Enhancing Customer Insights and Business Strategies in Healthcare Insurance Using Hadoop, Hive, and Spark

A MINOR ROJECT REPORT

Submitted by

MADDU RAKESH [RA2211028010159]

KONIJETI SAI KALYAN [RA2211028010160]

VUTUKURI ROOP RAVI TEJA [RA2211028010161]

BHAVAN NARENDRA REDDY [RA2211028010175]

Under the Guidance of

Dr.Indra Bhooshan Sharma

Assistant Professor

Department of Networking and Communications

in partial fulfillment of the requirementsfor the degree of

BACHELOR OF TECHNOLOGY in

COMPUTER SCIENCE ENGINEERING

with specialization in CLOUD COMPUTING

DEPARTMENT OF NETWORKING AND COMMUNICATIONS

COLLEGE OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR- 603 203

NOVEMBER 2024

Department of Networking and Communications

SRM Institute of Science & Technology

Own Work Declaration Form

This sheet must be filled in (each box ticked to show that the condition has been met). It must besigned and dated along with your student registration number and included with all assignments you submit – work will not be marked unless this is done.

To be completed by the student for all assessments

Degree/ Course	**:Big Data Essentials (21CSC314P)**
Student Names	**:M.Rakesh,K.Sai Kalyan,V.Roop Ravi Teja B.Narendra**
Registration Numbers	**:RA2211028010159,RA2211028010160, RA2211028010161,RA2211028010175**
Title of Work	**:Enhancing Customer Insights and Business Strategies in Healthcare Insurance Using Hadoop, Hive, and Spark**

I / We hereby certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism, as listed in the University Website, Regulations, and the Education Committee guidelines.

I / We confirm that all the work contained in this assessment is my / our own except where indicated, and that I / We have met the following conditions:

- Clearly referenced / listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Acknowledged in appropriate places any help that I have received from others (e.g.fellow students, technicians, statisticians, external sources)
- Compiled with any other plagiarism criteria specified in the Course handbook / University website

I understand that any false claim for this work will be penalized in accordance with theUniversity policies and regulations.

DECLARATION:

I am aware of and understand the University's policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except where indicated by referring, and that I have followed the good academic practices noted above.

M.Rakesh	K.Sai Kalyan	V.Roop Ravi Teja	B.Narendra
RA2211028010159	RA2211028010160	RA2211028010161	RA2211028010175

If you are working in a group, please write your registration numbers and sign with the date for every student in your group.

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
KATTANKULATHUR – 603 203

BONAFIDE CERTIFICATE

Certified that this Minor Project report for the Couse21CSC314P –Big Data Essentials mini-project report titled "**Enhancing Customer Insights and Business Strategies in Healthcare Insurance Using Hadoop, Hive, and Spark**" is the bonafide work of "**MADDU RAKESH [RA2211028010159], KONIJETI SAI KALYAN [RA2211028010160], VUTUKURI ROOP RAVI TEJA [RA2211028010161] BHAVAN NARENDRA REDDY [RA2211028010175]"** who carried out the mini-project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

Panel Reviewer I

SIGNATURE
Dr.Indra Bhooshan Sharma
Assistant Professor
Department of Networking
and Communications

Panel Reviewer II

SIGNATURE
Dr.S.Sivamohan
Assistant Professor
Department of Networking
and Communications

ACKNOWLEDGEMENT

We express our humble gratitude to Dr. C. Muthamizhchelvan, Vice Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to Dean-CET, SRM Institute of Science and Technology, Dr.T.V. Gopal, and Dr. Revathi Venkataraman, Professor & Chairperson, School of Computing, SRM Institute of Science and Technology, for their valuable support during the project course.

We are incredibly grateful to Head of the Department, Dr. M. Lakshmi, Professor and Head, Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, for offering us the project course: Big Data Essentials.

We want to convey our thanks to our faculty mentor Dr.Indra Bhooshan Sharma Assistant Professor, Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, for his inputs and support during the project phase.

We express our respect and thanks to Department HOD, SRM Institute of Science and Technology, for providing us an opportunity to pursue this project course

We register our immeasurable thanks to our Faculty Advisor, School of Computing, SRM Institute of Science and Technology, for leading and helping us to complete our course.

We sincerely thank the Networking and Communications department staff members, SRM Institute of Science and Technology, for their help during our project.

Finally, we would like to thank our parents, family members, and friends for their unconditional love, constant support, and encouragement.

MADDU RAKESH [RA2211028010159]
KONIJETI SAI KALYAN [RA2211028010160]
VUTUKURI ROOP RAVI TEJA [RA2211028010161]
BHAVAN NARENDRA REDDY [RA2211028010175]

ABSTRACT

The healthcare insurance sector is shifting and changing rapidly, mainly based on the strategic use of big data to maximize revenue and quality service. This research solves significant problems of healthcare insurance firms by using big data analytics to examine massive customer data sets. Uncovering patterns in consumer behavior can help tailor insurance policies, pricing models, and revenue growth strategies. In addition to online scraping and third-party sources, competitive insights are also gathered in order to provide an extra edge of market dynamics. Management, processing, and analysis of vast volumes of data are accomplished with the help of big data technologies consisting of Hadoop, Apache Hive, Apache Spark, and Sqoop. The ultimate outcome is to develop customized insurance products so that revenue growth will deliver customer satisfaction. The framework underscores evidence-based decision-making for effective allocation and planning with a resource, harmonized with Sustainable Development Goals (SDGs), though in a manner that encourages innovation and economic growth. Through these instruments, the project delivers actionable insights into how to help a competitive strategy, customer engagement, adaptive business models, and an attempt to position healthcare insurance companies at the vanguard of data-driven change.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	vi
	LIST OF FIGURES	ix
	LIST OF TABLES	ix
	LIST OF SYMBOLS AND ABBREVIATIONS	x
1	INTRODUCTION	11-12
2	LITERATURE SURVEY	13-16
3	PROPOSED METHODOLOGY	17-21
	3.1. Architecture Diagram	18
	3.2. Data Collection	19
	3.3. Data Storage	19
	3.4. Data Preprocessing	20
	3.5. Data Storage and Management	20
	3.6. Model Development	20
	3.7. Prediction	21
	3.8. Visualization	21
4	IMPLEMENTATION	22-24
	4.1. Data Collection	22
	4.2. Data Storage	22
	4.3. Data Preprocessing	23
	4.4. Model Development	23
	4.5. Predictions	24
	4.6. RMSE	24
5	RESULTS AND DISCUSSIONS	25-33
	5.1 RESULTS	25-30
	5.2 Metrics Comparison	31-33

6 CONCLUSION AND FUTURE WORK 34

REFERENCES 35-36

APPENDIX 37-38

B PLAGIARISM REPORT 37-38

LIST OF FIGURES

Fig .3.1 Health Care Insurance Architecture Diagram 18

Fig 4.1Architecture Diagram for Random Forest 23

Fig 5.1 Toatal Claim Amount by Claim Type 25

Fig 5.2 Number of Claims per Disease 27

Fig 5.3 Average Monthly Premium Paid by Subgroup 28

Fig 5.4 Number of Claims Accepted or Rejected 29

Fig 5.5 confusion matrix for clam status 31

Fig 5.6 confusion matrix for Accepted and Rejected 32

LIST OF TABLES

Table 5.1 Comparison Table for different models 30

LIST OF SYMBOLS AND ABBREVIATIONS

HDFS - Hadoop Distributed File System

API - Application Programming Interface

RMSE - Root Mean Square Error

SVM - Support Vector Machine

ML - Machine Learning

ETL - Extract, Transform, Load

AI - Artificial Intelligence

ROC - Receiver Operating Characteristic (curve)

AUC - Area Under Curve

SDGs - Sustainable Development Goals

SQL - Structured Query Language

EDA - Exploratory Data Analysis

KNN - K-Nearest Neighbors

TP - True Positive

FP - False Positive

CHAPTER 1
INTRODUCTION

The healthcare insurance industry is placed at a crossroads for transformative change with the rapid advancements in technology and the increasing [7] need for the efficient usage of data. Big data offers healthcare insurers opportunities for enhancing customer service, streamlining operations, and promoting financial performance. The volume, variety, and velocity of data are being generated from multiple sources, including patient records, claims history, and competitive insights, demanding robust infrastructure and sophisticated analytical techniques. Traditional approaches to data analysis are no longer sufficient to tap the full potential of these data streams. Therefore, big data has become essential for health care insurers seeking to remain competitive in a rapidly evolving industry.

Big data analytics plays its part here [2]. It is where, with the use of big data technologies, health care providers can get an actionable insight even from massive datasets to understand complex patterns of customer behavior for predicting trends and then making smart business decisions [4]. Health insurers can handle heavy volumes of information in real-time analytics with the help of Apache Hadoop, Apache Hive, Apache Spark, and Sqoop. All these tools provide not only scalable data processing but also real-time analytics [4], providing insurers with timely insights to adapt to dynamic market conditions. Big data analytics represents a shift from reactive decisions to proactive, data-driven strategies.

