Exploring the Potential of Predictive Analytics Using Big Data in Healthcare Industry

Vedant Marwadi
Data Warehousing and Big Data
Master of Computer and Information Sciences
Auckland University of Technology

***Abstract*—The healthcare industry is dealing with a massive surge of data, creating a pressing need for advanced analytics solutions. This study explores the application of predictive analytics in healthcare, focusing on classifying patient admissions as elective, urgent, or emergency. Using big data techniques, the research utilized Apache Hive for data exploration. Decision Tree and Random Forest were subsequently employed to construct predictive models. The accuracy attained by the Decision Tree model was 88.17%, while the Random Forest classifier demonstrated a slightly higher accuracy of 88.91%. Both models exhibited promising capabilities in accurately predicting patient admission categories. The findings highlight the potential of predictive analytics to improve hospital resource management, reduced wait times, and better patient outcomes. This research supports the growing consensus on the revolutionary impact of predictive analytics in the healthcare industry.**

***Index Terms*—Predictive Analytics, Machine Learning, Decision Tree, Random Forest, Patient Admissions**

I. Introduction

The recent years have witnessed a notable rise in data production within the healthcare industry, largely fueled by the rapid adoption of digital technologies [1]. This shift has resulted in an unprecedented volume of data produced daily, from patient records to operational metrics, fundamentally transforming how healthcare is managed [2]. This explosion of data is linked to the idea of big data. It consists of a large amount of both well-structured and miscellaneous data that traditional data processing software finds challenging to manage effectively [3].

In healthcare settings, large data sets consist of a range of information types, including health records, medical images, genetic data, and much more [4]. When leveraged effectively, this wealth of data can provide valuable insights through predictive analytics [5]. Predictive analytics is a powerful application of big data [6]. It is introduced as a transformative approach that harnesses past records and current data to project future happenings [7]. The process of predictive analytics is illustrated in the figure 1.

Breakthroughs in technology such as machine learning, data processing, and Internet of Things are considered fundamental enablers of this methodology [7]. By applying machine learning frameworks and statistical methods, healthcare providers can predict patient admissions. They

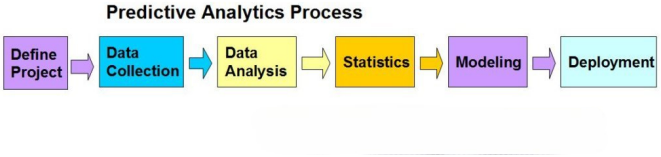

Fig. 1. The process of predictive analytics. Adapted from [8].

can also anticipate disease outbreaks and forecast treatment responses. Such capabilities enhance patient care standards and improve the healthcare facility operations [9].

This research focuses on the application of big data and predictive analytics to classify patient admissions into three distinct categories: elective, urgent, or emergency, using historical patient data obtained from Kaggle. To achieve this, the study employs Decision Tree and Random Forest algorithms due to their inherent advantages. Decision trees are straightforward and intuitive, as they simulate the way humans make decisions by splitting the data into subsets based on the most significant predictor variables, thereby constructing a hierarchical tree structure [10]. Random forests, on the other hand, improve upon Decision Trees by generating an ensemble of multiple trees during the training process and aggregating their predictions to produce a more robust and accurate outcome [11].

In addition to model development, the research utilizes Apache Hive for data exploration. Hive functions as a Hadoop-based data warehouse solution that supports the effective retrieval and evaluation of large, distributed datasets [12]. It provides a SQL-like interface to process vast amounts of data, making it a powerful tool for data preprocessing, exploration, and aggregation [13]. By integrating Hive with predictive analytics, the study ensures a scalable and efficient workflow for handling extensive healthcare dataset, enabling comprehensive data exploration before model training.

Through the combined use of Decision Tree and Random Forest algorithms, alongside Hive for data exploration, this study seeks to provide a reliable tool for predicting patient admission types. The findings of this research can potentially contribute to more effective resource management, reduced

wait times, and improved patient outcomes, thereby improving overall healthcare efficiency. Furthermore, it can also serve as a foundation for future studies in predictive healthcare analytics, promoting the development of data-driven strategies for operational improvements and patient care.

II. Literature Review

The traditional healthcare approach waits for patients to become ill before providing care, resulting in late and uncontrolled diagnoses of disease progression [14]. However, with the introduction of big data and predictive analytics, this paradigm is shifting [7]. Instead of waiting for symptoms to appear, healthcare providers can now utilize these advanced technologies to predict disease onset, track patient health trends, and intervene before conditions worsen [15]. For instance, Dinov and colleagues demonstrated how big data-driven predictive analytics can precisely forecast the progression of Parkinson's disease [16]. Likewise, Sharma and the team developed the BHARAT framework to identify Alzheimer's disease at an early stage. It achieved impressive results in just five minutes using PySpark and deep learning techniques [17].

