance of feature i in an image d;

n_d - Total number of features in an image d

n_i - Number of images that contain the feature i

N - Number of images

By substituting the above formula and the overall weighting, t_i is calculated. From the outcome, the similarity is observed by generating the Cost Matrix.

COST MATRIX GENERATION

The cost matrix is the matrix of all possible histogram comparisons (every image is compared with every image). This is one of the effective ways of comparison of images. The entries are the distance between the histograms. The entries in the cost matrix reflect the similarity. The Euclidean distance [10] and Cosine distance (1- cosine smilarity) are derived and higher cost separation between similar and dissimilar images is observed.

EUCLIDEAN DISTANCE AND COSINE DISTANCE

The Euclidean distance is one way to generate cost matrix. The Euclidean metric is given as:

$$\surd\sum(xi - yi)^2 \tag{1}$$

This is a well-known distance measure, which generalizes our notion of physical distance in two- or three-dimensional space to multidimensional space.

Another method is finding the Cosine distance, the cosine distance is calculated as:

$$1 - \frac{x \cdot y}{||x||||y||} \tag{2}$$

Where $\frac{x \cdot y}{||x||||y||}$ is termed as Cosine similarity.

When comparing both, the difference observed is for the unit length vectors, squared euclidean distance differs from cosine similarity up to a constant.

Consider having vector x and y with unit length, then:

$$||x - y||^2 = (x - y)^T(x - y) = x^Tx - 2x^Ty + y^Ty \tag{3}$$

Since ||x|| = ||y|| = 1 and so,

$$||x - y||^2 = 2 - 2x^Ty = 2 - 2\cos\theta \tag{4}$$

The relationship is defined as [5],

$$\text{Euclidean Distance} = \sqrt{2 - 2\text{cos\_similarity}}.$$

CHAPTER 5

A Glance Review on Data Science and its Teaching: Challenges and Solutions

Srinivasa Rao Gundu[1], **Charanarur Panem**[2,*] and **J. Vijaylaxmi**[3]

[1] *Department of Digital Forensics, Malla Reddy University, Dhulapally, Hyderabad, Telangana, India*

[2] *Department of Cyber Security and Digital Forensics, National Forensic Sciences University Tripura Campus, Tripura, India*

[3] *PVKK Degree & PG College, Anantapur, Andhra Pradesh, India*

Abstract: The word "data science" has become more popular in recent years, with a growing number of people embracing it. Only a small minority of people, on the other hand, are able to offer a clear explanation of what the term refers to when it is used in context. With no defined term to communicate and understand one another, it is difficult for organizations that are devoted to the collaboration, utilization, and application Data Science to communicate and understand one another.

As a result of technological advancements, it has become increasingly difficult to define and execute Data Science in a way that is compatible with how it was previously considered and understood in the past.

Specifically, we could now set out to develop definitions of Data Science that are representatives of current academic and industrial interpretations and perceptions, map these perspectives to newer domains of Data Science, and then determine whether or not this mapping translates into an effective practical curriculum for academics. Aspects of data science that differentiate it include how it is now used and how it is projected to be used in the future. Data science is also characterized by its ability to forecast the future.

Keywords: Curriculum, Data science, Quality, Practices, Problem solving.

INTRODUCTION

Despite the increasing relevance of data science, there is still no commonly accepted method of teaching it in both academia and industry.

[*] **Corresponding author Charanarur Panem:** Department of Cyber Security and Digital Forensics, National Forensic Sciences University Tripura Campus, Tripura, India; E-mail:panem.charan@gmail.com

Biswadip Basu Mallik, Kirti Verma, Rahul Kar, Ashok Kumar Shaw & Sardar M. N. Islam (Naz) (Eds.)

Students in data science courses at universities and colleges around the nation are growing more diverse in terms of their backgrounds, including practitioners, academic scholars, and data scientists. Interviews with twenty data scientists who teach in a variety of settings ranging from small-group workshops to massive online courses were conducted in order to gain a better understanding of how these practitioner-instructors transfer their expertise and how this differs from teaching conventional forms of programming, such as Python. To be effective, teachers must be sensitive to a diverse range of student backgrounds and expectations, teach technical workflows that integrate authentic practices around code, data analysis, and communication, and overcome challenges such as choosing authenticity over abstraction in software setup, finding and curating relevant datasets, and preparing students to live with uncertainty in data analysis, among other things. It is feasible that as a consequence of this research, more effective ways of teaching data science will be created, as well as an increase in the number of persons who are data-savvy [1].

“Data Science” (DS) has been a well-known word for quite some time now, and with good reason. It is difficult to define the background, skills, and collection of talents required by a Data Scientist, and it is much more difficult to find a definition that accurately describes these characteristics. This has resulted in the absence of any academic topic or programme that would allow certified Data Scientists to be trained at any academic institution or university. While some individuals believe that DS is a legitimate academic topic, others believe that it is more correctly defined as a mixture of traits and knowledge that would be impossible to uncover in a single person rather than an academic subject. According to the results of this study, data science is characterized in a number of ways by professionals from both the commercial and academic sectors [2]. The study has culminated in the proposal that a Master's Degree in Data Science is developed at the university level as a consequence of the findings.

As described by Wing (2019), the phrase data science is defined as the study of extracting value from data and the extraction of knowledge from data to meet business difficulties. Alternatively, “data science theory, method, and technology” or “data science technique and technology” or “data science technique and technology” have all been used to refer to this concept. In accordance with Irizarry, it is thought that the acronym DS stands for “data extraction”, which refers to “the entire complex and multi-step procedures required extracting value from data”. One of the fundamental premises of data science is the notion that “all procedures required” should be used in order to extract the most amount of value possible from massive, filthy, and unorganized datasets. Nonetheless, it is possible to construct an overarching definition from this broad notion that spans a wide variety of diverse sub-disciplines while keeping the overall concept in mind [3].

Upon closer inspection, we will be able to see that the definitions of what it takes to be a Data Scientist are much too broad and unclear in their application. To be able to provide degree programmes in developmental disabilities and define curricular requirements in this field, it is essential to have a clear definition of what constitutes a developmental disability.

According to certain circumstances, the needs of businesses and the leadership of academic institutions may be critical in determining what constitutes fundamental knowledge in data science and, as a result, the curriculum that will prepare data scientists to make contributions to businesses and society as they progress through their careers. Collaboration between industry and academia would be beneficial for designing a DS curriculum since it would enable the curriculum to be implemented immediately after development.

Data science is developing at a fast speed, and this will continue to be the case in the foreseeable future. It is vital to acknowledge that data scientists must be adaptive and versatile in order to be successful in their industry [4]. When learning new skills and abilities understanding the vocabulary and knowledge necessary to browse accessible resources, search for critical information, and judge the quality of a given resource is an important part of the process of learning new skills and abilities. Understanding the vocabulary and knowledge necessary to browse accessible resources, search for critical information, and judge the quality of a given resource is an important part of the process of learning new skills and abilities. It is necessary for students to get instruction and examples of appropriate behaviour from their instructors in order to comprehend how to make use of what is currently available and develop general approaches for gaining the next talent, which is not always obvious. It is preferable for teachers to utilize conventional tools in order to educate their students rather than outdated, proprietary, or specialized technology. As a direct consequence of these concerns, we strongly recommend all new course authors to take into consideration tools while developing their courses.

The possession of a technical background alone may not be enough to ensure long-term success in a data science position, particularly in the early stages of one’s professional career. A lot of data science careers have been automated, resulting in a decrease in the total number of data science experts. Among the suggestions is the notion that data scientists should focus on building talents that are more difficult to automate [5]. Examples include business insight; explanation; and storytelling, to name a few examples. Specifically, we

CHAPTER 6

Optimization of Various Costs in Inventory Management using Neural Networks

Prerna Sharma[1,*] and **Bhim Singh**[1]

[1] *Department of Basic Science, Sardar Vallabh Bhai Patel University of Agriculture and Technology, Meerut (U.P.), India*

Abstract: The process of maintaining the right quantity of inventory to meet demand, minimising logistics costs, and avoiding frequent inventory problems, including reserve outs, overstocking, and backorders is known as inventory optimisation. One has a finite capacity and is referred to as an owned warehouse (OW), which is located near the market, while the other has an endless capacity and is referred to as a rented warehouse (RW), which is located away from the market. Here, lowering the overall cost is the goal. Neural networks are employed in these works to either maximise or minimise cost. The findings produced from the neural networks are compared with a mathematical model and neural networks. Findings indicate that neural networks outperformed both conventional mathematical models and neural networks in terms of optimising the outcomes. The best way to understand supervised machine learning algorithms like neural networks is in the context of function approximation. The benefits and drawbacks of the two-warehouse approach compared to the single warehouse plan will also be covered. We investigate cost optimisation using neural networks in this chapter, and the outcomes will also be compared using the same.

