Design and Implementation of FAIR principles in digital health data in Uganda, Africa

Basajja, M.

Chapter 8

Machine Learning Models in FAIR Health Data from Uganda

This chapter is based on the following publication:

Mariam Basajja, Wolstencroft Katy, Fons Verbeek; A comparative analysis of the performance of Machine Learning Models in FAIR Health Data from Uganda, Africa. [Accepted, publication pending]

Abstract

In this chapter, we discuss the application of machine learning (ML) and the FAIR data principles - Findable, Accessible, Interoperable, and Reusable - within the healthcare sector. It utilizes anonymized datasets from VODAN-Africa, specifically from Uganda. The data include Uganda's Outpatient Department (OPD) records and Uganda's integrated antenatal care (ANC) information. These data sets are structured and standardized according to the FAIR principles and are used to train various machine learning models for tasks such as predicting pregnancy risk factors, assessing fetal conditions, detecting anomalies, and performing clustering analyses. This chapter highlights the effective integration of Machine Learning with FAIR data in a federated architecture in healthcare analytics and underscores the significance of high-quality, accessible, and reusable data to support collaborative research and enhance public health outcomes.

Keywords: FAIR data, machine learning, federated learning, federated architecture, outpatient register, antenatal care register, Allegrograph triple store.

The source code for the scripts in this chapter can be found here: `https://github.com/mariambasajja/CH8-THESIS-CODE`

8.1 Introduction

In recent years, the combination of machine learning (ML) and advanced data science techniques in digital health research has revolutionized the field. ML provides powerful tools for predictive analytics that enable the extraction of meaningful patterns from complex data. The patterns provide insights that inform the strategies used in the implementation of health interventions and the allocation of resources to reduce the prevalence and death of diseases, thus saving lives [153]. Despite the magnitude of this opportunity to improve healthcare systems, there are existing concerns about data privacy, data quality, and ethical issues that arise when applying machine learning to solve the problems faced in healthcare. Since the application of data science driven by advancements in big data and artificial intelligence (AI) has increased, the need to implement robust and ethical practices in data management has also grown [154]. Strict ethical and legal administration must be put in place to ensure that sensitive data in healthcare is protected. This will reduce the negative impact of malpractices in the management of health data [155].

An advantage of ML in healthcare is its utility in analyzing information in multiple domains, as valuable patterns often require large amounts of diverse data to emerge. A single source of data, such as from individual healthcare organizations or specific industries, is rarely adequate in scale or diversity to reveal robust and generalizable insights. However, integrating data from different sources presents challenges in data integration, standardization, and collaboration. In addition, it increases concerns about data privacy and security, as sensitive health information must now be shared or accessed across institutional and sometimes national boundaries. These challenges present the need to utilize an approach that supports collaborative data analysis in healthcare. The FAIR data principles provide disruptive potential in the biomedical sector since they critically improve data accessibility, interoperability, and reusability, which are determinants in the improvement of healthcare systems. Biomedical research and the digital health framework apply to FAIR (Findable, Accessible, Interoperable, and Reusable) data, which introduces a standardized approach to managing and sharing demanding data. Findability is the principle that ensures that the search for metadata-tagged data is simple, so that researchers and healthcare facilities can identify data that applies to their field on various platforms. Accessibility enables relevant users to access information, and various stakeholders in healthcare research and organizations can effectively access data in a manner that captures the provisions of security and privacy issues.

The interoperability aspect of FAIR data ensures that there is no interruption when integrating data on different systems and platforms. This is particularly critical in the healthcare sector because patient information may come from a large number of sources, including hospitals, clinics, mobile health applications, and research studies. FAIR data allows these disparate sources to interoperate with each other through standardisation of data formats and protocols, hence allowing collaborative efforts across institutions and even countries. This provides medical personnel with access to the detailed picture of the past of a patient, allowing them to make more reasonable clinical decisions and positively affect patient outcomes.

The reusability principle makes it possible to repurpose datasets and use them in new research and analyses rather than collecting the same data repeatedly. With respect to the biomedical industry, this is particularly useful when funding for data acquisition is scarce. The FAIR data allow the utilization of existing datasets across several studies, empowers researchers to perform Longitudinal analyzes, enhances predictive model accuracy, and finds new ways to track

patterns of diseases, treatment reactions, and drugs.
Practically, based on FAIR data principles, data silos, which commonly occur in healthcare systems, are broken down, and different institutions, including local hospitals, global research organizations, and others, can share their experiences and collaborate more efficiently. To give an example, a rural hospital might not have easy access to the data of different facilities, but observing the principles of FAIR data accessibility, it would be able to safely share the anonymized data on patients and participate in collaborative research that would influence more general policies in healthcare. This introduces novel possibilities of using data in decision-making, which results in more effective treatment plans, rational use of resources, and healthier results among populations. In addition, privacy and data security are factors in the design of FAIR data. Improves the flow of information, but at the same time ensures that confidential data concerning patients remains secure. Access controls are clearly established within the system so that no unauthorized people or institutions can access specific datasets. This is especially decisive in the health care industry, where the issue of security of personal health information is not just mandated by the law, but is also a prime example of protecting health information on an ethical front. Its main strengths are that it allows sharing data, but at the same time ensures high control over access.
In general, the principles of FAIR data are critical in improving collaborations in healthcare systems at the local and global scales. These principles can advance the creation of more valid predictive models, enable collaborative work, and promote better decision-making and higher-quality care because they assist in the safe exchange of high-quality, standardized data. FAIR data will be essential in establishing a more connected, transparent, and efficient healthcare ecosystem as more healthcare systems move toward digital health technologies, where much more could be done with data to advance patient care, unlock operations efficiencies, and spur innovation. The implementation of FAIR data principles alongside machine learning emerges as a potential solution to these challenges [176]. Data can be FAIR, without being open. Strict accessibility criteria can be applied to individuals or institutions with permission to use them. Alternatively, Federated Machine Learning (FML) approaches can be applied more easily when data is FAIR and therefore standardised.