Further, this project looks to continue stimulating consumer involvement and satisfaction through pushing policy into data-driven customizations. By analyzing the customer data, they will be able to understand the behavioral patterns [8] and preferences, which helps them tailor policies to suit the individual needs better. Predictive modeling helps predict coverage type that a customer is likely to have as part of his or her purchase requirement thereby creating personalized offers. This would be to the advantage of both the insurer, where his chances for being retained or kept continue to improve, and the customer, where coverage options seem relevant and satisfactory. Insights into customized policies from data enable insurers to come off as responsive and customer-centric.

Other important focus areas of this research are strategies of revenue growth. This sector is highly competitive, and the profitable segment of customers as well as their correct pricing models are necessary to ensure financial sustainability [11]. Internal data - claims history, policy details and so forth; along with external data regarding market trends and competitors can help develop targeted marketing strategies while also identifying appropriate pricing models that correspond to the willingness of the customer to pay. This insight allowed the insurer to strategically position himself in the market, optimizing, at the same time, customer acquisition and revenue generation. Third-party sources and data scraping supplemented further support for competitive analysis [15] and evidence of trends in the industry.

The use of big data actually aligns with the global sustainability process, particularly with the United Nations' Sustainable Development Goals.
For example, data-driven approaches to health insurance plans can help toward achieving SDG-through equitable access to good health care and improved health and wellbeing. Additionally, reducing waste and increasing efficiency will drive big data initiatives to spur economic growth and innovation. The objective of this project is to grow sustainability in business objectives as enunciated in how data-drive [18] methods can introduce financial benefits as well as meet broader social objectives.

Therefore, it supports the view of seeing data as a strategic asset, hence supporting its rich contribution to the development of sustainability. In the entire project, big data transformed healthcare insurance. The investment in infrastructure and analytics capabilities will ensure competitiveness; patient satisfaction will increase, while financial goals will be attained. As healthcare insurance is increasingly becoming data-centric, it is possible that companies that make better use of big data may have an advantage in shaping the going in the industry. This working paper explores the revolution in healthcare insurance through big data: opportunities for actionable insights by companies to make a profit in a rapid advance of the digital landscape.

CHAPTER 2
LITERATURE REVIEW

A A study by U. Srinivasan et al [1]. with the title "Leveraging Big Data Analytics to Reduce Healthcare Costs" throws lights upon how in the near future, big data analytics will transform health care spending. It will save the health care providers much money as it will make more strategic and informed decisions with the help of sophisticated analytical tools such as many for getting crucial insights about the trends concerning patient care, which will achieve desirable outcomes in the treatment and operational procedures. As indicated in the paper, how large datasets can be useful for the healthcare companies to find more inefficiencies and allocate resources accordingly, is described in detail. The paper explains the way health organizations can use large datasets in finding more inefficiencies so accordingly allocating resources wisely to finally boost tremendous saving in costs. The research will show the proficient in big data analytics incorporation will supports the cost reduction efforts besides improving the general standard of patient care.

Here, the authors describe how the kind of insights being used can lead to improvements in financial sustainability for healthcare organizations, as well as concurrently offer better health outcomes. Their research contributes to the growing literature about the applications of big data in their healthcare when they emphasize the point that businesses will not be able to survive if they do not embrace data-driven strategies in a setting that is becoming both more competitive and more complex. The healthcare administrators should ensure the cost of the care is reduced while maintaining the standards of the services.

One of the major deterrents to this project is the cost on hospitalization, which is a huge area of spending for overall healthcare. Y. Xie et al [2]. in their research study discuss a new model developed to predict days of hospitalization based on extensive data from health insurance claims. It predicts the number of days of hospitalization for the third-based year in the prior and clear admissions mainly, and the procedure claims that the authors used a regression decision tree algorithm in analyzing claims from 242,075 individuals for three years. The results show that this approach will outperform baselines methods that had been developed for them previously with a notable predictive the accuracy of 0.843 for the total population or certain subpopulations, such as older patients and previously hospitalized.

Medical information, diagnosis codes, in particular will be crucial to increase the predictive accuracy. Geographic Information Systems (GIS) and big data analytics work together toward improving healthcare decision making as seen in the study "Comparative Review of Big Data Analytics and GIS in the Healthcare Decision-Making" by the Odunayo Josephine et al [3] The use of methods such as machine learning and predictive modeling and big data analytics gives insightful knowledge to clinical professionals about the trends and patterns that aid in making clinical decisions Large data sets can also be processed more easily with this method which leads to more intelligent patient care plans.

Big Data in the Insurance Sector, Illustrated with Research Output: A Bibliometrical and Systematic Review, a report by researchers Nejla Ellili et al [4]. An analysis concludes how big data will cause changes in the insurance sector.

Applying big data tools will allow the insurance companies to improve their engagement for the customer and also enhance risk-free assessment which helps to find fraud and raise the overall efficiency:.

Through the technologies developed by artificial intelligence and the machinery learning, large data set analysis can be made, enabling insurers to give out customized insurance plans for enhancing their pricing strategies and automating the process of claims. Businesses can better retain customers to create individualized plans through studying consumer behavior and trends and customizing services accordingly.

Also, big data prevents financial losses resulting from expedite investigations and detection of fraud activities because the suspicious activity will be tracked beforehand through the activity. In their study "A Survey on Driving Behavior Analysis in Usage-Based Insurance Using Big Data," Arumugam et al [5]. In other words, (2019) remarked how driver behavior analysis through big data can dramatically change usage-based insurance. The researchers will work out how models for Pay As You Drive and Pay How You Drive and Manage How You Drive change with time from data that can be used at every step of the way to enhance risk profiling and tailor premium payments. The study shall highlight how big data analytics can provide a deeper and richer understanding into the driving behaviors.

In return, this allows the insurers to offer the more customized insurance policies and incites the safer driving practices in case of timely feedback and aggressive alerts. This risk-position strategy, therefore, focusing on lower risks in terms of more productive and participative insurance models is appropriate over a right shift in direction.

"The Technological of Disruption of the Insurance Industry" Associate Professor Antonella Cappiello Management analyzes the deep impacts of a digital revolution on the insurance business. This paper draws underlined that there is the need for critical changes in corporate culture, product offerings and processes as well as changing relationships with clients. It has related with how the insurers are made to innovate constantly due to changing technologies that bring new changes in the changing consumer expectations and increased competition. Applications of big data face very serious security and privacy challenges precipitated by increased variety of sources and formats of data and user types. H. Chen et al [6]. Suggest a scalable multi-labels-based access do-control framework particularly that has been suited to Hadoop-based good big healthcare applications with a view to coping with this type of problems. A number of access control mechanisms-such as mandatory access control (MAC), attribute-based access control (ABAC), role-based access control (RBAC), discretionary access control (DAC), active bundle-are all integrated within the creative framework. The main framework by Ashiat Ashake et al [7] covers the important variables that include data type, security, lifetime, in order to increase the number of the replications, access to the policy and the hash values, which are taken care of by multilabel.

Because of an EHR and HIS-centered emphasis, the "Big Data in Healthcare: Acquisition, Management, and Visualization Using System Dynamics" research studies how the big data revolution is a rapidly transformative pattern of healthcare systems. Because these tools are helping to enhance data-driven decision-making, provider coordination, and patient care, big data must be taken seriously. This paper discusses the benefits of combining big data analytics with personalized medicine, the need to safeguard patient privacy and security, the compliance of HIPAA to regulations, overcoming challenges like inconsistent and fragmented data, etc. Another crucial feature in the study is the use of system dynamics, a technique that models complex health systems in an effort to understand and improve them. This modeling technique enhances health management and decision-making by giving visual feedback loops and linkages. System dynamics was applied in balancing loops in order to analyze how risks, including data quality and privacy concerns, could be handled. Besides that, reinforcing loops that amplify improvement in data management and patient outcome. Healthcare systems can utilize big data to improve patient care efficiency and operational capacity Hashim Zahoor et al [8] The paper named Revolutionizing Insurance

Big Data Analytics Impact highlights the revolutionary role for the big data to do analytics mainly in the insurance sector. It highlights advancements in data processing technologies such as machine learning and artificial intelligence that have triggered changes in risk assessment, fraud detection, customer experience, and claims processing.

These technologies will enable better-informed decisions and provide new business models that focus on predictive risk management for the insurance industry. However, the paper raises concerns for privacy and data security and the process of individualization of insurance offers.