Keywords: Business environment, Inventory, Neural networks, Warehouse.

INTRODUCTION

Every businessman needs a warehouse as it is an important location in the trading. Due to the current state of the market and its globalisation, the business community is extremely competitive, and everyone works hard to satisfy their consumers' needs. As a result, wholesalers and merchants always retain a stock of the products in their stores. Suppliers provide certain discounts on full sale purchases throughout the holiday season as well as various trade credit financing plans to attract the attention of their merchants in this unforgiving business environment.

[*] **Corresponding author Prerna Sharma:** Department of Basic Science, Sardar Vallabh Bhai Patel University of Agriculture and Technology, Meerut (U.P.) India; E-mail: prernam2002@gmail.com

Biswadip Basu Mallik, Kirti Verma, Rahul Kar, Ashok Kumar Shaw & Sardar M. N. Islam (Naz) (Eds.)

Retailers need extra space to store the products they buy in bulk during the accessible period in order to take advantage of these supplier programmes, but due to limited space in busy markets, retailers struggle to find storage space at their own single warehouse and must therefore rent another storage room to house their excess product purchases. To get out of this predicament, they lease another storage unit for a short time. This rental warehouse is utilised as an additional storage facility offered by public, private, or governmental organisations, and these facilities are used as the storage space that results from their use. In order to gain space on a hiring basis for storage needs, the idea of two warehouses was introduced in the inventory modelling. When modelling an inventory. Hartley introduced the idea of a two-warehouse system for the first time, and many authors have since used it, referring to one as a "Own warehouse" with a restricted capacity and the other as a "Rented warehouse" with an unlimited capacity. In this idea, it is frequently stated that because the owner of an additional warehouse provides better protection and preservation facilities, the carrying cost of items in rented warehouses is higher than those in owned warehouses. As a result, it is wise to store the items in rented warehouses first in order to reduce the holding costs incurred in rented warehouses.

Neural networks were inspired by the type of computation done by an individual intellect because they are simple and straightforward representations of the organic nervous system. Broadly speaking, a neural network is a highly unified network of several neurones, which are processing components in planning inspired by the brain. A neural network is believed to reveal parallel distributed distribution since it can be highly parallel.

Neural networks exhibit characteristics such as pattern relationships or mapping properties, generalisation, sturdiness, fault tolerance, and quick and simultaneous information processing.

RELATED WORK

In the year 2010, the authors [1] studied about multi-product inventory optimization using uniform crossover genetic algorithm. In other studies [2] and [3], authors have studied inventory optimization supply chain management and efficient supply chain management using genetic algorithms. Authors have studied about inventory analysis using genetic algorithm in supply chain management [4, 5, 6]. Kannan *et al.* discussed a genetic algorithm approach for a model for closed loop supply chain [7]. Jawahar and Balaji [8] studied for the two-stage supply chain distribution problems with a fixed charge using a genetic alogorithm. Authors rase the problem of modified Pareto genetic algorithm and multi-criterion optimization genetic algorithm [9]. Yimer and Demirli [10] presented same approach of dynamic supply chain scheduling. Similar types of studies were performed by various authors such as Wang *et. al.*. [11], Ram Kumar *et. al.* [12], Sherman *et. al.* [13]. Some authors have presented inventory model for breakable items [14]. A study [15] solved constrained knapsack problem in fuzzy environment for improved genetic algorithm. Two storage inventory problems involing dynamic demand and interval valued lead-time over finite time horizon were propoed by Dey *et al.* [16] Jawahar and Balaji [17] proposed "A genetic algorithm-based heuristic to the multi-period fixed charge distribution problem". Yadav *et al.* [18-22] studied various aspects on warehouse inventory model having deteriorating items with time-dependent demand, shortages and also with variable holding cost. Auto-warehouse model for deteriorating items were studied [23, 24] and after that authors have proposed it holding cost under particle swarm optimization. A focus on optimal ordering policy for non-instantaneous deteriorating articles with tentatively permissible delay in payment under two storage management was shown by Sharma *et al.* [25].

Table **1** represents the comparison between optimized results with neural network and the proposed sytem.

ASSUMPTION AND NOTATIONS

B_I: Maximum amount of inventory backlogged.

C_A: Ordering cost.

k: A Definite time interval to which holding cost remains constant.

h_w: It is holding cost per unit time in Ownware-house.

H_C: holding inventory cost.

h_r: The holding cost in rented ware house.

$I^r(x)$:Inventory level in rented warehouse.

$I^i(x)$:Inventory level in own ware-house.

$I^s(x)$:When the product has shortages, it denotes the inventory level.

L_c: The opportunity cost.

L_C: lost sale cost.

CHAPTER 7

Cyber Security in Data Science and its Applications

M. Varalakshmi[1,*] and **I. P. Thulasi**[1]

[1] *Marudhar Kesari Jain College for Women, Vaniyambadi, Tirupattur(dt), Tamilnadu, India*

Abstract: The implementation of data science in cyber security to help preserve against attacks and improve approach to better conflict cyber warning has many welfares. Honestly, data science has changed cyber security and the reaction has been profound and transformed. Cyber security uses data science to keep digital devices, services, systems, and software Safe from cyberattacks. Here, we talk about cyber security data science, present day uses for the cyber security field and data guide quickwitted managerial systems that can safeguard our system from cyber-attacks.

Keywords: Cyber security, Data science, Hack, Mathematics, Research, Statistics, Warning.

INTRODUCTION

Data science is a multifaceted field which works with the study methodology, algorithms, action, and method to understand commotion, correct and incorrect data, and apply the data across a wide-ranging application sector.

Combining statistics, data analysis, informatics, and their pertinent processes to identify and evaluate real words is a good concept. It differs from engineering and is used to approach and for ideas derived from diverse domains of mathematics, statistics, data science, and domain knowledge.

DATA SCIENCE TODAY

In the last 3 decades, data science has conservatively developed to carry establishment and management worldwide. It's currently employed by administration, biogenetics, and conjointly even cosmologists. According to the analysis, data science using vast amounts of data wasn't only about gathering data; it also included updating existing systems for managing data and the methods used to gather and analyze it.

[*] **Corresponding author M. Varalakshmi:** Marudhar Kesari Jain College for Women, Vaniyambadi, Tirupattur(dt), Tamilnadu, India; E-mail: varalakshmijanaki93@gmail.com

Biswadip Basu Mallik, Kirti Verma, Rahul Kar, Ashok Kumar Shaw & Sardar M. N. Islam (Naz) (Eds.)

Data science has developed into a vital component of commercial and academic analyses. Artificial intelligence, machine learning, voice recognition, the digital economy, and search engines are all components of technology. The organic sciences, health care, medical information studies, compassion, and public sciences have all become part of data science's analytical domain. Business science, government, business, and finance are presently influenced by data science.

One peculiar—and maybe harmful—result of the information science uprising has been a gradual shift towards writing more robust programmes. The data scientists have decided to extend unnecessary complex algorithms by investing an excessive quantity of flow and energy, when simpler special jobs may be done just as effectively. As a consequence, “novel” modifications that are noticeable happen less often.

Many data scientists today think that extensively revising is just too hazardous, therefore they instead try to shatter ideas into a tiny portion. Each component is tested before being properly phased into the information flow. Although more conservative programming is faster and more cost-effective, it also discourages exploration and prevents creative, “out-of-the-box” thinking and discoveries.

By using these safe-playing strategies, one may save time and money, avoid making costly errors, reduce the chance of running into serious obstacles, and get around real progress. Google stated: “One topic we tend to spend a lot of time discussing is how we prevent incrementalism when more significant changes are needed. It's difficult since these testing tools will greatly inspire the engineering staff, but they might also end up providing them with a strong incentive to comprehend just minor improvements. We definitely want those little improvements, but we also tend to want to think beyond the box.

MOTIVE AND SIGNIFICANCE OF DATA SCIENCE

Information is mostly used to look for patterns in statistical data. It uses several statistical techniques to explore and get knowledge from the specifics. The data should be completely surveyed by a data researcher from the data production, conflict, and preparation. They then have the responsibility of creating forecasts using the data. Data science is designed to draw conclusions from the data. With these outcomes, he will always be qualified to help businesses make well-dressed business decisions.

IMPORTANCE OF DATA

Data are crucial for model evaluation, characterisation, verification, activity, calibration, validation, and prognostication of the long-term structural robustness and presentation of materials in harsh environments. A lot of models would be useless if there were no reliable data to estimate and test them.