Federated Machine Learning facilitates the development of ML models using decentralized data sources without compromising data privacy [156]. This approach addresses ethical concerns in data management and security while simultaneously improving the accuracy of prediction and generalizability of learning models [157]. The application of federated learning and the adoption of the FAIR data principles guaranty that smaller healthcare facilities, including resource-constrained environments, can easily contribute to participating in data-driven patient health research without any significant emphasis on expenditure-heavy infrastructure support. The architecture can be scaled since health facilities only have to store and process local datasets, so they are easy to use. Moving toward the implementation of this architecture, hospitals and clinics will be able to collaborate safely and ensure privacy about their patients, therefore, enjoying the potential of machine learning systems, which are not only highly efficient but also safe to use. By enabling the construction of ML models using decentralized sources of data, Federated Machine Learning(FML) offers a promising path forward to harness the power of diverse datasets while maintaining stringent privacy standards. [315]

The emergence of federated machine learning allows models to be trained together across multiple organizations without sharing raw data, thus addressing some of the privacy concerns of centralized data processing [160]. If decentralized data sources are used, federated machine

learning reduces the risks associated with data breaches and facilitates secure, scalable, and privacy-preserving machine learning [163]. FML provides an outline consistent with the FAIR principles to ensure that data is interoperable and reusable while maintaining data privacy [164].

Furthermore, federated machine learning frameworks have demonstrated their versatility and robustness in a variety of applications, such as real-time edge intelligence and mobile packet categorization [165], [166]. Through the combination of machine learning and federated learning, analytical capabilities continue to evolve, allowing sophisticated models to tackle challenges in healthcare, such as the complexity of biomedical image reconstruction [167].

In this study, we show the feasibility of using FAIR data resources in Ugandan Digital Health by training several traditional machine learning approaches on different datasets. AllegroGraph is a semantic graph database and RDF triple-store designed to store, query, and integrate complex linked data. In this study, it serves as the back end for managing anonymized VODAN-Africa health datasets, enabling structured storage and efficient retrieval of patient and maternal health information. FAIR data are organized and standardized, making it easier to work with. We are able to demonstrate different use cases that are possible in a federated architecture. The use cases include: FAIR Outpatient department diagnoses, clustering of K-means, detection of anomalies in FAIR Integrated Antenatal Care Uganda, clustering of pregnANCy risk factors and determinants of fetal states, and modeling of FAIR ANCfetal state prediction. For the FAIR Uganda ANC data, K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), and Gradient Boosting models are employed for both pregnancy risk factor prediction and fetal state prediction. The performance of these algorithms is evaluated after training to determine the effectiveness of the implementation.

This chapter also presents a federated learning implementation for fetal state prediction in a simulated environment using ANC Uganda data. This approach demonstrates how federated learning can be applied to train models across geographically distributed datasets without sharing raw patient information. The simulated implementation mimics a decentralized setting in which data from different health facilities are represented as local data sources. Local models are trained independently on these subsets, which were generated by splitting the Ugandan ANC data set. A central server coordinates the training process and aggregates the locally trained models, ensuring data privacy throughout. Together, these implementations demonstrate the potential of applying FAIR data principles and machine learning in healthcare to enable collaborative model development while protecting sensitive patient data.

Among the primary findings of this study is the idea that the concepts of FAIR data can be implemented in the real world, especially in resource-limited environments, in healthcare settings. Most low-resource health care environments (e.g., rural hospitals, developing world clinics) have reduced access to sophisticated data infrastructure and sophisticated technologies. Such environments also find it difficult to embrace the full potential of digital health innovations due to the fact that they may not have the funds to invest in complicated systems of data management or have the financial capability to acquire costly infrastructure.

Nevertheless, FAIR data principles present a chance to transform this dynamic and introduce a case-flexible, scalable means to data management by providing a consistent way of managing data. FAIR data makes data access and reuse among various institutions easy without regard to the technical systems available at each institution. For example, health data in a rural hospital can be represented on a simple electronic health record (EHR) system that lacks advanced data-sharing capabilities; however, following the FAIR principles, the hospital would still be able to render its data interoperable with the other healthcare institutions, opening new possibilities

of collaboration and research.

Our suggested architecture is an extension of these FAIR data concepts, integrated with machine learning, which enables multiple groups to build end-to-end models together without sharing individual raw patient data. Federated learning enables hospitals to train their own machine learning models on their locally stored sensitive data without outsiders learning more about said data. In so doing, by sharing sensitive information, hospitals have had an opportunity to utilize shared insights to develop predictive models collectively; moreover, quality of care also changes positively, and this helps in early detection and evaluation of disease risks. In particular, federated learning also maintains privacy-preserving access to patient data since model parameters (weights and gradients) are exchanged and not raw data.

The proposed architecture is expected to be scalable, i.e., applicable in a great number of care facilities, including small local hospitals and larger regional healthcare facilities. The secret behind such scalability is that the system does not require enormous infrastructure upheavals. Local hospitals are not required to invest in costly computing hardware and software to integrate new systems with existing ones and data sources, and can utilize their current systems without having to alter them. The federated learning strategy allows data access and cooperation between hospitals without the need to physically combine or transfer patient information. This is more so in areas where resources are sparse, since the process of reworking the whole data process might be costly and logistically impossible.

The simplicity in implementation means that even health facilities in other regions, such as those located remotely or in developing nations, can easily employ the architecture. With the necessary expertise and infrastructure in place, the facilities will easily implement federated learning and machine learning models on their local data.

The architecture helps to decentralize the process of learning, allowing hospitals to build predictive models based on their own history, taking the specifics of the health problems relevant to the area into account (e.g., outbreaks of a disease affecting the region, health trends unique to the region). With the help of this initiative, hospitals can exchange their experience, but do not infringe on the privacy of their patients. In contrast, this collaborative practice helps to improve the level of accuracy and generalizability of predictive models, since they are constructed on data sets of various environments. Furthermore, through such a collection of data, hospitals can learn about each other, have access to new knowledge, and improve healthcare outcomes in all regions.