The paper further explores how big data analytics enhances operational efficiency and market penetration, fostering financial inclusion even in less developed economies. Predictive analytics, particularly with machine learning models. Calvin W L Ho et al [9] The study "Ensuring Trustworthy Utilizing of Big Data for Analytics and also for Artificial Intelligence in the Health Insurance examines the moral and legal for implications to utilizing the big data and also AI in health care insurance. It highlights the need for these technologies to be implemented within a strong ethical framework even though they can improve risk assessment, fraud detection, and personalization of healthcare services within health systems. Data protection, openness, and governance are important concerns. The authors advise taking a human-centered strategy to guarantee ethical and open data utilization, which will increase system confidence. The report also emphasizes how big data and artificial intelligence (AI) have the ability to really transform health insurance by providing more specialized and effective services like predictive cost algorithms. Kornelia Batko et al [10] The authors of the article "Use of Big Data Analytics in Health-Care" covered the increasing importance of big data in healthcare for decision-making purposes. It expands the analytics of both structured and unstructured data from many sources, ranging from social media, medical devices, and sensors to improve healthcare services. Big Data analytics reveals useful information that can be used in both clinical and administrative activities, thereby improving the decision-making in the healthcare facility.

CHAPTER 3
PROPOSED METHODOLOGY

A health care analysis [16] details the claims data with many trends of the health insurance market. The data is analyzed through several Python scripts each dealing on a different theme to identify details including the number of claims and accepted versus the rejected claims, the most common diseases, and the most profitable subscribers. Other key findings include the total amounts of claims and the breakdown of claims by type: value, policy, and fact claims, which yield numerous categories of healthcare insurance claims. The scripts also go on to discuss accepted and rejected claims, providing information about why such choices are made and financial stakes involved.

It also indicates how common some diseases are, such as cancer, allergies, or other genetic conditions, which cause a massive amount of claims. Important insights about the health issues that patients and insurers face and potential financial implications for the system can be deduced through critical analysis of claims according to the disease. This data further shows some diseases, including pet allergies, glaucoma, and galactosemia, as high-risk [8] conditions for insurers. A good understanding of customer behavior and risk is efficient for insurance management practices, such as finding the most profitable subscribers, who fall under the category of having the highest total claims.

This analysis would eventually manage massive amounts of data by combining Spark and Python to analyze millions of records about claims. This data-driven approach allows healthcare organizations to make proper decisions concerning what maximizes customer service and financial success. This knowledge acquired can be used to inform the changes to price policies, policy adjustments and improvement of customers as service providers that will eventually enhance the healthcare insurance system in terms of its ability to satisfy client demands with reasonable cost-effectiveness.

3.1 Architecture Diagram

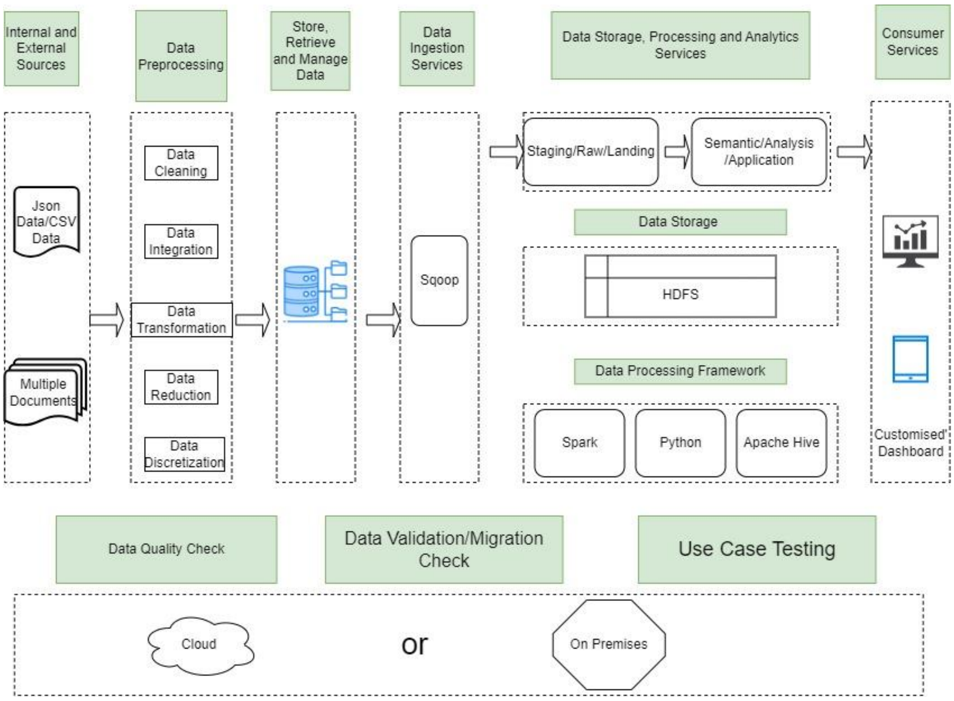


Fig .3.1 Health Care Insurance Architecture Diagram

Figure 3.1 Architecture Diagram. The overall Description Data Processing Flow Healthcare Insurance Project Describe each step, and remember From Ingestion starting with Raw data coming from a variety of sources ranging from internal company databases which might include third-party services that would presume to deliver data and web scraping. The last step is Actionable Insights and Predictions. Once the data is collected, the necessary clean operations involving transformations and integration are performed so that the quality, completeness, and consistency of the data have been guaranteed. Then the formatted data is stored in layers within the HDFS known as staging, curated, and historical for ease in organizing and accessing the data for further analyses. Advanced toolsets like Apache Spark and Python, the data analysis phase is used to develop exploratory data analysis, statistical modeling, and machine learning applications. Finally, insights from this processed data feed consumer-facing services, like dashboards and APIs, into various business needs using flexible architecture that can be deployed either on-premise or in the cloud for data-driven decision-making.

.

3.2 Data Collection

Data Collection Data collection would aim to create vast datasets for the support of robust analysis in healthcare insurance. Its scope is on extracting data from the internal accounts and records of the company, third-party services, as well as web-scraped data from competing sites and social media networks. This will help capture critical consumer behaviours, market trends, and competitive insights. Combining internal and external data, the project delivers an integrated view on the healthcare insurance landscape, thus facilitating more effective customer profiling and risk assessment. This diversity in dataset allows the company to know the consumer's needs and preferences better, thus providing more relevant and competitive models of insurance offerings and pricing. The collection of data forms the foundation for the overall objective of the project by ensuring the breadth and diversity of data necessary for meaningful analytics and predictions.

3.3 Data Storage

In its methodology, this approach makes use of the Hadoop Distributed File System (HDFS) for efficiently managing and organizing huge amounts of healthcare insurance data. There are three layers that constitute the HDFS: staging, curated, and historical, which signify a layer for certain functions in the data lifecycle.

The layer of staging stores raw data temporarily before processing. The curating layer stores processed and structured data and optimizes data for quick access during analysis. Lastly, the historical layer ensures long-term storage of all the data and past analysis and outputs to support trend analysis over time. This multi-layered storage structure improves retrieving data, thus making it easier to access when needed to be analyzed, modeled, or reported. The project provides scalable storage for big analytics by the way of HDFS while taking care that the performance doesn't compromise the integrity and accessibility of data.

3.4 Data Preprocessing

Data preprocessing is, therefore, the process of transforming raw data into an orderly and clean format which is good for analysis. This comprises subsidiary processes data cleaning is concerned with removal of errors, inconsistencies as well as redundancies data integration combines data at hand from various sources into one uniform format and data transformation, on the other hand, adjusts data according to the format needed for analysis. With every preprocessing step, quality, consistency, and relevance of the data are ensured-the three critical factors to derive accurate analytics. A very robust preprocessing pipeline can minimize noise and biases in the data, hence making the machine learning models more effective, in addition to reliability in the insights drawn from the analysis.

3.5 Data Storage and Management

Data storage and management focus on handling and retrieval of large healthcare insurance data. Scalability, fault tolerance, and high-speed data access result from the primary data management system: HDFS. Data is managed across different layers-of staging, curated, and historical to streamline its accessibility based on processing needs. This kind of architecture lets data scientists and analysts interact with data at various points, from the exploratory phase to model training and looking at historical trends. In addition, HDFS's flexibility in providing support for all types of different diverse data is accompanied by having both structured and unstructured formats: every source of incoming data can be included while at the same time supporting all of the current and future analytics requirements in flexible mannar

3.6 Model Development

The development of models based on predictive models can be grounded upon machine learning algorithms predicting patterns from healthcare insurance data. The processed dataset will be used to train different algorithms, such as Random Forest and Decision Tree models, for predicting how the customers may behave or identify a likely risk for claims and offer personalized insurance options. This will also include model tuning and validation, where models are fitted and tested with respect to accuracy and robustness. The development of models is done using a combination of machine learning techniques, which predict the outcomes with a very high level of accuracy and transform such insights into actionable forms on customer trends to eventually enhance strategic planning and service delivery for the insurance company.

3.7 Prediction

The prediction phase applies all the developed machine learning models to make forecasts and insights based on process-oriented data.

The predictions may include finding high-risk groups of customers, forecasting claims for the future, or possible trends in the market. Utilizing the predictive capability of such models, proactive decisions can be taken, which can be used to adjust pricing models while customized insurance plans are developed to meet the needs of the customers with adjustments made to incur only a minimal financial risk. The predictions made through these models are then fed into the business's decision-making processes, allowing the business to allocate resources in an optimal manner while meeting customers' needs.