Further illustrating its transformative potential, predictive analytics has shown remarkable success in predicting numerous medical diseases. As an example, Koppad and Kumar used the J48 Decision Tree algorithm to predict chronic obstructive pulmonary disease with precision of 91. 22% [18]. In addition, Chen and his fellow researchers conducted an experiment to predict the risk of cerebral infarction by analyzing big data from healthcare communities in central China. They introduced a novel convolutional neural network based approach, which demonstrated a high prediction accuracy of 94.8% [19]. For diabetes management, Kumar and peers utilized Hadoop and MapReduce to predict diabetes types, complications, and treatments, especially in remote areas [20]. Similarly, Rallapalli and Suryakanthi used a Hadoop-configured environment and a MapReduce-based Random Forest algorithm for diabetes risk assessment, achieving an accuracy of 87.5% [21]. Furthermore, Mir and Dhage tested various machine learning algorithms in the "Pima Indians Diabetes Database", finding that Support Vector Machine achieved the highest precision at 79.13% [22].

In cardiovascular care, Ismail and colleagues developed a heart disease prediction model using a Support Vector Machine classifier on Microsoft Azure, achieving a 90.6% accuracy rate and real-time alerts from wearable devices [23]. Additionally, Ustun and co-researchers used big data analytics and the Supersparse Linear Integer Models to improve predictions for Obstructive Sleep Apnea [24]. Personalized disease risk prediction is another emerging approach in healthcare. Chawla and Davis contributed to this field by introducing Collaborative Assessment and Recommendation Engine (CARE) and its iterative version, Iterative Collaborative Assessment and Recommendation Engine (ICARE). These models use collaborative filtering to predict future diseases based on patient similarities [25].

Neurocritical care also benefits from big data and predictive techniques. Alkhachroum and colleagues used these methods to manage patients with aneurysmal subarachnoid hemorrhage, employing multimodal monitoring and machine learning to customize personal treatments and care [26]. Moreover, these analytical approaches have been used to improve operational aspects of healthcare too. Harris and his team developed a model to optimize outpatient clinic scheduling by forecasting patient no-shows. By analyzing historical attendance data and employing regression and exponential functions, they demonstrated how past behavior influences future attendance patterns [27]. Another noteworthy contribution comes from Chennamsetty and research team, who developed a Hive-based data management system to process and analyze electronic health records from 1000 hypothetical patients. This study focused on chronic conditions such as diabetes, blood pressure, and cholesterol levels. By leveraging Hadoop's scalability and Hive's SQL-like querying capabilities, they generated predictive analytics reports to evaluate the effectiveness of medications like Lipitor and Onglyza [28].

The adoption of predictive analytics in healthcare, however, is not without its challenges. One of the foremost issues is that, healthcare data comes from a variety of sources [29]. Combining data from these diverse sources is technically demanding [7]. In addition, given the highly sensitive and confidential nature of healthcare data, robust privacy and security measures are essential. However, even with these security measures in place, the persistent threat of data breaches and cyberattacks still remain a substantial challenge [29]. Big data's role in healthcare also brings forth ethical issues, especially in relation to patient consent and the ownership of data [30].

Furthermore, healthcare institutions that are venturing into predictive analytics using big data must acknowledge the significant upfront costs associated with this endeavor. These costs include not only the technology and infrastructure necessary for data acquisition, but also essential investments in training and development programs [29]. Many healthcare organizations also grapple with challenges stemming from inadequate infrastructure, which complicates the integration of new technologies into existing healthcare systems [30].

In conclusion, although there are several hurdles to address, the potential benefits of implementing predictive analytics in healthcare are considerable and far-reaching. Improved disease prediction, personalized treatment plans, and enhanced operational efficiency are just a few of the numerous ways in which big data and predictive analytics can significantly transform healthcare delivery. By harnessing the power of these technologies, healthcare providers can not only anticipate

patient needs with greater precision but also optimize resource allocation and reduce overall costs.

III. Opinion

Reflecting on recent advances, I can see that predictive analytics using big data offers tremendous potential. The integration of this technology into healthcare signifies a major shift from reactive to proactive care. Studies such as those by Dinov and Sharma show how the use of big data and machine learning can predict diseases like Parkinson's and Alzheimer's with high precision. Envisioning a future where healthcare care manages diseases before they fully manifest and offers care tailored to each patient's unique profile is both exciting and promising to me. Such a proactive strategy could greatly lower the occurrence of chronic rates and improve patient health outcomes.