IMPORTANCE OF DATA SCIENCE

Data can work wonders. Industries want data to help them make informed decisions. Data science transformed recent data into comprehensive comprehension. Ultimately, information science is desired by enterprises. A data scientist is a magician who understands how to create magic with data. Every data a good data scientist comes across will be able to be mined for meaningful information. It benefits the business in the right manner. He is a guru and the organisation needs a stable data steering solution. The foundational fields of statistics and computer science are all strong points of data science. The ability to solve problems logically in business.

Use up the responsibility of data science focuses on the examination and direction of details; it depends on the industry's expertise in each sector, so data scientists must have lofty knowledge of the field.

MOTIVATION OF DATA IMPORTANT INDUSTRIES

Businesses need data. They need it in order to represent recommendations based on his or her data and provide a better customer experience. Now, allow me to walk you through the specific area where these businesses want to develop well-dressed data handling determination [1].

DATA SCIENCE FOR PREFERABLE TRADE

By offering a useful understanding of client preferences and behaviours, data science in trading may be utilised for channel optimisation, client segmentation, lead targeting and professional lead grading, real-time interactions, and other purposes. Never before has information been more readily available or necessary for managing a firm.

Industries need data to research their marketing strategy and create effective advertisements. Businesses often spend an enormous amount on the retailing of their goods. This sometimes may not provide the expected results. Hence, researching and looking at the client report industries will result in fantastic adverts.

CHAPTER 8

Artificial Neural Networks for Data Processing: A Case Study of Image Classification

Jayaraj Ramasamy[1,*], **R. N. Ravikumar**[2] and **S. Shitharth**[3]

[1] *Department of IT, Botho University, Gaborone, Botswana*

[2] *Department of Computer Engineering, Marwadi University, Gujarat, India*

[3] *Department of Computer Science, Kebri Dehar University, Kebri Dehar, Ethiopia*

Abstract: An Artificial Neural Network (ANN) is a data processing paradigm inspired by the way organic nervous systems, such as the brain, process data. The innovative structure of the information processing system is a crucial component of this paradigm. It is made up of a huge number of highly linked processing components (neurons) that work together to solve issues. Neural networks handle data in the same manner that the human brain does. The network is made up of several densely linked processing units (neurons) that operate in parallel to solve a given problem. They are unable to be programmed to execute a specific activity. ANN, like humans, learns by example. Through a learning process, an ANN is trained for a specific application, such as pattern recognition or data categorization. In biological systems, learning includes changes to the synaptic connections that occur between neurons. This is also true for ANNs. Artificial Neural Networks are used for classification, regression, and grouping. Stages of image processing are classified as preprocessing, feature extraction, and classification. It can be utilized later in the process. ANN should be provided with features and output should be classified. This paper provides an overview of Artificial Neural Networks (ANN), their operation, and training. It also explains the application and its benefits. Artificial Neural Network has been used to classify the MNIST dataset.

Keywords: Artificial neural network, Biological neural network, Neurons, Classification.

INTRODUCTION

Artificial neural network relates to a biological subfield of artificial intelligence that looks similar to a human brain model. The structure and composition of a

[*] **Corresponding author Jayaraj Ramasamy:** Department of IT, Botho University, Gaborone, Botswana; E-mail: jayaraj.ramasamy@bothouniversity.ac.bw

Biswadip Basu Mallik, Kirti Verma, Rahul Kar, Ashok Kumar Shaw & Sardar M. N. Islam (Naz) (Eds.)

biological neural network are used to design ANN architecture. Artificial neural networks [1], like the brain [2], are made up of neurons that are linked together at different network levels. These neurons are referred to as nodes. In Artificial Neural Networks, inputs are represented by dendrites from Biological Neural Networks, nodes are represented by cell nuclei, weights are represented by synapse, and output is represented by axon. An Artificial Neural Network is a sort of neural network that attempts to mimic the structure of neurons that make up the human brain in order for computers to interpret things and make decisions in a biological manner. Machines are designed to act simply as connected cells in the brain in a way to build an artificial neural network. There are around 100bn neurons in the human brain [3 - 6]. So every neuron has a connection between 1,000 to 100,000 points. The brain stores information in a way that it might be dispersed and can be recovered more than one piece of this knowledge from memory at the same moment if necessary. The brain may be thought of as a collection of incredibly strong multicore processors, as shown in Fig. (**1**) and Fig. (**2**) showing a typical BNN and an ANN diagram, respectively.

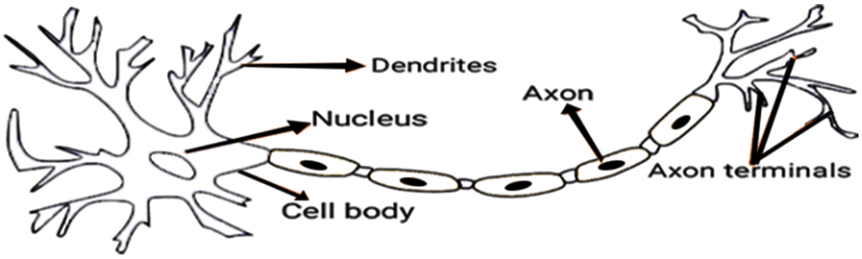

Fig. (1). Biological Neural Network (BNN).

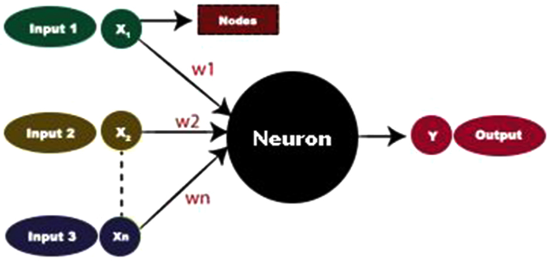

Fig. (2). Artificial Neural Network (ANN).

ARCHITECTURE OF ANN

We must first describe a neural network in order to appreciate the design about an artificial neural network. A neural network is defined by the placement of a great number of artificial neurons, referred as units, together in sequence of layers as shown in Fig. (**3**).

Let's take a glance at the many levels that a neural network might have.

- In the input layer, there are three input nodes.

- In the hidden layer, there are five hidden nodes.

- In the output layer, there are two output nodes.

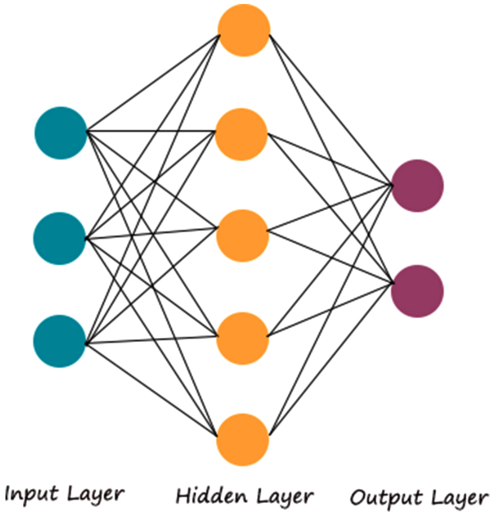

Fig. (3). Architecture of Artificial Neural Network.

Input Layer

It takes input in a variety of formats provided by the programmer, as the name indicates. Input nodes are indeed the outside-world inputs/information that the system use in addition to learning and making inferences. The information is sent from the input nodes to the next layer.

Hidden Layer

We have hidden layer located in between the visible and outgoing layer. It does all of the calculations required to identify hidden traits and patterns. The hidden layer is a cluster of neurons which do all of the operations on the information. There can be any number of hidden layers in a neural network. A single hidden layer is present in its most basic network.

Output Layer

The hidden layer transforms the input, leading in output that is sent over the same layer. When an input is received, the weighted total of the values, as well as a

CHAPTER 9

Carbon Emission Assessment by Applying Clustering Technique to World's Emission Datasets

Nitin Jaglal Untwal[1,*]

[1] *Maharashtra Institute of Technology, Aurangabad, India*

Abstract: The greenhouse gas emissions mostly include carbon-dioxide as the major component. The CO_2 level is increasing day-by-day which is a great cause of worry for the future world's environment. The reason why greenhouse gases' level increases in the environment is to be assessed and controlled. The greenhouse gases have heat-trapping capacity. A rise in numerous activities, including transportation, power production, agriculture, business, and residential, which are the main drivers of the increase in GHG levels in the atmosphere, is to blame for the rise in GHG emissions. Nitrous oxide, Methane, and Carbon Dioxide are all part of the GHG portfolio. Deforestation, traffic, and soil degradation all contribute to an increase in CO_2. As a result of burning biomass and urban trash, methane levels are also rising. The chlorofloro carbons are also rising due to refrigeration and industrial operations; so keeping the above concern in mind, the researcher had decided to conduct the study title. Carbon Emission Assessment by Applying Clustering Technique to World Emission Datasets using Python Programming. The study considers a period of 169 years (1750-2019). The study is carried out in five steps data fetching in python programming, feature engineering, standardization, clustering. The study generates 6 clusters. Cluster one contains 220 countries, cluster two includes Russia, France, Germany, China, Europe (others). America (others), Asia Pacific. Cluster three includes the United Kingdom. Cluster four includes the United States. Cluster five includes EU-28. Cluster six includes Malawi.

