RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ ECHAHID HAMMA LAKHDAR D'EL-OUED

Faculté des Sciences Exactes

Département d'Informatique

THÈSE

Pour l'obtention du grade de DOCTEUR 3ème CYCLE LMD EN INFORMATIQUE

Option: Intelligence artificielle

Présenté par: **Kamel MAALOUL**

Titre

L'analyse des Big Data pour les villes intelligentes

Jury:

Président	Meftah Mohammed Charaf Eddine	Professeur	Université d'Eloued
Rapporteur	LEJDEL Brahim	Professeur	Université d'Eloued
Co-Rapporteur	Nedioui Mohamed Abdelhamid	Assistant Prof.	Université d'Eloued
Examinateur	Korichi Ahmed	Professeur	Université d'Ouargla
Examinateur	Laouer Mohamed Ridda	Professeur	Université de Tebessa
Examinateur	Boucherit Ammar	Professeur	Université d'Eloued

2022/2023

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

UNIVERSITY ECHAHID HAMMA LAKHDAR OF ELOUED

Faculty of Exact Sciences

Department of Computer Science

Thesis

Presented with a view to obtaining the LMD Doctorate degree in Computer Science

Option: Artificial intelligence

Presented by: **Kamel MAALOUL**

Title

Big Data Analysis for Smart Cities

Jury:

President	Meftah Mohammed Charaf Eddine	Professor	University of Eloued
Supervisor	LEJDEL Brahim	Professor	University of Eloued
Co-Supervisor	Nedioui Mohamed Abdelhamid	Assistant Prof.	University of Eloued
Examiner	Korichi Ahmed	Professor	University of Ouargla
Examiner	Laouer Mohamed Ridda	Professor	University of Tebessa
Examiner	Boucherit Ammar	Professor	University of Eloued

2022/2023

Acknowledgements

First of all, I thank God, who helped me and condemn it for my presence to provide me with this opportunity and gave me the ability to go forward successfully.

I would like to express my sincere gratitude to Professor Brahim Lejdel to obtain the continuous support for doctoral study and relevant research, his patience, motives and tremendous knowledge. His guidance helped me all the time research and write this thesis. I couldn’t imagine a better consultant and teacher to study.

We would like to thank the members of the jury, who have done us the honor of accepting to be reviewers and for kindly devoting some of their time to reviewing our work.

We would like to thank all the people with whom we had the joy of collaborating, directly or indirectly, on our project. They were all able to share their knowledge and skills with us.

Finally, I express my gratitude to all those who have contributed in one way or another to the development of this work.

O Allah, send your blessings on your noble messenger, his family, and companions, and bless us in our life.

Dedication

I dedicate this work to my parents:

May they find here the testimony of my deep gratitude and acknowledgment

To my brothers and my sisters, my grandparents and my family who give love and liveliness.

To all those who have helped me - directly or indirectly - and those who shared with me the emotional moments during the accomplishment of this work and who warmly supported and encouraged throughout my ýourney.

To all my friends who have always encouraged me, and to whom I wish more success.

Thanks!

Kamel MAALOUL

Abstract

The emergence of smart cities aims in one to solve the problems posed by the continuous development of urbanization and the increasing density of the population in cities. To address these challenges, governments and policymakers are undertaking smart city projects aimed at sustainable economic growth and improved quality of life for residents and visitors alike. Information and Communication Technology (ICT) is the main technology for developing cities. Big data refers to the vast amount of information generated from various sources, such as sensors, social media, and internet of things (IoT) devices. Extracting hidden information and correlations from Big Data is a growing trend in information systems aimed at providing better services to citizens and supporting decision-making processes. But, to extract useful information for the development of smart services in the city, the datasets must be generated, integrated, and analyzed in a smart system. This process is generally referred to as big data analysis. In this thesis, we use big data technologies and methods to support the development and operation of smart cities. This integration of big data and smart cities enables cities to collect and analyze vast amounts of data from various sources to make in- formed decisions and optimize their operations.
Key-words: *Big Data, Big data analysis, IOT, ICT, Smart Cities.*

ملخص

يهدف ظهور المدن الذكية إلى حل أحد المشكلات التي يطرحها التطور المستمر للتوسع العمراني وزيادة الكثافة السكانية في المدن. لمواجهة هذه التحديات ، تقوم الحكومات وصناع القرار بتنفيذ مشاريع المدن الذكية التي تهدف إلى النمو الاقتصادي المستدام وتحسين نوعية الحياة للمقيمين والزوار على حد سواء. تعتبر تكنولوجيا المعلومات والاتصالات (ICT) هي التكنولوجيا الرئيسية لتطوير المدن. تشير البيانات الضخمة إلى الكم الهائل من المعلومات المتولدة من مصادر مختلفة ، مثل أجهزة الاستشعار ووسائل التواصل الاجتماعي وأجهزة إنترنت الأشياء(IoT) . يعد استخراج المعلومات والارتباطات المخفية من البيانات الضخمة اتجاهًا متزايدًا في أنظمة المعلومات . تهدف إلى تقديم خدمات أفضل للمواطنين ودعم عمليات صنع القرار . ولكن لاستخراج معلومات مفيدة لتطوير الخدمات الذكية في المدينة ، يجب إنشاء مجموعات البيانات ودمجها وتحليلها في نظام ذكي . يشار إلى هذه العملية عمومًا باسم تحليل البيانات الضخمة. في هذه الأطروحة ، نستخدم تقنيات وطرق البيانات الضخمة لدعم تطوير وتشغيل المدن الذكية. هذا التكامل بين البيانات الضخمة والمدن الذكية يمكّن المدن من جمع وتحليل كميات هائلة من البيانات من مصادر مختلفة لاتخاذ قرارات مستنيرة وتحسين عملياتها

كلمات مفتاحية : البيانات الضخمة ، تحليلات البيانات الضخمة ، إنترنت الأشياء ، تكنولوجيا المعلومات والاتصالات ، المدن الذكية.

Résumé

L'émergence des villes intelligentes vise à résoudre les problèmes posés par le développement continu de l'urbanisation et la densité croissante de la population dans les villes. Pour relever ces défis, les gouvernements et les décideurs entreprennent des projets de villes in- telligentes visant une croissance économique durable et une meilleure qualité de vie pour les résidents et les visiteurs. Nous savons que les technologies de l'information et de la communication (TIC) sont la technologie clé pour le développement des villes. En outre, le Big Data fait référence à la grande quantité d'informations générées à partir de diverses sources, telles que les capteurs, les médias sociaux et les appa- reils de l'Internet des objets (IoT). L'extraction d'informations cachées et de corrélations à partir du Big Data est une tendance croissante dans les systèmes d'information visant à fournir de meilleurs services aux ci- toyens et à soutenir les processus décisionnels. Mais, pour extraire des informations utiles au développement de services intelligents dans la ville, les ensembles de données doivent être générés, intégrés et analysés dans un système intelligent. Ce processus est généralement appelé ana- lyse de mégadonnées. Dans cette thèse, nous utilisons les technologies et méthodes du big data pour soutenir le développement et le fonc- tionnement des villes intelligentes. Cette intégration des mégadonnées et des villes intelligentes permet aux villes de collecter et d'analyser de grandes quantités de données provenant de diverses sources pour prendre des décisions éclairées et optimiser leurs opérations.

Mots clés : *Données massive, analyse de big data, fouille de données, TIC, villes intelligentes.*

Contents

List of Figures

List of Tables

List of Abbreviations

IoT: Internet of Things
RF: Random Forests
AI: Artificial Intelligence
HAR: Human Activity Recognition
LR: Logistic Regression
SVC: Support Vector Classifier
DTC: Decision Tree Classifier
RFC: Random Forrest Classifier
GPS: Global Positioning System
RSSI: Received Signal Strength Indication
KNC: K neighbors Classifier
GBC: Gradient Boosting Classifier
AP: Access Points
ETC: Extra Trees Classifier
CSV: comma-separated values
PCA: Principal component analysis
BLE: Bluetooth Low Energy
CO: Monoxyde de carbone
SO2: Sulfur dioxide
O3: Ozone

List of publications

A. International journals

1. Maaloul, K., Lejdel, B., and Abdelhamid , N. M. (2023). Real-time human activity recognition from smartphones using linear support vector machines. TELKOMNIKA (Telecommunication Computing Electronics and Control), 21.3 (2023): 574-583. DOI: http://doi.org/10.12928/telkomnika.v21i3.24100

2. Maaloul, K., Lejdel, B., Clementini, E., and Abdelhamid, N. M. (2023). Bluetooth beacons-based indoor positioning in a shopping malls using machine learning. Bulletin of Electrical Engineering and Informatics, 12(2), 911-921. DOI: https://doi.org/10.11591/eei.v12i2.4200.

3. Kamel Maaloul, Lejdel Brahim, "Comparative Analysis of Machine Learning for Predicting Air Quality in Smart Cities," WSEAS Transactions on Computers, vol. 20, pp. 248-256, 2022, DOI: 10.37394/23205.2022.21.30

B. Chapters

1. Maaloul, K., Abdelhamid, N. M., and Lejdel, B. (2021, January). Machine Learning Based Indoor Localization Using Wi-Fi and Smartphone in a Shopping Malls. In International Conference on Artificial Intelligence and its Applications, El-Oued, pp. 1-10, Springer, Cham.Link

2. Maaloul, K., and Lejdel, B. (2022, January). Big Data Analytics in Weather Forecasting using Gradient Boosting classifiers Algorithm, International Conference on Advances in Artificial Intelligence Developments and Applications, Algiers, Springer, Cham.(In Press)

C. International conferences

1. Maaloul, Kamel, and Brahim Lejdel. "Weather Forecasting and Prediction in Smart Cities using Machine Learning Algorithm." 5th Artificial Intelligence, Doctoral Symposium (AID'2022).Bab Ezzouar, Algiers, Algeria.2022.

2. Maaloul, Kamel, and Brahim Lejdel. "Predicting the Smart Health in smart cities Using machine learning Algorithm." 3rd INTERNATIONAL SCIENTIFIC RESEARCH AND INNOVATION CONGRESS. Turkey, 2022.

3. Maaloul, Kamel, and Brahim Lejdel. "Predicting The Direction Of Stock Market Prices In Smart Cities Using Machine Learning Algorithm." 2nd INTERNATIONAL MEDITARIAN SCIENTIFIC RESEARCH AND INNOVATION CONGRESS. Turkey, 2022. Link

4. Maaloul, Kamel, and Brahim Lejdel. "Comparison And Evaluation Of Advanced Machine Learning Methods For Performance Prediction Of A Vehicle CO2 Emission, 4th INTERNATIONAL PALANDOKEN SCIENTIFIC STUDIES CONGRESS, Turkey, 2022.

General introduction

Context

Big data is rapidly changing the way cities operate, and smart cities are at the forefront of this transformation. Smart cities use a wide range of data driven technologies to improve the quality of life for citizens and increase the efficiency of cities services. They use data from a variety of sources, including sensors, cameras, and social media, to gain real-time insights into cities operations and make data-driven decisions.

On the other side, big data analytics is a key technology for smart cities, allowing them to process and analyze large amounts of data from various sources. This data can be used to improve transportation systems, optimize energy usage, and enhance public safety. For example, traffic data can be used to improve traffic flow and reduce congestion, while energy usage data can be used to optimize building energy consumption.

Smart cities also use data to improve the delivery of public services. For example, by analyzing data from social media, cities can identify areas where residents are experiencing problems and quickly deploy resources to address those issues. Similarly, by analyzing data from sensors, cities can identify areas where there is a high risk of crime and take steps to increase public safety.

Therefore, a smart cities relies heavily on big data. IoT devices and sensor data can beprocessed by cities to identify patterns and needs. The study can make roads less congested and accidents while also assisting with parking. Additionally, data can improve energy systems, smart cities lighting, and criminal activity. Numerous practical uses exist for big data analytics. For instance, marketers can mine, aggregate, and analyze both types of data in almost real-time by using big data approaches and analysis methods. This can assist them in identifying hidden patterns, such as the interactions between various client groups and how these affect purchase behavior. With the use of this information, businesses can create customized marketing strategies that cater to the tastes of each customer.

Utilization and effective analysis of big data are key factors for success in many services and business domains. In this thesis, we review different works for smart cities based on the application of big data. We also compare and discuss the different definitions, challenges, benefits, and applications of big data and smart cities. In addition, We provide a platform for big data analytics to help data scientists and researchers for using it to resolve the different problems in Big data analytics.

However, with the benefits of big data come privacy concerns. Smart cities must ensure that they are collecting, storing, and using data in a way that respects citizens' privacy. This includes ensuring that data is collected and used only for specific, authorized purposes and that appropriate security measures are in place to protect the data. Overall, big data is a powerful tool for smart cities, providing real time insights that can improve cities operations and enhance the quality of life for citizens. As cities continue toadopt data driven technologies, we can expect to see even more innovative uses of big data in the future.

From this, we expect that predictive analytics will be a trend in big data technologies. Predictive analytics requires a significant skill of dynamic adaptation given partly already attained outcomes by nature. In this trend, we assert that the analytics must adapt (or, better yet, self adapt) to what is occurring, with the knowledge that there are, of course, certain previously user defined rules that specify which adaptation it may be appropriate to decide to initiate.

The use of big data analytics techniques is one way to design new processes to comprehend the vast amount of data gathered. Big data can be mined and modeled using analytics approaches to improve the functionality of smart cities and gain better insight. Big Data analytics can be applied in a variety of ways to enhance businesses and organizations. Examples include forecasting future trends to improve company decisions and employing analytics to comprehend customer behavior to optimize the customer experience. When it comes to big data analytics, the options are almost limitless. It depends on how you want to use it to advance your company.

In conclusion, smart cities use big data analytics to analyze and make sense of the large amounts of data collected from various sources, such as sensors, cameras, and social media. This data can be used to improve cities services, such as traffic management, publicsafety, and energy efficiency. Big data analytics can also be used to gain insights into the behavior and needs of cities residents, which can help to inform urban planning and policy decisions. Additionally, big data analytics can help cities to identify patterns and trends that can be used to predict and prevent problems, such as crime or infrastructure failures.

Problematic

These days, the data that needs to be examined is not only large but also includes streaming data and other different kinds of data. Big data may alter statistical and data analysis methods due to its distinctive characteristics of being "massive, high dimensional, heterogeneous, complicated, unstructured, incomplete, noisy, and erroneous". Many of the available analytics solutions today are capable of addressing ongoing big data analytics difficulties and delivering good outcomes for an affordable price. With the recent rise in smart, wearable, and other measuring devices for ambient applications, we are just beginning to address all the facets of this new big data. For smart cities, the ability to retrieve data from sensors is a big advantage in resolving the problems and barriers preventing their development.

Therefore, these questions must be answered:

- ➢ What are the insights of big data analytics to support smart cities by focusing onhow big data can fundamentally change urban populations at different scales?
- ➢ What are the most important big data analysis platforms applied in smart cities?
- ➢ How can we exploit integrated and scalable big data analytics frameworks for smart cities?

Objectives

The main objective of this thesis is:

- ➢ Review the applications of big data in the smart cities and explore the opportunities and challenges of using big data in the smart cities. In addition, research the general requirements for designing and implementing big data applications for smart cities applications and services.
- ➢ Presenting the most important best frameworks that allow accurate analysis of big data for smart city management. As well as providing a comprehensive survey of big data in smart cities.

Contributions

In our thesis, we propose three contributions to solving many problems using big data analytics in smart cities.

- ➢ Contribution 01: to assess common and crucial human activity, we will integrate many cutting edge machine learning algorithms and will present a thorough frame- work for Kafka based on real time to predict human activity behaviors. We will conduct a comparison examination for different approaches using Logistic Regression Classifiers. We choose a Tree classifier, and a Random Forest classifier on actual wearable sensor information to confirm the efficacy of the classification algorithms for human activity detection. [1]

- ➢ Contribution 02: we use various machine learning methods to get indoor localiza- tion using Bluetooth beacon. In this contribution, we analyzed and compared the performances of 6 individual predictors using machine learning algorithms in in- door localization features. We have validated the performance of the system using Bluetooth beacons on smartphones. To find clients' devices in shopping centers, we deployed machine learning techniques. [2]

- ➢ Contribution 03: we compared various algorithms, including linear regression, de-cision trees, support vector machines, random forests, and gradient boosting to forecast the air quality index using machine learning approaches. We want to fundamentally alter the dynamics of air quality forecasting. In smart cities, envi- ronmental quality may also be predicted quickly. to put into practice preventative measures for smart cities. [3]

Thesis structure

This thesis is divided into four main chapters.

The first chapter presents an overview of the definition of "big data", its characteristics, challenges, and applications, as well as the role of big data in smart cities.
The second chapter starts with a brief introduction to Big data analytics, followed by discussions about Big data analyses in smart cities. We have cited some important open issues and some tools and frameworks that use big data analytics.
The third chapter introduces a structure based primarily on open view technology, and platforms consisting of Apache Kafka, stream messages over the Internet, systematize them and provide a form of real time existing facts. The problem of identifying human activities using a Smartphone tool has also been formulated as a kind of trouble using the collection of statistics by phone sensors.
The fourth chapter presents indoor localization based on Bluetooth signals using many different machine learning technologies to locate customer devices in shopping centers and solve the problems of finding the data used to analyze the spatial behavior of individuals using Bluetooth technology.
In the five chapters, we solve the different issues of smart cities by discovering knowledge in Big data collections to predict air quality in smart cities. This chapter presents background, relevant works, data extraction, and large data to make predictions of a large amount of data using machine learning technologies and provide a comparative study to determine the best model for predicting air quality.
In the conclusion, we discuss the different proposed works for smart cities using big data analytics and we cite some future works related to the presented frameworks.

Chapter 1

Chapter 1: Big Data and Smart Cities Concepts and Definitions

1.1 Introduction

Big data and smart cities together form a powerful combination, as big data analytics provides several valuable insights to decision makers in designing and implementing smart cities initiatives. The goal is to make the living environment more developed, efficient, and conducive to the well being of all residents. In this chapter, we introduce some concepts and definitions related to big data and smart cities.

1.2 Big Data Definition

There are many definitions of "big data," including but not limited to:

- The Big Data is defined by Gartner [4] as high volume, high velocity, and/or high variety data that require new processing paradigms to enable insight discovery, improved decision making, and process optimization. According to this definition, Big Data is characterized by specific size metrics, but rather by the fact that traditional approaches are struggling to process them due to their size, velocity, or variety.

- Sha et al.[5] defined big data as the management and processing of any dataset that has a huge size for a direct individual explanation, which is considered a limitation that might cause conflict in dataset interpretation. The interpretation may be grouped into two categories or meanings: from a computing point of view, it is the quantity of the gathered data that pushes new techniques to be developed; and from the human point of view, it is the human-guided development necessary to declare the actual data meaning. This second category means that the main concern is about who will manage and use the data and not about the size of the big data.

- Gandomi et al.[6] discussed two definitions that are gathered from the literature review that big data is mainly referring to high volume, velocity, and variety of information assets that demand cost effectiveness, innovative forms of information processing to support enhanced future vision, and a decision making process. The other definition is stating that big data is a term that describes large volumes of high complexity, variability, and velocity of data that requires capable tools, techniques, and technologies to enable the capturing, storing, distributing, managing, and analyzing of information.

On other hand, most data scientists and experts define Big Data by the following Five main characteristics (called the 5Vs):

1.2.1 Volume

Volume alludes to the enormous volumes of data produced every second by social media, mobile devices, automobiles, credit cards, photos, videos, and other sources. Currently, data is stored using distributed systems, which connect data stored in several places with a software framework like Hadoop. [7]

1.2.2 Velocity

Compared to the other factors, velocity is crucial; there is no use in making a significant investment just to have to wait for the results. Consequently, the primary function of big data is to deliver data more quickly and on demand. Velocity refers to the speed at which the data is being accumulated. [8]

1.2.3 Variety

The term "variety" is a representation of the proliferation of information that has led to a variety of data kinds (textual, numeric, etc.), formats, structures (structured, semi-structured, unstructured), encodings, syntaxes, semantics, etc. [9]

1.2.4 Value

Value is the quantity of worthwhile, trustworthy, and reliable data that must be processed, saved, and examined to derive insights. It is not limited to the amount of data we store or process. Even if we collect a lot of data, it is useless unless we can learn something from it. Value describes how useful the information is while making decisions. We must use the right analytics to extract value from big data. [10]

1.2.5 Veracity

Veracity means the degree of reliability that the data has to offer. Since a major part of the data is unstructured and irrelevant, big data needs to find an alternate way to filter or translate it. The data is collected from multiple sources, we need to check the data for accuracy before using it for business insights. [11]

Figure 1.1 illustrates the features of big data that typically have five development characteristics: size, speed, variety, value, and honesty.

Figure 1.1: 5V's of Big Data [12]

1.3 Big Data Challenges

The challenges in big data are the real implementation hurdles. The challengesexist at several levels of data collection, storage, sharing, searching, administration, and visualization [104]

. The main Big Data issues and their solutions are listed below:

1.3.1 Big Data Tool Selection Confusion

When choosing the right technology for analyzing and storing big data, confusion often arises as to which data storage technology is best. The best course of action is to seek professional help. Either hire seasoned experts who have more knowledge of these tools, or another option is to use big data consulting. [13]

1.3.2 Technical Challenges

Among the technical challenges in big data is development scalability. This problem has led many to turn to cloud computing. It creates several difficulties, such as how to manage and implement so many tasks so that the objectives of each workload can be met at a reasonable cost. It also calls for effective management of system failures[14]. Another technical challenge is fault tolerance. The calculation of fault tolerance is very difficult and requires complex algorithms. Nowadays, some emerging technologies, such as big data and cloud computing, are designed with the idea that in the event of a failure, the damage should be less than an acceptable level. [15].

1.3.3 Big Data Aggregation

The use of data aggregation has been increasingly common and significant throughout the era of distributed computing and the IoT when enormous amounts of data are continuously collected from devices and services all over the world and subsequently evaluated. The difficulty lies in integrating distributed big data platforms with internal infrastructures. The analysis of the data produced within businesses is insufficient. It is vital to go further and combine internal information with information from external sources to glean insightful knowledge. Third party sources, data on weather forecasts, traffic forecasts and feedback, and customer and citizen comments are examples of external data. It may help improve the models used in the examination. [16]

1.3.4 Big Data Analytics

Big data analytics is the act of spotting patterns, trends, and correlations in vast quantities of unprocessed data to support data driven decision make development. The difficulty of making this data usable despite its many sources, formats, resolutions, or characteristics[128]. To study data and comprehend the links between features, advanced data analysis is necessary. An organization might, for example, examine patterns that may have a beneficial or negative impact on the business and extract insightful information from data by using data analysis. Real time analysis is required by other data driven applications as well, such as navigation, social networks, finance, biology, astronomy, and intelligent transportation systems[17].

1.3.5 Big Data Analysis

Big Data analysis refers to the process of examining, interpreting, and making sense of large and complex data sets in order to uncover valuable insights and information. This process often involves the use of advanced analytics techniques such as machine learning, statistical analysis, and data visualization. The goal of big data analysis is to help organizations make informed decisions, improve operations, gain a competitive advantage, and achieve strategic goals by uncovering new patterns and relationships in their data. Big data analysis requires the use of specialized tools and technologies that can handle the volume, velocity, and variety of big data. It also requires a deep understanding of the data and the business context in which it is being analyzed. Effective big data analysis involves a combination of technical skills, business acumen, and creative thinking to uncover new insights and opportunities hidden within the data.