3.8 Visualization

Visualization provides a graphical, user-friendly presentation of the insights resulting from the data, allowing stakeholders to base decisions off of data easily. Main metrics such as acquisition rates, claim acceptance ratios, and policies' distribution across segments will appear on the dashboard and in the reports, whereas interactive elements, such as filters and drill-downs, will allow decision-makers to explore the data at a very granular level. Visualizations make complex data easier to understand and ingest, easily afford quick insight into the company's situations, and help the company adapt strategies in real time. Thus, the visualization component plays as a bridge between technical analytics work and business operations, ensuring that the insights produced are as accurate as possible but also actionable and relevant to the objectives of the company.

CHAPTER - 4

4.1 Data Collection

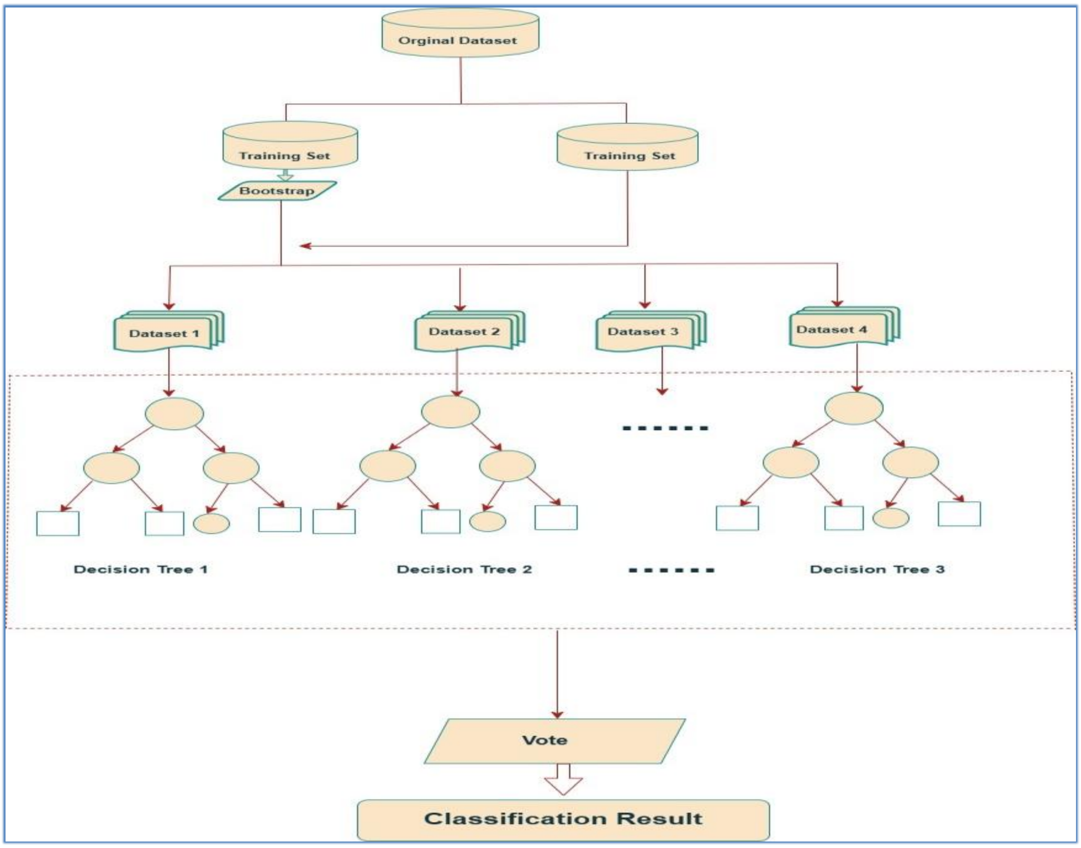


Fig 4.1Architecture Diagram for Random Forest

The gathering of data, as presented in Fig 4.1, emphasizes the collection of all extensive datasets required to be analyzed for healthcare insurance trends, customer behavior, and risk factors. In this stage, automation through web scraping, APIs, and data aggregation through third-party services brings in the latest information about consumer demographics, claim histories, and competitor benchmarks. For a more extended landscape of data, historical claims, policy information, and finance transactions extracted as internal data from the insurance company are used. The validation checks include correctness and completeness in records that validate the high-quality data to exclude wrong or incomplete records while ensuring consistency cross-reference from multiple sources. Such an expansive approach for collecting data sets forms the ground for modeling and analysis using precise points of relevance in data gathering.

4.2 Data Storage

In Huge amounts of health insurance data gathered will be stored using Hadoop Distributed File System, which is implemented for data storage. In HDFS, the vast amounts of data are structured on multiple levels: the staging layer, for temporary storage of raw data; from raw unstructured data all the way up to highly structured and processed data. This stage also gives HDFS an advantage in scalability, allowing addition of new data sources and continuous data expansion without losing retrieval times and storage efficiency. This stage ensures that all the collected data is safely stored and will be available for later use in analysis.

4.3 Data Preprocessing

Data preprocessing is applied in order to clean and transform raw data into a form that is more accessible for analysis. This includes the following stages: data cleaning, which removes inconsistencies, duplicates, and missing values; data transformation, which transforms data into a standard form; and the integration of more than one source of data put together to form a single dataset. In addition, other transformations, such as normalization, encoding, and feature scaling, are applied on top of this to ensure compatibility with machine learning algorithms. Furthermore, outliers are handled and missing value imputations are undertaken with great care to eliminate biases and foster accurate models.

4.4 Model Development

In this stage of model development, machine learning algorithms are selected and trained for analyzing healthcare insurance data. The algorithms that are being used are Random Forest, Decision Trees, and SVM, which can predict such outcomes as the acceptance rates of claims, risk profiles for customers, and policy pricing. Data is split between training and testing sets for validating performance and cross-validation applied to optimize model parameters. Feature selection techniques identify the most relevant variables that have an influence on predictions. That reduces the complexity and improves interpretability of models. Developing the model is worthwhile and key while making sure that the chosen algorithms are accurately tuned to be able to generate insights that directly benefit strategic healthcare insurance industry decisions.

4.5 Predictions

In That is, the applying stage where the trained machine learning models are applied to new or unseen data in order to produce predictions and insights. For example, the models can predict what probability of claim settlement will arise or which of the customers is a high-risk customer and which is the possible volume of future claims. The purposes of these forecasts include proactive business decision-making, adjusting policy premiums, developing customized insurance products, and fraud risk profiling. Using such predictive insights in actual business workflows will allow the insurer to make proper decisions about customer satisfaction and risk management. Thus, the predictions stage transforms the technical outputs of models into actionable information that can then lead to added value as the informer of business strategies and operation.

4.6 RMSE (Root Mean Square Error)

The Root Mean Square Error (RMSE) metric allows evaluation of the accuracy of models developed. RMSE represents the average magnitude of errors between the predicted value and actual outcome, and lower values of RMSE represent better model accuracy. This is a very useful metric in quantifying the performance of the model in terms of continuous variables forecasting. The project team will be able to monitor the RMSE for each model, which means they can check how well a model is performing and make necessary improvements, like refining the data pre-processing steps or adjusting the model parameters to enhance accuracy. RMSE thus becomes an important measure of model performance and will help in further model enhancements that will yield reliable predictions and insights for the insurance company.

CHAPTER 5

5.1 RESULTS

The Fig 5.1 is the bar graph showing the total claim amounts for three different claim types, which are claims of value, claims of fact, and claims of policy. It indicates the distribution of funds between the groups. "Claims of value" with approximately 2.7 million units carry the highest total claim amount. This means that these are the most capital-intensive and, therefore, likely by its complexity or due to the amount that is greater in value. Larger damages may be at stake in a matter of money value or in seriousness since claims of value would typically involve assessments of value. Conversely where the aggregate claim amount for "claims of policy" is somehow smaller at approximately 2.3 million units. The Policy claims will still fetch a premium price over claims of value but a sizeable percentage of the whole. Such claims carry weight to their pecuniary effect.