Looking ahead, I believe that the continuous evolution of predictive analytics promises even greater advancements. As technology becomes more sophisticated and accessible, I anticipate its integration into everyday healthcare practices will increase significantly, transforming the landscape of patient care. Additionally, I also find that predictive analytics will enhance the operational efficiency of hospitals. Personalized medicine is another exciting frontier, and I think frameworks like CARE and ICARE will allow for individualized risk assessments, making treatment strategies more effective and patient specific.

In many parts of the world, access to quality healthcare is limited due to geographical, financial, and infrastructural constraints. I see predictive analytics as a bridge for these gaps, providing valuable insights that help in early disease detection and timely intervention, even in resource-limited settings. For instance, telemedicine platforms enhanced with predictive analytics can offer remote diagnostics and personalized treatment plans, reducing the need for physical visits and making healthcare more accessible to remote populations.

Nonetheless, I recognize that achieving this potential demands a thoughtful balance of ethical, technical, and practical factors to build a more resilient healthcare system. From my point of view, it is important to tackle ethical issues like patient consent and the privacy of data to ensure that predictive analytics is used responsibly. Ensuring transparency in how predictive models are developed and applied will be crucial in building trust among patients and healthcare professionals.

Furthermore, in my opinion, to handle technical challenges effectively, continuous education and training programs are essential for equipping healthcare professionals with the skills needed to interpret and act on predictive insights. I also believe that collaboration between technology experts and healthcare practitioners is vital for developing user-friendly tools that seamlessly integrate into clinical workflows. Moreover, while practical considerations such as the cost of implementing predictive analytics solutions are important, I acknowledge that the initial investment may be high. However, I argue that ongoing benefits, such as better patient outcomes and lower healthcare expenses, can justify these expenses.

In conclusion, I see the future of healthcare powered by predictive analytics, holding great promise. The ability to anticipate and manage health issues proactively, coupled with personalized care tailored to individual needs can transform patient care. Nevertheless, achieving this goal will require overcoming ethical, technical, and practical issues.

IV. Experiment

A. Data and Fundamental Statistical Analysis

This experiment aimed to leverage predictive analytics to classify patient admissions into three categories: elective, urgent, or emergency using a healthcare dataset which was created synthetically to mimic real-world healthcare data. The dataset was downloaded from Kaggle and is available at: Healthcare Dataset. It comprises 15 columns and contains a total of 55,500 rows. Each entry includes the patient's name, age, and gender, as well as their blood type and primary medical condition. The dataset also features the date of admission and discharge, the doctor responsible for care, the hospital and room number, insurance provider, and billing amount. Additionally, it contains information about the nature of the admission, records of administered medications and test results.

To improve the quality of the analysis, several steps were first undertaken. Given the relatively modest size of the dataset, each row was duplicated, increasing the total number of rows from 55,500 to 111,000 to improve the model's performance. In addition, a new feature, "Length of Stay" was introduced by calculating the difference between the "Date of Admission" and "Discharge Date". Afterward, the original columns used in the calculation were removed. Along with this, the columns "Hospital", "Insurance Provider", and "Room Number" were removed due to the presence of gibberish data. Finally, a thorough check for null values was performed across the dataset, confirming that no null values were present.

A thorough exploration of the dataset was then conducted. This exploration was performed using Apache Hive 3.1.3 through Hue 4.11.0 on AWS Elastic MapReduce (EMR). A Hive table was created to structure the dataset and the dataset was ingested into the Hive table from an S3 bucket.

Fundamental statistical analysis revealed a relatively balanced distribution of admission types within the dataset. Elective admissions were slightly more common, accounting for 37,310

cases, followed closely by urgent admissions with 37,152 cases, and emergency admissions with 36,538 cases.

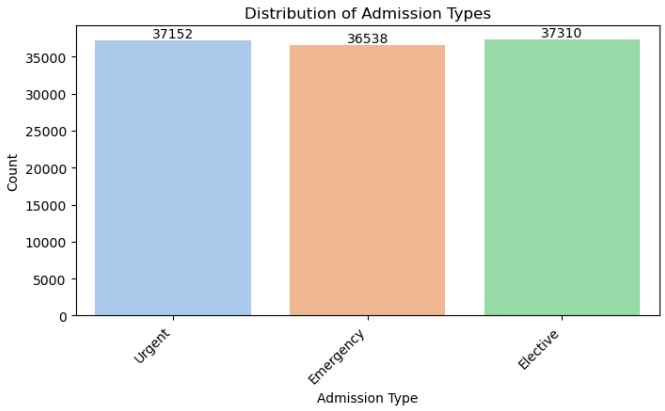

Fig. 2. Comparison of Admission Types

The near-equal distribution among these categories suggested that the dataset comprehensively covered different scenarios under which patients sought medical care, from planned procedures and treatments to more immediate and critical situations.