FAIR data are also based on the concept of collaborative access to data without compromising patient privacy. Hospitals and clinics can use the force of multilateral information; however, it should be implemented while complying with ethical principles and privacy regulations. Under a federated learning approach to computation, sensitive patient information is not shared directly. Only the local model training results, e.g., updated parameters, are aggregated and shared. This gives institutions the opportunity to be involved in global machine learning activities and still retain data protection regulations (e.g., HIPAA or GDPR).

Finally, FAIR data allow hospitals in resource-restricted environments to take advantage of the combined potential of machine learning and big data analytics without having to constantly invest heavily in the infrastructure or in their data privacy. It offers an implementable, scalable approach to the application of machine learning to the delivery and outcomes of healthcare within low-resource settings.

8.1.1 Research Questions and Objectives

The research questions and sub-questions that we consider most relevant to the evaluation of the ease of use of FAIR data in digital healthcare in Africa are presented in this section. The main objective of this study is to assess the ease of use of FAIR data in VODAN-Africa by providing analytical information using machine learning to help drive operational decision-making and strategic planning. The study also aims to show the possibility of investigating the patterns revealed by applying ML prediction and clustering algorithms to the FAIR VODAN-Africa data. This will inform policies and other administrative decisions in the development, implementation, and maintenance of health interventions. These objectives will be achieved using data from the Outpatient Department Uganda, Antenatal Care Register Uganda, and data made available for retrieval from Allegrograph triple-store databases according to the signed data agreements between VODAN-Africa and the different health facilities. OPD Uganda data contain anonymized information on patients, including their age, sex, district of residence, blood pressure and blood sugar levels, body mass index (BMI), alcohol use, date of visit, and diagnosis. The FAIR ANC data contains anonymized data on pregnant women. The ANC Uganda dataset includes the woman's age, gestational age, height, weight, and mid-upper arm circumference (MUAC) measurement.

Research Questions

Main Research Question: Does the use of FAIR data support and strengthen the accessibility of different sources of health data in machine learning?

Sub-Research Questions:

Sub-Research Question 1:
what advantages does FAIR data provide for using ML algorithms for clustering and prediction? What patterns emerge when clustering the OPD Uganda Dataset, and how can these patterns inform healthcare practices?

Sub-Research Question 2:
Is the FAIR data sufficiently rich to predict pregnancy risks and fetal states across different age groups? How do pregnancy risk factors and fetal state correlate with the data obtained from the ANC Uganda datasets in predictive modeling?How can we apply ML in a federated architecture?

8.1.2 Our contributions

This research assesses the use of FAIR Data to support and strengthen privacy-preserving accessibility of different types of health data for machine learning. It investigates the use of FAIR data with machine learning to offer new insights into geographic patterns in disease and infection. This is achieved using FAIRified health data from VODAN-Africa, governed by shared Data Use Agreements between VODAN-Africa and health facilities in Uganda. Data anonymisation and these agreements ensure ethical data access to key health statistics, enabling the training of machine learning models for predictions and cluster analysis. The standardized

and interoperable nature of the data within the VODAN network further enhances the quality and reliability of the models. These include:

1. Generation of visualizations that reveal patterns in FAIR OPD Uganda data patterns using K-Means clustering analyses.

2. Anomaly detection among pregnant women using the traditional isolation forest using FAIR Antenatal care(ANC) from Uganda.

3. Clustering analyzes on determinants of pregnancy risk factors and fetal states in FAIR ANC data from Uganda using K-Means clustering algorithms.

4. FAIR ANC Uganda data fetal state prediction modelling using traditional Machine learning and federated learning with a simulated dataset split.

Ethical clearance was obtained from Uganda's Ministry of Health, health facility directors, and other key stakeholders to access the healthcare data used. The paper registers acquired containing OPD Uganda, ANC Uganda data were FAIRified using the VODAN-Africa architecture. [162] This makes data more accessible and reusable by researchers under well-defined terms, and enhances its integration capabilities, facilitating the application of machine learning algorithms. The use cases aim to show the possibility of providing analytical information showing trends that can be used for operational decision-making and strategic planning in healthcare settings.

This chapter explores related work in machine learning, federated machine learning, data science, and big data in health under Section 8.2. This is followed by the methodology implemented to train the machine learning models for the clustering of patient data, prediction of pregnancy risk factors, and fetal states in Section 8.3. The results and consequent recommendations are discussed in the same section. The limitations and future outlook are outlined in Section 8.4. The chapter's conclusion details how the chapter answers the research questions and achieves the objectives set in Section 8.5.

8.2 Related Work

In this section, we discuss the application of machine learning in healthcare.

8.2.1 Machine Learning in Healthcare

Machine learning (ML), a subset of artificial intelligence, has become increasingly integral to healthcare systems, offering innovative solutions to various challenges in the sector. ML involves the development of algorithms and statistical models that enable computer systems to improve their performance on a specific task through experience, without being explicitly programmed [202]. In healthcare, ML has found applications in diagnosis, treatment planning, patient monitoring, and drug development, among others.

Machine learning can be broadly categorized into three types: Supervised Learning: The algorithm learns from labeled training data to predict outcomes for unforeseen data. In healthcare, this is often used to predict diagnoses based on patient data. Unsupervised learning: The algorithm finds hidden patterns or intrinsic structures in the input data without labels. This is

useful in healthcare to segment patients and identify unknown disease subtypes. Reinforcement Learning: The algorithm learns to make decisions by interacting with an environment to achieve a goal. In healthcare, this can be applied to optimize treatment strategies over time.