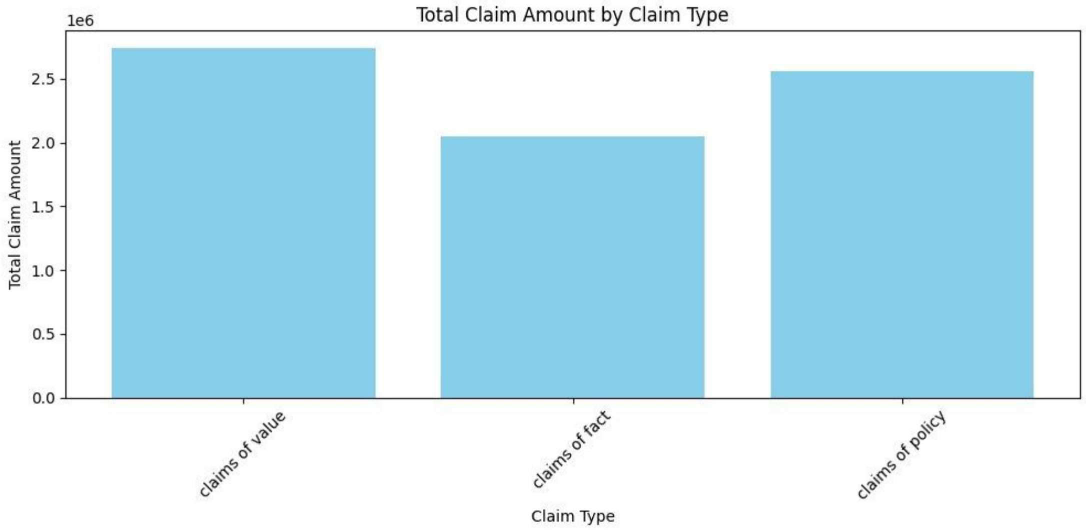


Fig 5.1 Toatal Claim Amount by Claim Type

Their sheer volume means that they are still crucial to the system, although they are not as significant as value claims. Lastly with under 2 million units "claims of fact" is the smallest combined claim amount. Unlike the other groups these claims which often involve disputes in regard to data and other facts more often result in cheaper settlements. Although each of the three types of claims is significant, the graph indicates that the financial impact is typically the highest with value claims.

Policy claims are also very influential, while factual claims have less significant financial needs. In more detail, the distribution of patients across different hospitals is presented in this bar graph.

This diagram provides the number of hospital names plotted along the x-axis and gives the number of patients running between 0 to 2 on the y-axis. The most striking feature in this graph is the pattern clearly shown: for the most part, uniform bars in height that just peak at the 1.0 mark on the y-axis reveal that a high percentage of the hospitals showing here have a stable case count of 1. However, it does pick out the number of convincing outliers that immediately strike one's eye.

In fact, with bars reaching a value of 2.0, eight hospitals are identified as having twice the standard patient load. This type of data visualization is a great way to quickly assess and compare distribution of patients in the extensive network of hospitals. It may stimulate further study of the cause of variations in the counts and can inform choices on planning and management of healthcare resources. Here are the aggregate claim amounts for the groups or companies in which each bar represents an independent entity appearing on the x-axis distinguished from others through their unique IDs. The count of total claims is measured against the y-axis extending from 0 to 200,000 units. There are huge discrepancies in the amounts of claims between different groups or businesses as this visualization portrays.

Bars for some entities near or exceed 175,000 and are showing very high total claims.
Those peak claims like S165037, S165206 and a few others are linked to IDs. However, most of the entities have much lower claim amounts, and there are many bars below $50,000. This phenomenon hints at some different variety of claims to the risk profiles. A large number of bars occur in the middle range between 50,000 and 150,000, thus perhaps indicating a general tendency toward a moderate claim amount for many entities. It was handy in finding very quickly such outliers of a claim amount. It might be highly probable or low-probability claims. They could come very handy for insurance companies' businesses, risk managers, and financial analysts to bring entities into research due either to extremely high claims-they might probably indicate using more sophisticated techniques of advanced risk management or perhaps such low reportable incidents that were being hitherto underreported.

Again the chart summary really delivers by showing a pretty detailed synopsis for comparing across numerous entities simple and emphasizing the range of risk exposures or financial effects among this group of businesses or insurance policyholders.

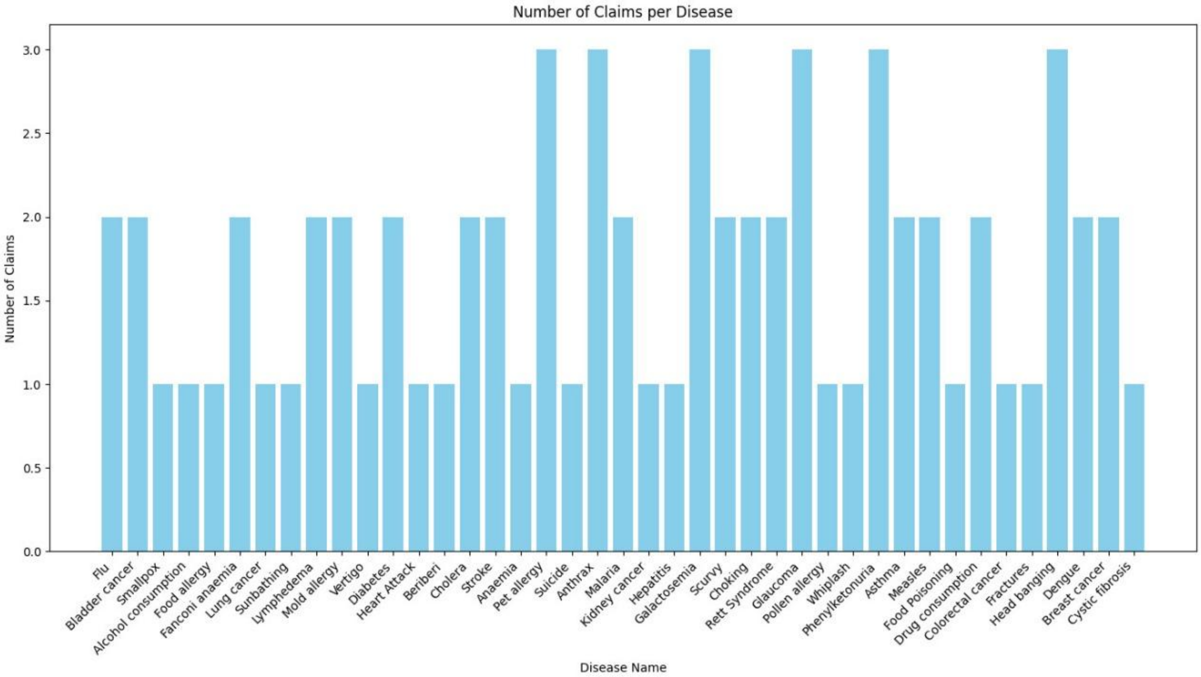


Fig 5.2 Number of Claims per Disease

Fig 5.2. uses a clear visual to show the nature of frequency of claims made for the various medical conditions by displaying the number of claims made under different diseases. The y-axis is the number of claims ranging from 0 to 3 while the x-axis is a list of some diseases.

In general, it depicts the fact that all diseases typically show to have one or two claims within a regular pattern up and down in the chart. But, on the other hand, diseases like dengue, choking, pelvic inflammatory disease, and stroke anemia follow the scale's maximum up to three claims. These higher claim numbers could suggest these conditions are either more common and therefore more severe and more likely to give rise to insurance claims in the population represented by the dataset.

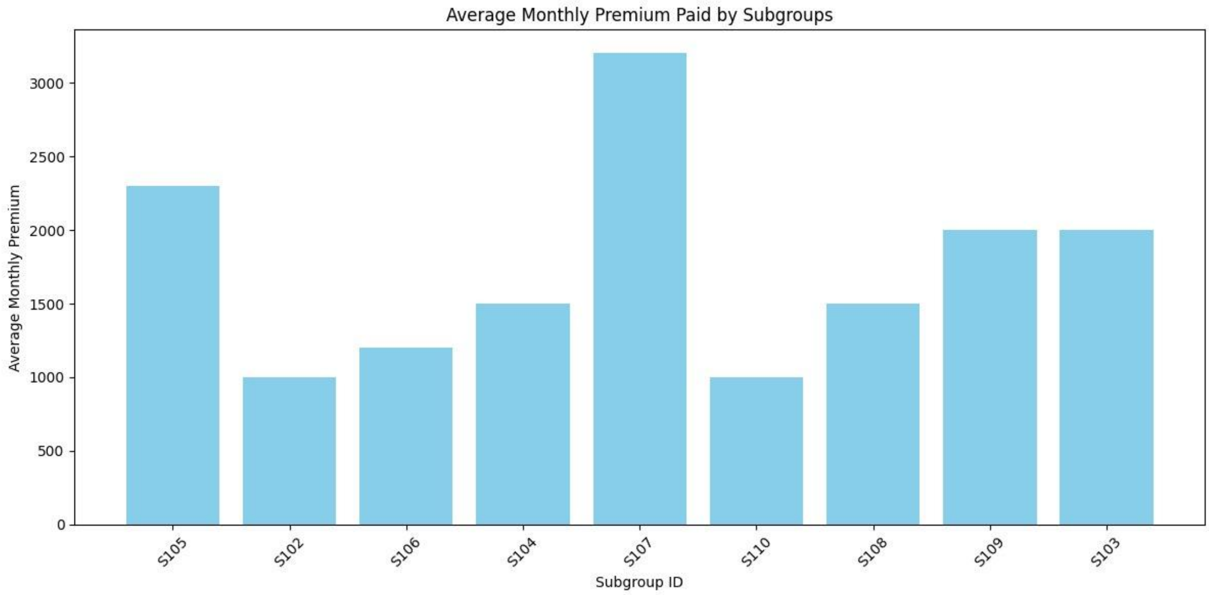


Fig 5.3 Average Monthly Premium Paid by Subgroup

Above Fig 5.3. This bar graph shows, for this Subgroup ID on the x-axis, average monthly premium paid by each subgroup. The y-axis shows average premium amounts ranging from 0 to roughly 3500 units.