1.3.6 Big Data and Machine Learning

Before using data, its integrity, accuracy, and structure must be validated. The output of a data validation procedure can be used for further analysis, or to train a machine learning model. Machine learning algorithms focus exclusively on how real time can use data to learn strategies and behaviors within specific contexts. Within the discipline of machine learning, you'll find subdisciplines like deep learning and reinforcement learning. Machine learning is the ability of computer systems to learn to make predictions from observations and data[121]. Machine learning can use the information provided by the study of big data to generate valuable business insights. An introduction will be given to tools and algorithms that can be used to create machine learning models that learn from data and to scale such models to big data problems[18].

1.3.7 Big Data Management

The design of big data must coincide with the supporting infrastructure of enterprises. The development of an information wellspring has arisen from several areas that are unorganized and riotous due to different types of information. Previously, most organizations could not capture or store this information, and the devices that could access it could not handle the information in a reasonable amount of time. Be that as it may, the new big data innovation enhances implementation, promotes development in business action plan elements and departments, and provides dynamic support. Big data innovation expects to reduce costs for equipment, setup, and validation of big data estimation before dispatching critical enterprise assets. Proper stewardship of big data is open, robust, secure, and reasonable. [19]

1.3.8 Imbalanced Big data

A classification data set with skewed class proportions is called unbalanced. The major categories make up a large portion of the dataset. The multilayered imbalance problems present new difficulties. It is more difficult to handle jobs that include several classes with varying class misclassification costs than jobs that include only two classes. Several approaches have been developed to address this problem. [20]

To avoid all big data problems, we highly recommend that you analyze solutions and identify the above issues. It is also generally summarized in Figure 1.2.

Figure 1.2: Big Data Challenges[21]

1.4 Big Data Application

Through the analysis of vast amounts of data, big data applications can assist businesses in making better business decisions. The areas listed below are where big data can be used:

1.4.1 Healthcare

Big data technologies are playing a vital role in the development of personalized medicine and prescriptive analytics in the healthcare industry. Big data in healthcare refers to collections of electronic health data that are so large and complex that they are difficult to manage. This data consists of electronic patient records as well as clinical data and clinical decision support systems (doctor's written notes and prescriptions, medical imaging, etc.). [22]

1.4.2 IoT and Smart Cities

IoT devices are constantly generating the data that they send to the server every day. This data is extracted to enable device communication. The large and diverse volume of data is increasing exponentially as a result of the development of the Internet of Things (IoT) and its applications in many other fields [23]. The huge volume of data generated by smart cities comes from a wide variety of applications, including healthcare, traffic management, and so on. As it has been obtained from a variety of sources, the quality of the data provided by smart cities is also critical. The source of the information must be trustworthy to ensure the quality of the information, and data collection in real time must be carried out from several sources. [24]

1.4.3 Government Monitoring

The government can achieve efficiencies in terms of cost, output, and originality by implementing big data technologies. Since identical data collection is utilized by numerous applications, numerous departments can collaborate. By intervening in each of these areas, the government contributes significantly to innovation. Governments could also employ big data frameworks to optimize the use of important resources and infrastructure. [25]

1.4.4 Media and Entertainment

The media and entertainment sectors are using new business models to produce, market, and distribute their content. Due to user demands, digital information may now be seen at any time and from any location. The development of online TV shows and other content demonstrates that new users are interested in both viewing TV and accessing data from any location. The media companies target audiences by figuring out what people want to view, how to target the advertisements, how to monetize the material, etc. By examining viewer behavior, big data technologies are thereby improving the income of such media companies. [26]

1.4.5 Industry and Energy

From an industry perspective, big data will be crucial to the fourth industrial revolution. The goal is to create smart factories where resources and machines interact like a social network. Such a smart factory will create smart items (smart products), which will collect and transmit data as they are used. These massive amounts of data, which will be collected and evaluated in real time, are conveyed by smart products. As a result, new information will be generated and used to develop smart factories at a higher level. [27]

Big data applications are applied in different fields such as banking, agriculture, data mining, cloud computing, marketing, stocks, healthcare, etc. as shown in Figure 1.3.

Figure 1.3: Big Data Applications
[28]

1.5 Smart cities

1.5.1 Definition

There are several definitions of a "smart cities," and there is no universal consensus on which definition should be adopted. One of these is the notion that a city's intelligence lies in its capacity to offer services to the populace, improving their daily lives. It is representatireal timeity that "realizes the individual benefit and at the same time the social advantage," in the words of Giuliano Dali. [29]

1.5.2 Elements of a Smart cities

The smart cities has been a component. With the help of these six resources, you can assess the smartness of the cities more accurately and know what a "smart cities" entails. They are as follows:

1.5.2.1 Smart Economy

An economy that relies on technology innovation, resource efficiency, sustainability, and high social welfare as success factors is known as a "smart economy." To raise productivity and competitiveness with the overarching objective of enhancing everyone's quality of life, it adopts innovation and new entrepreneurial ventures. A smart cities must have a competitive economy to qualify as one, making innovation, entrepreneurship, internalization, and flexibility essential tools. The cities will need to foster cooperation between the many moral businesses, both public and private, by supporting them and turning them into a hub for innovation and production. [30]

1.5.2.2 Smart Governance

Citizens' active participation is key to smart governance. The citizen is the leader of change and advancement in smart cities. The idea behind "smart governance" is to involve the public in the decision-making process. Public administrations not only advocate working together with individuals to provide valuable and effective services, but they also advance society by using open processes and sharing data. It is the process of utilizing contemporary technology and information and communication technology to guarantee a sustainable environment for individuals and governments that is collaborative, transparent, participatory, and communication based. [31]

1.5.2.3 Smart Living

Smart living solutions are altering our daily lives at home, from managing lighting and heating to creating a comfortable environment to wirelessly powering our phone chargers. Since they are at the core of any avant garde design, it might be challenging to tell them apart from modern furniture options. However, you can design a fantastic living area by implementing a smart house concept. [32]

1.5.2.4 Smart Mobilty

Smart mobility is a new and revolutionary way of thinking about how we get around—one that is cleaner, safer, and more efficient. Put another way, Neckermann calls this new vision “Zero Emissions, Zero Accidents, and Zero Ownership." The concept of smart mobility includes a wide range of modes of transportation: kick scooters, bicycles (regular, electric, foldable), buses, light rail trains, subways, streetcars, taxis, autonomous vehicles, and walking. Moreover, users have the option to own or share. [33]

1.5.2.5 Smart Environment

The concept of a "smart environment" is to design a space with integrated sensors, displays, and computing tools so that people may better comprehend and manage their surroundings[111]. It is a collection of sensors, actuators, and several computing components that offers services to enhance human life. The interplay between various gadgets and computing systems is what gives this environment its intelligence. [34]

112

1.5.2.6 Smart People

The term "smart people" refers to the human and social capital that must be as connected and coherent as feasible. Thus, managing diversity, citizen engagement, discourse, and cosmopolitanism become crucial. With the help of information or the delivery of services, Smart People aspires to change how citizens engage with the public and commercial sectors as individuals or companies.[35]

The elements of a smart cities are interdependent, as shown in Figure 1.4, and each element, whether addressed separately or in combination with the others, helps the city become smarter in terms of digitalization as well as social and individual well being.

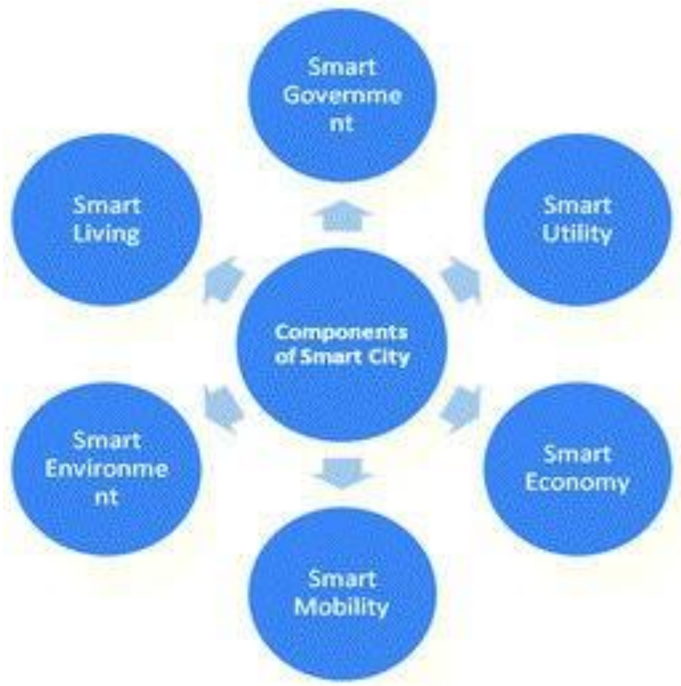

Figure 1.4: Elements of a Smart Cities
[36]

1.6 Big Data Analytics

1.6.1 Definition

Big data analytics is the act of spotting patterns, trends, and correlations in vast quantities of unprocessed data to support data driven decision making. These procedures employ well known statistical analysis methods, such as clustering and regression, on larger datasets with the aid of more recent instruments. It is the act of spotting patterns, trends, and correlations in vast quantities of unprocessed data to support data driven decision making [37]. Big data analytics are crucial because they enable businesses to use their data to find areas for growth and progress. Across all corporate sectors, improving efficiency results in more shrewd operations overall, more profits, and happier customers.

1.6.2 Benefits of Big Data Analytics

The benefits of big data and analytics include better decision making, greater innovation,and improved product pricing. The most important benefits of big data analytics include:

- Cost optimization: big data can lower expenses by centralizing all commercial data storage. The cost advantages big data systems like Hadoop and Spark provide for storing, processing, and analyzing massive amounts of data are among their most alluring features. [38]

- Acquisition and Retention of Consumers: Information about customers' preferences, demands, purchasing patterns, etc. can be found in their digital footprints. Big data is used by businesses to track consumer trends and then customize their goods and services to meet the needs of individual clients. This significantly increases consumer satisfaction, brand loyalty, and, eventually, sales. By providing a better customer experience, data driven algorithms support marketing initiatives (targeted advertisements, for instance) and boost consumer satisfaction. [39]

- Focused and Targeted Promotions: Thanks to big data, firms may deliver personalized items to their target market without spending a fortune on ineffective advertising campaigns. By tracking POS transactions and internet purchases, businesses can use big data to study consumer patterns. Using these insights, focused and targeted marketing strategies are created to assist businesses in meeting consumer expectations and fostering brand loyalty. [40]

- Potential Risks Identification: Because businesses operate in high risk settings, they need efficient risk management solutions to deal with problems. The creation of efficient risk management procedures and strategies depends heavily on big data. By optimizing complicated decisions for unforeseen occurrences and prospective threats, big data analytics and tools quickly minimize risks. [41]

- Complex Supplier Networks: Businesses that employ big data provide business-to-business communities or supplier networks with greater accuracy and insight. Big data analyticscan be used by suppliers to get over the limitations they often experience. Higher degrees of contextual intelligence, which are essential for success, can now be usedby suppliers thanks to big data. [42]

- Enhance Efficiency: Big data tools can enhance operating efficiency. Your in- teractions with consumers and their insight free back assist in the collection of significant quantities of priceless customer data. Analytics can then uncover sig- nificant trends in the data to produce products that are unique to the customer.To give employees more time to work on activities demanding cognitive skills, the tools can automate repetitive processes and tasks. [43]

- Innovate: The secret to innovation is the insights you uncover utilizing big data analytics. Big data enables you to innovate new products and services while up- dating ones that already exist. The vast amount of data gathered aids firms in determining what appeals to their target market. Product development can be aided by knowing what consumers think about your goods and services. [44]

1.6.3 Big Data Analytics Tools

This big data analysis is performed using big data analysis tools. We explain these tools in detail:

1.6.3.1 Apache Hadoop

Apache Hadoop is an open source software platform for data storage and application ex- ecution. The two main parts of Hadoop are MapReduce and the Hadoop Distributed File System, or HDFS. The software creates a distributed storage framework and processes massive data using the MapReduce programming methodology. Hadoop is regarded asa top big data analytics solution because of its excellent capacity to store and disperse large data sets across hundreds of cheap computers. Without any downtime, its users can increase the cluster's size by adding more nodes as needed. [45]

The Apache Hadoop consists of four main modules:

- Hadoop Distributed File System (HDFS): A distributed file system that runs on standard or low end hardware. HDFS provides better data throughput than traditional file systems, in addition to high fault tolerance and native support for large datasets. [105]

- Yet Another Resource Negotiator (YARN): manages and monitors cluster nodes and resource usage. It schedules jobs and tasks. [106]

- MapReduce is a framework that helps programs do parallel computations on data. The map task takes input data and, converts it into a dataset that can be computed in key value pairs. The output of the map task is consumed by reduced tasks to aggregate output and provide the desired result. [107]
- Hadoop Common: Provides common Java libraries that can be used across all modules. Each of these modules covers different capabilities of the Hadoop frame- work. The following diagram (See Figure 1.5) depicts their positioning in terms ofapplicability for Hadoop 3. X releases. [108]

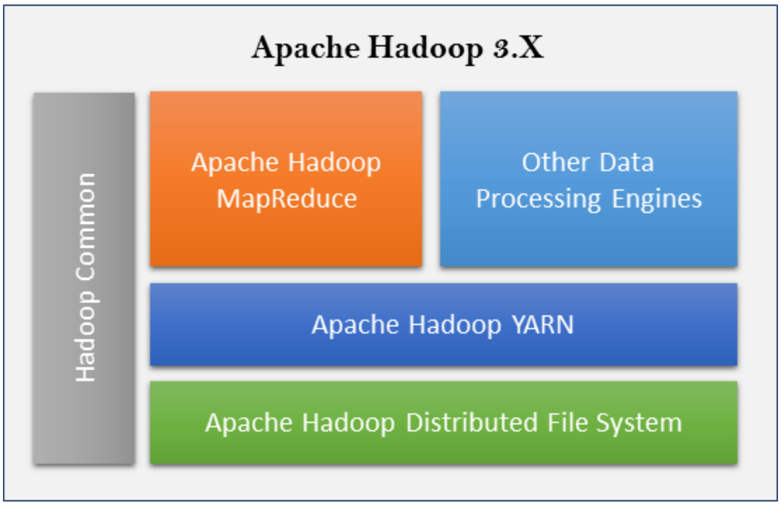

Figure 1.5: The main modules for Apache Hadoop

1.6.3.2 Apache Spark

One of the most potent open source big data analytics tools is Apache Spark. It is free data processing that can access very large data sets quickly. Additionally, it can be used independently or in conjunction with other distributed computing tools to divide data processing duties among several computers. Apache Spark is the fastest and most popular generator for big data transformation because it has built in support for streaming, SQL, machine learning, and graph processing. [46]

1.6.3.3 MongoDB

A NoSQL document oriented database called MongoDB is used to store large amounts of data. Hadoop and MongoDB are distinct from one another due to MongoDB's reputation for resilience. Instead of employing rows and columns like conventional rotating databases, MongoDB uses collections and documents. Key value pairs, which are regarded as the fundamental unit of data in MongoDB, make up these documents. Every

MongoDB database has collections, which in turn have documents. However, each document has a different size, content, and number of fields. [47]

1.6.3.4 Apache Cassandra

Apache Cassandra is another free tool to use. The open-sourced software can manage massive volumes of data thanks to the NoSQL DBMS that was constructed. It can even handle data that has been distributed across a variety of different commodity servers. It has its CQL (Cassandra Structure Language) to interact with its database. It is designed to provide high availability while managing enormous volumes of data spread across several commodity servers.[48]

1.6.3.5 Altamira LUMIFY

A platform for large data fusion, analysis, and visualization is called Lumify. It also helps you comprehend connections and investigate the interaction between your data, much like all other big data analytics tools. Because it enables street access analytics choices including graph visualizations, full text faceted search, dynamic histograms, interactive geospatial views, and collaborative workspaces that can be shared in real time, Lumify is regarded as a strong big data analytics solution. [49]

1.6.3.6 Cloudera

The platform distribution is fully free, and the code is open sourced. With Cloudera, open source software may be used throughout the entire platform. The package contains fantastic technologies and tools, including Apache Hadoop, Spark, Impala, and others. You can process, gather, manage, administer, discover, share, and model an infinite quantity of data using this option! Just be aware that the licensing cost for the Hadoop cluster is based on a per node basis. [50]

1.6.3.7 RapidMiner

The software platform RapidMiner was created for analysts who enjoy integrating data preparation, machine learning, and the deployment of predictive models. The cherry on top is that it is a free, open source software tool for data and text mining. The most effective and simple graphical user interface for the design of the analytical process is provided by RapidMiner. RapidMiner supports Macintosh, Linux, and Unix operating systems in addition to Windows opera systems. [51]

1.6.3.8 Data warehouses

Using specified schemas, storage for huge amounts of data is collected from numerous sources. The data warehouse is a central collection of data that can be examined to help decision makers become more knowledgeable. Transactional systems, relational databases, and other sources all provide data to the data warehouse, which receives it frequently. Through business intelligence (BI) tools, SQL clients, and other analytics software, business analysts, data engineers, data scientists, and executives have access to data. [52]

Figure 1.6 shows the most important and best big data analysis tools used by data scientists.

Figure 1.6: Big Data Analytics Tools
[53]

1.6.4 Challenges in Big Data Analytics for Smart Cities

Problems arise when evaluating big data for many computing algorithms that excel at processing fewer data. I have categorized the difficulties associated with big data analytics into four main groups:

1.6.4.1 Data Storage and Analysis

Therefore, one of the current big data issues is big data management, where the difficulty is in the phases of gathering and storing data with few criteria (hardware and software).

Data must be collected from many sources, cleaned for dependability, and then prepared for security and privacy considerations as part of the big data management process[141]. The complexity of the data—its pace, amount, and variety—presents a big data cleaning issue[130]. The challenges of big data aggregation require integrating disparate big data platforms (such as apps, repositories, sensors, networks, etc.) and external data sources into a cohesive system [54].

1.6.4.2 Knowledge Discovery and Computational Complexities

Large scale dataset analysis necessitates increasingly complicated computational tasks. Handling inconsistencies and uncertainty in the datasets is the main challenge. Finding a complete mathematical framework that is universally applicable to big data should be challenging. However, site specific data analytics can be completed quickly, provided the true difficulties are understood. Such developments could be used to emulate big data analytics in numerous fields. The least memory intensive machine learning algorithms have been used in a lot of research and surveys in this direction. This research's main goal is to reduce computing complexity and processing costs. [55]

1.6.4.3 Scalability and Visualization of Data

Giant Data Analysis approaches to face the biggest challenges in terms of scalability and security. Over the past few decades, academics have focused on accelerating data analysis and the processors that follow Moore's Law. The development of sampling, online, and multi dimensional analysis methods is required first. Within the context of large scale data analysis, incremental approaches have high scalability properties. There is a natural and dramatic change in processor technology that is incorporated with an increasing number of cores because data sizes are rising far faster than CPU speeds. The development of parallel computing is the result of this change in processors. [56]

1.6.4.4 Information Security

A vast amount of data is correlated, examined, and mined for useful patterns in big data analysis. Big data poses a serious security threat. Consequently, data analytics is creating a huge challenge with information security. By utilizing the procedures of authentication, authorization, and encryption, the security of massive data may be improved. Big Data applications must contend with several security obstacles, including network size, a wide variety of devices, real time security monitoring, and a lack of intrusion detection systems. Data security has focused on the security issues brought on by big data. [57]

1.7 Machine Learning Algorithms

We providean overview of the various machine learning algorithms:

a. **Logistic Regression (LR):** Logistic regression is a simple method. It belongs to the linear classifier organization and has special similarities to linear regression and polynomial regression. Logistic regression is the preferred tool for learning binary type algorithms. It goes much further than linear regression with a binomial reaction variable. Other benefits over Mantel Haenszel or include the simplicity of addressing numerous ex- planatory variables at once and the flexibility to apply explanatory variables constantly. If the dependent variable is dichotomous, LR is the optimal regression technique to use (binary). Like many other types of regression, LR is a predictive analysis. One or more LR models may be used to describe statistics and the relationship between a single binary dependent variable and one or more independent variables, such as nominal, ordinal, c programming language, or independent variables [93]. In Figure 1.7, on X-axis is the variable and on Y-axis is the output. The line with the most unique form for a model is the regression line. And finding this line with the optimum shape is the main goal of this method. Over-becoming can also result via linear regression, however, it can be prevented by using go validation, regularization algorithms, and dimensionality discounting techniques[117].It over simplifies real international issues via assuming a linear courting of a few of the variables, therefore now not encouraged for realistic use cases (See Figure 1.7).

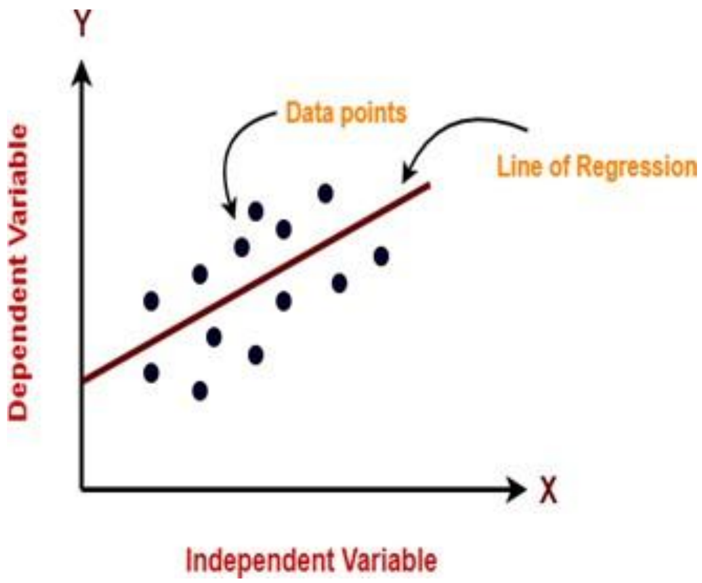

Figure 1.7: Graph of Linear regression

b. Linear Support Vector Classifier (Linear SVC): The "best appropriate" super- script that splits or categorizes your records is returned by linear SVC based on how well the given records fit together. Once you've reached the hyper stage, you can feed the unique classifier a few functions to determine what the "anticipated" class is. A category known as linear SVC can carry out binary and multi magnitude categories on a dataset[119]. The Linear SVC has additional parameters, such as a loss feature and penalty normal- ization that applies "L1" or "L2," as compared to the SVC version. Since linear SVC is dependent on the kernel linear approach, the kernel technique cannot be modified.