Several machine learning (ML) models have found significant applications in healthcare. Neural networks, inspired by biological neural networks, are utilized for complex pattern recognition tasks such as image analysis in radiology and pathology, and prediction of patient outcomes based on electronic health records (EHRs) [203]. K-Nearest Neighbors (KNN), a simple, instance-based learning algorithm, is used for classification and regression and has been applied in cancer detection and heart disease risk classification [204]. Support Vector Machines (SVM), effective in both classification and regression tasks, especially in high-dimensional spaces, have been used in cancer classification, neuroimaging for Alzheimer's disease, and prediction of hospital readmissions [205].

K-means Clustering, an unsupervised learning algorithm, is used to partition data into clusters and has applications in patient segmentation, identifying disease subtypes, and analyzing gene expression data [206]. Hierarchical clustering creates a tree-like structure of clusters without requiring a prespecified number of clusters. In healthcare, it's used for patient stratification, gene expression analysis, and disease subtype identification. This method is valuable when the data structure is unknown or when exploring different levels of clustering granularity [207].

Gradient Boosting, an ensemble technique that combines weak learners, is applied to predict disease progression, patient risk stratification, and drug response prediction [208].

The Isolation Forest algorithm is an unsupervised learning method used for anomaly detection and can help identify unusual patient cases, rare diseases, or flag potential errors in medical records [197]. Lastly, Decision Trees, a non-parametric supervised learning method, are valued for their interpretability and are used in clinical decision support, diagnosis, and treatment planning [209].

These models are often combined or used in ensemble methods to improve performance. For example, random forests, which combine multiple decision trees, have been successful in various healthcare applications, including the prediction of patient survival, disease diagnosis, and drug response [211].

Moreover, deep learning, a subset of machine learning that uses artificial neural networks with multiple layers, has shown remarkable success in healthcare, particularly in medical imaging. Convolutional neural networks (CNN) have been used for tasks such as detecting diabetic retinopathy on retinal images or identifying cancerous lesions on radiological images [212].

The choice of ML model depends on the specific healthcare problem, the nature and amount of available data, the need for model interpretability, and computational resources. As healthcare data continues to grow in volume and complexity, the application of these ML models is expected to expand, potentially revolutionizing various aspects of healthcare delivery and medical research.

Applications of Machine Learning in Healthcare

The adoption of ML in healthcare systems has had positive impacts in solving different problems in the healthcare industry. ML has been most effective in cases in predicting disease occurrence, probable outbreaks, and patient outcomes. This capability has enabled a clearer definition of best practices in the treatment of diseases, increasing patient health status through earlier and more effective diagnosis of diseases.

Epidemic Forecasting

Forecasting the epidemiological patterns of infections like the flu or COVID-19 using ML models has been proven effective in providing more insights into preventing and controlling the diseases. A recent study by Qian et al. (2020) presents an example of an ML model used for the prediction of the daily number of new cases of COVID-19 in China, demonstrating the potential of ML in the prediction of epidemics [195]. Estimates of the size, the peak, and the duration of outbreaks have also been done using machine learning techniques. Epidemic forecasting approaches include Artificial Neural Networks (ANN), time series, hybrid, and data mining to predict epidemic outbreaks. This is demonstrated in a study by Datilo et al. (2019) that focused on the utilization of ANNs to forecast the emergence of diseases, including influenza, SARS, HIV / AIDS, dengue, tuberculosis, malaria, and cardiovascular diseases [196].

COVID-19 Management

A review by Lalmuanawma et al. (2020) investigated the real-world use of various ML technologies to combat the SARS-CoV-2 outbreak. The study showed the results of using various machine learning models for the prediction of SARS-CoV-2:

1. The support vector regression and stacking assembly models were used for clinical data forecasting, showing an error range of 0.87%-3.51% for one day ahead, 1.02%-5.63% for three days ahead, and 0.95% - 6.90% for six days ahead, indicating strong predictive capabilities.

2. XGBoost Classifier achieved an accuracy of 90% when using clinical data and blood samples with 75 features, demonstrating its potential for accurate prediction.

3. Deep learning using the LSTM network was used for demographic data to predict the end point of the pandemic in Canada.

4. The Hybrid Wavelet-Autoregressive Integrated Moving Average Model and Regression Tree analysed demographic data from various countries to provide real-time forecasts and forecasts for 10 days, identifying seven key features associated with COVID-19-related deaths.

The results show the applicability of using machine learning as a solution to provide insights into how to manage widespread outbreaks. Machine learning technologies provide the alternative methods needed for rapid screening and prediction, as well as contact tracking, forecasting, and developing vaccines or drugs with more accurate and reliable operation [201].

Clinical Decision Support

ML algorithms have been applied to Electronic Health Records (EHR) to predict the probability that a patient's condition worsens, thus improving clinical decision making and diagnosis [210]. A study by Tomaev et al. (2019) showed how deep learning models applied to EHRs could predict acute kidney injury up to 48 hours in advance, potentially allowing earlier interventions [213].

The examples explored in this section demonstrate that applying machine learning is feasible for improving the strategies used to prevent, manage, and monitor diseases, particularly in areas with low human resources compared to the diseases' prevalence. By applying machine

learning methodologies, diagnostic activities can be conducted automatically in healthcare systems, treatment plans can be enhanced, and general patient control can be improved. It enables early intervention and accurate resource allocation for areas struggling with a lack of healthcare workers.

8.2.2 Federated Machine Learning

Security and privacy issues related to centralized data storage are addressed by federated machine learning, as it ensures that sensitive information remains locally available. This reduces the risk of data breaches and increases compliance with data privacy laws by enabling ML models to train ML models on decentralized data [160]. Various healthcare-related applications, including individual therapy and collaborative research between organizations, have demonstrated the feasibility of this approach. For example, FML could support collaborative research in multiple hospitals without jeopardizing patient confidentiality, as Rieke et al. mention [171]. Feki et al. (2021) demonstrated the use of FAIR principles in conjunction with federated learning for COVID-19 chest X-ray classification, showing how this combination can facilitate collaborative research while preserving data sovereignty [172].