B. Aggregation and Visualizations

Understanding the gender distribution of the dataset was the first step carried out during this stage. The result indicated almost equal distribution between the two genders in the dataset.

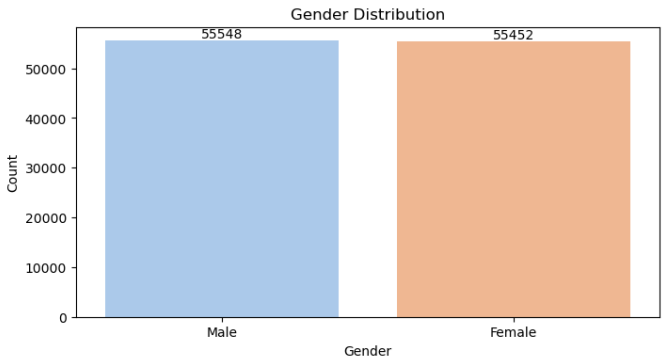

Fig. 3. Comparison of Gender Distribution

Specifically, there were 55,548 records for males and 55,452 records for females. This close parity in the number of records for each gender suggested that the dataset was well balanced in terms of gender representation.

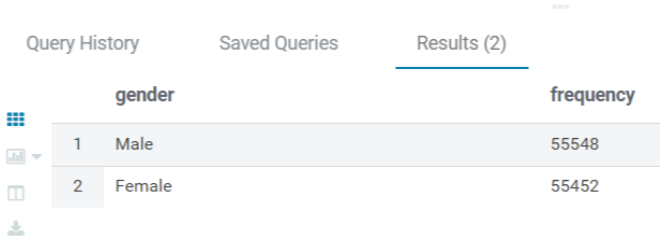

Fig. 4. Hive Query Output for Gender Distribution

The next step involved collecting statistical insights related to ages of patients. The average age of the patients was approximately 51.54 years. The oldest patient was Heather Dawson, who was 89 years old, while the youngest patient was Kristin Ortiz, who was 13 years old. These statistics highlighted the diversity in the age of patients included in the dataset, ranging from adolescents to the elderly.

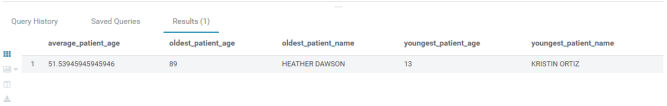

Fig. 5. Hive Query Output for Age Analysis of Patients

To gain a broader perspective, a Hive query was then executed to investigate the distribution of patients by blood type. This revealed a relatively even distribution across the various groups.

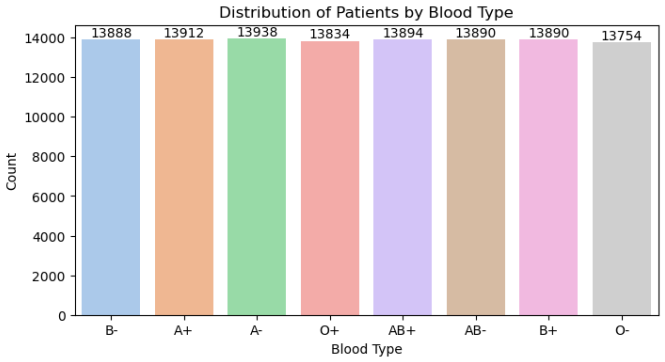

Fig. 6. Comparison of Blood Type Distribution

The result revealed each blood type having a count close to 14,000. The most common blood type in the dataset was A-, followed closely by A+ and AB+. The least common blood type was O-, but it still had a substantial representation with 13,754 records.

Query History | Saved Queries | Results (8)

	blood_type	count
1	A-	13938
2	A+	13912
3	AB+	13894
4	B+	13890
5	AB-	13890
6	B-	13888
7	O+	13834
8	O-	13754

Fig. 7. Hive Query Output for Blood Type Distribution

Following this, attention was shifted to examine the distribution of various medical conditions among the patients. The result exhibited that the dataset included six prevalent medical conditions, each with a significant number of cases.

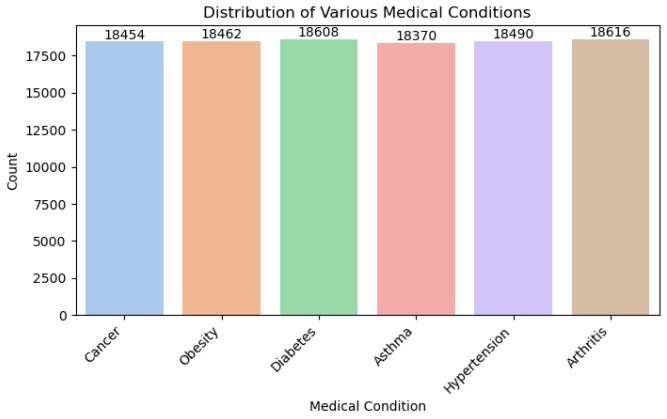

Fig. 8. Comparison of Medical Condition Distribution

Arthritis was the most common condition, affecting 18,616 patients, closely followed by Diabetes with 18,608 cases. Hypertension, Obesity, and Cancer also had substantial representations, each with over 18,000 cases. Asthma, while the least common among these, still affected 18,370 patients.