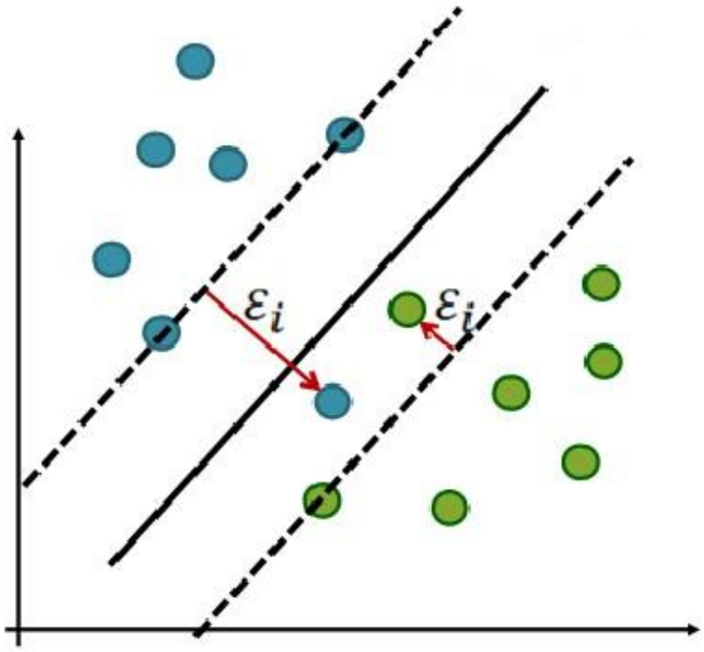

Figure 1.8: Graph of Linear SVC

Figure 1.8 demonstrates how Linear SVC aims to maximize the space between training sessions [94]. Support vector machines (SVMs) are super- vised learning models with corresponding learning algorithms used in machine learning for regression and classification analyses. An SVM training method creates a model that categorizes fresh data measurements into one of two categories as a non probabilistic bi-nary linear classifier [138].(see Figure 1.9).

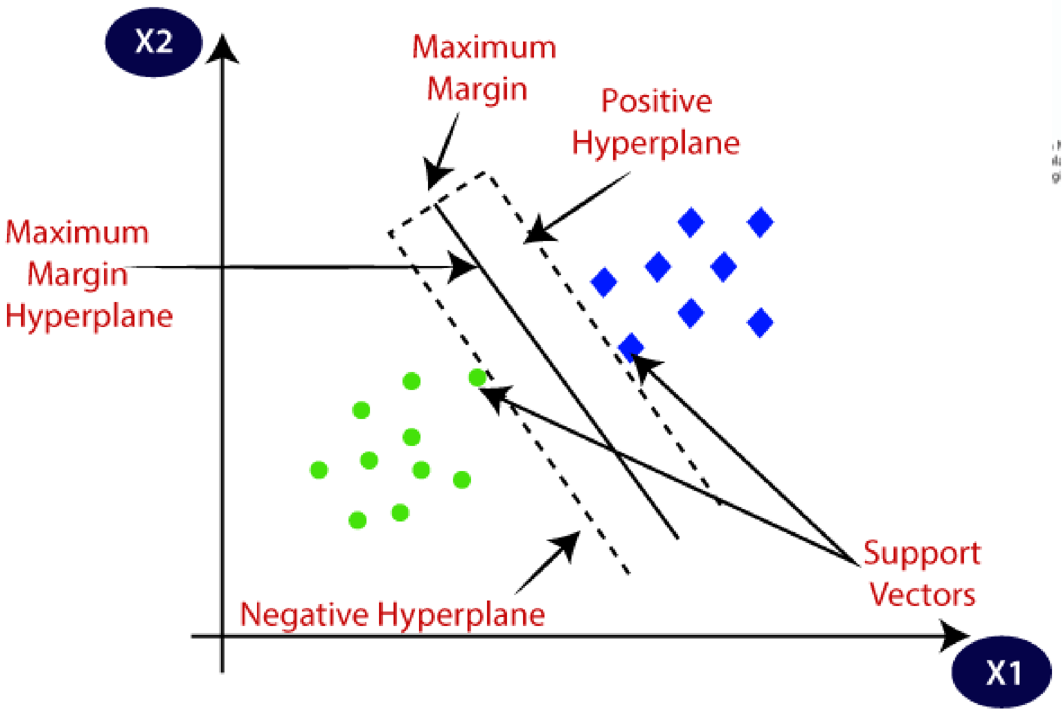

Figure 1.9: Support Vector Machine Algorithm

c. Decision Tree (DT):

Making decisions is a popular form of machine learning that uses a tree like shapeto represent choices. They can be constructed utilizing metrics such as gene impurity and facts in a top down manner. Both class and regression issues are modeled using selection timber. Although the selection tree is simple, it oversplits into tendencies and takes training with academic records seriously. They are often clipped to prevent them from growing further to avoid this. Nodes in a decision tree include the choice node and the leaf node, for example. While Leaf nodes record the results of those decisions and no longer contain any similarly branching Choice nodes, Choice nodes are used to make any choice and have multiple branches. [95]

Classification and regression trees are utilized for the quantitative outcome variable, but they do not compute the sets of decision rules[135]. A decision tree decides by carrying out a series of tasks. A decision node is one, but a leaf node is the opposite. Selection nodes have multiple branches, and leaf nodes are the outcomes or conclusions of the choices made to get there [136]. These decisions are based on a predetermined dataset (See Figure 1.10).

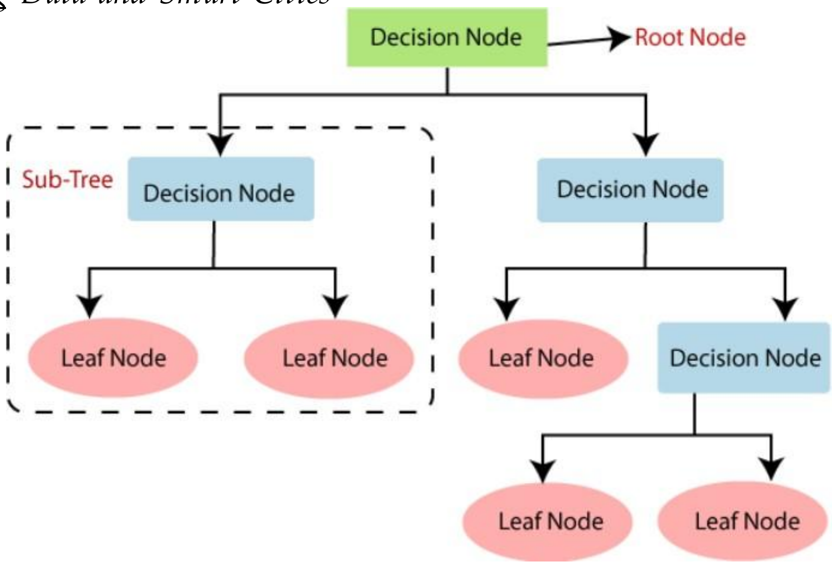

Figure 1.10: Decision Tree Classification Algorithm

d. Random Forrest Classifier (RFC):

A set of selection trees make up the random forest classifier, which is building a forest. A hard and fast of selection bushes is created using a random subset of the training data, and the outcomes from all of the bushes are merged. This tactic works because some trees refuse to have any bearing on the outcome. Prejudice is also removed by averaging all forecasts. Random forest is different from selection trees in that it does not look for the best feature when splitting a node. The set of rules successfully models both type and regression issues. [96]

A random forest is a meta-estimator that aggregates many decision trees and enhances them, it integrates the outcomes of many forecasts. A percentage of the total number of functions (referred to as the hyperparameter) can only be split among each node [137].(See Figure 1.11).

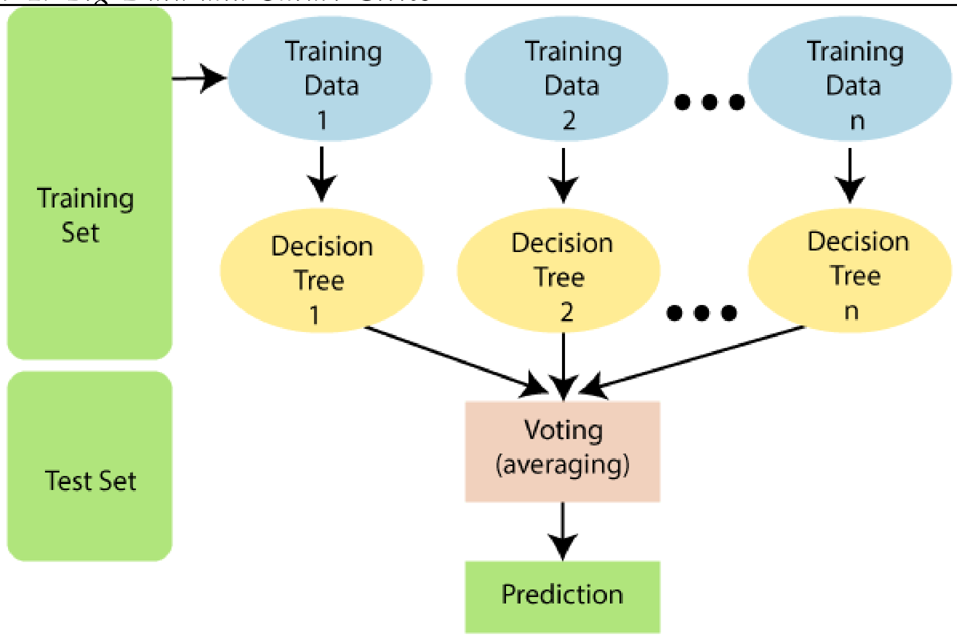

Figure 1.11: Random Forest Algorithm

e. K neighbors Classifier (KNC):

The basic idea behind K-neighbors classification is to find the k training samples that are physically nearest to a new sample that needs to be classified. The neighbors will decide on the labeling of several samples. The number of neighbors that must be found in k-nearest neighbor classifiers is controlled by user defined constants. Radius based neighbor learning methods have a variable number of neighbors dependent on point density; each sample inside a given radius has a variable number of neighbors. The kNN algorithm is a significant machine learning algorithm. It has excelled in a variety of regression and classification applications. [116]

f. Extra Trees Classifier (ETC):

Both ETC and RFC are extremely similar methods of grouping. But there is a distinction. The RFC employs bootstrap replicas, which implies it subsamples the input data with replacement, as opposed to the ETC, which uses the whole original sample. Users can bootstrap replicas using an optional parameter in the Extra Trees Scikit learn implementation, although by default it uses the entire input sample. Bootstrapping diversifies the data, which could lead to an increase in variance. The use of cut points for dividing nodes is another differentiator. The Random Forest Classifier chooses the split at random, while the Extra Trees Classifier determines the best split. After selecting the split points, the two methods compare which subset of features is best. [118]

g. Gradient Boosting Classifier (GBC):

Gradient boosting classifiers are a type of machine learning algorithm that combines multiple weak learning models to produce a potent predictive model. Gradient improve- ment typically makes use of decision trees. Using a weighted version of the training data, the gradient boosting method gradually creates fundamental models to precisely determine the best tree configuration. Correcting the mistakes made by the previous base models is the aim of each basic model addition phase. As a result, the gradient boosting method may produce more accurate projections. [120]

As a way to assess the capacity to catego- rize complicated information, gradient boosting models are becoming more and more popular. It is a machine learning technique for regression and classification that creates a prediction model from a group of weak prediction models, frequently decision trees[134]. The GRBT stands for the generalization of boosting to any arbitrary differentiable loss function. It represents a precise and useful solution that can be applied to regressionand classification issues alike [139].(see Figure 1.12).

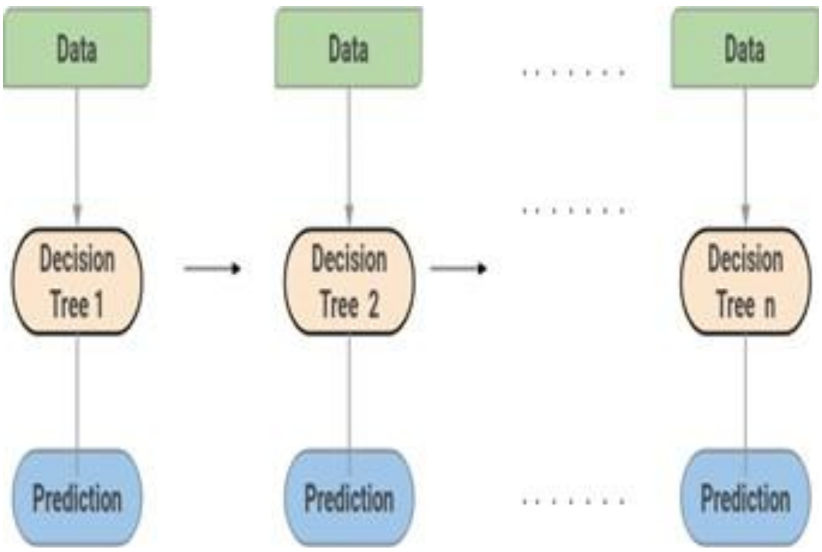

Figure 1.12: Gradient boosting Algorithm

1.8 Conclusion

In this chapter, we discussed several concepts related to big data and smart cities. We have provided an overview of the definition of big data, its characteristics, challenges, and applications along with a comprehensive definition of the smart cities and its basic elements. We have carefully examined Big Data characteristics and issues brought on by Big Data computing platforms. Additionally, we have discussed the importance of big data across several industries. We also focused on the components and technology utilized in each layer of the big several forms. In the next chapter, we will present several works that use big data to develop smart cities.

Chapter 2

Chapter 2: Big Data Analytics in Smart Cities-State of the Art-

2.1 Introduction

The integration of big data and smart cities has been a rapidly evolving field in recent years. Advances in technology, including the rise of IoT devices and cloud computing, have enabled cities to collect and analyze vast amounts of data from various sources. This has led to a growing number of innovative solutions and applications in various domains such as transportation, energy, public safety, and urban planning.

In conclusion, the state of the art in the field of big data and smart cities continues to evolve, with new technologies and applications emerging all the time. The goal is to create smarter and more sustainable cities that provide better services and improve the quality of life for all residents.

In this chapter, we will present different works, which use big data to create smart cities.

2.2 Related Works

We discovered numerous works in the literature that suggested various solutions to the problems of decision quality improvement and execution time reduction. All of these efforts aim to increase Big Data decision quality or decrease execution time. We provide a business overview on our topic that Big Data Analytics in Smart Cities summarizes the main research topics used, such as big data, smart cities, and big data analysis:

- Bibri in [58], contributed to identifying and analyzing the latest sensor based big data applications provided by the Internet of Things for environmental sustainability and models for data processing and computing platforms and proposes an analytical framework for data centric applications to enhance environmental sustainability in the context of smart sustainable cities.
- Osman in [59], provides a framework for smart cities big data analytics called "Smartcities Data Analytics - SCDAP". SCDAP introduces new big data analysis functions represented in managing and collecting data models as compared to traditional knowledge discovery methods.
- Ju et al., in [60], proposed a framework to take advantage of big data analytics for citizens to provide information in smart cities in two ways, one of which is urban management issues and the other is data analysis algorithms so that the framework consists of three layers (data integration layer, knowledge discovery layer, and decision making layer).
- Rathore et al., in [61], they used real time analysis of the big data generated in the smart cities based on the Internet of Things. The proposed system includes data generation, collection, filtering, classification, manipulation, storage, and decision making using the Hadoop ecosystem in a real environment, and the system has also been evaluated for efficiency in terms of scalability and real time data processing.
- Allam in [62], intended for those who want to integrate artificial intelligence and big data in smart cities. The framework also supports the optimal use of technology in smart cities with the introduction of big data and artificial intelligence using the three main dimensions of culture, metabolism, and governance that aim to ensure an advanced community life by following and introducing technology. As a primary feature, artificial intelligence is suggested to process, analyze, and interpret new data.
- Rathore in [63], focused on developing a secure, real time smart cities system with robust intrusion detection and access control at the system level and a secure and efficient communications security protocol at the data transmission level. The

proposed protocol includes the registration phase, the session key exchange phase, the session key revocation phase, and the data transfer steps. The focus is on the security of remote intelligent systems transfers to the Smart cities Building (ICB) and from the ICB to users.

- Sun in [64], launched a study on the role of the smart big data platform and chose the development of smart cities In Hefei as an experimental analysis. The blockchain government information resource exchange and exchange model was also designed with three parts: the network layer, blockchain infrastructure layer, and business application layer, TOPSIS based overall evaluation model; an experimental evaluation of Hefei Intelligent Development from 2012 to 2017; and a Merkle tree proposal that uses data segmentation to create a recurring tree shaped node.

- Lim et al., in [65], presented the results of analyzing different uses of big data in cities, classified urban data use cases into four reference models, and identified six challenges in transforming data into information for smart cities. This paper also proposes five considerations for addressing challenges in implementing reference models in real world applications. Collectively, the reference models, challenges, and considerations form a framework for using data for smart cities.

- Gohar in [66], proposed a big data analytics architecture for ITS that has built in storage and analysis capacity to work with ITS data and consists of four modules (Big Data Acquisition and Preprocessing Unit, Big Data Processing Unit, Big Data Analytics Module, Visualization Module data). A detailed analysis of the huge ITS data is done to monitor the average vehicle speed. Encouraging experimental results were obtained and the concept was validated with the Hadoop system.

- Song in [67], proposed a framework for analyzing big data for healthcare and social services, which is then applied to the prevention and proactive detection of suicide in Korea. Unstructured data are collected through aggregation engines and then subject to text and opinion mining. So, the unstructured data collected using the network was classified and converted into structured data.

- In the article [68], the authors (Yu and Wang) collected real time tweets from U. soccer fans' emotional responses to their tweets, particularly, the emotional changes after goals. This project revealed that sports fans use Twitter for emotional purposes and that the big data approach to analyzing sports fans' sentiment showed results generally consistent with the predictions of the disposition theory when the fanship was clear and showed good predictive validity.

- Khan in [69], proposed a cloud based analytical service. They designed a prototype, implemented it using Hadoop and Spark, and then compared the results to measure its effectiveness for analyzing big data. Therefore, they analyzed Bristol Open data to analyze the extent of correlation between cities specific indicators.

- Alsunaidi in[70], presented an overview of several COVID-19 data analysis applications and provided a taxonomy structure that categorized COVID-19's prospective applications into four groups: diagnostic, estimate or prediction of risk score, healthcare decision making, and pharmaceutical. The report described each tool's key features and introduced several data analysis tools. They also gave crucial information on several difficulties that would make it difficult for COVID-19 to employ data analytics tools.

- DAshaari in [71], proposed a methodology that involved a dual stage study of partial least squares structural equation modeling and deep learning, a developing form of artificial intelligence. The proposed model can be predicted with an accuracy of 83% using deep ANN architecture. The results of data driven choice-making from the interaction between big data analytic capability and data driven decision making towards the performance of HEIs also have important discoveries. Results showed that the association between big data analytic capabilities and the performance of HEIs might positively play a vital role in data driven decision making.

- Big data analytics are addressed by Mostafa in [72], for renewable energy power plants. A framework was created for the possible use of big data analytics for smart grids and renewable energy power utilities. The stability of the smart grid is predicted in five steps, utilizing five distinct machine learning techniques. Through the use of three different machine learning techniques, the stability of the system was predicted using data from a decentralized smart grid data system that contained 60,000 instances and 12 attributes.

- The authors of the paper [73], provided a high level overview of large data handling and technological advancements, along with examples of how they might be used in next generation wireless networks (NG). They used cutting edge analytics to assess the demands of mobile users and then applied that understanding to boost the effectiveness of "community wireless communication channels." The framework that has been proposed is capable of generating the logic, problem formulations, and methodology for robust computational models within the context of wireless networks by relying on massive data that reflects both the spectrum and other pressing user demands.

- The authors of the article [74], suggested electric vehicle networks in smart cities using big data analysis technology. They built an objective function based fuzzy mean clustering algorithm theory using K-means and fuzzy theory (FCM). The electric car network is simulated, and the FCM algorithm is enhanced. The findings indicate that when the likelihood of successful propagation is 100 percent and the value is between 0.01 and 0.05, it is the closest to the actual outcome and has the least data delay in the examination of network data transmission performance. When encountering congested road sections, the route guidance approach can effectively halt the expansion of congestion and achieve timely traffic congestion evacuation, according to the analysis of the route guidance effects.

- Lian in [75], reviewed recent research studies that used big data to analyze traffic safety in the ITS and CAV environment by discovering or predicting faults, discovering causes, analyzing driving behavior, and identifying the fault point. The traffic safety big data application integrates and addresses multi source big data, problem detection, and resolution.

- Bilal et al., in [76], proposed a big data architecture for construction waste analysis. The graph database (Neo4J) was adopted to store large and varied data, and the Spark graph processing system was used. The proposed big data architecture is based on a clear lifecycle for waste analysis, which is the backend for the intended construction waste simulation tool (which will be usable with current BIM authoring tools like Autodesk Revit and Micro station for designing building waste).

- Zeng in [77], maked two theoretical contributions. First, the results work at the top of expanding the concepts about using BDA in the context of a smart cities, especially in the field of smart tourism, based on the theoretical lens of technical exemptions. We depicted the use of BDA for smart tourism as a sequential cost-realization process. Secondly, it is linked to the theory of technology granting by expanding theorizing about the process. Incorporation of costs by proposing the concept of cascading benefits.

- Zhang in [78], proposed an adaptive, task level MapReduce framework. This framework extends the overall MapReduce architecture by designing each Map and Reduce task as a consistent run loop server. To further refine, this framework, we applied a method for predicting a workload flow, as well as smoothing and a Kalman filter, to estimate unknown workload characteristics. This document also included a diagram. Adaptive MapReduce at task level as well as runtime map and minimized task computation than in Adaptive MapReduce simulation.

- Khan in [79], suggested a paradigm for evaluating the big data analytics (BDA) adoption barrier associated with the growth of smart cities. The suggested approach helps policymakers and decision makers examine the obstacles to BDA adoption in the planning and development of smart cities. As they examined the primary obstacles to the adoption of BDA in the development of smart cities, it is useful to approach these obstacles in an optimum manner to hasten the growth of smart cities. The findings indicate that the main obstacles to incorporating BDA into the creation of smart cities are data complexity, acceptance frameworks for BDA, and a shortage of BDA technology.

- For the future's smart cities, Aljehane in [80], created an effective vehicle routing protocol based on metaheuristic optimization. The model that is being given shows VANET characteristics alongside big data characteristics. Paths on vehicles are simultaneously chosen to utilize the independent distributed framework Hadoop Map Reduce and VANET's OMFOVR algorithm. by doing a difficult set of simulations to demonstrate the useful outcome of the OMFOVR method that has been provided.

- Rahman in [81], suggested a new machine learning fusion method for real time precipitation forecasting for smart cities. The four commonly used supervised machine learning methods used in this suggested framework are decision trees, Naive

Bayes, K-nearest neighbors, and support vector machines. Fuzzy logic technology, often known as "fusion," has been integrated into the framework for efficient rainfall prediction to integrate the predictive accuracy of machine learning approaches.

- Kahveci et al., in [82], unveiled an end-to-end IoT-based big data analytics platform with five interconnected levels and several components for data gathering, integrating, storing, analyzing, and visualizing functions. Utilizing the most recent technology, the platform architecture integrates them in a methodical, compatible manner with a clear information flow. The developed platform has been implemented in the Automation Systems Group's Electric Vehicle Battery Module Assembly Automation System created at the University of Warwick in the UK. The created proof of concept solution illustrates how various tools and techniques can be coordinated to cooperate in intelligent manufacturing environments to support decision making and enhance the process and product quality.

- In a study by Jayasri et al., in [83], they combined the Innovative Hierarchical Decisions Interest Network, Association Rules (AR), and multi layer classification with the MapReduce architecture to the diabetic medical database. Health data is considered in the MapReduce framework's apriori algorithm for the association rule, which generates the rules. The correlation between an illness and its symptoms has also been found using this method.