8.2.3 Federated Architecture in Healthcare Application

Federated architecture has been extensively studied in the healthcare sector due to the sensitive nature of health data and the growing need for privacy-preserving machine learning models. Several studies have explored how federated learning (FL) and federated architectures can be employed to develop machine learning models that preserve the privacy of patient data while leveraging data across multiple institutions. These efforts are particularly important in developing countries, where healthcare data is often fragmented, and local hospitals or clinics may not have the computational resources to run advanced machine-learning models.
A notable example is the Federated Learning for Health (FL4H) initiative, which seeks to enable collaboration across different healthcare institutions without transferring sensitive patient data. In their research, [295] proposed a federated learning framework where data remains on local servers, and only model updates are shared with a central aggregator. This approach was tested using data from multiple hospitals in different regions, showing that federated learning can improve model accuracy while maintaining data privacy and confidentiality.
Another significant project, authored in [296], focuses on developing a federated architecture specifically for healthcare, integrating multiple data sources and machine learning models across hospitals in Europe. This project demonstrated the scalability and flexibility of federated learning in real-world healthcare scenarios, particularly in predicting patient outcomes using various forms of clinical data. It highlighted that federated learning's ability to collaborate across institutions, while ensuring data remains localized, provides significant benefits in terms of data diversity and model robustness.

In Africa, the application of federated learning in healthcare has started to gain traction, especially in the context of improving access to healthcare services and enhancing decision-making processes. Teo et al. [297] explored the feasibility of applying federated learning in resource-limited settings in Africa, with a focus on countries like Uganda. Their work emphasized the challenges posed by limited computational infrastructure and the need for privacy-preserving

models when dealing with sensitive health data. They found that federated architecture, which allows hospitals and clinics to collaborate without sharing patient data, offers a promising solution to these challenges.
In Uganda, a study [298] was conducted on the implementation of federated learning within the context of maternal health data, particularly for predicting pregnancy complications. This study focused on how federated learning could be used to pool data from rural health facilities without transferring patient data across boundaries. It demonstrated that federated learning could provide accurate predictions of pregnancy outcomes even with the limited data available in rural areas. The study also noted the benefit of federated learning in improving model accuracy through the aggregation of data from multiple, geographically diverse healthcare centers, helping overcome the issue of data sparsity.

While federated learning offers significant benefits, several challenges persist in its implementation in countries like Uganda. One major challenge is the technological infrastructure of healthcare institutions. Most hospitals in Uganda, particularly in rural areas, face constraints such as unreliable internet connectivity and limited access to high-performance computing resources. These challenges were highlighted in [299], where they noted that while federated learning could be beneficial, the lack of reliable data transmission and storage capabilities would likely hinder its widespread adoption in Uganda's healthcare system.

Another study [300] focused on Uganda's health data interoperability challenges, suggesting that the FAIR data principles—Findable, Accessible, Interoperable, and Reusable—are fundamental in addressing issues related to data fragmentation and standardization. The study noted that federated learning could only be effective in Uganda if data from multiple institutions could be integrated and processed. This highlighted the need for both technical and organizational alignment to ensure that health data from disparate systems could be combined for effective federated learning without compromising privacy or security.

Despite these challenges, there have been several successful implementations of federated architecture in Uganda, particularly in maternal and child health data. For example, [301] explored a pilot project that implemented a federated learning model for predicting maternal health risks, using data from a network of Ugandan hospitals. They found that federated learning enabled the creation of a more accurate and generalized model by pooling data from different regions of Uganda, which otherwise would not have been possible due to the fragmented nature of the data. This project also demonstrated how federated architecture can facilitate collaboration between institutions while ensuring patient privacy through secure data-access protocols.

The integration of federated learning in healthcare requires that the data used across institutions be standardized and interoperable, which ties directly into the FAIR principles. Federated architecture is highly effective when data is structured in a way that makes it findable, accessible, interoperable, and reusable across different healthcare institutions. In this regard, [302] emphasized the role of standardized healthcare data formats like HL7 and FHIR to ensure that data from different sources can be integrated effectively.

A key challenge in the African context, especially in Uganda, is that many healthcare institutions use proprietary or non-standardized data formats. This lack of interoperability can

hinder the success of federated learning models, as data from different sources cannot be easily merged. For federated architecture to work effectively, data standards must be established, and healthcare institutions must collaborate on data integration efforts. This need for standardization has been highlighted in several studies, such as [303], which recommend that healthcare institutions in Africa move towards adopting global interoperability standards to facilitate federated data access.

Moreover, the use of federated learning models with FAIR health data allows the pooling of diverse health data across various African countries, facilitating cross-border health collaborations. For instance, [304] found that federated learning could enhance predictive models for disease outbreaks across multiple African regions by leveraging data from different healthcare systems without compromising patient privacy.

8.2.4 Data Science and Big Data in Health

The rise of data science and big data has enabled the processing and analysis of large volumes of health data, completely transforming health research. These developments have improved patient care and resource utilization by facilitating faster and more accurate decision-making in the healthcare industry [162]. On the other hand, FAIR principles and FML aim to solve the issues of privacy-preserving accessibility, quality, standardisation, and interoperability that are brought up by the integration of big data into healthcare, as maintaining high standards in managing data is important to ensure success in applying big data analytics [174].

8.3 Use cases, Data and Methods

This study demonstrates the use of FAIRified data in practical applications in machine learning modelling for data analysis and prediction. The use cases include:

1. **Use case 1:** FAIR Outpatient Department Uganda diagnoses with K-means clustering analysis.
2. **Use case 2:** FAIR Integrated Antenatal Care (ANC) Uganda Data Anomaly Detection.
3. **Use case 3:** FAIR ANC Uganda Pregnancy risk factors and fetal state determinants' clustering analysis.
4. **Use case 4:** FAIR ANC Uganda Data Fetal State Prediction Modelling using traditional ML and federated learning algorithms.