Keywords: Carbon emission, Data extraction, Engineering, Feature extraction, K-mean clustering, Python programming, Standardizing and scaling.

INTRODUCTION

The greenhouse gas emissions mostly include carbon-dioxide as a major component. The CO_2 level is increasing day-by-day which is a great cause of worry for the future existence of the world's environment. The reason why green-

[*] **Corresponding author Nitin Jaglal Untwal:** Maharashtra Institute of Technology, Aurangabad, India; E-mail:nitinuntwal@gmail.com

Biswadip Basu Mallik, Kirti Verma, Rahul Kar, Ashok Kumar Shaw & Sardar M. N. Islam (Naz) (Eds.)

house gases' level increases in the environment is to be assessed and controlled. The greenhouse gages have heat-trapping capacity.

The reasons for the increase in GHG emissions are due to an increase in various activities like transportation, electricity generation, agriculture, commercial and residential which are the major contributors to the growth of GHG levels in the atmosphere. The GHG portfolio includes carbon dioxide, Methane, and Nitrous oxide. CO_2 is increasing because of deforestation, vehicles, and soil degradation. Methane levels are also rising because of urban waste, and biomass burning. Cholorofloro carbons are also increasing because of refrigeration and industrial processes hence keeping the above problem in mind, and the researcher decided to conduct a study titled Carbon Emission Assessment by Applying Clustering Technique to World Emission Datasets.

Machine learning includes unsupervised learning under which models are trained for unlabeled data sets and are allowed to act without any supervision. Unsupervised learning is applied to understand the meaningful patterns, a grouping inherent in data, and extracting the generative features. Unsupervised learning is an algorithm that learns patterns from untagged data or unlabelled data [1, 2, 3, 4].

Cluster analysis is used to determine similarities and dissimilarities in a given data set or objects. Data usually have some similarities, which enable us to categorize or group them into clusters. The k-mean clustering is non-hierarchical. The reason for the popularity of k-means clustering is its simplicity. K means clustering is a type of partitioning method having objects as data observations with the nearest location and distance from each other. The nearest objects form mutually exclusive clusters. Each cluster has its centroid which makes clusters distinctive [5, 6, 7, 8].

Clustering is one of the important machine learning algorithms. Clustering is a technique of grouping elements; it is an important method for classification and grouping. K-mean clustering is used to classify elements into different categories based on the nearest distance from the mean. The main objective of K-mean clustering is creating a partition of n objects into k-clusters. Objects belonging to different clusters are considered based on the nearest mean. The method produces exactly k different clusters of greatest possible difference, which is known as a priori. K-mean clustering reduces the total intra-cluster variance or the squared error function [9, 10].

It is represented by the equation:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad \textbf{(1)}$$

where J is the objective function, k is the number of clusters, n is the number of cases, x is the number of cases i, cj is the number of the centroid.

Research Methodology

Data source

Data taken for the study is from the Kaggle database

Study period

The study period commences in 1750 and ends in 2019. The data selected for analysis is yearly data country-wise.

Software used for Data Analysis

Python Programming

Model applied

For purpose of this study, we applied K-mean clustering

Limitations of the Study

The study is restricted to only cluster analysis for Green House Gases country-wise.

Future scope

A similar kind of cluster analysis can be done for different continents.

Research is carried out in Five steps:

- Feature extraction and engineering
- Data extraction
- Standardizing and Scaling
- Identification of Clusters

CHAPTER 10

A Machine Learning Application to Predict Customer Churn: A Case in Indonesian Telecommunication Company

Agus Tri Wibowo[1], **Andi Chaerunisa Utami Putri**[1], **Muhammad Reza Tribosnia**[1], **Revalda Putawara**[1] and **M. Mujiya Ulkhaq**[2,3,*]

[1] *Department of Consumer Service, PT Telekomunikasi Indonesia, Jakarta, Indonesia*

[2] *Department of Industrial Engineering, Diponegoro University, Kota Semarang, Indonesia*

[3] *Department of Economics and Management, University of Brescia, Brescia BS, Italy*

Abstract: This study aims to develop a churn prediction model which can assist telecommunication companies in predicting customers who are most likely subject to churn. The model is developed by employing machine learning techniques on big data platforms. Customer churn is one of the most critical issues, especially in high investment telecommunication companies. Accordingly, the companies are looking for ways to predict potential customers to churn and take necessary actions to reduce the churn. To accomplish the objective of the study, it first compares eight machine learning techniques, *i.e.*, ridge classifier, gradient booster, adaptive boosting, bagging classifier, *k*-nearest neighbour (kNN), decision tree, logistic regression, and random forest. By using five evaluation performance metrics (*i.e.*, accuracy, AUC score, precision score, recall score, and the F score), kNN is selected since it outperforms other techniques. Second, the selected technique is used to predict the likelihood of customers churning.

Keywords: Customer churn prediction, Churn, *k*-nearest neighbour, Machine learning, Telecommunication company.

INTRODUCTION

In the era of advanced technology, recent studies found that telecommunication sector has evolved and emerged as one of the brightest businesses due to the current needs of customers [1]. It has become one of the key sectors in the developed countries; hence, the level of competition increased as a result of technological advancement and growth in telecommunication providers [2].

[*] **Corresponding author M. Mujiya Ulkhaq:** Department of Industrial Engineering, Diponegoro University, Kota Semarang, Indonesia & Department of Economics and Management, University of Brescia, Brescia BS, Italy; E-mail: ulkhaq@live.undip.ac.id

Biswadip Basu Mallik, Kirti Verma, Rahul Kar, Ashok Kumar Shaw & Sardar M. N. Islam (Naz) (Eds.)

The telecommunication providers could perform several strategies to generate additional revenue, such as: upsell the current clients, obtain new clients, and lengthen customer retention [3]. Comparing these strategies based on return-on-investment value found that the last strategy is the most beneficial one [3]. It shows that keeping an existing customer is considerably less expensive than obtaining a new customer [4], and also is substantially more effortless than the upselling strategy [5]. This third strategy requires businesses to reduce possible customer churn, or the migration of clients from one service provider to another. Statistics showed that 53% of all causes of customer churn are due to three leading causes, *i.e.*, 23% of poor onboarding, 16% of weak relationship building, and 14% of poor customer service [6].

There are several telecommunication providers in Indonesia; hence, businesses are arranging measures to survive in this cutthroat market. The phenomenon of *churn* obviously affects telecommunication providers in Indonesia and enforces them to create a new business strategy focusing on customer orientation, which puts the needs of the customers over the needs of the business. These providers could implement customer relationship management (CRM) to study customer satisfaction, loyalty, profitability, and customer retention [7]. While CRM's objective is to make a strong engagement with the customers, this approach is widely acknowledged and applied in many industries. Regarding the telecommunication industry, CRM can be used as a tool to gather information on the organization's marketing efforts, customers, competitors, contracts, and agreements [8]. Subsequently, CRM can also be implemented in the telecommunication industry where customers might switch their providers due to a variety of reasons, such as more comprehensive services, better pricing plans and connections, and so forth [9]; hence, it is necessary to define adequate models to accurately forecast the likelihood of customers to churn.

This study aims to develop a model to predict customers churn that can help telecommunication provider to forecast customers who are most likely to churn; and then predict the likelihood of customers to churn. To do so, it first compares several machine learning algorithms to predict customer churn, *i.e.*, gradient booster, ridge classifier, *k*-nearest neighbor (kNN), adaptive boosting (AdaBoost), bagging classifier, random forest algorithms, logistic regression, and decision tree. This is to show how each machine-learning technique can be implemented to model customer churn. The comparison is performed in five performance evaluation scores, including AUC score, accuracy, F score, recall score, and precision score. Second, the selected technique is employed to predict the likelihood of customers to churn.

The remaining parts of this work are structured as follows. In the next section, we present a literature review discussing previous studies about implementing machine learning in telecommunication company in Indonesia. It shows that this research area is under-studied, especially among scholars in Indonesia. The research design is discussed in Section 3. Section 4 discusses briefly machine learning techniques. Section 5 shows how to compare and evaluate the machine learning techniques used. Section 6 shows the results, while the last section is the concluding remarks.