2.3 Taxonomy of literature review

We compare many studies in terms of many important points (see Table 2.1), and the taxonomy categorizes reviews according to the:

2.3.1 Objective(Focus)

When conducting applied research, the researcher goes above and beyond to pinpoint a problem, formulate a research hypothesis, and then move on to test these hypotheses through an experiment. This research strategy frequently uses empirical techniques to address real-world issues. There are 4 types of Objective applied research. These are evaluations: applications, research outcomes, research methods, and theories.

2.3.2 Types of Big Data

There are primarily three types of data in big data:

- Structured: Structured data refers to the data that you can process, store, and retrieve in a fixed format. It is highly organized information that you can readily and seamlessly store and access from a database by using simple algorithms.
- Unstructured: Data with an unknown structure is termed "unstructured data." Its size is substantially larger than that of structured data, and it is heterogeneous.
- Semi structured: As the name suggests, semi structured data contains a combination of structured and unstructured data. It is data that hasn't been classified into a specific database but contains vital tags that separate individual elements within it.

2.3.3 Type of Big Data Analytics

Big data analytics has been divided into four categories: descriptive, predictive, diag- nostic, and prescriptive[140]. They use several technologies for operations like data mining, purification, integration, visualization, and many more to improve the process of data analysis and ensure the company benefits from the data they get.

- **Descriptive analytics:** Descriptive analytics is one of the most often used forms of analytics by companies to track current trends and operational performance. One of the first steps in the analysis of raw data is to use simple mathematical operations to provide conclusions about samples and measurements. The other sorts of analytics can be used to discover more about the causes of the trends you've discovered with descriptive analytics. When working with finances, production, and sales, you must employ descriptive analytics. The creation of financial reports and metrics, surveys, social media campaigns, and other work related assignments are among the tasks that call for this kind of analytics.
- **Predictive analytics:** As the name implies, the focus of this sort of data analytics is on creating predictions about potential outcomes based on data insight. It employs a variety of sophisticated predictive techniques and models, including machine learning and statistical modeling, to get the best results. One of the most

popular categories of analytics in use today is predictive analytics. Predictive analytics' key advantage is its ability to foretell the future with greater accuracy and dependability. Companies may manage shipment schedules, stay on top of inventory needs, and identify methods to save and make money by using the predictions generated by this kind of analytics.

- **Diagnostic analytics:** One of the most sophisticated sorts of big data analytics you may employ to examine data and content is diagnostic analytics. With the help of this kind of analytics, you can use the knowledge you obtain to explain why something occurred. By studying data, you understand the causes of customer and employee behavior and events related to organizations. Although product promotion does not change, sales can change dramatically. To find the causal relationship for such a change, diagnostic analyses will be used.

- **Prescriptive analytics:**Prescriptive analytics uses numerous simulations and methodologies to find ideas for improving company procedures based on the findings from the descriptive and predictive analysis. It makes recommendations for the company's best next move based on data insights. One of the numerous businesses that employ this kind of analytics is Google. When creating their self driving automobiles, they used it. These vehicles use prescriptive analytics to analyze data in real time and make decisions.

2.3.4 Audience

Big data provides real insights into who people are. The more marketers can understand, interpret, and analyze consumers, the more leverage they have to deliver personalized, valuable campaigns to the right people, in the right place, at the right time.
The various reviews' target audiences may be different from one another. Reviews can be written for the general public, practitioners, policymakers, specialist research groups, general research groups, or practitioners. Less language and detail are used by reviewers as they address the general public rather than expert scholars, and reviews prepared for more specialist audiences are sometimes condensed and simplified by popular writers.

2.3.5 Coverage

The coverage is most likely what makes literature reviews unique. May understand the causes of specific behaviors and occurrences related to the firm you work for, their clients, staff, goods, and more processes that are specific to this type of scholarship when they conduct literature searches and decrial timer or not a causal work is appropriate and of high quality. They include: exhaustive, exhaustive with selective citation, representative, central, or pivotal.

The first level, exhaustive coverage, denotes the reviewer's intention to present all works pertinent to the subject at hand. In the second type of coverage, conclusions are also based on the entire body of literature, but only a chosen selection of works is discussed in the paper. Either of the patterns to be briefly described may be used as the strategy for choosing which works to cite. The reviewer focuses on works that have been essential or influential to a topic area when using the coverage strategy.

2.3.6 Big Data Platform

Many popular big data platforms can help manage and analyze data and provide actionable insights, including Apache Hadoop, Cloudera, Apache Storm, Amazon Web Services, Oracle, and Snowflake. Companies and decision makers are looking for ways to leverage big data and generate useful insights to improve decision making. For this reason, they are using big data platforms, which offer a one stop shop for all of their data requirements. They support gathering, arranging, storing, searching, sharing, analyzing, and reporting data insights.

2.4 Comparison between different works in big data analytics

In Table 2.1, we summarize the related works that were previously defined:

Table 2.1: Comparison between different works in big data analytics

Source	**Objective**	**Types of big data**	**Type of Analysis**	**Audience**	**Coverage**	**Platform**
[58]	Application	Structured	Prescriptive analytics	scholars, researchers	Coverage exhaustive	Hadoop MapReduce
[59]	Application	Unstructured	Descriptive analytics	Specialized scholars	Coverage exhaustive	Hadoop
[60]	Application	Unstructured	Prescriptive analytics	practitioners, policy makers	central or pivotal	Hadoop distributed file system (HDFS)
[61]	Research outcomes Application	Unstructured	Descriptive analytics	specialized scholars, policy-makers	exhaustive with selective citation	Hadoop, Spark, and Giraph technology
[62]	Research methods	Structured	Prescriptive analytics	Policy Makers, Data Scientists and Engineers	Policy Makers, Data Scientists and Engineers	/
[63]	Application	Semi-structured	Prescriptive analytics	specialized scholars,practitioners	central or pivotal	Hadoop

[64]	Research outcomes Application	Semi-structured	Predictive analytics	specialized scholars,policymakers	exhaustive with selective citation	blockchain technology
[65]	Research Methods Theories	Unstructured	Prescriptive analytics	specialized scholars, practitioners	exhaustive	/
[66]	Research outcomes Application	Unstructured	Descriptive analytics	practitioners, policymakers,general public	central or pivotal	Hadoop,Spark and MapReduce
[67]	Research outcomes	Structured and Unstructured	Prescriptive analytics	specialized scholars, practitioners,policymakers	representative	/
[68]	Research outcomes Application	Unstructured	Diagnostic analytics	specialized scholars,policymakers ,general public	central or pivotal	Python and the Natural Language Toolkit (NLTK)
[69]	Research outcomes	Unstructured	Descriptive analytics	specialized scholars,practitioners,policymakers	exhaustive	Hadoop and Spark

[70]	Research outcomes Application	Unstructured	Predictive analytics	specialized scholars, practitioners,policymakers	representative	Apache Hadoop
[71]	Research outcomes Application	Unstructured	Predictive analytics	specialized scholars,general scholars,practitioners,policymakers	exhaustive with selective dictating	MapReduce, Hadoop
[72]	Research outcomes Application	Unstructured	Predictive analytics	specialized scholars,policymakers	central or pivotal	Amazon S3, BigQuery, google drive, Microsoft Azure, and Hadoop
[73]	Research methods Theories	Semi-structured	Diagnostic analytics	specialized scholars,policymakers	representative	Hadoop, Neural Networks and Deep Learning
[74]	Research methods Theories	Unstructured	Prescriptive analytics	specialized scholars,general scholars,practitioners,policymakers	exhaustive with selective dictating	Apache Spark, Hadoop
[75]	Research outcomes	unstructured	Predictive analytics	Research outcomes	specialized scholars,practitioners,policymakers	/

[76]	Research Methods Research outcomes	Unstructured	Descriptive analytics	specialized scholars, practitioners,policymakers	exhaustive	Apache Hadoop and Berkeley Big Data Analytics Stack (BDAS)
[77]	Research outcomes Application	Unstructured	Diagnostic analytics	specialized scholars,policymakers	central or pivotal	/
[78]	Research methods Theories	Unstructured	Diagnostic analytics	general scholars,practitioners,policymakers	exhaustive with selective citation	Hadoop MapReduce
[79]	Application	Unstructured	Descriptive analytics	specialized scholars,practitioners	exhaustive	Apache Hadoop
[80]	Research methods	Unstructured	Descriptive analytics	specialized scholars, policymakers	exhaustive with selective citation	Hadoop Map Reduce
[81]	Research methods Theories	Unstructured	Predictive analytics	scholars, researchers	representative	machine learning techniques

[82]	Research methods Theories	Unstructured	Prescriptive analytics	smart manufacturing environments	representative	Apache Spark, Hadoop
[83]	Research outcomes Application	Unstructured	Predictive analytics	practitioners, policymakers	representative	MapReduce

2.5 Discussion

The works presented above cover various topical areas where the quantity and variety of data are important aspects characterizing the manipulation and analysis of processed data. This challenge is solved using the means and capabilities offered by big data tools (Hadoop, HDFS, MapReduce, HBase, Hive, MongoDB, etc.). The need to overcome the heterogeneity of different types of data generated from a multitude of resources can provide intelligent data that is easy to share and reuse because it is standardized and harmonized. To achieve the correct findings, new methodologies, technologies, and solutions in terms of human computer interactions must be developed. Big data privacy, safety, and security are the key topics that will be debated more in the future.

It turns out that big data applications use big data analysis methodologies to effectively evaluate enormous amounts of data, as shown by our comparison of various works and earlier studies. Due to the difficulties in big data processing and application, it might be challenging to choose the best big data tools based on big data processing, streaming, and analysis technologies.

In our study of historical and contemporary literature, we examine the application of big data tools and big data analytics techniques in fields like health and medical care, social networks, the Internet, government and the public sector, natural resource management, and economics. Additionally, the business sector prompted us to go through the following crucial phases:

- Recognize the direction of big data research and the present big data technology frameworks.
- Spotting patterns in the application or study of big data tools based on batch processing, data streaming, and big data analysis methods.
- To aid in the proper development of new research activities in this sector and to offer new researchers and practitioners the necessary resources.

Insights and knowledge about current big data platforms and their application domains, big data analysis, and utilization techniques, and new research prospects in the development of big data systems were therefore offered by this comparison.

2.6 Conclusion

In this chapter, we present many works which use Big data analytics for smart cities. Several tools and frameworks for big data analytics exist, as well as several significant outstanding concerns. Big data has revolutionized the way organizations operate and make decisions. The state of the art big data applications in smart cities have proven to be valuable across a range of industries, including finance, healthcare, retail, and many others. These applications leverage the vast amounts of data generated by businesses and individuals to deliver insights and recommendations that would have been impossible to uncover just a few years ago. In the next chapter, we will present our contributions.

Chapter 3

Contribution 01: Real-Time Human Activity Recognition from Smart Phone

3.1 Introduction

A wide range of sensors is incorporated into the current generation of phone mobile devices, including smartphones, music players, smart watches, and fitness trackers, which can be used to analyze human activity and behavior. With several applications in the fields of healthcare, sports, interactive gaming, and general monitoring systems, human activity recognition (HAR) has emerged as a significant study area in ubiquitous computing and human computer interaction[109]. Numerous motion sensors have been positioned at various body positions to collect data and deduce specifics about human activity as a result of the rapid technological improvement of mobile phones and other wearable devices.

To validate our contribution, this chapter is devoted to the proposed approach. We offer a smart solution to deal with urban challenges in the real lives of citizens. Human activity recognition and indoor positioning are among the most important and relevant areas for the development and sustainability of smart cities [1].Our contributions are composed of three following points:

- We propose an architecture that combines several layers to extract supervised local features along with statistical features that encode global properties of the time series.

- We use the Kafka System to reduce the time of response.
- We compare our model with others works of literature.

3.2 Problematic

Human Activity Recognition (HAR) is a difficult problem, yet it must be solved because it is mainly used for elderly care and healthcare as an assistive technology when combined with other technologies such as the Internet of Things (IoT). HAR in smartphones is seen as essential not only to better understand human behavior in everyday life but also to provide context for other applications in the smartphone. The problem is how to improve the quality of HAR statistical and logical models and reduce the execution time of data mining algorithms based on machine learning algorithms. Finding a way to apply this technology to big data in the form of a multi agent system that interacts with local analytics in a self organizing way. A major challenge in this area is shifting from data collection that can address the persistent difficulties of big data analysis, processing of models, feature extraction, and activity classification.

We provide insight into the complexity and multidimensionality of HAR using smart-phones, the types of data collected, and the methods used to translate numerical mea-surements into human activities. We discuss the generalizability and reproducibility of the approach, features that are essential and apt to large and diverse groups of study participants. Finally, we identify challenges that need to be addressed to accelerate the wider use of smartphone based HAR in public health studies.

A smart cities use of cell devices integrated with its widely dispersed sensors enables it to provide residents with guidance, instructions, and care. Shifting from data collection to data analysis, focusing on the methods used for data acquisition, machine learning can be used to investigate statistical models of human behavior. Due to statistics series, many studies have not examined type overall performance in real time.

3.3 Global System

3.3.1 Layers of System

Architecture is a fundamental organization of a system made up of its components, their interrelationships with the based environment, and the principles that guide its design and evolution. These definitions describe architecture as structure[129]. This struc- ture is made up of the components or elements and the relationships or connectors be- tween them. In the scientific literature, many research works are proposed. We make a classification of architectures: general architecture, referential architecture, application- oriented architecture, and architecture based on Cloud Computing.

The main objective of the proposed system is to enable effective control of the devices of smart buildings (air conditioning, heating, doors, etc.) and indoor spaces as well as to identify the activities of the residents to present an appropriate strategy and ensure the convenience of the residents of smart cities.

In this section, we describe the different layers that make up our approach and the relationships among them. The smart system for managing real time smart cities applications is based on a multi layered architecture. Each layer contains agents with specific behavior and specific roles. Other layers are added to meet the requirements of smart cities, such as the high number of serviced homes (processing speed and storage space), real time interaction, autonomy services, etc. It also provides scalability with big data infrastructure, autonomy, resilience, and predictability.

The global architecture is divided into four essential layers; each layer contains multiple based modules that play a complementary role with each other and exchange data with adjacent modules to achieve the goals. (Table 3.1 shows the layers and associated technologies.)

Table 3.1: The layers of the architecture and the associated technologies used

The layer	**Tools and Technologies**
The user layer	sensors, devices
Data processing layer	Kafka, MongoDB, Hadoop HDFS, MapReduce.
Security layer	security audit
Platform layer	clusters of hardware, operating systems, and communication protocols

Big data tools are deployed to process thousands of gigabytes of data generated by different devices installed in smart cities. These tools offer promising solutions for managing the volume, speed, and diversity of data. Likewise, the multi layered approach provides the robustness, capacity, and scalability needed for data processing and analysis. In addition, it allows the management of both real time data and archived data [84].

The big data architecture of smart cities consists of the platform layer, the security layer, and the data processing layer. In the following, we describe each layer, and a brief explanation is also included of why each technology is selected and its role.

- The User layer: Data generated by different sensors is collected and aggregated everywhere. It is then sent to Kafka for further redirection for processing [85].

- Data processing layer: This is a central data processing mechanism that provides all data processing functions from data acquisition to data extraction. This layer supports the processing of real time data and historical data respectively. This layer also provides important features such as model management and aggregation. [86]

- Security Layer: The following security measures should be considered during physical design, especially for critical analytics. Restricted access to the framework should also be granted to critical and sensitive data as well as multi level user authentication. [87]

- Platform layer: is a horizontally scalable platform comprising clusters of hardware, operating systems, and communication protocols. In horizontally scalable platforms, additional compute nodes can be added as needed. [88]

3.3.2 Proposed Architecture

Human activities are continuously tracked in the smart home by transferring the data produced by the sensors of mobile phones to obtain the information in real time. Powerful devices with parallel processing of record evaluation and real time decision-making are needed for the Metropolis smart cities devices. The key feature of this design is that it uses Kafka as a bridge between the many data sources used to compile the characteristic statistics, the environment used to develop the model, and the utility used to offer predictions.

The proposed architecture is composed of two layers: the smart cities layer and big data layer. We describe each layer, and present a brief explanation of them in the following points (see Figure 3.1):

- **Smart city layer:** The era of massive amounts of information and the rapid growth of the cellular network, which is the foundation of vast records, are both results of the cellular network's quick development. The information produced by the sensors is accumulated, collated, and then delivered to Kafka for processing. Through Kafka, he receives lower backorders. It real time source layer is responsible for generating the data. It includes all aspects of building a vehicle network and deploying roadside sensors to generate real time information. The generated data can thus be transmitted over the Internet using relay nodes, coordinators, and gateways. This layer represents the user interface agent, who serves as the system's and users' interface (such as a supplier) and whose main function is to interact with users and respond to their queries while assisting them in carrying out specified tasks. Collects data generated by different sensors and collated at a supervised agent in each house who sends it to Kafka to then be redirected for processing. Also, it receives commands that are returned by the service provider layer through Kafka to activate and change the state of devices to modify environmental conditions.

- **Big Data layer:** The core of our strategy is this second layer, which is made up of several service agents and Hadoop storage structures (such as HDFS). To give services to users, the service agents make use of the processing engines. To accomplish the system's overall objectives, this layer enables service agents to co-operate and communicate with one another. The data pipeline is designed using the different requirements of a batch processing or data storage system by the big data layer. Two layers in this arrangement guarantee the secure flow of data:

  - Data Streaming and Storage: The mobile app uses JSON to communicate with

Kafka. The message is handled by the predictor, who then publishes it to Kafka with the prediction result, where the application will finally receive it. The predictor will post a "retrain subject" message after several messages. The "retrain subject" message will be received by the trainer, who will then begin retraining the algorithm, and the predictor will continue to provide forecasts. The trainer publishes a message to the predictor indicating that the algorithm has been trained, at which point the predictor downloads the updated model. The Hadoop layer stores the data processes it quickly, and manages storage, availability, speed, scalability, and other factors well (via HDFS and MapReduce).

Once Kafka receives the data, it forwards each of them to the one responsible for processing, through communication with the storage layer. At this layer, Kafka acts as an event flow delivery platform for building real time data pipelines and streaming applications. It runs as a cluster on multiple servers that can span multiple data centers and can handle trillions of events a day. Their use in our architecture is to create real time data pipelines that reliably move data between layers and turn it into data streams.

- Machine Learning (ML): Kafka receives data from the numerous apps and databases that house feature data, this data is used to build models. The environment will alter based on the skills needed and the toolset used. The model might be created using a data warehouse or a big data environment like Hadoop. The model can be made accessible so that it can be used to process incoming examples by any production software that receives the same model parameters (perhaps using Kafka Streams to help index the feature data for easy usage on demand). The production application may be a pipeline that merely gets data from Kafka, or it could be a Kafka Streams application. Kafka serves as the ML architecture’s brain when it comes to generating, building, and analyzing models.

Figure 3.1: The architecture of the implemented system

3.4 Method

In this study, we presented a comprehensive framework for Kafka based real time human activity recognition structures and implemented several machine learning algorithms to analyze popular and essential human activity. To compare the various methods used to classify data from wearable sensors, we used Logistic Regression Classifier, Linear SVC Classifier, Decision Tree Classifier, and Random Forrest Classifier.

The main objective is to provide machine learning methods for employing smartphone sensors to identify human activity. We describe the environments, tools, and datasets utilized to implement our suggested architecture in this section. Further, we display the outcomes that were accomplished.

3.4.1 Hardware and tools

We have studied and compared the training times of several machine learning algorithms in a specified environment. As we applied the simulation of our framework to the supposed cluster (5 machines: a master with 4 GB of RAM and 100 GB of hard disk, and 4 slaves with 2 GB of RAM and 250 GB of the hard disk).

According to the objective of the project, a programming language is chosen over another language in the field of big data. However, Python is the most widely used language and is especially effective in the field of big data analysis and development.

- **Python:** Nowadays, Python is the most widely used programming language. Several hints of programming languages like PYPL. (see Figure 3.3)

PYPL PopularitY of Programming Language

Worldwide, Nov 2022 compared to a year ago:

Rank	Change	Language	Share	Trend
1		Python	28.44 %	-1.2 %
2		Java	17.03 %	-0.3 %
3		JavaScript	9.51 %	+0.3 %
4		C#	7.08 %	-0.1 %
5		C/C++	6.51 %	-0.4 %
6		PHP	5.12 %	-1.0 %

Figure 3.2: Most used programming languages for data science by PYPL index.

Python is one of the most powerful and flexible languages, and it is also widely used in data science. The main reason is its simple and elegant syntax, as well as its large collection of third party libraries. A tool you'll find everywhere in the field of data science is Jupyter. With Jupyter Notebooks, you can quickly see the results of the code you're working with, plot data, and create documentation of your code through markdown blocks. is not a Python only tool, but the most common combination is Python and Jupyter. [97]

Any data science task you can do with Python. This is mainly thanks to its rich ecosystem of libraries. With thousands of powerful packages backed by its huge community of users, Python can perform all kinds of operations, from data pre-processing, visualization, and statistical analysis to the deployment of machine learning and deep learning models. Here are some of the most used libraries for data science and machine learning purposes:

- NumPy is a popular package that offers an extensive collection of advanced mathematical functions. Many packages are based on Numpy objects, like the famous NumPy arrays. NumPy is the fundamental package of Python that makes possible scientific computing. [98]

- Pandas: is a key library in data science, used for performing all kinds of manipulation of databases, also called DataFrames. [99]

- Matplotlib: the standard Python library for data visualization. Matplotlib is a Python library that helps in 2D plotting for hardcopy publication formats with an interactive environment across platforms. Matplotlib allows the generation of plots, bar charts, histograms, error charts, power spectra, scatter plots, and more. [100]

- Scikit learn: built on top of NumPy and SciPy, it has become the most popular Python library for developing machine learning algorithms. [101]

- TensorFlow: developed by Google, is a powerful computational framework for developing machine learning and deep learning algorithms. [102]

- Keras: an open source library designed to train neural networks with high performance. [103]

3.4.2 Datasets used

A group of 30 individuals between the ages of 19 and 48 participated in the experiments. Each participant completed six tasks while carrying a Samsung Galaxy S II smartphone around their waist (WALKING, WALKING UPSTAIRS, WALKING DOWN-STAIRS, SITTING, STANDING, LAYING). With the use of the device's internal accelerometer and gyroscope, we captured 3 axial linear acceleration and 3 axial angular velocity at a constant frequency of 50 Hz. To enable manual labeling of the data, the tests were recorded. The collected dataset was divided into two sets at random, with 30% of the volunteers chosen to produce test data and 70% of the volunteers chosen to provide training data.

Before sampling, noise filters were applied to the sensor data (accelerometer and gyroscope), which were then sampled in 2.56 second fixed width sliding windows with

50% overlap (128 readings/window). The body motion and gravitational components of the sensor acceleration data were separated into body acceleration and gravity using a Butterworth low pass filter. The gravitational force is thought to only have low-frequency components, hence, a filter with a cutoff frequency of 0.3 Hz was used. A vector of properties from each frame was produced by calculating variables in the time and frequency domains. We received 563 attributes:

- These time domain signals, which start with the letter "t" to denote time, were gathered. They were filtered with a median filter and a 20 Hz corner frequency Butter-worth filter to reduce noise.
- The acceleration signal was divided into body and gravity acceleration signals using a low pass Butterworth filter with a corner frequency of 0.3 Hz (tBodyAcc-XYZ and tGravityAcc-XYZ).
- To generate Jerk signals, the body's linear acceleration, and angular velocity were then determined (tBodyAccJerk-XYZ and tBodyGyroJerk-XYZ). The magnitude of these three dimensional signals was also calculated using the Euclidean norm (tBodyAccMag, tGravityAccMag, tBodyAccJerkMag, tBodyGyroMag, tBodyGyroJerk-Mag).