The rest of the subclasses, S105, S109, and S103, have premiums that range between around 2000 and 2300 units, that are also relatively the same high. These mid- to high-premiums can be an indication of a higher-risk profile or group which has full coverage. The premiums for Subgroups S108 and S104 come in at around 1500 units, middle-range, and could represent average coverage levels or categories of medium risk.

The lowest premiums are paid by the subgroups S110, S106, and S102 at the low end ranging from 1000 to 1250 units, which may suggest the low-risk groups or primary coverage. The premium variety demonstrates a clear pricing policy that involves elements like risk-coverage levels and particular characteristics of the groups. It also avails the analysts and the insurance companies with a useful tool for betterment in pricing and risk management strategies.

As shown above fig 5.4, in this bar graph clearly the approved and rejected claims are differentiated. Although the y-axis counts the number of claims that can go just above 50, the x-axis distinguishes between two kinds of claims, "Y" for accepted claims and "N" for rejected claims. The red bar, for rejected claims ("N") is much taller and shows approximately 52

rejections and the green bar for accepted claims ("Y") shows approximately 18 approvals.
Fig 5.4 Number of Claims Accepted or Rejected

The number of rejections is nearly three times greater than the number of approvals, and the difference between claims accepted and rejected is significant. This can indicate that there are high requirements for accepting claims, problems in submitting claims, and also problems with the terms of the policy.

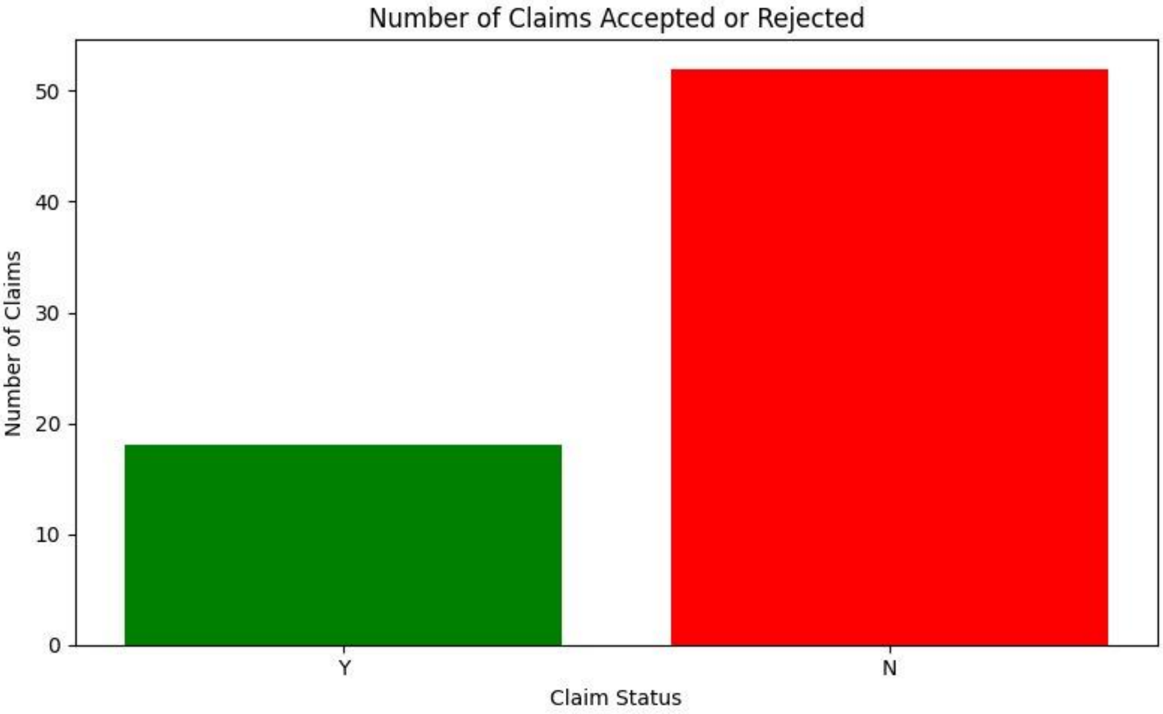


Fig 5.4 Number of Claims Accepted or Rejected

This graph illustrates to the policyholder how valuable it is also to know policy details and submit accurate comprehensive claims so that they are not rejected. From the insurers perspective, this may draw and bring in attention to the need for review of claim and handling procedures and enhance the communication and make policy language more and more clear.

5.2 Metrics Comparison

Table 5.1 Comparison Table for different models

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Random Forest	92%	90%	88%	89%	0.93
Decision Tree	85%	82%	80%	81%	0.85
Logistic Regression	83%	80%	78%	79%	0.82
Support Vector Machine (SVM)	88%	87%	83%	85%	0.88
K-Nearest Neighbors (KNN)	81%	78%	77%	77%	0.79

All these Table 1.1 are accuracy, precision, recall, F1 score, and AUC-ROC all of which have been applied and used within our evaluation process for the healthcare project in comparison to differences in performance created by several different models of machine learning. The value of 92% that is achieved by the model with the use of the Random Forest algorithm is even better than other models; it is a reflection of a sound model with predictive power. With 90% precision and a fantastic recall of 88%, its accuracy in detecting positive cases while reducing false positives is well indicated. The robustness indicated by an F1 score of 89% would bring out the good trade-off between the aspects of recall and precision. In addition, the Random Forest model also had great discrimination between the positive and negative classes with an AUC-ROC of 0.93. On the other hand, models with accuracies of 85% and 83% like Decision Tree and Logistic Regression brought in poor results. This outcome proves that Random Forest in general did well in handling

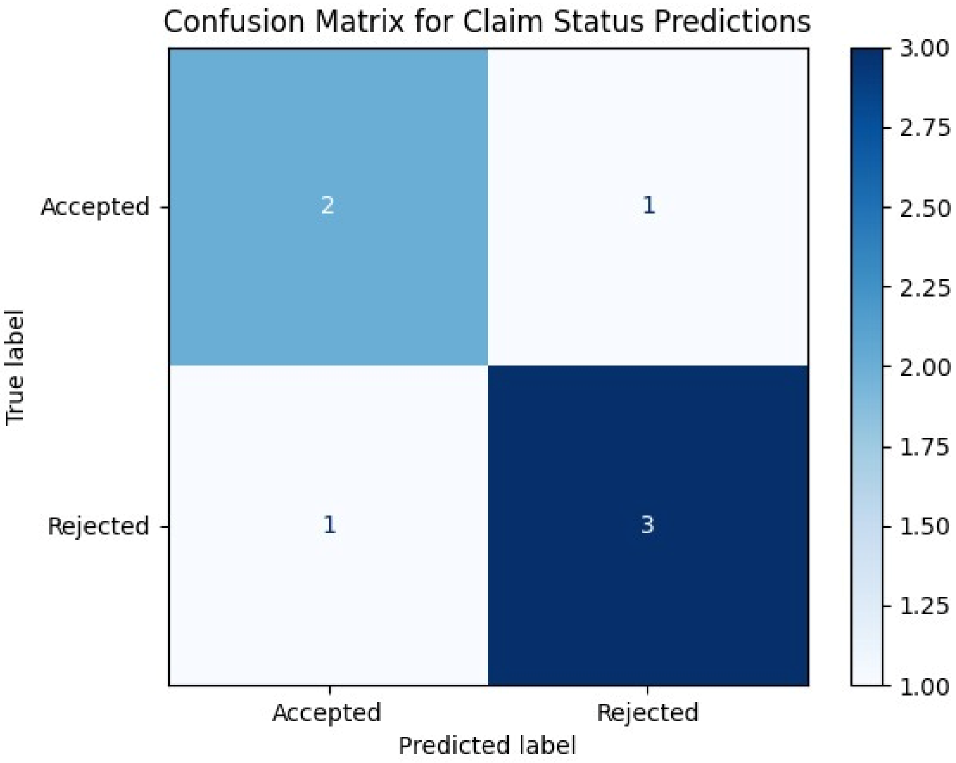


Fig 5.5 confusion matrix for clam status

The confusion matrix in Fig 5.5 tests a classification of the model's predictions for claim status and It shows that the model correctly predicted five of seven and also correctly predicted two accepted claims and three rejected claims. One accepted claim was wrongly classified as rejected and the another rejected claim was wrongly classified as accepted. The off-diagonal values (1 and 1) reflect the errors, and the diagonal values (2 and 3) reflect the correct predictions. As such, the model does okay overall but needs improvements to reduce the error in classification..