LITERATURE REVIEW AND CONTRIBUTION

Literature about implementing the machine learning (or data mining) technique in telecommunication companies in Indonesia is quite limited. To formally verify this claim, we conduct a literature review in the Scopus database (https://www.scopus.com/), following Mongeon and Paul-Hus [10] who mentioned, "Scopus includes most of the journals indexed in WoS [Web of Science]." This database provided access to scientific articles and a wide-ranging of journals from various fields. First, we used the following search terms: TITLE-ABS-KEY(("machine learning" OR "data mining" OR "knowledge discovery") AND ("customer*" OR "client") AND ("churn*" OR "evasion" OR "dropout") AND "telecom*").[1] It means that the articles which contained those search terms in the title, abstract, or keywords were extracted. The period of time was not limited. From a pragmatic point of view, only articles published in English were included. This search yielded 359 articles. In the second refinement, we added the term "Indonesia" into the previous search terms. This second search yielded only three articles. This low yield indicated that this research area was under-studied especially among scholars in Indonesia. All these three articles are discussed as follows.

Hartati *et al.* [11] investigated how to handle imbalanced data problem using a combination of synthetic minority over-sampling (SMOTE) and random under-sampling (RUS); and how to assess the performance of the model using only one classifier (*i.e.*, C4.5 classifier) with bagging approach. Result showed that a higher performance was obtained by implementing the SMOTE and RUS sampling methods. SMOTE was used to generate the synthetic data from the churn class to upsurge the probability of drawing the churn data; while RUS was employed to decrease the probability of an overfitting problem. Next, the authors also showed that the implementation of bagging approach (with number of bags equals to 7) in the classification was able to improve the F score.

Alamsyah and Salma [12] attempted to search for the best model for employee churn prediction using three widespread prediction models, *i.e.*, decision tree,

CHAPTER 11

A State-Wise Assessment of Greenhouse Gases Emission in India by Applying K-mean Clustering Technique

Nitin Jaglal Untwal[1,*]

[1] *Maharashtra Institute of Technology, Aurangabad, India*

Abstract: India is a vast country with variations in geography as well as in population density. The pollution in India is increasing day by day. The Greenhouse gas emission is on the rise due to various activities like agriculture, industry, power generation, transportation, *etc.* Carbon dioxide (CO_2), Carbon Monoxide (CO), and Methane (CH_4) are the major elements in greenhouse gases. The emission of greenhouse gases causes various threats to the environment and health. The states in India have been under development since independence. Various activities are on the rise. The states are not having balanced growth as far as the industrial and agriculture sectors are concerned. The powerhouse of industrial growth is the state of Maharashtra and Gujarat. The population density is also scattered in India. The states contribute differently to greenhouse gases emission and it is difficult for the government to make policy category-wise for the control of greenhouse gases emissions. The classification of states into different categories will help in the strategic formulation of policy and strategy for different states depending on their greenhouse gases emission and per capita analysis of these emissions. The per capita greenhouse gas emission is calculated by dividing the total emissions by the total population. After analyzing the above problem, the researchers have decided to conduct the study titled A state-wise Assessment of greenhouse gas emission in India by applying the K-mean Clustering Technique using Python Programming. Research is carried out in Five steps -Feature extraction and engineering, Data extraction, Standardizing and Scaling, Identification of Clusters, Cluster formation. The study period is 2020. The data selected for analysis is yearly data state-wise of different Indian states. Data taken for the study is from the Kaggle database. Findings - The k- mean algorithm (cluster analysis using Python Programming) classifies the states of India into three clusters. Cluster one includes 16 states of India *viz.* Arunachal, Assam, Bihar, Himachal Pradesh, Jammu & Kashmir, Jharkhand, Madhya Pradesh, Manipur, Meghalaya, Mizoram, Odisha, Rajasthan, Sikkim, Tripura, Uttar Pradesh, Uttarakhand. Cluster two includes 8 states of the India. *Viz* Andhra Pradesh, Goa, Gujarat, Karnataka, Kerala, Maharashtra, Tamilnadu, West Bengal. Cluster three includes 4 states of India *Viz* Haryana, Nagaland, Punjab, Chhattisgarh. The major contributors to greenhouse gase emission are in cluster three.

[*] **Corresponding author Nitin Jaglal Untwal:** Maharashtar Institute of Technology, Aurangabad, India; E-mail: nitinuntwal@gmail.com

Biswadip Basu Mallik, Kirti Verma, Rahul Kar, Ashok Kumar Shaw & Sardar M. N. Islam (Naz) (Eds.)

The medium-range emission for greenhouse gases emission are grouped in cluster two and Minimum Range greenhouse gase emission states are included in cluster one.

Keywords: Carbon emission, Data extraction, Feature extraction and engineering, K-mean clustering, Python programming, Standardizing and scaling.

INTRODUCTION

India is a vast country with variations in geography as well as in population density. The pollution in India is increasing day by day. The Greenhouse gase emission is on a rise due to various activities like agriculture, industry, power generation, transportation, *etc.* Carbon dioxide (CO_2), Carbon Monoxide (CO), and Methane (CH_4) are the major elements in greenhouse gases. The emission of greenhouse gases causes various threats to the environment and health. The states in India have been under development since independence. Various activities are on the rise. The states are not having balanced growth as far as the industrial and agriculture sectors are concerned. The powerhouse of industrial growth is the state of Maharashtra and Gujarat. The population density is also scattered in India. The contribution of states contributes differently to greenhouse gases emission and it is difficult for the government to make policy category-wise for the control of greenhouse gase emissions. The classification of states in different categories will help in the strategic formulation of policy and strategy for different states depending on their greenhouse gase emission and per capita analysis of greenhouse gase emissions. The per capita greenhouse gase emission is calculated by dividing the total emissions by the total population. After analyzing the above problem the researcher hat decided to conduct the study titled A state-wise Assessment of greenhouse gases emission by Applying the Clustering Technique to different states of India. Research is carried out in Five steps -Feature extraction and engineering, Data extraction, Standardizing and Scaling, Identification of Clusters, Cluster formation.

Introduction to Cluster Analysis

Machine learning includes unsupervised learning under for which models are trained for unlabeled data sets and are allowed to act without any supervision. Unsupervised learning is applied to understand the meaningful patterns, the grouping inherent in data, and extraction of the generative features. Unsupervised learning is an algorithm that learns patterns from untagged data or unlabelled data.

Cluster analysis is used to determine similarities and dissimilarities in a given data set or objects. Data usually have some similarities which enable us to categorize or group them into clusters. The k-mean clustering is non-hierarchical. The reason for the popularity of k-means clustering is its simplicity. K means clustering is a type of partitioning method having objects as data observations with the nearest location and distance from each other. The nearest objects form mutually exclusive clusters. Each cluster has its centroid which makes clusters distinctive [1, 2].

Clustering is one of the important machine learning algorithms. Clustering is a technique of grouping elements; it is an important method for classification and grouping. K-mean clustering is used to classify elements into different categories based on the nearest distance from the mean. The main objective of K-mean clustering is creating a partition of n objects into k-clusters [3, 4, 5, 6]. Objects belonging to different clusters are considered based on the nearest mean. The method produces exactly k different clusters of greatest possible difference which is known as a priori. K-mean clustering reduces the total intra-cluster variance or the squared error function [7, 8, 9, 10].

It is represented by the equation:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad \textbf{(1)}$$

Where J is the objective function, k is the number of clusters, n is a number of cases, x is a number of cases i, c_j is a number of the centroid.

Drawbacks of k mean clustering:

a. The early-stage clusters affect the overall results.
b. When dataset size is small clustering is not accurate.
c. As we give variables the same weightage we do not know which variable occupies more relevance in the clustering process.
d. The noise can reduce the accuracy of the mean which further pulls the centroid away from its original position.

How to overcome the above drawbacks:

a. To increase the accuracy data set should also be increased
b. Use median to prevent outlier (Noise)

CHAPTER 12

Data Mining Techniques: New Avenues for Heart Disease Prediction

Soma Das[1,2,*]

[1] *Life Science, B.Ed. Department, Syamaprasad Institute of Education and Training, Kolkata, India*

[2] *Honorary Guest Faculty, Sports Science Department, University of Calcutta, Kolkata, India*

Abstract: The medical management sector assembles a large volume of unexposed data on the health status of patients. At times this hidden data could be useful in diagnosing diseases and making effective decisions. For providing an appropriate way out and planning a diagnostic system based on this information, now-a-days, the newest data mining strategies are in use. In this study, a thorough review has been done on the identification of an effective heart disease prediction system (EHDPS) designed by neural network for the prediction of the risk level of cardiovascular diseases. The study focused on the observation of various medical parameters, namely, age, height, weight, BMI, sex, blood pressure, cholesterol, and obesity. Based on this study, a concept map has been designed on the prediction ways for individuals with heart disease with the help of EHDPS. The study assembled considerable information about the multilayer perceptron neural network with rear proliferation as the algorithm for data analysis. The current review work may be significant in establishing knowledge of the association between health factors related to the risk level of heart disease. The study also suggests means of early intervention and prevention of medical emergencies posed by the late detection of cardiovascular diseases, especially in the context of post COVID 19 complications.