This implementation employs different data sets for each of the use cases mentioned to evaluate the performance of the models trained on Outpatient Registry and Antenatal Care register data from Uganda.

8.3.1 Data Sources and Code

The FAIR data used for the clustering and training of machine learning models were retrieved from the VODAN Africa AllegroGraph triple store databases in Uganda. These databases pro-

vide access to data governed by shared Data Use Agreements, ensuring ethical and compliant data usage. AllegroGraph, a graph database, specializes in managing RDF triple data structures for Linked Data applications [175]. Data retrieval utilized AllegroGraph version **[7.3.1], available at** [1]. Access to the repositories was established after setting up SPARQL queries using the SPARQLWrapper library to communicate with the Allegrograph SPARQL endpoint. Authentication credentials were provided via the setCredentials method, enabling secure access to the repositories. These data are not open to the public, but are accessible to authorized members of the VODAN-Africa network, including researchers affiliated with participating health facilities and other approved collaborators. The queries retrieved relevant data points essential for building and validating the models, while maintaining strict access control to protect sensitive health information.

These data are stored and retrieved from Allegrograph triple-store database repositories with clear protocols and usage rights, allowing authorized users to easily access, query, and analyse whenever needed under well-defined terms. The data templates and graph models used to structure the FAIR data across all repositories and the code used to train different machine learning models on the FAIR data are available on this repository containing this work's documentation.

The Antenatal Care Register(ANC) dataset contains anonymized data on pregnant women. The ANC Uganda dataset comprises the woman's age, gestational age, height, weight, and mid-upper arm circumference (MUAC) measurement. The shape of the data in this dataset included 909 patient records.

To increase the diversity and volume of model training data, the FAIR OPD Uganda datasets (Outpatient Department), as well as the FAIR ANC Uganda datasets (Antenatal Care), were augmented using the synthetic minority oversampling technique (SMOTE). SMOTE is a data augmentation technique that is used to address class imbalance by generating synthetic samples for underrepresented classes. This method enhances the robustness of machine learning models by preventing bias toward majority classes, ensuring better generalization and predictive performance. The FAIR OPD Uganda datasets were expanded from 200 instances each to 2,114 instances.

8.3.2 Federated Learning Architecture for Local Hospital Systems

In a Federated Learning (FL) architecture for healthcare, the data remain decentralized and stay at local sources, such as individual hospital databases, while the model updates (not the raw data) are shared with a central server. This design preserves patient privacy and ensures compliance with data protection regulations by eliminating the need to transfer sensitive medical data outside of the local institution.

How Hospitals Act as Local Clients

In this federated architecture, each hospital acts as a local client with its own dataset, often containing patient information, diagnostic records, or medical imaging data. The hospital does

[1] `http://102.130.124.31:8081/`

not share its data with other institutions or a central server. Instead, it runs machine learning (ML) tasks, such as training a model, on its local data set.

The following outlines the process in detail.

- **Local Model Training:** Each hospital trains a machine learning model using its local data. For example, hospitals might use ML algorithms to predict disease outcomes, identify patient risk factors, or cluster patients based on similar characteristics. The model is trained solely on the local data.
- **Model Updates:** After the local training, each hospital sends the model updates (such as the weights and parameters adjusted during the learning process) to a central server. These updates do not contain sensitive patient data, thus preserving privacy.
- **Global Model Aggregation:** The central server aggregates all the received model updates and combines them to form a global model. This is done using the Federated Averaging technique, which calculates the average of the model updates received from each hospital. The global model benefits from the collective learning of all participating hospitals without the need to share data.
- **Distributing the Global Model:** The global model, which now incorporates insights from multiple hospitals, is sent back to each local hospital. The hospitals can then apply the updated global model to make predictions on new patient data or improve their local model further.

This decentralized approach ensures that while the knowledge from multiple hospitals contributes to the overall model, each hospital retains control over its own data, which remain local and private.

Challenges of Implementing Federated Learning in Local Hospitals

1. Limited Computing Power:
    - Many hospitals, especially in rural or resource-constrained settings, may not have the computational infrastructure necessary to run sophisticated ML algorithms.
    - Solution: Light-weight models and edge computing can be used to process smaller datasets locally. Additionally, computation-intensive tasks could be offloaded to a central server or cloud system, especially during aggregation of the model updates.
2. Data Heterogeneity:
    - Data across different hospitals may vary in format, structure, or quality, making it difficult to train uniform models.
    - Solution: Adopting standardized data formats (such as those defined by FAIR data principles) can help ensure data interoperability and facilitate smooth integration across different systems. Hospitals might need to pre-process and clean the data to ensure that it can be used effectively for ML tasks.

3. Network and Connectivity Issues:

    - In rural areas, internet connectivity can be unreliable, complicating the transmission of model updates.
    - Solution: Implementing periodic synchronization could help mitigate this issue. Hospitals could collect model updates locally and then transmit them to the central server during stable connectivity windows.

4. Data Privacy and Security:

    - Even though FL does not transfer raw data, transmitting model updates can still pose privacy risks if not properly encrypted.
    - Solution: Encryption and secure multi-party computation (SMPC) methods can be used to ensure that model updates are transmitted securely. These techniques allow hospitals to collaborate on machine learning models while maintaining the confidentiality of patient data.

Transitioning from Centralized Machine Learning to Federated Systems

The transition from a centralized machine learning model, where all data is stored and processed in a central database, to a federated architecture involves several steps:

- Data Preparation: The first step is to ensure that the data from various hospitals is standardized, accessible, and compliant with FAIR principles. Hospitals must ensure that data are well structured, with clear metadata and standardized formats to enable easy aggregation and model training.
- Model Development: The hospitals can then apply machine learning techniques to local data. The ML models will initially be trained on centralized datasets as a proof of concept, but the final models could be adapted to work in a federated setting.
- Infrastructure Setup: The technical infrastructure supporting the federated architecture must be established. This includes setting up secure communication channels, cloud or edge computing resources for model aggregation, and ensuring that local hospitals have the necessary hardware and software to run ML algorithms.
- Model Deployment and Monitoring: After federated learning is implemented, the models must be continuously updated and monitored. Hospitals will need to participate in periodic model updates, and the central server must manage the aggregation of new model updates and ensure that the system remains secure and efficient.
- Scaling and Support: For a wider deployment, it is crucial to ensure that the federated system can scale to accommodate more hospitals. This requires careful planning around data integration, secure communication protocols, and cloud or edge computing resources to support computational demands.