3.5 Results and Discussions

3.5.1 Obtained results

We used a comparative analysis of four classifiers after choosing the feature to improve the model. The four classification algorithms that were examined were the logistic regression classifier, the linear SVC classifier, the decision tree classifier, and the random forest classifier. We used L1 based feature selection techniques to condense the size of the feature dataset, and we contrasted the results with a measure of the model's correctness and the length of time it took to build and train the model. Out of the 7352 features in the original dataset, the L1 based feature selection technique selected 563 features. The Random Forrest Classifier model was created by evaluating "n estimators" with values of 100, 50, 200, 250, and 350. After 200, it was discovered that the precision did not improve. (see the result in Table 3.2)

Table 3.2: Using performance analysis to rank the classifiers

Model Name	**Accuracy**	**Precision**	**Recall**	**Rank**
Logistic Regression	95.83%	95.83%	95.83%	2
Linear SVCr	99,06%	96,45%	95,42%	1
Decision Tree	87,21%	82,74%	80,94%	4
Random Forest	92,09%	87,38%	86,71%	3

We can use Linear SVC as our best model while using the Machine Learning Classical Model. The model accuracy scores in Figure 3.3 offer more context and a more thorough analysis of the results.

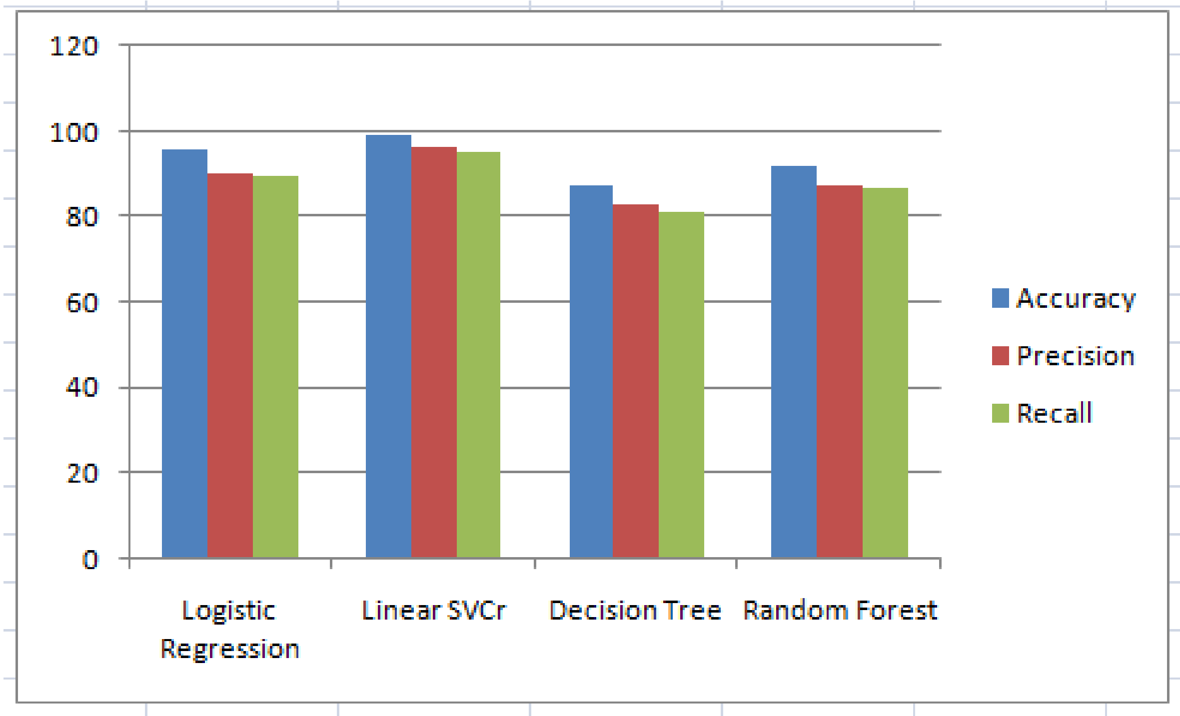

Figure 3.3: Model accuracy scores

3.5.2 Discussions

The results of the different ML classifiers are displayed in Table 3.2. The results of each classifier's findings are shown by their ranks. The rank of the classifier is determined by the degree of recognition accuracy. Ranking 1 goes to the classifier with the highest recognition accuracy, followed by ranking 2, which goes to the classifier with the lowest recognition accuracy. Based on their respective recognition accuracy scores of 99.06%, 95.83%, 92.09%, and 87.21%, the classifiers SVM, LRC, RFC, and DT are ranked from 1 to 4 in Table 3.2. It is amply shown that SVC outperforms the classifiers selected for this investigation.

As seen in the preceding table, linear SVC enabled the classification of human activities to have an accuracy of up to 99%. The Linear SVC technique is simple to use because it is based on calculating the distance (often Euclidean distance) between the features retrieved from the new window to be categorized and the data in the training set. It has low memory and processing needs, which allows for an even further decrease in the necessary qualities. We want HAR to be carried out in real time for the vast majority of applications. This necessitates the development of the classifier on a wearable device's microcontroller and the prompt processing of the gathered data.

Since it is flexible, offers a superior solution, and unifies different regions, the linear SVCr algorithm is a great tool for handling challenging problems. In the formulation, many regularizations (L1, L2) might be utilized. The convex optimization problem known as SVC has workable solutions (a global optimum can be found). With its near-perfect ability to fit linearly separable datasets, linear regression is commonly used to determine the nature of the relationship between variables. The linear SVM tends to converge more quickly the more samples there are. This is so that Liblinear, which is designed for specific cases like the linear kernel, can perform better than Libsvm. Because of this, the results are excellent and the best for these aspects and attributes when compared to other algorithms.

3.6 Conclusion

In this study, we presented a comprehensive framework for Kafka based real time human activity recognition structures and implemented several new machine learning algorithms to analyze popular and essential human activity. To verify the effectiveness of the classification algorithms for body activity recognition, we carried out a comparative analysis of the various methods used by the Logistic Regression classifier, the Decision Tree Classifier, and Random Forrest Classifier on actual wearable sensor statistics from the database. The Linear SVC approach produces acceptable and close to current results for challenging classification problems while being significantly faster than any other algorithm.

Chapter 4

Contribution 02: Bluetooth Beacons Based Indoor Positioning

4.1 Introduction

Using the Global Positioning System (GPS) to find mobile devices outside has become common practice. However, it might be difficult to measure position indoors as opposed to outdoors. Due to the attenuation of satellite communication signals through walls, roofs, and other obstructions, it is challenging to receive GPS signals. Using wireless technologies like Bluetooth, WiFi, or ultra wideband (UWB) communication is one of the most widely used techniques for identifying an indoor position. [110]

WiFi and Bluetooth technologies are so pervasive that on so many devices, locations can be determined without the need for new infrastructure. Fingerprinting technology using Received Signal Strength Indicator (RSSI) has been used in several WiFi and Bluetooth studies. This method calculates the position by aggregating RSSI from multiple room transmitters[112]. Due to the multipath, noise, and reflection caused by external condi- tions, even at a fixed position in a wireless connection, the range of RSSI is inconsistent. This is one of the main problems with indoor positioning using wireless communication[113]. An approach involving machine learning has been proposed as a solution to this prob- lem. However, the accuracy may be affected when multiple factors are involved as it only depends on RSSI[2].

GPS signals cannot pinpoint locations indoors, and since walls greatly reduce signal strength, they are unable to pinpoint across tall buildings and inside structures. How does an indoor location system pinpoint a user's location accurately and permanently and quickly?, and what are the modern technologies that work on internal positioning?

4.2 Proposed Architecture

By combining data from smartphone sensors and Bluetooth RSSI values, our positioning method introduces a system for quickly and accurately locating locations inside buildings. Our technology employs a model derived from machine learning methods to determine the position. We use the RSSI values of Bluetooth beacons dispersed throughout the study area, measurements from the mobile accelerometer, and the user's previous estimated location as input parameters in our model. [2]

The suggested indoor positioning model's primary objective is to gather data on the variables involved in indoor locations, such as distance and position, and build the model with this data. The suggested model for indoor positioning based on Bluetooth signals in shopping malls is depicted in Figure 4.1 below and uses several machine learning methods.

There are three primary components to the proposed indoor positioning system. A data gathering module first gathers the beacon RSSI values that mobile devices have received. After the data has been corrected by the filter algorithm, a data processing module applies a positioning approach to the filtered findings. The third essential component, the data management module, stores and manages the data processing outcomes in a database.

Our system architecture contains multiple models that work together to enable the ability to detect intrusions in big data environments. These modules are:

- Data Collection Module: The data collection module gathers a range of data from the user's mobile device. Utilizing Bluetooth connectivity, a mobile application on the user's device collects information from a beacon and sends it to the data-collection module of the indoor positioning system. During communication with the mobile device, the mobile application sends the message authentication code (MAC) value of the device and the beacon's identifier to the data acquisition module.

- Data Processing Module: The data processing module corrects the RSSI data collected from the data acquisition module using a filtering technique to reduce the error range. Using the RSSI data of the beacons that were collected by correlating with those devices in the collection module, this module calculates the separation between each beacon and the mobile device. The position of the user within the interior space is then calculated using those distances.

- Data Management Module: The data management module locates a user's mobile device and manages various data required by the indoor positioning system after the data processing module has examined the information gathered by the acquisition module.

Figure 4.1 shows the general structure of the system. The system has been proposed for the display of works of indoor positioning in shopping malls.

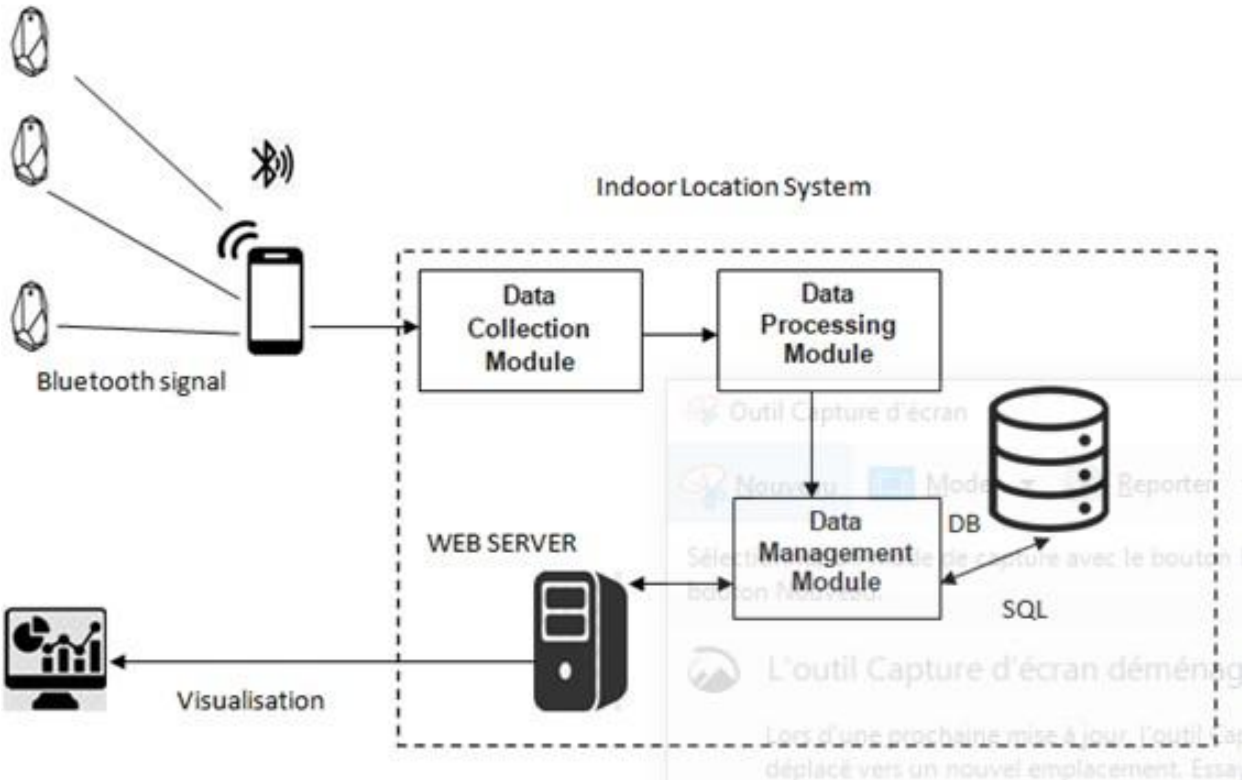

Figure 4.1: Architecture of the proposed indoor positioning system

4.3 Method

In our contribution, we use various machine learning methods to provide Bluetooth beacon based indoor localization. We analyzed and compared the performances of six individual predictors using machine learning algorithms for indoor localization features to characterize places on Earth. We have validated the performance of the system using Bluetooth beacons on smartphones. To find clients' devices in shopping centers, we deployed machine learning techniques. More than 90% of the time, the Extra Trees Classifier and KNeighborsClassifier correctly located the device. Other algorithms had less success pinpointing the site. The outcomes also demonstrated that Bluetooth technology is a reliable method for collecting the information needed to study people's spatiotemporal activity.

To obtain RSSI values and location coordinates (x, y) from stationary signals, the method for indoor positioning was investigated and put into practice. Using performance measures, the chosen models were evaluated and compared for their effectiveness. In this study, we use Bluetooth signal localization as a classification problem to address the smartphone indoor positioning issue. The goal is to localize people and help them navigate an intelligent environment using machine learning algorithms. This can help citizens work more effectively and save time.

Figure 4.2 details the overall structure and workflow of the proposed classification system. Most of the time, one chooses a subset of the relevant population whose values for the target attribute are known or, if necessary, creates the data. Concurrently with this procedure, a machine learning algorithm that will work best for the anticipated target number is chosen. Creating, identifying, and cleaning the data to ensure its accuracy, consistency, etc. take up the majority of the labor. The second step is to choose an algorithm and figure out how to map the system's properties—the model's input—into a suitable format.

Wireless signals are sent by iBeacon devices via communication with smart devices. A store is represented by the individual identifier on each iBeacon device. It automatically saves consumer information, such as distinctive identity and time information. This data allows for the tracking of a customer's movement through a mall. Based on the distance and RSSI values, which differ, the results are studied and contrasted to identify the algorithm that performs the best and is the most accurate. With the aid of Bluetooth and machine learning, we suggest a positioning algorithm for an indoor positioning system. We assessed model performance using measures of accuracy, precision, and recall in machine learning and compared model performance in terms of the mean error.[114]

To do this, the raw data must be transformed into particular attributes that will be used as algorithmic inputs. After this process is finished, the model is trained by maximizing performance, which is frequently determined using some kind of cost function. Typically, this entails altering the hyperparameters that control the model's internal structure, training process, and properties. The data are separated into many sets. A validation dataset that is different from the test and training sets should be used to optimize the hyperparameters. [115]

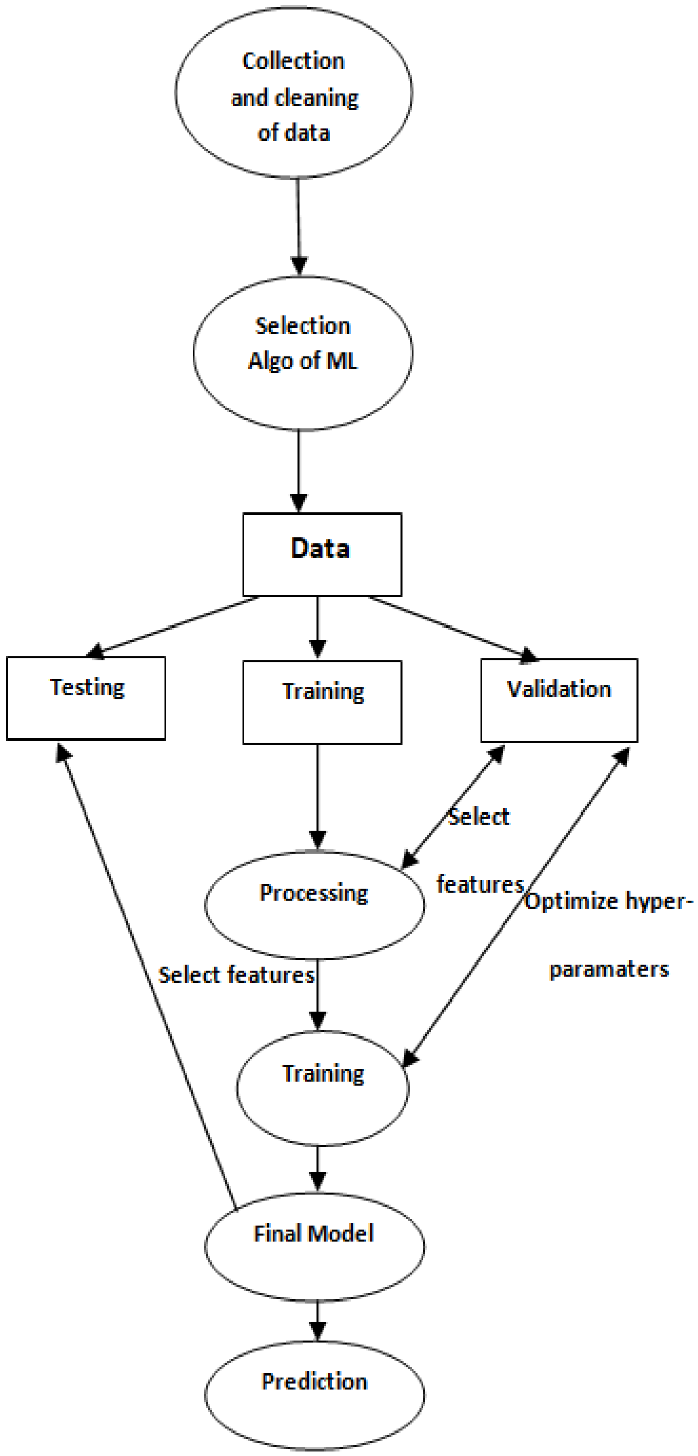

Figure 4.2: Workflow of the proposed system

We have selected some algorithms: KneighborsClassifier (KNC), RandomForestClassifier, ExtraTreesClassifier, Support Vector Machine Classifier, Gradient Boosting Classifier, and Decision Tree algorithms. They are used with continuous variables that are supposed to be feature independent, which is why this is the case. One or more output values can be predicted by varying input attributes. Low bias/high variance techniques like KNC, decision trees, or SVMC were also used due to the size of the data set we will train on and the presence of notes over features.

We used Bluetooth beacons in our proposed method to use smartphones to forecast the indoor position. The first step was to collect data from various Access Points (AP) on the building's floor. In our study, we use the area of 2.74 m (width) and 4.38 m (length), and 3 Kontakt Bluetooth beacons are mounted. The transmit power of the three beacons is -12 dBm. The data is recorded using a Sony Xperia E3 smartphone with Bluetooth enabled as the receiver. Recordings are done in several positions in the building an intervals of 30-60 seconds in the same position. The test PC was an Intel i3-9100 with 16 GB of RAM and an LGA 1151 card. We trained six models based on KNN, RFC, ETC, SVM, GBC, and DT. Each of the six ML models received testing data sets for prediction.

4.4 Results and Discussions

In this section, we will focus on the implementation phase. We will discuss the practical aspect of our system; this is to explain the hardware environment in which our system was developed, the programming languages, and the tools and technologies used. We describe the experiment and evaluation data that we employed, along with some alternative strategies. It displays the configuration used to put the model into practice as well as compare results.

4.4.1 Datasets used

After data gathering, a dataset is produced that includes various signal levels at various places. The data has been transformed using the CSV format. Initially, CSV files featured features like (Distance A, Distance B, Distance C, Position X, Position Y, Date, and Time).

Data preprocessing in machine learning is the process of cleaning and organizing raw data to make it acceptable for building and training machine learning models. The running average is used to eliminate noise from the dataset, and if there is a gap in the

RSSI values, the filter fills it with the lowest RSSI value it can find. Additionally, the data set is checked for null values, which are then replaced with either the data row's highest or lowest value. An example of the dataset is shown in Table 4.1.

Table 4.1: Example of the dataset.

Distance A	**Distance B**	**Distance C**	**Position X**	**Position Y**	**Date**	**Time**
0.877462	0.768608	1.457214	122	180	Feb 09 2017	12:20:22
1.201608	1.03122821	1.893498	122	180	Feb 09 2017	12:20:23
1.614344	1.098873	2.112560	122	180	Feb 09 2017	12:20:24
1.513634	1.135640	2.296293	122	180	Feb 09 2017	12:20:28
1.517499	1.148356	2.388172	122	180	Feb 09 2017	12:20:29
1.489097	1.163176	2.498214	122	180	Feb 09 2017	12:20:31
1.533095	1.174496	2.605271	122	180	Feb 09 2017	12:20:32

We describe the different fields of the Table 4.1 in the following:

- The RSSI of this Bluetooth beacon is used to calculate the separation in meters between beacon A and the apparatus (Smartphone), so that we call it Distance A.
- The RSSI of this Bluetooth beacon is used to calculate the separation in meters between beacon B and the apparatus, so that we call it Distance B.
- The RSSI of this Bluetooth beacon is used to calculate the separation in meters between beacon C and the apparatus, so that we call it Distance C.
- Position X is the X coordinates in centimeters.
- Position Y is the Y coordinates in centimeters.

4.4.2 Obtained results

To find customers' devices in shopping centers, we used machine learning algorithms. More than 90% of the time, the Extra Trees Classifier and K-Neighbors Classifier correctly located the device. Other algorithms had less success pinpointing the location.

When regular graphs are stretched to higher dimensional data sets, even graphs are generated. This is useful for investigating correlations between multidimensional data sets. We have extracted data from a high dimensional space by projecting it onto a low-dimensional subspace, which is a linear dimensionality reduction technique (PCA). You can use this to shorten the training and testing periods for machine learning algorithms because the data has a lot of features and learning takes a while. Scikit primary Learn's component analysis is shown in Figure 4.3.