Keywords: Data mining, Disease detection, Heart diseases, Neural network, Risk factors.

INTRODUCTION

The noncommunicable diseases (NCDs) generally include various types of cancers, chronic respiratory diseases, diabetes, cardiovascular diseases (CVD) and so on. These diseases together contribute to 60% of all deaths. CVDs are the leading cause among NCDs which account for 17.7 million deaths [1]. According to the World Health Organization (WHO), CVD related death in India contributes

[*] **Corresponding author Soma Das:** Life Science, B.Ed. Department, Syamaprasad Institute of Education and Training, Kolkata, India; & Honorary Guest Faculty, Sports Science Department, University of Calcutta, Kolkata, India; E-mail: somad.cal@gmail.com

Biswadip Basu Mallik, Kirti Verma, Rahul Kar, Ashok Kumar Shaw & Sardar M. N. Islam (Naz) (Eds.)

to one-fifth of global mortality specifically in immature age groups. The Global Burden of Diseases reported that nearly a quarter (24.8%) of all deaths in India are associated with CVD [2]. The global average age-specific mortality rate on account of CVDs is 235 out of 100 000 population, whereas it is much higher (272 out of 100 000 population) in India [2]. The disease burden due to CVDs in ratio expression at India and Global level is shown below in Fig. (**1**). The age-adjusted CVD mortality rate is found more in males (255–525 out of 100 000 population) than their females (225–299 out of 100 000 population) counterparts. The sharp rise in CVD cases has brought a profound change in the disease profile in India within a span of the last two decades. It is marked by the epidemiological transition from diseases caused by poor nutrition, maternal - childhood diseases and infectious diseases to noncommunicable diseases (NCDs) across the country [3]. Literature revealed that the disease overload contributed by undernutrition (like marasmus, kwashiorkor), diarrhoea, cholera, measles, maternal - childhood diseases decreased by 50% in the last two decades. In addition, there is a rise in the average life expectancy of individuals at 65.2 years which earlier was 58.3 years and slow pace of getting old of large number of people in the same time frame [3]. The NCDs specially CVDs have a direct association with aged people. Increased life expectancy is leading to rapid growth in NCDs burden, more specifically a sharp rise in CVD overload among the general population [3]. Presently, in India, around 66% of the total mortality cases are attributed to health issues linked with cardiovascular system [4]. The CVD-mediated severity has risen considerably in developing countries like India than in developed countries of the world [5, 6]. In comparison to the Western world, people in India got affected by CVD a decade earlier which unfortunately resulted in the appearance of moderate to severe health complications in the most productive Middle Ages of Indian population [4]. A recent study indicated the mortality cases due to CVD fall within 23% in Western populations below the age of 70 years whereas the similar cases in India rise up to 52% [7].

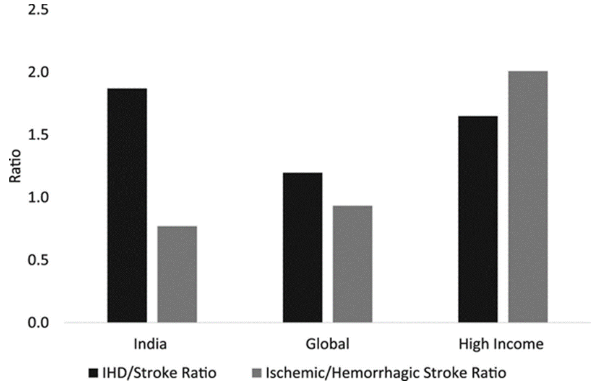

Fig. (1). Ratio Expression of CVDs at India and Global Level [2].

The World Health Organization (WHO) reported that India would pass through a huge economic burden of $237 billion due to massive spending on treatment and medical facilities for CVD patients in coming 10 years [7]. Prospective relevant studies reported myocardial infarction (MI) and cardiac arrest (CA) are the most common (83%) causes of CVD mediated complications which account for above one-fifth (21.1%) of mortality cases in India [2]. In this connection, these CVDs contribute to 1/10th of the premature death of the younger generation over the deaths of the older Indian population [8], and the value has risen by 59% from 1990 to 2010 [3].

ADVERSE IMPACT OF CARDIOVASCULAR DISEASES IN INDIA

A wide range of risk factors (RFs) having vast biological and social heterogeneity includes hypertension, diabetes mellitus, metabolic diseases, poor physical fitness, obesity, and other related risk factors like smoking, sedentary lifestyles, unbalanced food habits, poor dietary intake of fiber, mental stress, *etc.* as key players in the faster progression of CVDs, the high rate of fatalities, and the advanced propensity to develop CVD as the profound source of mortality across the country. Some of the prominent risk factors that must be addressed to improve awareness among common Indians are listed below:

Smoking

Presently, India surpasses China in the consumption of tobacco [9]. According to the Global Adult Tobacco Survey report, there has been a 6% decline in the consumption of tobacco among adult males in India [9]. However, the cases of tobacco smoking among Indian males (23.6%) are still higher than the cases worldwide (22%). The consumption of tobacco is a common threat to cardiorespiratory diseases; it can, however, be modified or detached from CVD.

Hyperglycaemia

Hyperglycemia, commonly known as diabetes, is one of the most common lifestyle diseases in India. It is so prevalent nowadays that it occurs in 1 out of 10 individuals within the age group of 18 years. The number of potent hyperglycemic patients is rising sharply in urban as well as rural Indian communities and is estimated to reach up to 8.8% between the age groups of 20 and 70 years in the near future [10].

Hypertension

Hypertension, or an increased blood pressure level, is another very common risk factor for CVD. One in every four Indians over the age of 18 has hypertension.

CHAPTER 13

Data Science and Healthcare

Armel Djangone[1,*]

[1] *Dakota State University, Business Analytics and Decision Support, Washington Ave N, Madison, United States*

Abstract: Data science is often used as an umbrella term to include various techniques for extracting insights and knowledge from complex structured and unstructured data. It often relies on a large amount of data (big data) and the application of different mathematical methods, including computer vision, NLP (or natural language processing), and data mining techniques. Advances in data science have resulted in a wider variety of algorithms, specialized for different applications and industries, such as healthcare, finance, marketing, supply chain, management, and general administration. Specifically, data science methods have shown promise in addressing key healthcare challenges and helping healthcare practitioners and leaders make data-driven decision-making. This chapter focuses on healthcare issues and how data science can help solve these issues. The chapter will survey different approaches to defining data science and why any organization should use data science. This chapter will also present different skills required for an effective healthcare data scientist and discusses healthcare leaders' behaviors that in impacting their organizational processes.

Keywords: Data science, Healthcare, Mathematical models, Machine learning, Natural language processing.

INTRODUCTION

So, What is Data Science?

Data Science is a crucial component of today's growing world. The reason is that data scientists have been reading and compiling data for a long time now. On average, data scientist spends 80% of their time collecting and cleaning data.

Therefore, the trends that an assembled data sets, the facts it produces, and all the new information gathered help develop organizations, healthcare, understanding a pattern, and overall building the world. Data science refers to the study of data. Data science allows the development of methods of collecting, storing, and analy-

[*] **Corresponding author Armel Djangone:** Dakota State University, Business Analytics and Decision Support, Washington Ave N, Madison, United States; E-mail: djangonearmel17@gmail.com

Biswadip Basu Mallik, Kirti Verma, Rahul Kar, Ashok Kumar Shaw & Sardar M. N. Islam (Naz) (Eds.)

zing data to find insightful information. Data science is essential for extracting knowledge and information from any type of structured or unstructured data.

Data science and computer science are two separate fields. Computer science is an area where algorithms and programs are developed to record and process data, while data science is about the analysis of data, with or without computers. However, fields of mathematics such as Statistics are related to data science as they can be applied for data collection, organization, analysis, and presentation.

Data science is not a building block for the IT industry, as modern companies and institutions deal with a massive amount of data. For example, someone with a massive amount of data would need to use data science so they can create useful approaches to collect, organize, and analyze the data.

Data Science Techniques *vs.* Data Mining

Data science and Data mining are often confused and used for one another. Although, data mining is actually a subset of the field of data science. Data mining is particularly aimed at analyzing large data (such as Big Data) to identify patterns, among other useful information. Data science, on the other hand, exclusively focuses on data collection and analysis.