Query History | Saved Queries | Results (6)

	medical_condition	count
1	Arthritis	18616
2	Diabetes	18608
3	Hypertension	18490
4	Obesity	18462
5	Cancer	18454
6	Asthma	18370

Fig. 9. Hive Query Output for Distribution of Medical Conditions

With a clearer understanding of these factors, an exploration of how admission types varied across different genders was carried out. The analysis showed that the distribution of admission types was fairly balanced across genders, with slight variations.

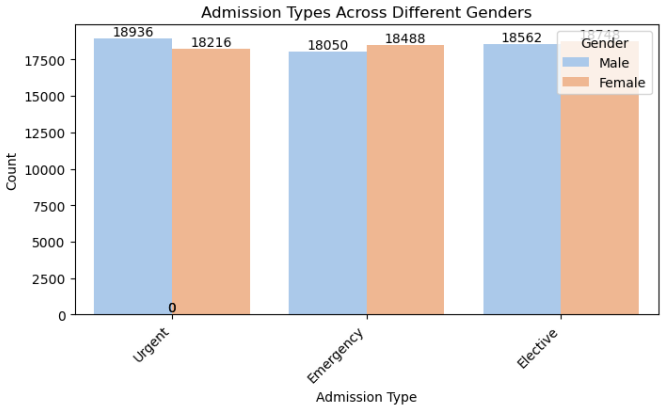

Fig. 10. Comparison of Admission Types Across Genders

For females, elective admissions were the most common, followed by emergency and urgent admissions. Males also exhibited a similar trend, with elective admissions being the most frequent, but urgent admissions were slightly more common than emergency admissions. The data indicated that while there were differences in the number of admissions by gender, the overall distribution of admission types remained relatively consistent.

Query History | Saved Queries | Results (6)

	gender	admission_type	count
1	Female	Elective	18748
2	Female	Emergency	18488
3	Female	Urgent	18216
4	Male	Elective	18562
5	Male	Emergency	18050
6	Male	Urgent	18936

Fig. 11. Hive Query Output for Admission Types by Gender

Following this, attention was focused on identifying the extremes in patient stay duration. The finding highlighted that Elizabeth Jackson had the longest hospital stay, lasting a total of 30 days. This notably prolonged duration of hospitalization suggested that her condition might have been quite intricate or required a more extended treatment and recovery process. In contrast, Whitney Ramirez had a remarkably brief hospital stay of just 1 day. This indicated either a very minor procedure or an outpatient visit.

Query History | Saved Queries | Results (1)

	max_stay_name	max_stay_age	max_stay_length	min_stay_name	min_stay_age	min_stay_length
1	ELIZABETH JACKSON	42	30	WHITNEY RAMIREZ	36	1

Fig. 12. Hive Query Output for Patient Stay Duration Analysis

Finally, the Hive query was executed to identify the top 10 doctors who handled the most cases. The result offered the distribution of cases among the most active doctors in the dataset.

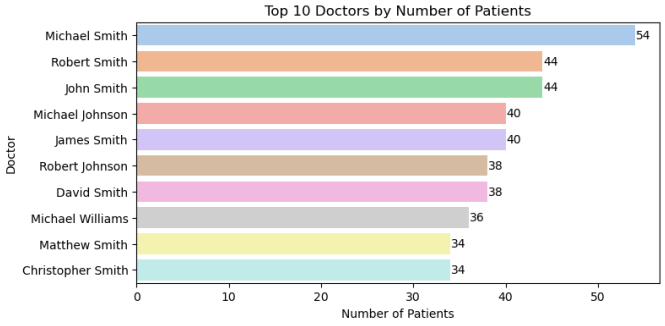

Fig. 13. Comparison of Case Distribution Among Top 10 Doctors

Michael Smith was the most frequently listed doctor, handling 54 cases, showing a high level of involvement. The next most frequent doctors, John Smith and Robert Smith, both managed 44 cases each. The distribution among the top 10 doctors showed a range of case volumes, from 54 to 34, reflecting varying levels of caseloads and potentially differing areas of specialization or patient loads.