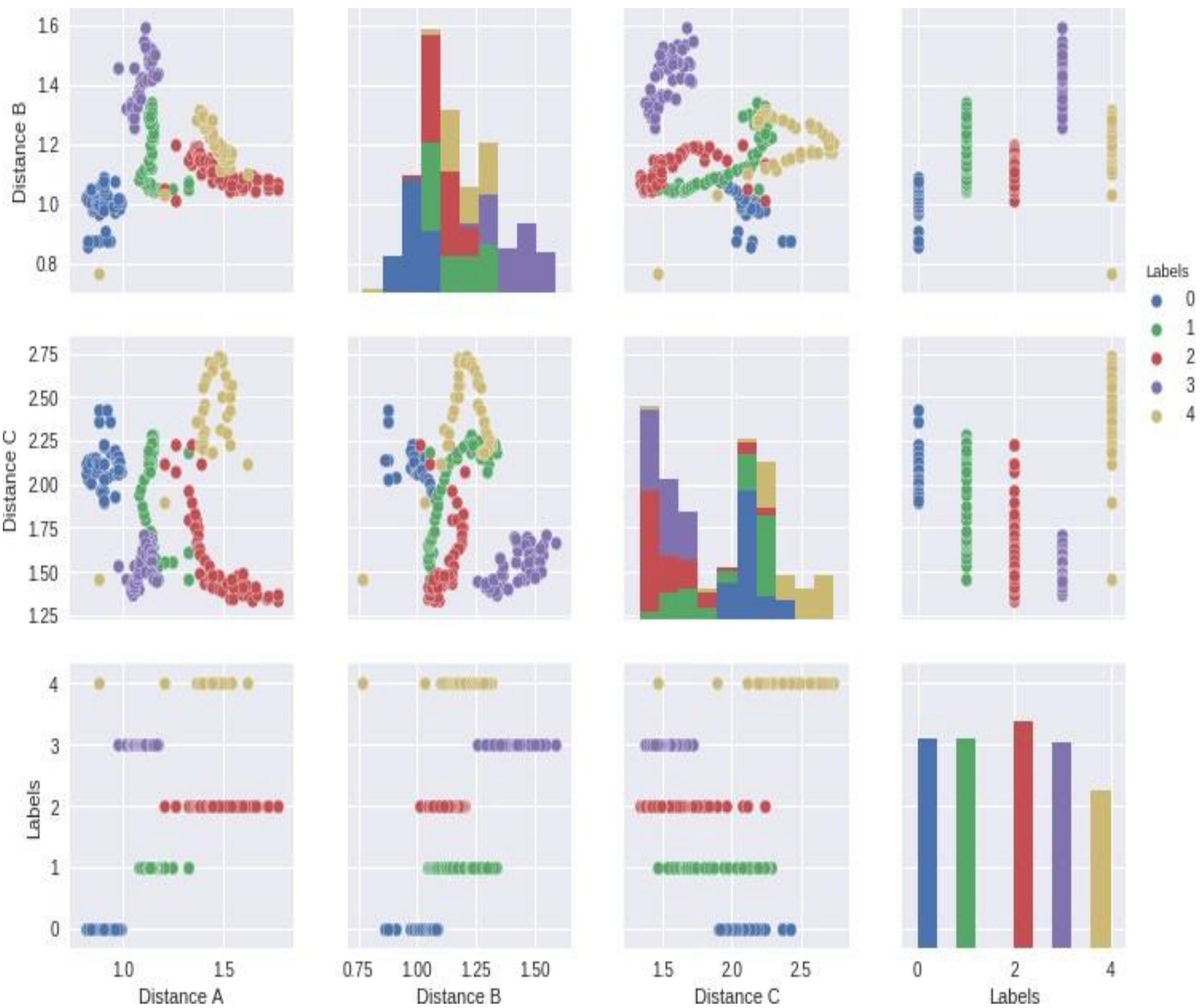

Figure 4.3: Principal component analysis using scikit learn

Then, a training set made up of 80% of the dataset and a testing set made up of 20% were produced. Although there are other ratios to divide a dataset. In the experiment,

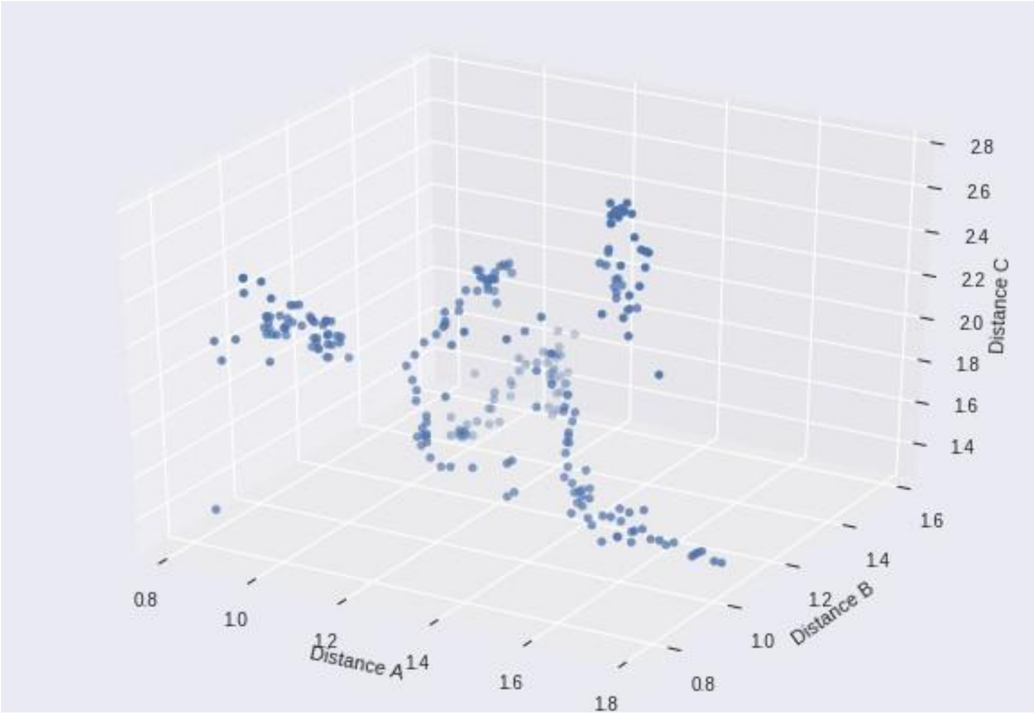

Figure 4.4: 3D plotting of Data visualization

We use the fairest splitting recommended for a small dataset. A 3D data visualization is shown in Figure 4.4. It was made using a three dimensional chart to represent the data in a better and more accurate way.

4.4.3 Discussions

We used ML metrics for accuracy, precision, and recall to assess model performance. Six methods were utilized in this experiment to improve location prediction according to the following criteria:

- Accuracy: The easiest performance metric to understand is accuracy, which is just the proportion of properly predicted observations to all observations. When we have similar data sets with values for false positives and false negatives the PCA is nearly equal, and accuracy is a fantastic metric. As a result, to assess the effectiveness of the suggested model, we must include other factors as: [122]

- Precision: In terms of positive observations, precision is the proportion of accurately anticipated observations to all predicted positive observations. A low false positive rate is correlated with high accuracy. How many actual correct locations exist? [123]

- Recall: Recall is the percentage of accurately predicted positive observations to all of the actual class's observations. [124]

Table 4.2 shows a model performance comparison and each model's detection rate based on the testing dataset:

Table 4.2: Performance comparison of Models

Model Name	**Accuracy**	**Precision**	**Recall**
K-Neighbors Classifier	92.24	91.31	91.20%
Random Forest Classifier	88.66	88.02	87.90%
Extra Trees Classifier	95.63	94.85	94.80%
Support Vector Machine	82.63	81.55	81.55%
Gradient Boosting Classifier	89.96	89.60	89.51%
Decision Tree	87.96	87.91	87.88%

With a detection rate of 95.63% through the histogramme in Figure 4.5, Extra Trees Classifier (ETC) is a powerful method that can locate the device. In most retail malls, the Extra Trees Classifier algorithm produced predictions that were more accurate than those of other algorithms.

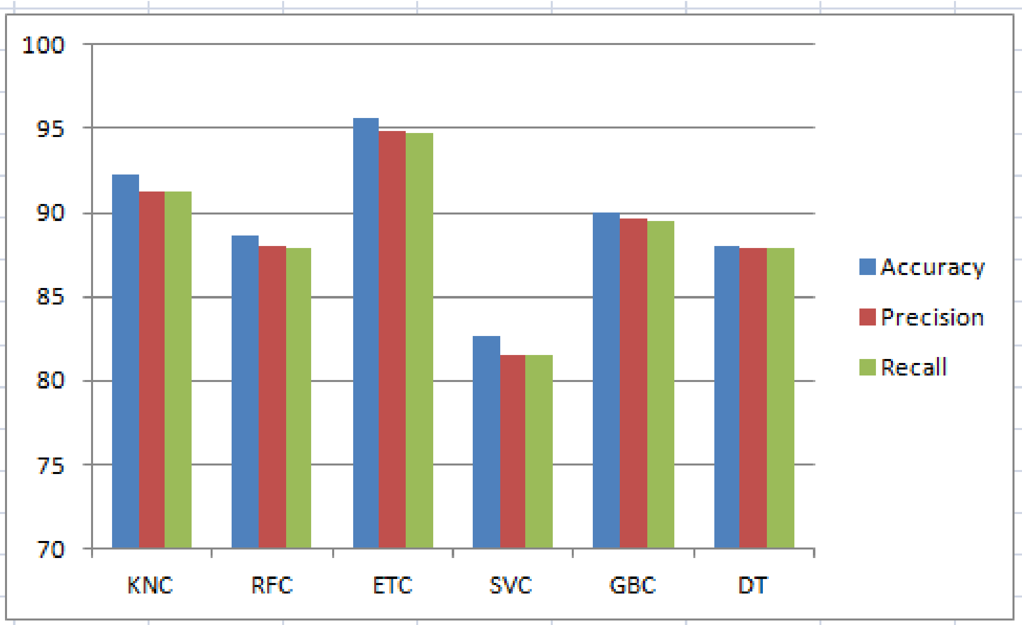

Figure 4.5: Prediction performance when using different Models

Based on the experiments carried out and the obtained results, we note that the six algorithms in particular were capable of correctly identifying the majority of the locations. Extra Trees Classifier can also be used to locate sampled microsites. The models for data preparation were successfully trained.

Because it incorporates more data dependent on the distance from the internet connection stage, ETC performs better than KNN, GBC, RFC, and DT. Because it requires more test samples for precise localization, SVM performs poorly. One of the worse classification results was obtained using the SVM classifier because of its superior performance with short training sample sets. The comparison of experiments discovered that using both the detection rate and the classification matrix can predict the location of the device with the highest degree of accuracy. This is clearly illustrated by Figure 4.5 and Table 4.2.

The outstanding performance of the ETC algorithm results from a variety of reasons. However, the primary contributor to Bluetooth Beacons' performance boost is the fact that they do not heavily rely on any one particular set of capabilities. When a lot of features are used in the data collection process, great outcomes are produced. When the methods under consideration can be improved to use fewer resources, excellent results are obtained. The experimental results demonstrated the need for a sophisticated path loss model for RSSI based distance estimation due to the high computational demands of machine learning algorithms.

Generally, technical components can be removed to greatly improve proximity accuracy localization convergence error range up to a few distances from the receiver, the experimental results significantly outperformed the homogeneous results. BLE signals are a promising solution because they are affordable, simple to set up, and demand little power.

4.5 Conclusion

To define locations on Earth, we examined and contrasted the performances of six distinct predictors utilizing machine learning techniques for indoor localization features. We have used Bluetooth beacons on smartphones to verify the system's performance. We discover that the majority of the places could be correctly identified by these methods. Based on the outcomes of the experiments, the Extra Trees Classifier model based on Bluetooth beacons provided us with the most precise locations. But there is still room for improvement in machine learning algorithms due to their complex calculations.

Chapter 5

Contribution 03: Predicting Air Quality in Smart Cities

5.1 Introduction

One of the biggest threats to human health is air pollution. The World Health Organization estimates that 7 million individuals are at risk for health problems as a result of air pollution. It is a major risk factor for many illnesses, including bronchitis, lung cancer, disorders of the respiratory system, heart problems, throat and eye conditions, infections of the skin, heart problems, and infections of the blood vessels. In addition to the health issues caused by air pollution, our planet is also seriously endangered. The greenhouse effect is primarily caused by pollution emissions from sources like industry and vehicle emissions, with CO_2 emissions ranking among the major contributors. [125]

Air quality is a crucial factor in the creation of smart cities designs. On the other hand, poor air quality is a significant problem in many places. The high rate of population growth resulted in increased use of energy, industry, transportation, and vehicles. Numerous studies have found that air quality is a significant challenge for smart cities and that, despite its underuse in many global cities, machine learning is a better and more effective way to predict air quality. In recent years there have been some machine learning methods proposed for solving air pollution prediction problems[126].

5.2 Problematic

These days, the data that need to be examined are not only large but also in- clude streaming data and other different data kinds. Big data may alter statistical and data analysis methods due to its distinctive characteristics of being "massive, high dimensional, heterogeneous, complicated, unstructured, incomplete, noisy, and erroneous. Many of the available analytics solutions today are capable of addressing ongoing big data analytics difficulties and delivering good outcomes for an affordable price. With the recent rise in smart, wearable, and other measuring devices for ambient applications, we are just beginning to address all the facets of this new big data. [127]

Because it exposes people to environmental toxins, air pollution affects their health. Rapid regulation of air pollution is necessary to prevent disaster[131]. Quick environmental quality forecasting is required in smart cities to implement preventative interventions. For smart cities, the ability to retrieve data from sensors is a big advantage in resolving the problems and barriers preventing their development. This is why air pollution is a major challenge for the public and governments around the world. Deployment of the Internet of Things based sensors has considerably changed the dynamics of predicting air quality. [3]

5.3 Proposed Architecture

Architecture layers for predicting Air Quality in Smart Cities have been proposed as shown in Figure 5.1. These layers are:

- Data Collection: This layer compiles information from various heterogeneous devices connected to smart cities. Sensors placed across the cities calculate the various levels of air quality. Collection and aggregation occur here because a lot of data is gathered from various sources. Data can be in a variety of formats, so this is where all the pre-processing and initial filtering happens. At this layer, pre-processing is done, and extra information is found and eliminated.

- Communication: This layer is in charge of moving every piece of data from the data gathering layer to the following layers. This layer is made up of various technologies. Here, all information is transferred from IoT devices to the data processing layer. The use of this layer is also possible for real time processing-capable gateways. Here, initial data processing and in the moment decisions can be handled.

- Data Management:This layer is charged to store and process data.For real time processing, tools like Spark, VoltDb, Storm, etc., can be employed. This layer can manage and store vast amounts of data in the HDFS system. For querying and analyzing historical data, various systems can be employed. At this layer, in-memory and offline data analysis are both performed. Additionally, it can be used to learn using various machine learning techniques. In this layer, predictions and pattern recognition are also performed.

- Applications: This layer serves as the interface for all information with full significance. Events generated here are sent to the real time devices because this layer is connected to them. This layer is used to display reports and data as charts and dashboards. Governmental organizations in charge of air quality monitoring are the layer’s final users. Then, crucial decisions are made using this data. All the information relating to air quality can be announced by this layer. People interact with it to track statistics on the air quality and make decisions.

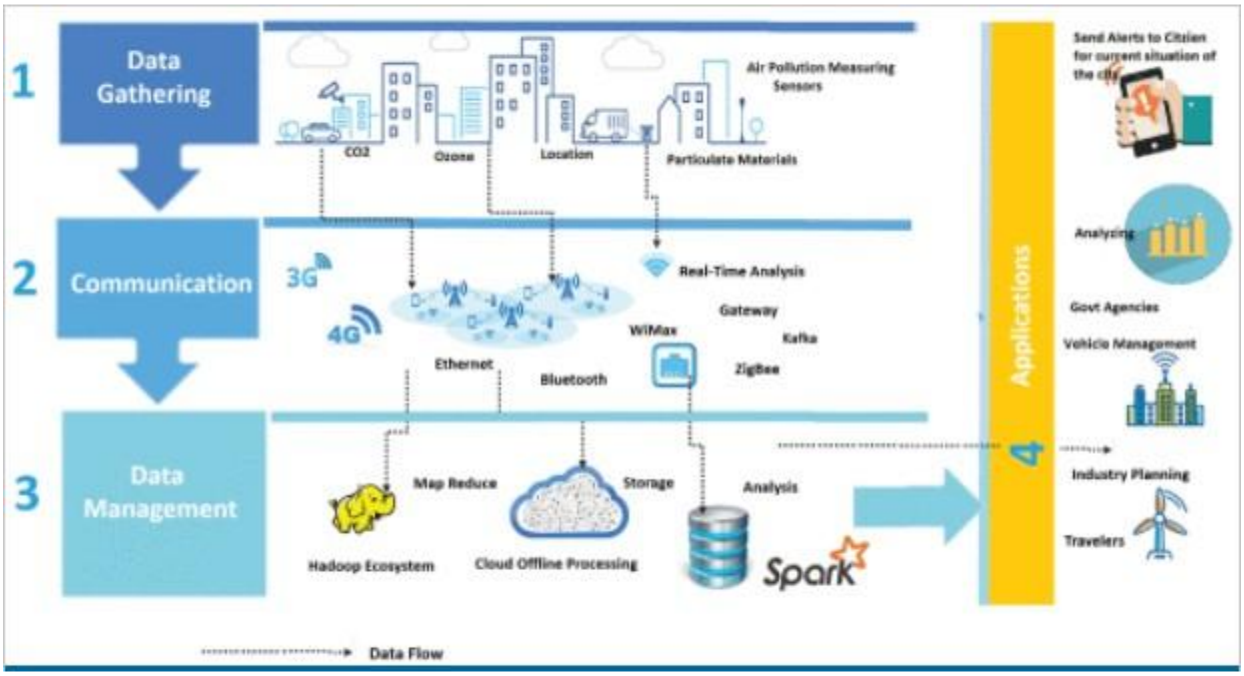

Figure 5.1: Smart cities air Quality monitoring architecture

5.4 Method

In our Contribution, to predict the air quality index, we employ machine learning techniques such as linear regression, decision trees, support vector machines, random forests (RF), and gradient boosting. Based on the data, we concluded that the RF method offers a better prediction of the air quality index. RF separately trains every tree with a random sampling of the data. This randomization makes the model less likely to change based on training data and more robust than a single decision tree. In RF, the number of trees and the number of features to be chosen at each node are typically the two parameters. The proposed model architecture is shown in Figure 5.2, in which sensors are used to gather data on air pollution, which is then preserved as a dataset. Preprocessing has already been done on this data collection. After that, the data set is divided into a training data set and a test data set. More supervised machine learning techniques are applied to the training dataset. The collected data is then evaluated by comparison with the test data set.

The workflow for the suggested architecture can be defined in some steps. To perform the analysis and data preparation process, the first step is to gather the data set from various sources, such as sensors. The preparation of the data is the next phase, one of the most crucial ones for improving the accuracy of machine learning models. Data cleaning, also known as data pre-processing, is the process of transforming raw data into a clean data set. Data transformation, filling in missing values, discarding missing values, machine learning, and outlier detection are a few preprocessing methods that can be used.

We separate the data into three categories for training the models: "Training Data," "Validation Data", and "Test Data". The "Training Data Set" is used to train the classifier, the "Validation Set" is used to validate parameters, and the "Test Data" is used to evaluate the classifier's performance. A training set is used in a data set to build a model, and a test (or validation) set is used to verify the model. The test (validation) set does not include any of the training set's data points. [132]

The model is then examined to determine which one best captures our data and will be successful going forward. We can change the hyper parameters of the model, to try to increase accuracy while also taking the confusion matrix into account to try to increase the proportion of true positives and true negatives. Increasing the number of training sessions after the model has been evaluated can produce more accurate results.[133]

After completing each of these steps, we use predictive analytics to determine and forecast airquality in smart cities. [142]

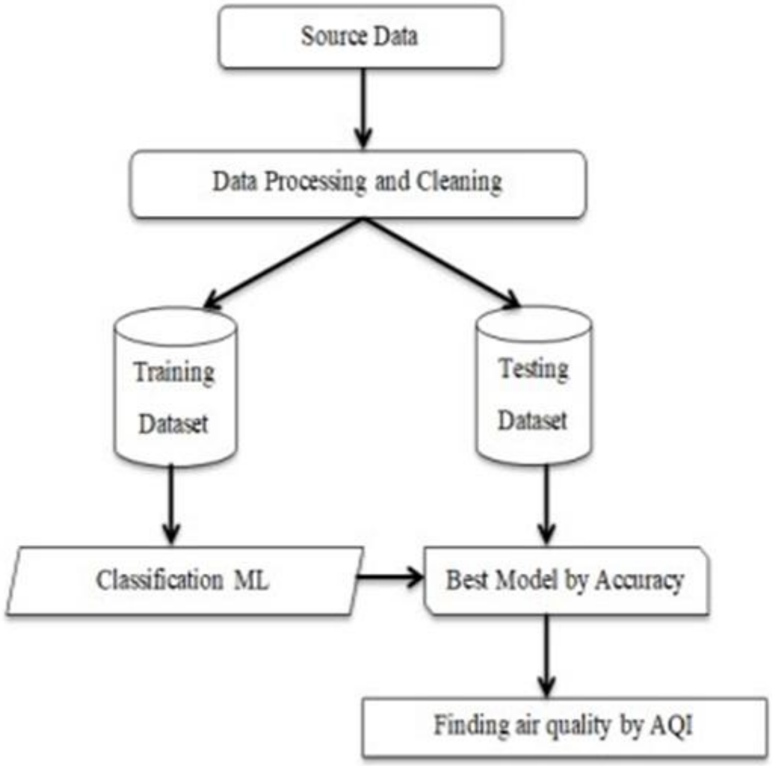

Figure 5.2: Smart cities air Quality monitoring model

The phenomena of air quality are influenced by various elements. Correct identification of these factors impacting air pollution is required for an exact prediction. Before beginning a machine learning analysis of the data, we more thoroughly assess the high level characteristics of the data. As can be shown, PM2.5 concentrations are higher on weekdays. Therefore, it is evident that PM2.5 is affected by human activity and is particularly related to traffic levels. [3]

To determine the exceptional performing set of rules, the test results were analyzed and compared. The test computer had an Intel i3-9100 processor, 16GB of RAM, and an LGA 1151 card. We primarily used decision trees, linear regression, random forests, SVM, and GBC to train 5 models. After training the five ML models, we gave them a set of test facts for making predictions. In the experiment, we use a Jupyter notebook for implementing machine learning algorithms in Python.

5.5 Results and Discussions

5.5.1 Datasets used

The pollutant data was used in this study in order to train the algorithm to detect air quality. This data was gathered from a data set of 74 cities in China obtained each day from Oct. 2013 to April.2015. The parameters of the meteorological data which used to train the system are temperature, wind speed, humidity, and wind direction. [3]

5.5.2 Obtained results

We used machine learning algorithms to assess model performance using measures of accuracy, precision, and recall. In this experiment, six algorithms were applied to improve location prediction.

a. CO prediction

The Random Forest method is the best among the other classifiers, as presented in Table 5.1. It benefits from being the easiest of the five machine learning algorithms evaluated to understand and apply. We achieved a forecast accuracy of more than 90% for the CO (Carbon monoxide). We get prediction accuracy as presented in Table 5.1.

Table 5.1: Performance comparison of Model prediction probability of CO

Model Name	**Accuracy%**	**Precision%**	**Recall%**
Linear Regression	6.00	5.89	5.77
Decision Tree	69.00	68.96	68.50
Random forest	91.00	90.84	90.12
Support Vector Machine	80.00	79.55	79.50
Gradient Boosting	36.00	35.41	35.30

Figure 5.3 illustrates the degree of difference across the various employed machine learning techniques and helps to further explain the results.