Data science finds patterns within a set of data. With data extraction, we can record history, predict future possibilities, and understand our behavior. These closely similar interests are why users often may not accurately distinguish the role of a data scientist.

Now, Why is Data Essential?

Data is an essential facet of every industry and organization, including healthcare. According to the internet, just about 2TB of data is generated daily by users. This data comes in several types that may be either structured or unstructured. The proper use of data can allow companies to make the best decisions.

Data Science turns raw data into meaningful insights. Thus, businesses should integrate data science approaches. Data science has helped big companies in obtaining great heights. It helps in better marketing, and a company needs to understand its customers as the market grows. Data science helps understand the pattern a customer follows, what is more, liked and disliked. This way, data science has helped the commercial sector and the whole world with its innovation and discoveries. The marks of data acquired from several sources have led to the most significant age of innovation. Data Science is allowing us to discover the greatest of secrets hidden within the data. Data Science is the play-acting mechanism behind our current evolution, and it is growing at a petrifying pace.

What is an Ideal Data Scientist?

A Data Scientist works extensively with Big Data applications. The routine roles and responsibilities of a Data Scientist can be predictable sometimes, and sometimes they deal with extraordinary data challenges. There are several requirements to fulfill to become a Data Scientist. If someone is keen to be a data scientist, in that case, they must have the skills for munching data, making new assumptions, and the ability to look at the same problem from different angles, and so on.

A Data Scientist's job is to make the most of collected data by analyzing and deriving actionable insights. Some of those tasks may include:

- Identifying the data analytics issues that can unlock the most outstanding value for the organization.
- Learning about the most suitable datasets and variables.
- Working with different unstructured data forms like video or images.
- Exploring the latest opportunities and solutions through data analysis.
- Collecting massive structured and/or unstructured data from various sources.
- Reviewing, filtering, and validating data to gain the highest accuracy and completeness.
- Mining big data by evaluating and applying various algorithms.
- Analyzing the data set to identify underlying patterns and trends.
- Communicating different findings to relevant stakeholders through visualizations.

Technical and Soft Skills for Healthcare Data Scientists

The U.S. Bureau of Labor Statistics expects a massive job expansion in the field of data science by 2026, along with an estimation of 11.5 million new jobs within just the next six years. But there is a possibility of an even larger gap between the number of job opportunities and the number of job seekers in the healthcare industry particularly.

Furthermore, a HIMSS survey covers that only 38% of healthcare organizations are adequately staffed in their IT departments. These departments consist of data science, analysis, and data management professionals. Therefore, for any aspirant healthcare data worker, the following soft and technical skills need to be an integral part of your profile and portfolio.

SUBJECT INDEX

A

Activation functions, sigmoid 124
Activities 66, 196
  economic 66
  financial 196
Age, numerical 193
Aggregate 69, 192
  occurrence 69
  volume 192
Agriculture 162, 163, 172, 174
  activities 172, 174
  sectors 162, 163
Air resistance 40
Algorithms 61, 132, 151, 166, 181
  action 61
  distance-based 132, 166
  machine-learning 151
  statistical 181
Analytics 56, 113, 197
  cutting-edge 197
  cyber security information 113
ANN cells 119
Applied math experiments 57
Applying 128, 129, 132, 162, 165, 166, 167, 183
  clustering technique 128, 129
  data mining 183
  domain knowledge 165
  K-mean clustering 162
  python code 132, 166, 167
Artificial neural network 120, 121
  constructing 120
  works 121
Atherogenic dyslipidaemia 180

B

Bayesian networks (BN) 182
Bayes' theorem 17, 51, 52, 61
Bernoulli distribution 57
Biases, algorithmic 90
Big data 24, 186, 187, 188, 190, 193, 195
  analysis 195
  applications 188
Binary classification problem 119
Biological neural network (BNN) 116, 117
Blood pressure 177
BMI, high 180
Boltzmann constant 37
Boosting algorithms 151
Business 55, 65, 92, 189
  environment 92
  growth 189
  planning 55
  sectors 65

C

Carbohydrates 180
Carbon 128, 162, 163
  emission assessment 128
  monoxide 162, 163
Carbon dioxide 128, 129, 142, 162, 163
  environment protection agency 142
Cardiac 94, 102, 179, 180, 197
  arrest (CA) 94, 102, 179, 180, 197
  diseases 180
  dysfunction 180
Cardiovascular system 178
Chronic infection 181
Churn prediction 151
Cloud storage systems 87
Cluster 129, 130, 131, 133, 134, 162, 163, 165, 166, 169, 175
  analysis 129, 130, 131, 162, 163, 165, 166, 169, 175
  formation 131, 133, 134, 162, 163, 165, 169
Clustering 128, 163, 164
  process 164
  technique 128, 163
Coactive neuro-fuzzy inference system (CANFIS) 182

Coconut oil 180
Cognitive learning factor 102
Computational algorithms 18
Cosine distances 68, 70, 77, 78
Customer 144, 145, 154, 156, 159
  churn 144, 145, 154, 156, 159
  relationship management (CRM) 145
Cyber security 111, 112, 113, 114, 115
  data analytics 111, 112, 113, 114, 115
  flexible 112
  systems 112, 115

D

Data 3, 12, 66, 86, 146, 177, 181, 182, 183, 187, 188, 189, 195
  analytics issues 188
  mining 3, 146, 177, 182, 183, 187, 189, 195
  mining tools 181
  technology 66
  transformation 3, 12
  warehousing 3
  wrangling 3, 86
Data analysis 1, 2, 17, 24, 55, 81, 85, 105, 114, 130, 165
  massive 114
  techniques 114
Data science 1, 3, 21, 53, 81, 82, 83, 84, 85, 86, 89, 90, 186, 187, 189, 193, 195
  approaches 187, 195
  contemporary 89
  courses 81, 84, 85
  education 1, 83, 84, 89, 90
  industry 86
  methods 3, 186
  opportunities for healthcare 195
  position 82
  processes 53
  skills 189
  techniques 81, 187, 193, 195
  theory 21, 81
  training 83
Databases, traditional 193
Decision tree fuzzy system (DTFS) 183
Deterioration cost 95
  for own warehouse 95
  for rented warehouse 95
Development, business intelligence 1
Diabetes mellitus 179
Differential calculus 55
Diseases 177, 178, 179, 180, 181, 183
  cardiorespiratory 179, 180
  cardiovascular 177, 179, 180, 181, 183
  chronic respiratory 177
  infectious 178
  metabolic 179
  noncommunicable 177, 178
Dyslipidaemia 180

E

EDA technique 148
Educational framework, traditional 89
Effective heart disease prediction system (EHDPS) 177
Electronic health records (EHR) 193, 194, 195
Emission, industrial process 171
Encyclopedia Britannica 53
Energy emission 173
Ensemble learning approach 152
Environment(s) 94, 107
  fuzzy 94
  harsh 107
Epidemiological transition 178

F

Facility 179, 198
  medical 179
  nursing 198
Fatty acids 180
Financial 2, 193
  projections 2
  transactions 193
Functions 8, 9, 17, 21, 27, 28, 55, 119, 120, 190, 196, 200
  probability density 8, 9
  transfer 119

transformational leadership 200

G

Game theory 21
Gaussian 9, 51, 52, 57, 63
distribution 57, 63
Naive Bayes 9
Gene sequencing 182
Genetic 64, 93, 180
alogorithm 93
mutations 64
risk factors 180
Genetic algorithm (GA) 93, 94, 182
based heuristic 94
technique 182
GHG emissions 128, 129
Global positioning system 33
Gradient boosting machines (GBMs) 151
Greenhouse gases 128, 141, 142, 143, 162, 163, 175
emission 141, 142, 143, 162, 163, 175
Growth 129, 144, 162, 163, 191
corporate 191
industrial 162, 163

H

Healthcare 191, 192, 193, 194, 195, 198, 199
data 191, 192, 193, 194, 195
managers 198, 199
system 195
Heart diseases 177, 182
High-density lipoproteins (HDL) 180
Hubble’s 41, 43, 45
constant 41, 43, 45
law 41
Human-computer interaction (HCI) 90
Hyperglycemia 179
Hypertension 179, 180

I

IBM software white paper 196
Identification of clusters by elbow method 132, 166
Illnesses, genetic bone 197
Image 4, 24, 33, 68, 69, 116, 124
analysis 68, 69
predictions 124
processing 4, 24, 33, 68, 116
Imaging data 193
Imbalanced data problem 146, 147
Industries 82, 83, 88, 89, 90, 107, 108, 142, 145, 162, 163, 186, 187, 197, 198, 200
telecommunication 145
Inputs, transform data 121
Intelligence 20, 106, 112, 114, 116
artificial 20, 106, 114, 116
system 112
Intelligent heart disease prediction system (IHDPS) 182, 183
Inventory system 95
Iron calorimeter 46