Query History | Saved Queries | Results (10)

	doctor	cases_handled
1	Michael Smith	54
2	John Smith	44
3	Robert Smith	44
4	James Smith	40
5	Michael Johnson	40
6	David Smith	38
7	Robert Johnson	38
8	Michael Williams	36
9	John Johnson	34
10	Christopher Smith	34

Fig. 14. Hive Query Output for Top 10 Doctors by Number of Cases

In conclusion, the data exploration of the healthcare dataset provided a comprehensive overview of key aspects such as patient demographics, medical conditions, admission types, lengths of stay, and doctor workloads. The dataset revealed a balanced distribution of genders and blood types, as well as a diverse range of prevalent medical conditions. The insights into admission types showed a nearly equal distribution across elective, urgent, and emergency cases, while the extremes in lengths of stay highlighted varied patient experiences. The analysis of doctor workloads identified the top-performing physicians and their respective case counts. With these insights in hand, the next phase of the analysis was predictive model development.

C. *Predictive Model Development*

Following data exploration, the next step involved developing two predictive models using the Decision tree and the Random forest. To begin with, the dataset was imported into a pandas DataFrame. This DataFrame included the columns ”Name” and ”Doctor”, which were considered irrelevant to the predictive task at hand. As a result, these columns were removed to concentrate on the relevant features.

Afterwards, for the two algorithms to function correctly, it was essential to convert the remaining categorical columns in the DataFrame into a numeric format. This was achieved using a combination of Label Encoding and One-Hot Encoding. Label Encoding was applied to the ”Gender”, ”Admission Type”, and ”Test Results” columns. One-Hot Encoding was employed for the columns ”Blood Type”, ”Medication”, and ”Medical Condition”.

After encoding the categorical variables, the dataset was divided into features and the target variable. All columns except ”Admission Type” were used as features, while ”Admission Type” was used as the target. The data was subsequently divided into 80% training and 20% testing. In the following step, feature scaling was performed using StandardScaler. This step was important for the chosen algorithms. To prevent data leakage, the scaler was applied to both the training and testing sets.

The Decision Tree classifier was first trained on the scaled training data. The model achieved an accuracy of 88. 17% in the test data. Next, Random Forest classifier was trained. It achieved an accuracy of 88.91%, slightly outperforming the Decision Tree classifier.

Through data preparation, appropriate categorical variable encoding and model evaluation, it became evident that the two algorithms were effective in predicting the types of patient admission.

D. *MapReduce pseudo-code*

The classification of patient admissions using the MapReduce framework is implemented through the following steps:

Map function:

The Map function is designed to extract meaningful features from the healthcare dataset for each patient record. It reads through each record and extracts key attributes, including the patient’s age, gender, blood type, medical condition, billing amount, medications, test results, and length of stay. These features are then combined into a structured feature vector that represents the patient’s profile. Once the feature vector is created, it is emitted as the key with a placeholder value of 1. The role of the Map function is to preprocess and transform raw patient data into a standardized format that can be utilized for classification in the Reduce function.

```
function Map(record):
    // Extract relevant features from the
        ↪ healthcare dataset
```

```
    age = ExtractAge(record)
    gender = ExtractGender(record)
    blood_type = ExtractBloodType(record)
    medical_condition =
        ↪ ExtractMedicalCondition(record)
    billing_amount = ExtractBillingAmount(
        ↪ record)
    medication = ExtractMedication(record)
    test_results = ExtractTestResults(record)
    length_of_stay = ExtractLengthOfStay(
        ↪ record)

    // Combine the extracted features into a
        ↪ feature vector
    features = [age, gender, blood_type,
        ↪ medical_condition, billing_amount,
        ↪ medication, test_results,
        ↪ length_of_stay]

    // Emit the feature vector as the key and
        ↪ a placeholder value of 1
    Emit(features, 1)

end function
```

Shuffle and Sort phase:

The Shuffle and Sort phase is automatically managed by the MapReduce framework. During this phase, the emitted feature vectors are grouped and sorted based on their content. Identical feature vectors are grouped together. This organization is crucial for the classification of patient admissions.

```
// Shuffle and Sort Phase (handled
    ↪ automatically by the MapReduce framework
    ↪ )
// Groups identical feature vectors together,
    ↪ organizing the data for the Reduce phase
```

Reduce function:

The Reduce function receives the feature vectors generated by the Map function and uses them to predict the admission type of each patient. Instead of simply aggregating counts or frequencies, the Reduce function leverages a pre-trained predictive model to classify the admission type (e.g., "Elective", "Emergency", "Urgent"). It uses the feature vector as input to the model and outputs the predicted admission type as the result. By applying the model, the Reduce function provides a classification for each patient based on their extracted features, thereby converting raw data into actionable insights.

```
function Reduce(key, values):
    // The key is the feature vector, and
        ↪ values are a list of 1s (
        ↪ placeholders)
    // No need to count the values since
        ↪ admission types are not being
        ↪ aggregated directly.

    // Extract the feature vector from the key
    features = key  // Example: [45, "Male", "
        ↪ O+", "Diabetes", 1500.75, "
        ↪ Paracetamol", "Normal", 3]

    // Apply the predictive model to classify
        ↪ the admission type
    admission_type = PredictAdmissionType(
        ↪ features)  // Example output: "
        ↪ Elective", "Urgent", "Emergency"

    // Emit the feature vector (key) with the
        ↪ predicted admission type
    Emit(features, admission_type)

end function
```