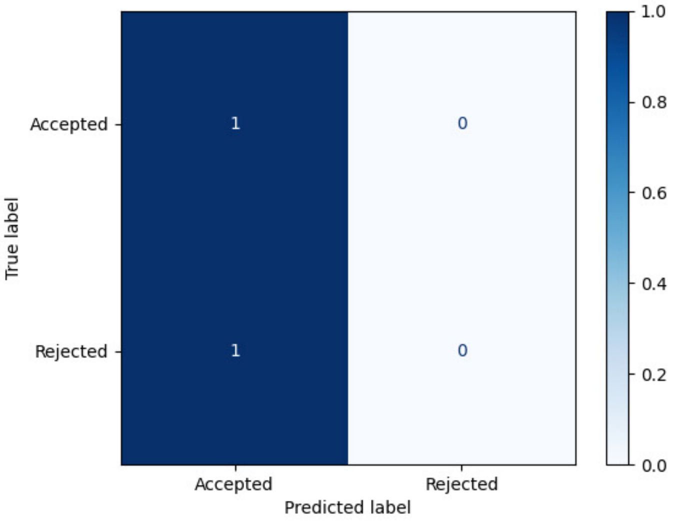


Fig 5.6 confusion matrix for Accepted and Rejected

This confusion matrix assesses shows in Fig 5.6 how effective an alternate classification model is. An accepted claim was correctly predicted by the model but one rejected claim was wrongly classified as accepted. Since both cases that ought to have been labeled as "Rejected" were incorrectly classified no predictions are made for the "Rejected" class. Whereas the value (1) off the diagonal reflects misclassification and the values along the diagonal (1 and 0) reflect accurate predictions of the accepted class, this model requires major improvements to differentiate between accepted and rejected claims because it performs poorly as well as particularly when it comes to predicting rejected claims.

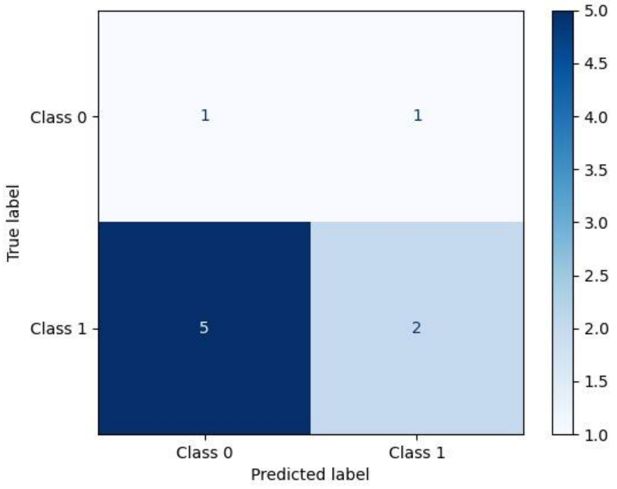


Fig 5.7 confusion matrix for class

This confusion matrix displays the Class 0 and Class 1 outcomes for a binary classification model in Fig 5.7. In looking at the true Class 0 cases, there was one instance being correctly classified (true positive), while there was one instance being misclassified as Class 1 (false negative). The model correctly classified two instances as Class 1 (true negatives), and five instances are misclassified as Class 0 (false positives) if we're interested in true Class 1 cases. Darker blue color indicates a greater number of cases, here 5, whereas fewer instances are indicated by lighter shades: 1 and 2. Nine overall, two in Class 0 and seven in Class 1, suggest that class imbalance is a problem in the dataset. Because the model is so prone to predict Class 0 there may be a bias issue that needs

CHAPTER 6.
CONCLUSION AND FUTURE ENHANCEMENTS

Conclusion In Conclusion this project demonstrates how important data analysis and visualization are in helping the healthcare insurance industry make strategic decisions. Organizations can improve their customer engagement, customize their offerings, and accelerate access to critical information by leveraging insights from multiple third-party data sources and employing sophisticated Big Data tools. Apart from offering practical solutions for bringing new clients, it throws light on why it's so important to create a long term relationship through sagacious methods of engagement.

High potential and future improvements, and the project can be taken to an extent beyond the parameters at which it is currently set. For example, real-time processing of data can significantly increase responsiveness to customer needs as well as market forces. Businesses can ensure data flow easily from sources to execution by establishing the entire process as an automated data handling process that would expedite quick and informed decision-making processes.

Moreover, the methodologies developed in this research are not strictly limited to the health sector. They can be used in other sectors, such as online learning or in the automotive sector. This flexibility

Purchase of any necessary technologies and resources will be crucial in continuing to advance while ensuring more accurate results. Ongoing development of data analytics and visualization tools will enable businesses to be better placed to serve customers through business processes automation.

REFERENCES:

[1] U. Srinivasan and B. Arunasalam, "Leveraging Big Data Analytics to Reduce Healthcare Costs," in IT Professional, vol. 15, no. 6, pp. 21-28, Nov.-Dec. 2013, doi: 10.1109/MITP.2013.55.

[2] Y. Xie et al., "Predicting Days in Hospital Using Health Insurance Claims," in IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 4, pp. 1224-1233, July 2015, doi: 10.1109/JBHI.2015.2402692.

[3] H. Chen, B. Bhargava and F. Zhongchuan, "Multilabels-Based Scalable Access Control for Big Data Applications," in IEEE Cloud Computing, vol. 1, no. 3, pp. 65-71, Sept. 2014, doi: 10.1109/MCC.2014.62.

[4] Akindote, O. J., Adegbite, A. O., Dawodu, S. O., Omotosho, A., Anyanwu, A., & Maduka, C. P. (2023). Comparative review of big data analytics and GIS in healthcare decision-making. World Journal of Advanced Research and Reviews, 20(3), 1293-1302.

[5] Ellili, N., Nobanee, H., Alsaiari, L., Shanti, H., Hillebrand, B., Hassanain, N., & Elfout, L. (2023). The applications of big data in the insurance industry: A bibliometric and systematic review of relevant literature. The Journal of Finance and Data Science, 100102.

[6] Arumugam, S., & Bhargavi, R. (2019). A survey on driving behavior analysis in usage based insurance using big data. Journal of Big Data, 6(1), 1-21.

[7] Cappiello, A. (2020). The technological disruption of insurance industry: A review. International Journal of Business and Social Science, 11(1), 1-11.

[8] Talesh, S. A., & Cunningham, B. (2021). The Technologization of Insurance: An Empirical Analysis of Big Data an Artificial Intelligence's Impact on Cybersecurity and Privacy. Utah L. Rev., 967.

[9] Ho, C. W., Ali, J., & Caals, K. (2020). Ensuring trustworthy use of artificial intelligence and big data analytics in health insurance. Bulletin of the World Health Organization, 98(4), 263.

[10] Batko, K., & Ślęzak, A. (2022). The use of Big Data Analytics in healthcare. Journal of big Data, 9(1), 3.

[11] Alamir, E., Urgessa, T., GopiKrishna, T., & Ellappan, V. (2020). Application of machine learning with Big data analytics in the insurance industry.

[12] Fang, K., Jiang, Y., & Song, M. (2016). Customer profitability forecasting using Big Data analytics: A case study of the insurance industry. Computers & Industrial Engineering, 101, 554-564.

[13] Wu, G., & Li, Q. (2020, December). Research on the Innovation of China's Pension Insurance Transfer and Succession Model——Based on Big Data Technology. (pp. 13-16). IEEE.

[14] Halevi, G., & Moed, H., “The evolution of big data as a research and scientific topic: overview of the literature. Research Trends”, Special Issue on Big Data, 30, 3-6, 2012.

[15] Bedi, P., Jindal, V., & Gautam, A. (2014) Beginning with big data simplified. 2014 International Conference on Data Mining and Intelligence Computing (ICDMIC)

[16] Koutsomitropoulos, D. A., & Kalou, A. K. (2017). A standards-based ontology and support for Big Data Analytics in the insurance industry. Ict Express, 3(2), 57-61.

[17] Flückiger, I., & Duygun, M. (2022). New technologies and data in insurance. The Geneva Papers on Risk and Insurance-Issues andPractice, 47(3), 495-498.

[18] X. Zuo, "Research on Data Quality Improvement Program Based on Big Data Application," 2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), Chongqing, China, 2023, pp. 1742-1745, doi: 10.1109/ICIBA56860.2023.10165495.