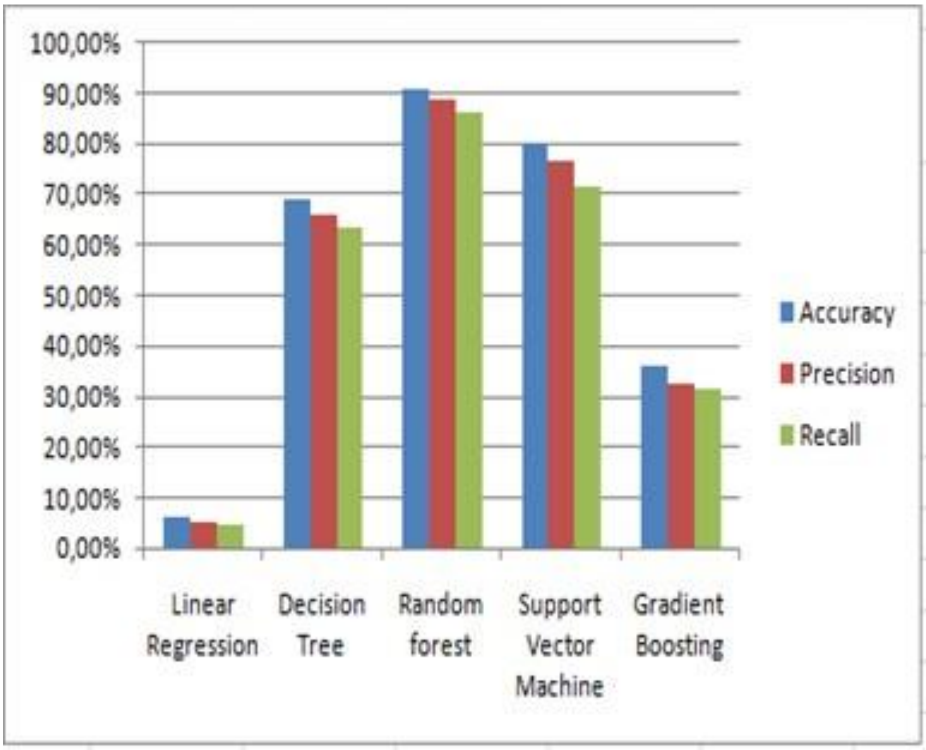

Figure 5.3: CO prediction probability

In this study, we use the Chi-Square test which is a statistical procedure for determining the difference between observed and expected data. Let us calculate the chi-square data points by using the following formula:

$Chi - squarepoints = (observed - expected)^2/expected$

We have calculated Chi-Square Test by the function "chi2-contingency" from the scipy package which returns the values of these parameters:

p, dof, expected = chi2-contingency(table dataset)

p: The p-value of the test.

dof: Degrees of freedom.

expected: The expected frequencies, based on the marginal sums of the table.

After the calculation process, we found that the value of the chi-square is: p-value = 0.00134877.

The p-value is = 0.00134877 or 0.13%, which means that we reject the initial hypothesis, which contains the possibility that there is independence between the machine learning algorithms. We conclude that there is a strong relationship between the machine learning algorithms chosen for the study. Also, the quality of prediction often depends on the accuracy of the machine learning algorithms.

b. SO2 prediction

The Random Forest method is the best among the other classifiers, as presented in Table 5.2. These findings shed light on the relationship between SO2 (sulfur dioxide) concentrations and climatic variables. Additionally, we saw that the Random forest algorithm consistently provided better forecasts, which was encouraging. We obtained a prediction accuracy of greater than 96% for the SO2 prediction. Table 5.2 present the accuracy.

Table 5.2: Performance comparison of Model prediction probability of SO2

Model Name	Accuracy%	Precision%	Recall%
Linear Regression	3.00	2.34	1.87
Decision Tree	79.00	73.54	70.01
Random forest	96.00	90.84	87.15
Support Vector Machine	87.00	81.55	79.40
Gradient Boosting	54.00	50.49	49.0

Figure5.4 illustrates the results of the different machine learning techniques used in this study.

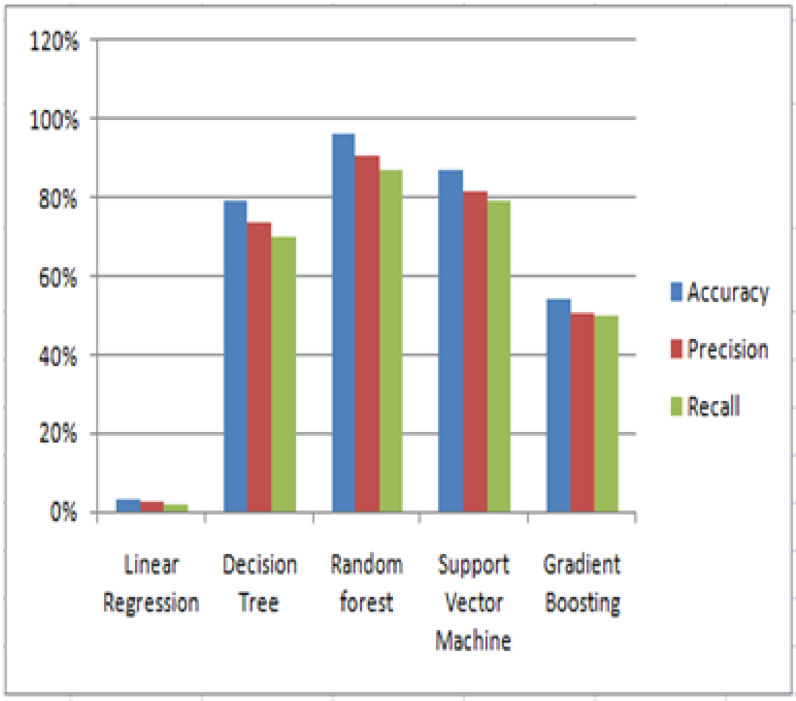

Figure 5.4: SO2 prediction probability

c. O3 prediction

The Random Forest method is the best among the other methods of machine learning. Table 5.3 presents this result.

The Random Forest algorithm is less prone to overfitting than Decision Tree and other algorithms. Random Forest algorithm also takes out the importance of features which is very useful. The Random Forest algorithm also devalues features, which is quite helpful. We achieved a prediction accuracy of more than 90% for the O3 (Ozone) forecast. Table 5.3 presents the performance and the accuracy of each model.

Table 5.3: Performance comparison of Model prediction probability of O3

Model Name	Accuracy%	Precision%	Recall%
Linear Regression	21.00	19.35	18.21
Decision Tree	55.00	53.90	52.86
Random forest	90.06	89.87	88.15
Support Vector Machine	63.00	63.55	60.40
Gradient Boosting	42.00	39.11	38.94

Figure 5.5 illustrates the degree of difference across the various employed machine learning techniques and helps to further explain the results.

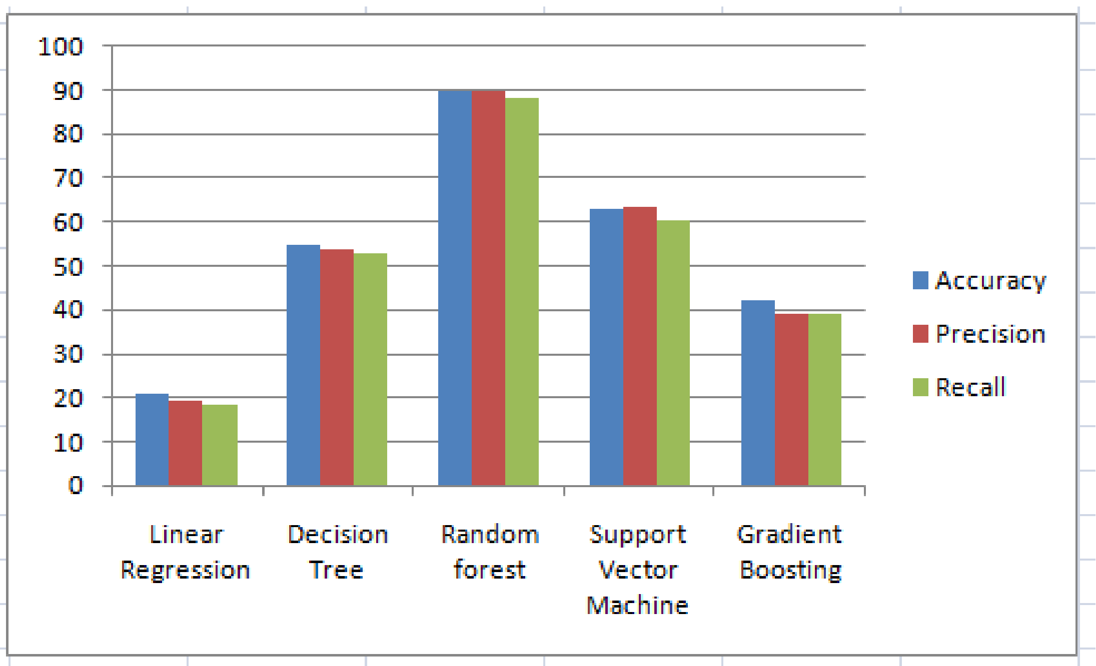

Figure 5.5: O3 prediction probability

5.5.3 Discussions

We conclude that the linear regression algorithm has much weaker indicators than other algorithms. The cause is the existence of numerous outliers in data. It produced the worst results because it presupposes a linear relationship between dependent and independent variables and because outliers can have a significant impact on the regression's

limits. We obtained a forecast accuracy for the CO and SO2 of less than 10%. Since linear regression is simpler to perform and the relationship between the variables is linear, it gives the worst results. Because linear regression is prone to overfitting when there is no association between the independent variables, it produced the worst results. It achieved a prediction accuracy of more than 20% for the O3 forecast.

The decision tree has an average prediction because it divides the data into columns; however, as the number of columns rises, so does its processing speed. We found that the CO prediction had a prediction accuracy of roughly 70%. It gives acceptable results. It achieved a prediction accuracy of almost 80% for the SO2 prediction. Techniques that use decision trees give intermediate results. We obtained an O3 prediction accuracy of almost 60%.

The Support Vector Machine method yields acceptable results. It is more effective in high dimensional areas and when there is a margin of difference between groups, as there is in our instance. We obtained a prediction accuracy of 80% for the CO. It achieved a forecast accuracy of over 80% for the SO2 prediction. We obtained an O3 prediction accuracy of almost 60%. SVM can be understood as the classification of various groups (labels).

The outcomes of the gradient boosting technique are average because trees are constructed sequentially, and training takes longer. We obtained a prediction accuracy for the CO of less than 40%. It achieved a forecast accuracy of more than 50% for the SO2 prediction. Gradient reinforcement models produce mediocre results because when an anomaly is initially missed by the model, it gives it more weight and preference in subsequent iterations, increasing its capacity to predict low participation categories. We found that less than 50% of the O3 prediction was accurate.

Random Forest algorithm has the advantage of being the easiest to use and doing the best in terms of processing speed and accuracy of results. See in Figure 5.3, Figure 5.4, and Figure 5.5.

5.6 Conclusion

To predict the air quality index, we employ machine learning techniques such as linear regression, decision trees, support vector machines, random forests (RF), and gradient boosting. Based on the results obtained, we concluded that the RF method offers a better prediction of the air quality index. RF separately trains every tree with a random sampling of the data. This randomization makes the model less likely to change based on training data and more robust than a single decision tree. We plan to examine the performance of these strategies in a multi core context in the future. We also plan to look at the other elements that contribute to air pollution.

General conclusion

Actually, Big Data is a key to developing intelligent cities using information technology. Indeed, standard technologies are insufficient beyond a few tens of terabytes; they are no longer able to assess the vast and divergent volumetrics of data.

Finding, using, and sharing these enormous types of data has become incredibly difficult in recent years due to the increasing expansion of data across all industries. This exponential growth is brought on by the expansion of several businesses and their operations, particularly the emergence and spectacular development of social networks (Facebook, Twitter, Instagram, YouTube, etc.).

Big Data platforms today are supported by a variety of processing, analytical, and dynamic visualization capabilities. These platforms enable the extraction of value and knowledge from a dynamic and complicated environment. They also support higher cognitive processes through recommendations and automatic detection of anomalies, abnormal behavior, or new trends.

We have carefully examined Big Data characteristics and issues brought on by Big Data computing platforms. Also, we have discussed the value of big data processing in several fields. In addition, we have concentrated on the parts and technologies used in each layer of big data platforms. The capacities, benefits, and limitations of various technologies and distributions are also contrasted. The big data analytics systems that supported the features and services offered to end users were likewise categorized.

For the first contribution, we introduced a thorough framework for Kafka based real time human activity detection behaviors in our first proposal under consideration, and we put several cutting edge machine learning algorithms into practice to examine common and significant human activity. A processing engine and a cross platform engine are needed to implement the system in Hadoop. But for this project, we went with Python as the most cutting edge programming language, using libraries for data analysis and machine learning techniques.

In the second contribution, we provide Bluetooth beacon based indoor localization, we employ a variety of machine learning techniques. To define locations of devices on Earth, we examined and contrasted the performances of six distinct predictors using machine learning techniques for indoor localization features. We used Bluetooth beacons on smartphones to verify the system's performance.Also, we implement six machine learning algorithms to locate clients' devices in shopping malls.

For the third contribution, we examined and contrasted six existing plans to address the problem of air pollution prediction. Thus, we have contrasted the methods in terms of processing speed and mistake rate. We use machine learning methods like gradient boosting, random forests (RF), decision trees, support vector machines, and linear regression. We concluded that the RF approach provides a better prediction for the air quality index.

In this thesis, we contracted an important field of research which is smart cities. In addition, we develop a real and relevant architecture that can solve the issues which can be met to develop smart cities. Users and policymakers should use big data technologies to update the intelligent systems which manage the cities. As a result, we require more advanced computer architectures, high tech storage, and data technologies like cloud computing that use more machine learning methods.

As future works, we can cite :

- Addressing the identification of more complicated activity numerous human activities, such as eating, studying, and watching television, don't provide notably unique acceleration effects. For these, alternative sensors built within modern mobile phones can be used to study techniques of ambient light and sound, and subsequent activities and system implementation can be thought of in real time on a smartphone.

- Locate people with exceptional accuracy indoors, we will enhance the data collecting validation process and combine this effort with a tracking system. As a result, the algorithm produces better results for monitoring everything required indoors. To improve accuracy, we also intend to identify access points and extend the experiment's scope to additional floors.

- Additionally, we seek to develop our algorithms for predicting air quality in smart cities by looking into the variables that affect air pollution. Because machine learning techniques are included in this work, more precise outcomes can be obtained with a collaborative strategy and extra implementation.

Bibliography

[1] Kamel Maaloul, Brahim Lejdel, and Nedioui Med Abdelhamid. Real-time human activity recognition from smart phone using linear support vector machines. *TELKOMNIKA Telecommunication Computing Electronics and Control*, 2023,in process.

[2] Kamel Maaloul, Brahim Lejdel, Eliseo Clementini, and Nedioui Med Abdelhamid. Bluetooth beacons based indoor positioning in a shopping malls using machine learning. *Bulletin of Electrical Engineering and Informatics*, 12(2):911–921, 2023.

[3] Kamel Maaloul and Brahim Lejdel. Comparative analysis of machine learning for predicting air quality in smart cities. *WSEAS Transactions on Computers*, 20:248-256,2022.

[4] Mark A Beyer and Douglas Laney. The importance of 'big data': a definition. *Stamford, CT: Gartner*, 2014-2018,2012.

[5] Xin Wei Sha and Gabriele Carotti-Sha. Big data. *Ai Society,Brighton University,Springer,1-4*, 2016.

[6] Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, Elsevier,35(2):137-144,2015.

[7] Saumya Gupta et al. Privacy protection in bigdata: A survey. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(2):562-568,2021.

[8] Pijush Kanti Dutta Pramanik, Saurabh Pal, and Prasenjit Choudhury. Green and sustainable high-performance computing with smartphone crowd computing. *Scalable Computing: Practice and Experience*, 20(2):259-284,2019.

[9] Abou el ela Abdou Hussien. How many old and new big data v's characteristics, processing technology, and applications (bd1). *International Journal of Application or Innovation in Engineering Management*, 9(9):15-27,2020.

[10] Royston Meriton, Rajinder Bhandal, Gary Graham, and Anthony Brown. An examination of the generative mechanisms of value in big data-enabled supply chain management research. *International Journal of Production Research*, Taylor Francis,59(23):7283-7310,2021.

[11] Manju Dahiya, Shikha Sharma, and Simon Grima. Big data analytics application in the indian insurance sector. In *Big Data Analytics in the Insurance Market*. Emerald Publishing Limited,145–164,2022.

[12] Kiran Ravi. Big data characteristics: Know the 5'vs of big data. https://www.edureka.co/blog/big-data-characteristics/, May 2022.

[13] Feng Zhu, Lijie Xu, Gang Ma, Shuping Ji, Jie Wang, Gang Wang, Hongyi Zhang, Kun Wan, Mingming Wang, Xingchao Zhang, et al. An empirical study on quality issues of ebay's big data sql analytics platform. Software Engineering in Practice,33-42,2022.

[14] Jun Zhang and Shah Nazir. Evaluating a's of big data for transformation of smart cities. *Scientific Programming,Hindawi*, 169-184,2022.

[15] Andreas Lund, Zain Alabedin Haj Hammadeh, Patrick Kenny, Vishav Vishav, Andrii Kovalov, Hannes Watolla, Andreas Gerndt, and Daniel Lüdtke. Scosa system software: the reliable and scalable middleware for a heterogeneous and distributed on-board computer architecture. *CEAS Space Journal*, Springer,14(1):2=161-171,2022.

[16] Joey Li, Munur Sacit Herdem, Jatin Nathwani, and John Z Wen. Methods and applications for artificial intelligence, big data, internet-of-things, and blockchain in smart energy management. *Energy and AI,Elsevier*, 100-118,2022.

[17] Muhammad Imran Razzak, Muhammad Imran, and Guandong Xu. Big data analytics for preventive medicine. *Neural Computing and Applications*, Springer,32(9):4417-4451,2020.

[18] Rajesh Kumar Dhanaraj, K Rajkumar, and U Hariharan. Enterprise iot modeling: supervised, unsupervised, and reinforcement learning. In *Business Intelligence for Enterprise Internet of Things*. Springer,55-79,2020.

[19] Bernhard Kühn, Marc H Taylor, and Alexander Kempf. Using machine learning to link spatiotemporal information to biological processes in the ocean: a case study for north sea cod recruitment. *Marine Ecology Progress Series*, 664:1-22,2021.

[20] William C Sleeman IV and Bartosz Krawczyk. Imbalanced big data oversampling: Taxonomy, algorithms, software, guidelines and future directions. *arxiv preprint*, 110-125,2021.

[21] Patil Sagar. Top7 big data challenges and howto solve them. https://www.linkedin.com/pulse/ top-7-big-data-challenges-how-solve-them-architect-/, May 2022.

[22] Murat Sariyar and Irene Schlünder. Challenges and legal gaps of genetic profiling in the era of big data. *Frontiers in big Data*, Frontiers Media SA,2:40-55,2019.

[23] Ning Chai, Chun Mao, Minjie Ren, Wenwen Zhang, Parthasarathy Poovendran, and P Balamurugan. Role of bic (big data, iot, and cloud) for smart cities. *Arabian Journal for Science and Engineering*, Springer,1-15,2021.

[24] S Muthuramalingam, A Bharathi, N Gayathri, R Sathiyaraj, B Balamurugan, et al. Iot based intelligent transportation system (iot-its) for global perspective: A case study. In *Internet of Things and Big Data Analytics for Smart Generation*. Springer,279-300,2019.

[25] Sanjay Kumar. Monitoring novel corona virus (covid-19) infections in india by cluster analysis. *Annals of Data Science*, Springer,7(3):417-425:2020.

[26] Paul Clemens Murschetz and Dimitri Prandner. 'datafying'broadcasting: Exploring the role of big data and its implications for competing in a big data-driven tv ecosystem. In *Competitiveness in emerging markets*. Springer,55-71,2018.

[27] Li Da Xu and Lian Duan. Big data for cyber physical systems in industry 4.0: a survey. *Enterprise Information Systems*, Taylor Francis,13(2):148-169,2019.

[28] Vidvan Tech. Top 15 big data applications across industries. https://techvidvan.com/tutorials/big-data-applications-across-industries/, May 2022.

[29] James Evans, Andrew Karvonen, Andres Luque-Ayala, Chris Martin, Kes McCormick, Rob Raven, and Yuliya Voytenko Palgan. Smart and sustainable cities? pipedreams, practicalities and possibilities, Taylor Francis,24(7):557-564,2019.

[30] Simon Elias Bibri. On the sustainability of smart and smarter cities in the era of big data: an interdisciplinary and transdisciplinary literature review. *Journal of Big Data*, SpringerOpen,6(1):1-64,2019.

[31] Zsuzsanna Tomor, Albert Meijer, Ank Michels, and Stan Geertman. Smart governance for sustainable cities: findings from a systematic literature review. *Journal of Urban Technology*, TaylorFrancis,26(4):3-27,2019.

[32] Patricia Baudier, Chantal Ammi, and Matthieu Deboeuf-Rouchon. Smart home: Highly-educated students' acceptance. *Technological Forecasting and Social Change*, Elsevier,153,119-355,2020.

[33] Nelvin Xechung Leow and Jayaraman Krishnaswamy. Environmental safety and smart mobility on road–an empirical study in malaysia. *International Journal of Environmental Policy and Decision Making*, Inderscience Publishers (IEL),3(2):139-153,2022.

[34] Jeannette Chin, Vic Callaghan, and Somaya Ben Allouch. The internet-of-things: Reflections on the past, present and future from a user-centered and smart environment perspective. *Journal of Ambient Intelligence and Smart Environments*, IOS Press,11(1):45-69,2019.

[35] Ramona Hosu and Ioan Hosu. 'smart' in between people and the city. *Transylvanian Review of Administrative Sciences*, 15(1):5-20,2019.

[36] Manan Bawa, Dagmar Caganova, Ivan Szilva, and Daniela Spirkova. Importance of internet of things and big data in building smart city and what would be its challenges. In *Smart City 360°: First EAI International Summit, Smart City 360°, Bratislava, Slovakia and Toronto, Canada, October 13-16, 2015.*, Springer,605-616,2016.

[37] R Divakar, P Sowmya, G Suganya, and T Primya. Prescriptive and predictive analysis of intelligible big data. In *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE,845-851,2022.

[38] Hani Al-Sayeh, Muhammad Attahir Jibril, Bunjamin Memishi, and Kai-Uwe Sattler. Blink: lightweight sample runs for cost optimization of big data applications. In *European Conference on Advances in Databases and Information Systems*, Springer,144-154,2022.

[39] Brent Kitchens, David Dobolyi, Jingjing Li, and Ahmed Abbasi. Advanced customer analytics: Strategic value through integration of relationship-oriented big data. *Journal of Management Information Systems*, Taylor Francis,35(2):540-574,2018.

[40] Heli Hallikainen, Emma Savimäki, and Tommi Laukkanen. Fostering b2b sales with customer big data analytics. *Industrial Marketing Management*, Elsevier,86,90-98,2020.

[41] Nadine Côrte-Real, Pedro Ruivo, and Tiago Oliveira. Leveraging internet of things and big data analytics initiatives in european and american firms: Is data quality a way to extract business value? *Information Management*, Elsevier,57(1):103-141,2020.

[42] Florian Kache and Stefan Seuring. Challenges and opportunities of digital information at the intersection of big data analytics and supply chain management. *International journal of operations production management*, Emerald Publishing Limited,7(1):10-36,2017.