K

Kaggle database 130, 162, 165
Kalman filter 32, 37, 40, 43, 45
process 32, 40, 43
technique 37, 45
Kalman filtering 24, 25, 35, 37, 43
process 35
technique 24
Kalman gain equation 32, 34
K-mean clustering 128, 129, 130, 162, 163, 164, 165
applied 130, 165
technique 162

L

Leadership 82, 190, 197, 198
health systems 190
transactional 198
transformative 198
Linear 5, 21, 26, 45, 47, 48, 53
dynamic systems 26, 47, 48

regression problem 21
systems 5, 45
transformation 53
Logistic regression 9, 13, 16, 53, 60, 144, 145, 147, 153, 156
Loss, standard squared-error 151
Low-density lipoprotein (LDL) 180

M

Machine learning 1, 4, 13, 20, 115, 144, 145, 146, 147, 150, 154, 155, 156, 159, 166
algorithms 1, 4, 13, 20, 145, 147, 166
method 115
techniques 144, 146, 147, 150, 154, 155, 156, 159
Malformation 109
Mastercard transaction 109
Mathematical 1, 21, 51, 52, 56, 61, 68, 69, 165, 186
analysis 1, 56
computation 69
methods 21, 68, 186
reasoning 51
Mathematics, applied 57, 59, 61, 84
MATLAB's matrices 3
Maximal frequent itemset algorithm (MAFIA) 183
Mean 26, 27, 28, 53, 54
absolute error (MAE) 54
squared error (MSE) 26, 27, 28, 53, 54
Measured doppler 49
Mechanism 24, 25, 120, 153, 182
computational 24
fuzzy logic 182
Medical emergency 177
Mental 179, 181
health problems 181
stress 179
Microarray chip information 182
MNIST 116, 122, 124
database 122, 124
dataset 116
Multilayer perceptron (MLP) 177, 181
Myocardial 179, 180
infarction (MI) 179
ischemia 180

N

Network's life expectancy 120
Neural network method 182
Neurological disorders 195
Neurons, artificial 117
Neutral hydrogen 49
Noisy data 24, 152
Nonlinear 47, 153
dynamics 47
relationships 153
Non-linear systems 46, 47
Nonparametric statistical methods 58
Norm, vector's 6

O

Obesity, abdominal 180
Obtrusion, networks-based 112

P

PCA 14, 15
algorithm 14
transformation 14, 15
Power, thermal 142
Principal component analysis (PCA) 14, 15, 112
Probability theory 3, 8, 12, 52
Problems 17, 18, 20, 56, 57, 60, 88, 90, 111, 146, 162, 163, 188, 189, 195, 196
decode 189
informational 111
overfitting 146
Process 2, 3, 10, 11, 25, 27, 28, 68, 69, 87, 89, 90, 129, 131, 190, 193, 195, 196
computational 89
filtration 25
industrial 129
problem-solving 87

Programming 66, 67, 81, 83, 85, 86, 87, 90, 128, 189
  computer 86, 87
  entry-level 85
  languages 87, 189
Properties, mapping 93
Python 69, 128, 130, 131, 162, 163, 165, 166, 169, 175
  programme 166
  Programming 69, 128, 130, 131, 162, 163, 165, 166, 169, 175
Python’s NumPy arrays 3

R

Radial velocity 42, 43, 45, 47
Radio equipment 24
Random forest features 150
Regression 13, 60, 61, 151
  logistical 60, 61
  polynomial 13, 60
  method 151
Remarks- kalman filter 47
Restrictions 21, 85, 86
  credit 86
Russia’s power emission 141

S

Scopus database 146
Sectors 144, 177, 193
  fitness 193
  medical management 177
  telecommunication 144
Security data 115
Service 87, 105, 111, 114, 115, 148, 149, 157, 158, 159
  intelligence security 115
  virtual private network 111
  web 87
Sigmoid functions 9
Skew distribution 64
Skills 54, 81, 84, 87, 88, 115, 186, 188, 189
  communication 189
  computational 84
  problem-solving 189
Software 61, 84, 89, 105, 111, 130, 165
  algorithmic 61
  data reliability 111
Soil degradation 128, 129
Support vector machine (SVM) 5, 6, 7, 61, 182

T

Techniques 10, 53, 145, 183
  deep learning 53
  machine-learning 145
  technology-based prediction 183
  transformation 10
Technology 12 ,67
  contemporary data 67
  data transformation 12
Telecommunication companies 144, 146, 159
Telemedicine 194
Test, chi-squared 59
Theory, chaos 21
Tobacco smoking 179

V

Vegetable oils 180
Virial theorem 49

Biswadip Basu Mallik

Prof. Biswadip Basu Mallik is presently a senior assistant professor of mathematics in the Department of Basic Sciences & Humanities at Institute of Engineering & Management, Kolkata, India. He has been involved in teaching and research for more than 21 years and has published several research papers in various scientific journals and book chapters with reputed publishers. He has authored five books at undergraduate levels in the areas of engineering mathematics, quantitative methods and computational intelligence. He has also published five Indian patents along with nine edited books. His fields of research work are computational fluid dynamics & mathematical modelling. Prof. Basu Mallik is a managing editor of Journal of Mathematical Sciences & Computational Mathematics (JMSCM), USA. He is also the editorial board member and reviewer of several scientific journals. He is a senior life member of Operational Research Society of India (ORSI) and a life member of Calcutta Mathematical Society (CMS), Indian Statistical Institute (ISI), Indian Science Congress Association (ISCA), and International Association of Engineers (IAENG).

Kirti Verma

Prof. Kirti Verma is an editor, mentor, author, supervisor and educationist. For the past 15 years, she has been working in the education sector. Presently, she works as dean academic associate professor & head of the Department of Mathematics, Lakshmi Narain College of Technology, Jabalpur. She received Ph.D. in mathematics (special function) from Barkatullah University Bhopal in 2015. She earned PGDCA from Maharishi University, Bhopal in 2008. She has 15 years of teaching experience. She has published several research papers in mathematics and on the Internet of Things (IoT). Her research interests lie in algebra, applied mathematics, fuzzy Logic, differential equations, linear algebra, abstract algebra, complex analysis, calculus & advanced calculus, biostatistics and research methodology, Internet of things (IoT) and mathematical modelling. She has published 27 research papers, 12 books and 7 book chapters and 15 patents. She has attended 30 plus conferences and is a reviewer and editorial board member of many reputed journals. She has published many research papers in international and national journals.

Rahul Kar

Prof. Rahul Kar did M.Sc. in pure mathematics and did Ph.D. from Bankura University and is currently working as a state aided mathematics faculty of Kalyani Mahavidyalaya, Kalyani, Nadia, West Bengal. He has worked as a guest faculty of Gurunanak Institute of Technology and other 2 colleges. He has 7 years of teaching experience and has published many research papers with Scopus index. He is also working as an editor in 4 national and international journals and a reviewer in 5 international journals. He is working as an editor in ten mathematics/computer books. He is an annual member of Indian Mathematical Society as well as honorary member of Carmels Research Institute, Dakar, Senegal.

Ashok Kumar Shaw

Prof. Ashok Kumar Shaw did M.Sc. in applied mathematics. He did Ph.D. from Indian Institute of Engineering Science and Technology (IIEST, Shibpur), India. Currently, he is working as a professor of mathematics. He is also the head of the Department of Mathematics and BSH, and the dean of R&D at The Budge Budge Institute of Technology. He is supervisor of Ph.D. research scholars at Maulana Abul Kalam Azad University of Technology (MAKAUT). He is currently working in the field of applied mathematics specially operations research, reliability optimization, maintenance, fuzzy mathematics, inventory and supply chain management etc. He has published twenty eight papers in the international journal of repute along with three books.

Sardar M. N. Islam (Naz)

Prof. Sardar M. N. Islam (Naz) Ph.D., LL.B. (Law) is a professor at ISILC, & Victoria University, Australia. He is also a distinguished visiting professor of artificial intelligence, UniSri, an adjunct professor of IT and Business, Armstrong Institute, Melbourne, Australia. He is the editor-in-chief of "International Transactions on Artificial Intelligence" an AI and modelling journal. He is a member of the following associations - Australian Computer Society, IEEE, the Computer Society, and other societies and several other computer science, data science associations. He has published 35 books and about 250 articles including 60 Q1 and Q2 grade journal articles. His current areas of interest and expertise are: AI, computer science & mathematics, and their applications in different areas.