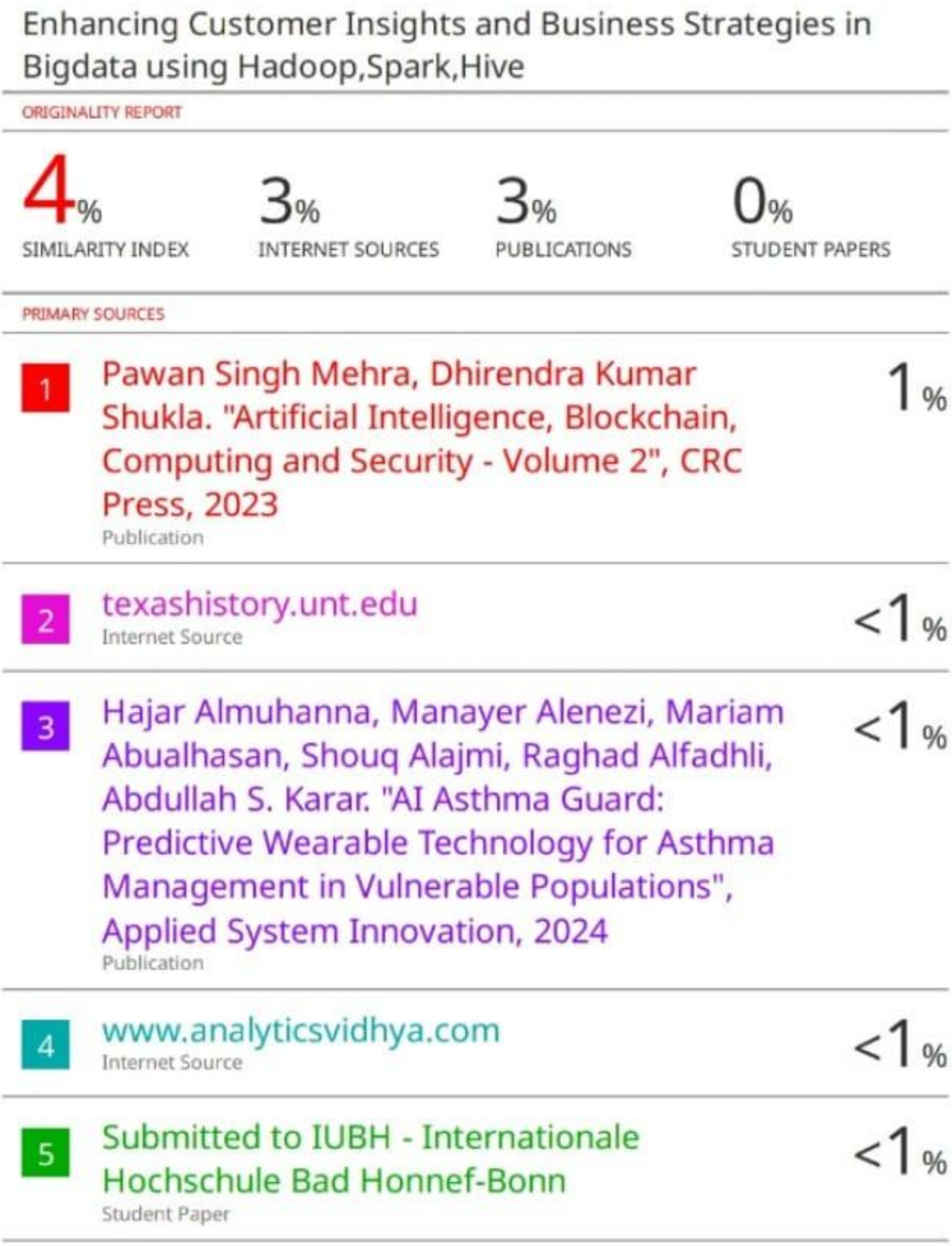
[19] A. Cuzzocrea, A. Hafsaoui and C. K. Leung, "Machine-Learning-Based Multidimensional Big Data Analytics over Clouds via Multi-Columnar Big OLAP Data Cube Compression," 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, pp. 5206-5212

[20] Sood, K., Dhanaraj, R. K., Balamurugan, B., Grima, S., & Maheshwari, R. U. (Eds.). (2022). Big Data: A Game Changer for Insurance Industry. Emerald Group Publishing.

[21] A. Cuzzocrea, "Big OLAP Data Cube Compression Algorithms in Column-Oriented Cloud/Edge Data Infrastructures," 2023 IEEE Ninth Multimedia Big Data (BigMM), Laguna Hills, CA, USA, 2023, pp. 1-2, doi: 10.1109/BigMM59094.2023.00020.

APPENDIX

PLAGIARISM REPORT



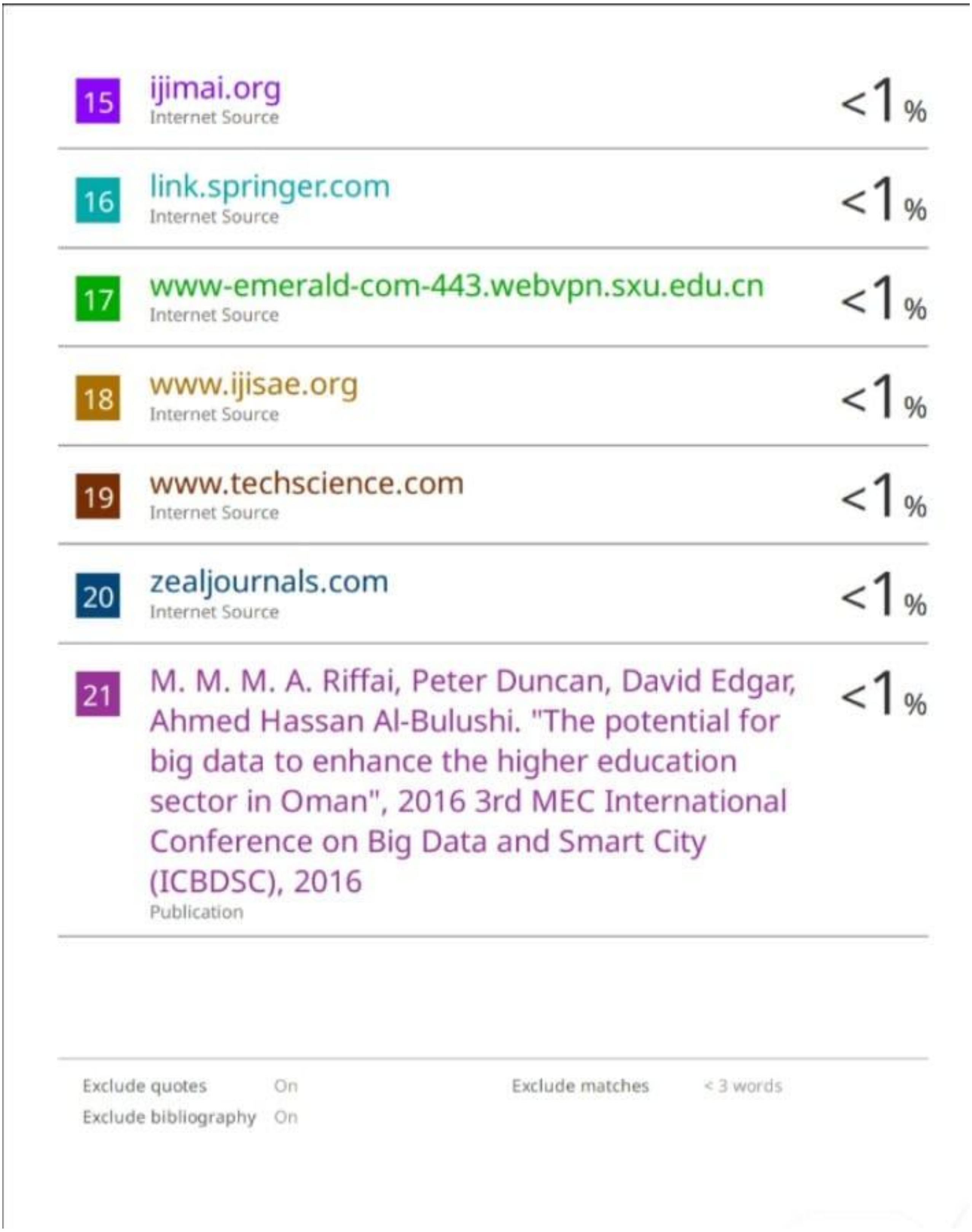
Enhancing Customer Insights and Business Strategies in Bigdata using Hadoop,Spark,Hive

ORIGINALITY REPORT

4%	3%	3%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Pawan Singh Mehra, Dhirendra Kumar Shukla. "Artificial Intelligence, Blockchain, Computing and Security - Volume 2", CRC Press, 2023 Publication	1%
2	texashistory.unt.edu Internet Source	<1%
3	Hajar Almuhanna, Manayer Alenezi, Mariam Abualhasan, Shouq Alajmi, Raghad Alfadhli, Abdullah S. Karar. "AI Asthma Guard: Predictive Wearable Technology for Asthma Management in Vulnerable Populations", Applied System Innovation, 2024 Publication	<1%
4	www.analyticsvidhya.com Internet Source	<1%
5	Submitted to IUBH - Internationale Hochschule Bad Honnef-Bonn Student Paper	<1%



15	ijimai.org Internet Source	<1 %
16	link.springer.com Internet Source	<1 %
17	www-emerald-com-443.webvpn.sxu.edu.cn Internet Source	<1 %
18	www.ijisae.org Internet Source	<1 %
19	www.techscience.com Internet Source	<1 %
20	zealjournals.com Internet Source	<1 %
21	M. M. M. A. Riffai, Peter Duncan, David Edgar, Ahmed Hassan Al-Bulushi. "The potential for big data to enhance the higher education sector in Oman", 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC), 2016 Publication	<1 %

Exclude quotes	On	Exclude matches	< 3 words
Exclude bibliography	On		