[43] Rameshwar Dubey, Angappa Gunasekaran, Stephen J. Childe, David J. Bryde, Mihalis Giannakis, Cyril Foropon, David Roubaud, and Benjamin T Hazen. Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: A study of manufacturing organisations. *International Journal of Production Economics*, Elsevier,226:107-119,2020.

[44] Christiana D von Hippel and Andrew B Cann. Behavioral innovation: Pilot study and new big data analysis approach in household sector user innovation. *Research Policy*, Elsevier,50(8):103-122,2021.

[45] Karwan Jameel Merceedi and Nareen Abdulla Sabry. A comprehensive survey for hadoop distributed file system. *Asian Journal of Research in Computer Science*, 11(2): 46-57,2021.

[46] Hoger Khayrolla Omar and Alaa Khalil Jumaa. Distributed big data analysis using spark parallel data processing. *Bulletin of Electrical Engineering and Informatics*, 11(3):1505-1515,2022.

[47] Alex Giamas. *Mastering MongoDB 6. x: Expert techniques to run high-volume and fault-tolerant database solutions using MongoDB 6. x*. Packt Publishing Ltd,Birmingham, UK,428-488,2022.

[48] J Jyothi. Cassandra is a better option for handling big data in a no-sql database. International Journal of Research Publication and Reviews,ijrpr,3(9):880-893,2022.

[49] Yassine Benlachmi and Moulay Lahcen Hasnaoui. Open source big data platforms and tools: An analysis. *Indonesian Journal of Electrical Engineering and Informatics*, IJEEI,9(3):732-746,2021.

[50] Alan Hong, Weidong Xiao, and JiangLing Ge. Big data analysis system based on cloudera distribution hadoop. In *2021 7th IEEE Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, IEEE,169-173,2021.

[51] Evaristus Didik Madyatmadja, Samuel Imanuel Jordan, and Johanes Fernandes Andry. Big data analysis using rapidminer studio to predict suicide rate in several countries. *ICIC Express Letters, Part B: Applications*, 12(8), 2021.

[52] Umar Aftab and Ghazanfar Farooq Siddiqui. Big data augmentation with data warehouse: A survey. In *2018 IEEE International Conference on Big Data (Big Data)*, IEEE,2785-2794,2018.

[53] BUSINESS APAC. Big data tools: What suits best for your company? https://www.apacbusinessheadlines.com/Big-data-Tools/, May 2022.

[54] Tareq Abed Mohammed, Ahmed Ghareeb, Hussein Al-bayaty, and Shadi Al-jawarneh. Big data challenges and achievements: applications on smart cities

and energy sector. In *Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems*, ACM,Jordan,1-5,2019.

[55] Melanie T Odenkirk, David M Reif, and Erin S Baker. Multiomic big data analysis challenges: increasing confidence in the interpretation of artificial intelligence assessments. *Analytical chemistry*, 93(22), ACS Publications,93(22):7763-7773,2021.

[56] Mahmoud A Mahdi, Khalid M Hosny, and Ibrahim Elhenawy. Scalable clustering algorithms for big data: A review. *IEEE Access*, IEEE,9:80015–80027,2021.

[57] GS Sriram and GS Sriram. Security challenges of big data computing. *International Research Journal of Modernization in Engineering Technology and Science*, irjmets,4(1):1164-1171,2022.

[58] Simon Elias Bibri. The iot for smart sustainable cities of the future: An analytical framework for sensor-based big data applications for environmental sustainability. *Sustainable cities and society*, Elsevier,38:230-253,2018.

[59] Ahmed M Shahat Osman. A novel big data analytics framework for smart cities. *Future Generation Computer Systems*, Elsevier,91:620-633,2019.

[60] Jingrui Ju, Luning Liu, and Yuqiang Feng. Citizen-centered big data analysis-driven governance intelligence framework for smart cities. *Telecommunications Policy*, Elsevier,42(10):881-896,2018.

[61] M Mazhar Rathore, Anand Paul, Won-Hwa Hong, HyunCheol Seo, Imtiaz Awan, and Sharjil Saeed. Exploiting iot and big data analytics: Defining smart digital city using real-time urban data. *Sustainable cities and society*, Elsevier,40:600-610,2018.

[62] Zaheer Allam and Zaynah A Dhunny. On big data, artificial intelligence and smart cities. *Cities*, Elsevier,89,80-91,2019.

[63] M Mazhar Rathore, Anand Paul, Awais Ahmad, Naveen Chilamkurti, Won-Hwa Hong, and HyunCheol Seo. Real-time secure communication for smart city in high-speed big data environment. *Future Generation Computer Systems*, Elsevier,83:638-652,2018.

[64] Minglin Sun and Jian Zhang. Research on the application of block chain big data platform in the construction of new smart city for low carbon emission and green environment. *Computer Communications*, Elsevier,149:332-342,2020.

[65] Chiehyeon Lim, Kwang-Jae Kim, and Paul P Maglio. Smart cities with big data: Reference models, challenges, and considerations. *Cities*, Elsevier,82:86-99,2018.

[66] Moneeb Gohar, Muhammad Muzammal, and Arif Ur Rahman. Smart tss: Defining transportation system behavior using big data analytics in smart cities. *Sustainable cities and society*, Elsevier,41:114-119,2018.

[67] Tae-Min Song and Seewon Ryu. Big data analysis framework for healthcare and social sectors in korea. *Healthcare informatics research*, Korean Society of Medical Informatics,21(1):3-9,2015.

[68] Yang Yu and Xiao Wang. World cup 2014 in the twitter world: A big data analysis of sentiments in us sports fans' tweets. *Computers in Human Behavior*, Elsevier,48:392-400,2015.

[69] Zaheer Khan, Ashiq Anjum, Kamran Soomro, and Muhammad Atif Tahir. Towards cloud based big data analytics for smart future cities. *Journal of Cloud Computing*, SpringerOpen,4(1):1-112015.

[70] Shikah J Alsunaidi, Abdullah M Almuhaideb, Nehad M Ibrahim, Fatema S Shaikh, Kawther S Alqudaihi, Fahd A Alhaidari, Irfan Ullah Khan, Nida Aslam, and Mohammed S Alshahrani. Applications of big data analytics to control covid-19 pandemic. *Sensors*, MDPI,21(7):22-38,2021.

[71] Mohamed Azlan Ashaari, Karpal Singh Dara Singh, Ghazanfar Ali Abbasi, Azlan Amran, and Francisco J Liebana-Cabanillas. Big data analytics capability for improved performance of higher education institutions in the era of ir 4.0: A multi-analytical sem & ann perspective. *Technological Forecasting and Social Change*, Elsevier,173:101-119,2021.

[72] Noha Mostafa, Haitham Saad Mohamed Ramadan, and Omar Elfarouk. Renewable energy management in smart grids by using big data analytics and machine learning. *Machine Learning with Applications*, Elsevier,9:100-136,2022.

[73] P Rohini, Sachin Tripathi, CM Preeti, A Renuka, José Luis Arias Gonzales, and Durgaprasad Gangodkar. A study on the adoption of wireless communication in big data analytics using neural networks and deep learning. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, IEEE,1071-1086,2022.

[74] Zhihan Lv, Liang Qiao, Ken Cai, and Qingjun Wang. Big data analysis technology for electric vehicle networks in smart cities. *IEEE Transactions on Intelligent Transportation Systems*, IEEE,22(3):1807-1816,2020.

[75] Yanqi Lian, Guoqing Zhang, Jaeyoung Lee, and Helai Huang. Review on big data applications in safety research of intelligent transportation systems and connected/automated vehicles. *Accident Analysis & Prevention*, Elsevier,146:105-117,2020.

[76] Muhammad Bilal, Lukumon O Oyedele, Olugbenga O Akinade, Saheed O Ajayi, Hafiz A Alaka, Hakeem A Owolabi, Junaid Qadir, Maruf Pasha, and Sururah A Bello. Big data architecture for construction waste analytics (cwa): A conceptual framework. *Journal of Building Engineering*, Elsevier,6:144-156,2016.

[77] Delin Zeng, Yenni Tim, Jiaxin Yu, and Wenyuan Liu. Actualizing big data analytics for smart cities: A cascading affordance study. *International Journal of Information Management*, Elsevier,54:102-126,2020.

[78] Fan Zhang, Junwei Cao, Samee U Khan, Keqin Li, and Kai Hwang. A task-level adaptive mapreduce framework for real-time streaming data in healthcare applications. *Future generation computer systems*, Elsevier,43:149-160,2015.

[79] Shahbaz Khan. Barriers of big data analytics for smart cities development: a context of emerging economies. *International Journal of Management Science and Engineering Management*, Taylor Francis,17(2):123-131,2022.

[80] Nojood O Aljehane and Romany F Mansour. Big data analytics with oppositional moth flame optimization based vehicular routing protocol for future smart cities. *Expert Systems*, Wiley Online Library,39(5):112-118,2022.

[81] Atta-ur Rahman, Sagheer Abbas, Mohammed Gollapalli, Rashad Ahmed, Shabib Aftab, Munir Ahmad, Muhammad Adnan Khan, and Amir Mosavi. Rainfall pre-

diction system using machine learning fusion for smart cities. *Sensors*, Multidisciplinary Digital Publishing Institute,22(9):04-35,2022.

[82] Sinan Kahveci, Bugra Alkan, Ahmad Mus'ab H, Bilal Ahmad, and Robert Harrison. An end-to-end big data analytics platform for iot-enabled smart factories: A case study of battery module assembly system for electric vehicles. *Journal of Manufacturing Systems*, ELSEVIER,63:214-223,2022.

[83] NP Jayasri and R Aruna. Big data analytics in health care by data mining and classification techniques. *ICT Express*, Elsevier,8(2):250-257,2022.

[84] Yong Zhong, Liang Chen, Changlin Dan, and Amin Rezaeipanah. A systematic survey of data mining and big data analysis in internet of things. *The Journal of Supercomputing*, Springer,1-49,2022.

[85] Elarbi Badidi and Karima Moumane. Enhancing the processing of healthcare data streams using fog computing. In *2019 IEEE Symposium on Computers and Communications (ISCC)*, IEEE,1113-1118,2019.

[86] Arezou Naghib, Nima Jafari Navimipour, Mehdi Hosseinzadeh, and Arash Sharifi. A comprehensive and systematic literature review on the big data management techniques in the internet of things. *Wireless Networks*, Springer,1-60,2022.

[87] Abeer Iftikhar Tahirkheli, Muhammad Shiraz, Bashir Hayat, Muhammad Idrees, Ahthasham Sajid, Rahat Ullah, Nasir Ayub, and Ki-Il Kim. A survey on modern cloud computing security over smart city networks: Threats, vulnerabilities, consequences, countermeasures, and challenges. *Electronics*, MDPI,10(15):11-18,2021.

[88] Sergio Trilles, Alberto González-Pérez, and Joaquín Huerta. An iot platform based on microservices and serverless paradigms for smart farming purposes. *Sensors*.

[89] Sabbir Ahmed, Md Hossain, M Shamim Kaiser, Manan Binth Taj Noor, Mufti Mahmud, Chinmay Chakraborty, et al. Artificial intelligence and machine learning for ensuring security in smart cities. In *Data-driven mining, learning and analytics for secured smart cities*. Springer,23-47,2021.

[90] Haithem Mezni, Mokhtar Sellami, Sabeur Aridhi, and Faouzi Ben Charrada. Towards big services: a synergy between service computing and parallel programming. *Computing*, Springer,103(11):2479-2519,2021.

[91] Cristian Martín Fernández, Peter Langendoerfer, Pouya Soltani Zarrin, Manuel Díaz-Rodríguez, Bartolome Rubio-Muñoz, et al. Kafka-ml: Connecting the data stream with ml/ai frameworks. Elsevier,126: 15-33,2022.

[92] Amin Shahraki, Mahmoud Abbasi, Amir Taherkordi, and Anca Delia Jurcut. A comparative study on online machine learning techniques for network traffic streams analysis. *Computer Networks*, Elsevier,207,108-836,2022.

[93] Azad Abdulhafedh. Comparison between common statistical modeling techniques used in research, including: Discriminant analysis vs logistic regression, ridge regression vs lasso, and decision tree vs random forest. *Open Access Library Journal*, Scientific Research Publishing,9(2):1-19,2022.

[94] Md Rahman, Shahadat Hossain, Mimun Barid, Md Hasan, et al. Inductions of usernames' strengths in reducing invasions on social networking sites (snss). In *International Conference on Intelligent Computing Optimization*, Springer,331-340,2021.

[95] Mark Eshwar Lokanan and Kush Sharma. Fraud prediction using machine learning: The case of investment advisors in canada. *Machine Learning with Applications*, Elsevier,8:100-269,2022.

[96] Qian Guo, Jian Zhang, Shijie Guo, Zhangxi Ye, Hui Deng, Xiaolong Hou, and Houxi Zhang. Urban tree classification based on object-oriented approach and random forest algorithm using unmanned aerial vehicle (uav) multispectral imagery. *Remote Sensing*, MDPI,14(16):38-85,2022.

[97] Mark Summerfield. *Programming in Python 3: a complete introduction to the Python language*. Addison-Wesley Professional,American,475-495,2010.

[98] Maxim Ziatdinov, Ayana Ghosh, Chun Yin Tommy Wong, and Sergei V Kalinin. Atomai framework for deep learning analysis of image and spectroscopy data in electron and scanning probe microscopy. *Nature Machine Intelligence*, Nature Publishing Group,1-12,2022.

[99] Phanwadee Sinthong and Michael J Carey. Exploratory data analysis with database-backed dataframes: A case study on airbnb data. In *2021 IEEE International Conference on Big Data (Big Data)*, IEEE,3119-3129,2021.

[100] Khac Sang Cao. Data mining for social network data. 2019.

[101] Igor Stančin and Alan Jović. An overview and comparison of free python libraries for data mining and big data analysis. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, IEEE,977-982,2019.

[102] Giang Nguyen, Stefan Dlugolinsky, Martin Bobák, Viet Tran, Álvaro López García, Ignacio Heredia, Peter Malík, and Ladislav Hluchỳ. Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review*, Springer,52(1):77-124,2019.

[103] Daniele Grattarola and Cesare Alippi. Graph neural networks in tensorflow and keras with spektral [application notes]. *IEEE Computational Intelligence Magazine*, IEEE,16(1à:99-106,2021.

[104] Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S Rellermeyer. A survey on distributed machine learning. *Acm computing surveys (csur)*, ACM New York, NY, USA,53(2à:1-33,2020.

[105] Pulkit A Misra, María F Borge, Íñigo Goiri, Alvin R Lebeck, Willy Zwaenepoel, and Ricardo Bianchini. Managing tail latency in datacenter-scale file systems under production constraints. In *Proceedings of the Fourteenth EuroSys Conference 2019*, USA,1-15,2019.

[106] Kamalakant Laxman Bawankule, Rupesh Kumar Dewang, and Anil Kumar Singh. Historical data based approach to mitigate stragglers from the reduce phase of mapreduce in a heterogeneous hadoop cluster. *Cluster Computing*, Springer,1-19,2022.

[107] PC Reddy and Alladi Sureshbabu. An adaptive model for forecasting seasonal rainfall using predictive analytics. *International Journal of Intelligent Engineering and Systems*, inass,12(5):22-32,2019.

[108] Yogesh Kumar, Kanika Sood, Surabhi Kaul, and Richa Vasuja. Big data analytics and its benefits in healthcare. In *Big data analytics in healthcare*. Springer,3-21,2020.

[109] Ohoud Nafea, Wadood Abdul, Ghulam Muhammad, and Mansour Alsulaiman. Sensor-based human activity recognition with spatio-temporal deep learning. *Sensors*, MDPI,21(6):21-41,2021.

[110] Masoud Zare, Rushikesh Battulwar, Joseph Seamons, and Javad Sattarvand. Applications of wireless indoor positioning systems and technologies in underground mining: A review. *Mining, Metallurgy & Exploration*, Springer,38(6):2307-2322,2021.

[111] Gonçalo Alegre de Matos Moura et al. Wireless sensor network for collecting environmental data and controlling several parameters in the perimeter of a vegetable farm. ipp,60-97,2019.

[112] Thai-Mai Thi Dinh, Ngoc-Son Duong, and Kumbesan Sandrasegaran. Smartphone-based indoor positioning using ble ibeacon and reliable lightweight fingerprint map. *IEEE Sensors Journal*, IEEE,20(17):10283-10294,2020.

[113] Francesco Pascale, Ennio Andrea Adinolfi, Massimiliano Avagliano, Venanzio Giannella, and Andres Salas. A low energy iot application using beacon for indoor localization. *Applied Sciences*, Multidisciplinary Digital Publishing Institute,11(11):4902-4903,2021.

[114] Xiaoqiang Zhu, Wenyu Qu, Tie Qiu, Laiping Zhao, Mohammed Atiquzzaman, and Dapeng Oliver Wu. Indoor intelligent fingerprint-based localization: Principles, approaches and challenges. *IEEE Communications Surveys & Tutorials*, IEEE,22(4):2634-2657,2020.

[115] Jenni AM Sidey-Gibbons and Chris J Sidey-Gibbons. Machine learning in medicine: a practical introduction. *BMC medical research methodology*, Springer,19(1):1-18,2019.

[116] Saurabh Tewari, Umakant Dhar Dwivedi, and Susham Biswas. A novel application of ensemble methods with data resampling techniques for drill bit selection in the oil and gas industry. *Energies*, MDPI,14(2):432-433,2021.

[117] Hindreen Rashid Abdulqadir, Adnan Mohsin Abdulazeez, and Dilovan Assad Zebari. Data mining classification techniques for diabetes prediction. *Qubahan Academic Journal*, 1(2):125-133,2021.

[118] Aakanksha Sharaff and Harshil Gupta. Extra-tree classifier with metaheuristics approach for email classification. In *Advances in computer communication and computational sciences*. Springer,189-197,2019.

[119] Kristen Jaskie and Andreas Spanias. Positive and unlabeled learning algorithms and applications: A survey. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, IEEE,1-8,2019.

[120] S Irin Sherly et al. An ensemble basedheart disease predictionusing gradient boosting decision tree. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(10):3648-3660,2021.

[121] Basim Mahbooba, Mohan Timilsina, Radhya Sahal, and Martin Serrano. Explainable 120intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, Hindawi,1-11,2021.

[122] Jennie Pearce and Simon Ferrier. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological modelling*, Elsevier,133(3):225-245,2000.

[123] Omri Allouche, Asaf Tsoar, and Ronen Kadmon. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (tss). *Journal of applied ecology*, Wiley Online Library,43(6):1223-1232,2006.

[124] Vijendra Singh, Vijayan K Asari, and Rajkumar Rajasekaran. A deep neural network for early detection and prediction of chronic kidney disease. *Diagnostics*, MDPI,12(1):116-117,2022.

[125] Richard Fuller, Philip J Landrigan, Kalpana Balakrishnan, Glynda Bathan, Stephan Bose-O'Reilly, Michael Brauer, Jack Caravanos, Tom Chiles, Aaron Cohen, Lilian Corra, et al. Pollution and health: a progress update. *The Lancet Planetary Health*, Elsevier,12:106-117,2022.

[126] Tiziana Campisi, Alessandro Severino, Muhammad Ahmad Al-Rashid, and Giovanni Pau. The development of the smart cities in the connected and autonomous vehicles (cavs) era: From mobility patterns to scaling in cities. *Infrastructures*, MDPI,6(7):100-102,2021.

[127] Pawan Sen, Rashi Jain, Vaibhav Bhatnagar, and Suhaib Illiyas. Big data and ml: Interaction & challenges. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE,16:939-943,2022.

[128] Syed Attique Shah, Dursun Zafer Seker, M Mazhar Rathore, Sufian Hameed, Sadok Ben Yahia, and Dirk Draheim. Towards disaster resilient smart cities: Can internet of things and big data analytics be the game changers? *IEEE Access*, IEEE,7:91885-91903,2019.

[129] AJH Redelinghuys, Anton Herman Basson, and Karel Kruger. A six-layer architecture for the digital twin: a manufacturing case study implementation. *Journal of Intelligent Manufacturing*, Springer,31(6):1383-1402,2020.

[130] Awais Ahmad, Muhammad Babar, Sadia Din, Shehzad Khalid, Muhammad Mazhar Ullah, Anand Paul, Alavalapati Goutham Reddy, and Nasro Min-Allah. Socio-cyber network: The potential of cyber-physical system to define human behaviors using big data analytics. *Future generation computer systems*, Elsevier,92:868-878,2019.

[131] Behailu Negash, Tomi Westerlund, and Hannu Tenhunen. Towards an interoperable internet of things through a web of virtual things at the fog layer. *Future Generation Computer Systems*, Elsevier,91:96-107,2019.

[132] Mucahid Mustafa Saritas and Ali Yasar. Performance analysis of ann and naive bayes classification algorithm for data classification. *International Journal of Intelligent Systems and Applications in Engineering*, 7(2):88-91,2019.

[133] Usman Mahmood, Zening Fu, Satrajit Ghosh, Vince Calhoun, and Sergey Plis. Through the looking glass: Deep interpretable dynamic directed connectivity in resting fmri. *NeuroImage*, Elsevier,194:119-137,2022.

[134] Ashwin Pajankar and Aditya Joshi. Supervised learning methods: Part 1. In *Hands-on Machine Learning with Python*. Springer,99-120,2022.

[135] Jasmin Zalonis, Frederik Armknecht, Björn Grohmann, and Manuel Koch. Report: State of the art solutions for privacy preserving machine learning in the medical context. *arxiv preprint*, 2201-2216,2022.

[136] Monalisa Jena, Ranjan Kumar Behera, and Satchidananda Dehuri. Hybrid decision tree for machine learning: A big data perspective. In *Advances in Machine Learning for Big Data Analysis*. Springer,223-239,2022.

[137] Manik Rakhra, Priyansh Soniya, Dishant Tanwar, Piyush Singh, Dorothy Bordoloi, Prerit Agarwal, Sakshi Takkar, Kapil Jairath, and Neha Verma. Crop price prediction using random forest and decision tree regression:-a review. *Materials Today: Proceedings*, Elsevier,123-139,2021.

[138] Ali Amkor, Kamal Maaider, and Noureddine El Barbri. An evaluation of machine learning algorithms coupled to an electronic olfactory system: a study of the mint case. *International Journal of Electrical Computer Engineering*, 12(4):25-34,2022.

[139] Mohammed AS Abourehab, Sameer Alshehri, Bader Huwaimel, Ali H Alamri, Rami M Alzhrani, Ahmed Alobaida, Hossam Kotb, Amal M Alsubaiyel, Sabina Yasmin, Kumar Venkatesan, et al. Enhancing drugs bioavailability using nanomedicine approach: Predicting solubility of tolmetin in supercritical solvent via advanced computational techniques. *Journal of Molecular Liquids*, Elsevier,365:120-133,2022.

[140] MARIANI, Marcello M., BORGHI, Matteo, et LAKER, Benjamin. Do submission devices influence online review ratings differently across different types of platforms? A big data analysis. Technological Forecasting and Social Change, 2023, vol. 189, p. 122296.

[141] Jie Chen, L Ramanathan, and Mamoun Alazab. Holistic big data integrated artificial intelligent modeling to improve privacy and security in data management of smart cities. *Microprocessors and Microsystems*, Elsevier,81:103-122,2021.

[142] Miguel Angel Moreno-Mateos and Diego Carou. A note on big data and value creation. In *Machine Learning and Artificial Intelligence with Industrial Applications*. Springer,1-182022.