8.3.3 Machine Learning Modeling Workflow

The machine learning models employed in these implementations primarily utilize supervised learning methods, such as KNN, MLP, and Gradient Boosting, driven by the availability of labeled datasets. The Isolation Forest and K-means algorithms, which are unsupervised learning techniques, are implemented for anomaly detection and patient data clustering tasks, respectively. The workflow used to demonstrate the application of FAIR data in machine learning is illustrated in Figure 8.1

1. Data Pipeline This step consists of data acquisition, data preparation, and data preprocessing. Data acquisition involves retrieving FAIR data from various sources, such as the Outpatient Department Register (OPD) Uganda and the Integrated Antenatal Care Register (ANC) Uganda Allegrograph databases. The data preparation step cleans and organizes the data by removing duplicates and handling missing values by removing rows that have missing values or converting data formats [177]. Data pre-processing involves selecting features that are relevant to the intended purpose of the models, such as 'age', 'blood pressure', and 'blood sugar ', to cluster patient pregnancy risk factors.[178]. This step also includes the scaling of features and the encoding of categorical variables [179].

2. Model engineering This step comprises the definition of the model, the training of the model, and the evaluation. The definition of a model involves choosing the type of model to use based on the problem and the available data. For example, a K-means model is selected and defined in the clustering of patient data. Multiple models can be chosen and trained so that the best-performing model is selected for deployment [180]. The chosen model is then trained on the preprocessed data. This involves feeding the data through the model, comparing the predictions with the actual values, and adjusting the model parameters to minimize error [217]. The performance of trained models is assessed using task-relevant metrics [220], such as silhouette scores for clustering algorithms and classification reports showing precision, recall, F1 score, and support for predictive models. In addition, accuracy scores are reported to provide an overall classification performance evaluation.

   Validation Techniques for Overfitting: To prevent overfitting, the following validation techniques are implemented where appropriate, based on the model type:

   - Early stopping: Monitoring validation loss during training of neural networks to stop before over-fitting occurs
   - Data augmentation: SMOTE (Synthetic Minority Oversampling Technique) [319] was subsequently applied to address data imbalances and compensate for any loss of diversity in the dataset due to missing value exclusion. Adding controlled Gaussian noise to training data was implemented to improve model robustness
   - Regularization: Applying L1 (Lasso) and L2 (Ridge) penalties to model parameters to prevent them from growing too large
   - Evaluation of the test set: Assessing model performance in a previously separated section of the data set to measure generalization ability

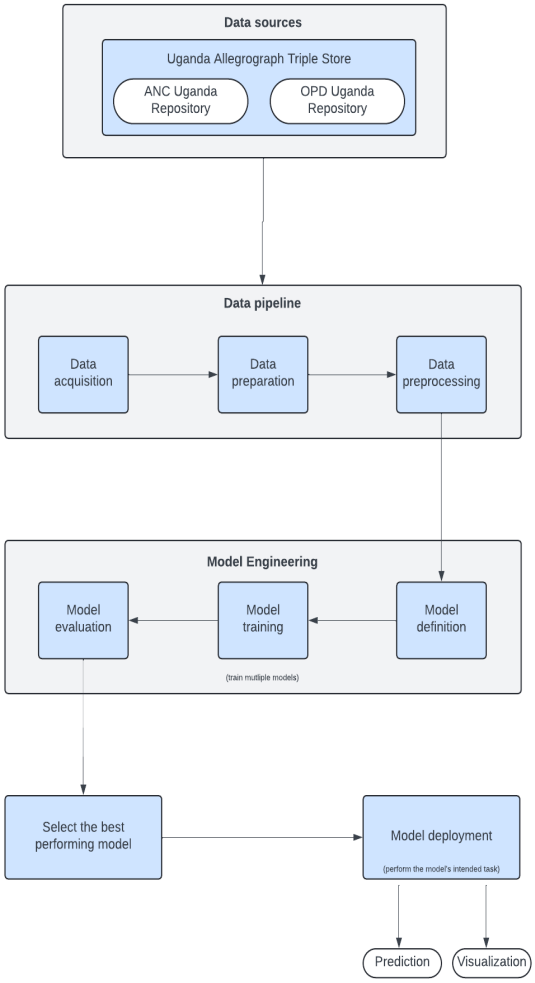


Figure 8.1: Implemented machine learning workflow

- Cross-validation: Evaluating model performance across multiple random train-test splits to ensure consistent behavior

3. Select the best performing model. Multiple models were trained, and the one with the best performance on evaluation metrics is selected. This decision also considers factors like model complexity and computational requirements

4. Model deployment. The selected model is implemented in a production environment where it can perform its intended task on new, unseen data. In the machine learning implementation carried out for this study, models that are used for prediction are presented with user input to predict the pregnancy risk factors and fetal state of a hypothetical patient.

5. Prediction and visualization. These are the outputs of the deployed model. Predictions are the model's results on new data, while visualization helps interpret results from the provided data.

The following sections discuss the implementation of this workflow that has been illustrated in the outlined case studies.