V. Conclusion

The study successfully demonstrated the effectiveness of predictive analytics in classifying patient admissions into elective, urgent, or emergency categories. The Decision Tree classifier achieved 88.17% accuracy, with precision, recall, and F1-scores around 0.88-0.89, particularly excelling in identifying urgent admissions. The Random Forest classifier performed slightly better, with 88.91%. Like the Decision Tree model, it also exhibited balanced performance across the three admission categories, with precision scores of 0.89 for all types. This consistency further emphasizes the robustness of the Random Forest model in a healthcare setting.

Overall, both models demonstrated high accuracy and reliability, supporting the notion that predictive analytics can significantly improve hospital preparedness and patient management. Although the study's outcomes are encouraging, numerous opportunities exist for further research to expand upon these results. For instance, testing these models in various healthcare settings, such as rural hospitals or specialized care centers, could provide more insights into their effectiveness across different environments. Further research could also explore other machine learning models, such as gradient boost or neural networks, to compare their performance against Decision Trees and Random Forests.

In conclusion, this research has demonstrated the potential of predictive analytics in improving patient admission classification within the healthcare sector. By accurately predicting patient admission categories, healthcare providers can allocate resources more effectively and ensure timely delivery of care.

References

[1] R. Ravikumar, A. Kitana, A. Taamneh, A. Aburayya, F. Shwedeh, S. Salloum, and K. Shaalan, "The Impact of Big Data Quality Analytics on Knowledge Management in Healthcare Institutions: Lessons Learned from Big Data's Application within The Healthcare Sector," *South Eastern European Journal of Public Health*, Jan. 2023. [Online]. Available: https://www.seejph.com/index.php/seejph/article/view/309

[2] D. Khanna, N. Jindal, H. Singh, and P. S. Rana, "Applications and Challenges in Healthcare Big Data: A Strategic Review," *Current Medical Imaging*, vol. 19, no. 1, pp. 27–36, 2022.

[3] H. E. Pence, "What is Big Data and Why is it Important?" *Journal of Educational Technology Systems*, vol. 43, no. 2, pp. 159–171, Dec. 2014, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.2190/ET.43.2.d

[4] A. Ahmed, R. Xi, M. Hou, S. A. Shah, and S. Hameed, "Harnessing Big Data Analytics for Healthcare: A Comprehensive Review of Frameworks, Implications, Applications, and Impacts," *IEEE Access*, vol. 11, pp. 112 891–112 928, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10274958/

[5] K. Jee and G.-H. Kim, "Potentiality of Big Data in the Medical Sector: Focus on How to Reshape the Healthcare System," *Healthcare Informatics Research*, vol. 19, no. 2, pp. 79–85, Jun. 2013, publisher: Korean Society of Medical Informatics. [Online]. Available: https://synapse.koreamed.org/articles/1075681

[6] A. A. H. de Hond, A. M. Leeuwenberg, L. Hooft, I. M. J. Kant, S. W. J. Nijman, H. J. A. van Os, J. J. Aardoom, T. P. A. Debray, E. Schuit, M. van Smeden, J. B. Reitsma, E. W. Steyerberg, N. H. Chavannes, and K. G. M. Moons, "Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review," *NPJ Digital Medicine*, vol. 5, p. 2, Jan. 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8748878/

[7] P. Amin, N. R. Anikireddypally, S. Khurana, S. Vadakkemadathil, and W. Wu, "Personalized Health Monitoring using Predictive Analytics," in *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*. Newark, CA, USA: IEEE, Apr. 2019, pp. 271–278. [Online]. Available: https://ieeexplore.ieee.org/document/8848236/

[8] imanuel, "What is Predictive Analytics ?" Apr. 2024. [Online]. Available: https://www.predictiveanalyticstoday.com/what-is-predictive-analytics/

[9] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems*, vol. 2, p. 3, Feb. 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341817/

[10] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar. 1986. [Online]. Available: https://doi.org/10.1007/BF00116251

[11] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[12] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, N. Zhang, S. Antony, H. Liu, and R. Murthy, "Hive - a petabyte scale data warehouse using Hadoop," in *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, Mar. 2010, pp. 996–1005, iSSN: 2375-026X. [Online]. Available: https://ieeexplore.ieee.org/document/5447738/?arnumber=5447738

[13] A. Thusoo, Z. Shao, S. Anthony, D. Borthakur, N. Jain, J. Sen Sarma, R. Murthy, and H. Liu, "Data warehousing and analytics infrastructure at facebook," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. Indianapolis Indiana USA: ACM, Jun. 2010, pp. 1013–1020. [Online]. Available: https://dl.acm.org/doi/10.1145/1807167.1807278

[14] K. Singhania and A. Reddy, "Improving Preventative Care and Health Outcomes for Patients with Chronic Diseases using Big Data-Driven Insights and Predictive Modeling," vol. 9.

[15] A. Merabet, A. Saighi, Z. Laboudi, and M. A. Ferradji, "Multiple diseases forecast through ai and iomt techniques: Systematic literature review," in *International Conference on Intelligent Systems and Pattern Recognition*. Springer, 2024, pp. 189–206.

[16] I. D. Dinov, B. Heavner, M. Tang, G. Glusman, K. Chard, M. Darcy, R. Madduri, J. Pa, C. Spino, C. Kesselman, I. Foster, E. W. Deutsch, N. D. Price, J. D. Van Horn, J. Ames, K. Clark, L. Hood, B. M. Hampstead, W. Dauer, and A. W. Toga, "Predictive Big Data Analytics: A Study of Parkinson's Disease Using Large, Complex, Heterogeneous, Incongruent, Multi-Source and Incomplete Observations," *PLOS ONE*, vol. 11, no. 8, p. e0157077, Aug. 2016. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0157077

[17] A. Sharma, D. Shukla, T. Goel, and P. K. Mandal, "BHARAT: An Integrated Big Data Analytic Model for Early Diagnostic Biomarker of Alzheimer's Disease," *Frontiers in Neurology*, vol. 10, p. 9, Feb. 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fneur.2019.00009/full

[18] S. H. Koppad and A. Kumar, "Application of big data analytics in healthcare system to predict COPD," in *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*. Nagercoil, India: IEEE, Mar. 2016, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/document/7530248/

[19] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7912315/

[20] N. S. Kumar, T. Eswari, P. Sampath, and S. Lavanya, "Predictive Methodology for Diabetic Data Analysis in Big Data," *Procedia Computer Science*, vol. 50, pp. 203–208, 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050915005700

[21] S. Rallapalli and T. Suryakanthi, "Predicting the risk of diabetes in big data electronic health Records by using scalable random forest classification algorithm," in *2016 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. Durban, South Africa: IEEE, Nov. 2016, pp. 281–284. [Online]. Available: http://ieeexplore.ieee.org/document/8073762/

[22] A. Mir and S. N. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. Pune, India: IEEE, Aug. 2018, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8697439/

[23] A. Ismail, S. Abdlerazek, and I. M. El-Henawy, "BIG DATA ANALYTICS IN HEART DISEASES PREDICTION," . *Vol.*, no. 11, 2005.

[24] D. Combs, S. Shetty, and S. Parthasarathy, "Big-Data or Slim-Data: Predictive Analytics Will Rule with World," *Journal of Clinical Sleep Medicine*, vol. 12, no. 02, pp. 159–160, Feb. 2016. [Online]. Available: http://jcsm.aasm.org/doi/10.5664/jcsm.5474

[25] N. V. Chawla and D. A. Davis, "Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework," *Journal of General Internal Medicine*, vol. 28, no. S3, pp. 660–665, Sep. 2013. [Online]. Available: http://link.springer.com/10.1007/s11606-013-2455-8

[26] A. Alkhachroum, J. Kromm, and M. A. De Georgia, "Big data and predictive analytics in neurocritical care," *Current Neurology and Neuroscience Reports*, vol. 22, no. 1, pp. 19–32, Jan. 2022. [Online]. Available: https://link.springer.com/10.1007/s11910-022-01167-w

[27] S. L. Harris, J. H. May, and L. G. Vargas, "Predictive analytics model for healthcare planning and scheduling," *European Journal of Operational Research*, vol. 253, no. 1, pp. 121–131, Aug. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0377221716300376

[28] H. Chennamsetty, S. Chalasani, and D. Riley, "Predictive analytics on Electronic Health Records (EHRs) using Hadoop and Hive," in *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. Coimbatore, India: IEEE, Mar. 2015, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/document/7226129/

[29] K. Batko and A. Ślezak, "The use of Big Data Analytics in healthcare," *Journal of Big Data*, vol. 9, no. 1, p. 3, Dec. 2022. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00553-4

[30] P. Galetsi, K. Katsaliaki, and S. Kumar, "Values, challenges and future directions of big data analytics in healthcare: A systematic review," *Social Science & Medicine*, vol. 241, p. 112533, Nov. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0277953619305271