8.3.4 Use Case 1: FAIR Outpatient Department Uganda Diagnoses with K-Means Clustering Analyses

Data Acquisition and Preparation

The FAIR OPD data used to train the K-means clustering algorithms were acquired from the Allegrograph repository of the Ugandan Outpatient Registry Department, which contains information collected from hospital visits, including age, sex, district of residence, blood pressure, and blood sugar. The following query was run on the OPD Uganda repository to retrieve the required data:

SPARQL Query 4

```
PREFIX opd: <http://dashboard.vodana.health/>

SELECT ?s ?age ?sex ?diagnosis ?prescription ?district
?blood_pressure ?blood_sugar
WHERE {
  ?s opd:Age ?age ;
     opd:Sex ?sex ;
     opd:Diagnosis ?diagnosis ;
     opd:Prescription ?prescription ;
     opd:District ?district ;
     opd:BloodPressureMmhg ?blood_pressure ;
     opd:BloodSugarMmL ?blood_sugar .
}
```

The results of running this query were exported, yielding 168 instances and seven features. After confirming the absence of duplicates, missing values in the dataset were initially handled by list-wise deletion. While this method is simple to implement, as Kang [223] notes, it can reduce statistical power due to data loss. Given the potential limitations of list-wise deletion, SMOTE was subsequently applied to address data imbalances and compensate for any loss of diversity in the dataset due to missing value exclusion. By generating synthetic samples for minority classes, SMOTE ensured that the final data set was balanced and robust for model training and validation [223].

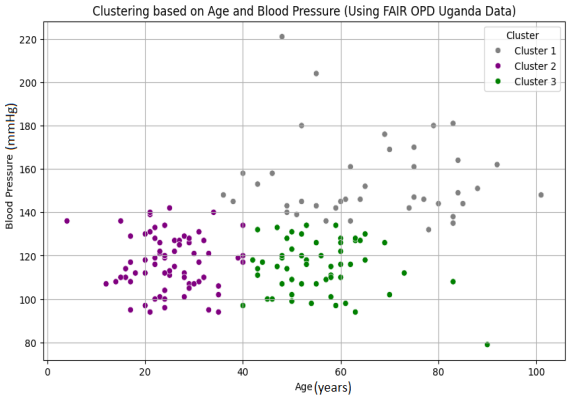
OPD Uganda K-Means Clustering Analysis based on Patient Age

Data Preprocessing
Three features, *age*, *blood pressure*, and *blood sugar*, were selected from the dataset for clustering. To ensure that the clustering algorithm treats all features equally, a `standardScaler` was used to standardize the data. This step transforms the features such that they have a mean of zero and a standard deviation of one, which is crucial for algorithms like K-Means that rely on distance-based measurements.
The K-Means Clustering Algorithm
A clustering algorithm was implemented to group patients according to their age, blood pressure, and blood sugar levels using data from the FAIR Ugandan Outpatient Register. Clustering is useful for identifying patterns or subgroups within the data that might not be immediately apparent.
The K-Means algorithm was applied with a predefined number of clusters, in this case 3 and 5. The number of clusters were chosen arbitrarily, but can be adjusted based on domain knowledge. The `fit_predict` method assigned a cluster label to each patient, which was stored in a new column called *cluster* in the data set.
Visualizations were generated for each variation of the data set to examine how the clusters are formed based on the features; age, blood pressure, and blood sugar. Seaborn's `scatterplot` was used for these visualizations, with different colors representing different clusters. These graphs help us to understand the distribution of the groups and their relation to age, blood pressure, and blood sugar levels.

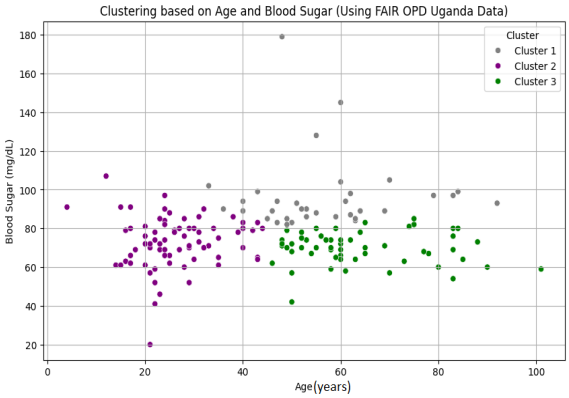
Cluster analysis was conducted using data derived from the FAIR Uganda Outpatient Register, which adheres to the FAIR (Findable, Accessible, Interoperable and Reusable) data principles. The FAIR-compliant structure of the data set allowed efficient retrieval and integration of patient-level variables such as age, blood pressure, and blood sugar for analytical purposes. Through the use of standardized and semantically structured health records, the data could be transformed into machine-learning-ready formats while maintaining consistency and interoperability. This facilitated reproducible clustering analyzes and demonstrated how FAIR data infrastructures can support advanced data-driven healthcare research, knowledge discovery, and future artificial intelligence applications.

Cluster plots from clustering using FAIR OPD Uganda data (3 clusters):
K-Means Clustering Plots (no. of clusters = 3)

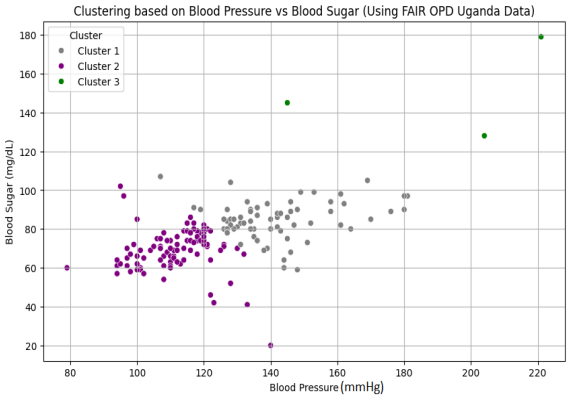
Figure 8.2 illustrates the K-Means clustering results obtained from the FAIR OPD Uganda dataset using three clusters ($K = 3$) and three different parameter pairings, namely age versus blood pressure, age versus blood sugar, and blood pressure versus blood sugar. The cluster-



(a) Clustering of OPD Uganda Patient Data Using K-Means: 3 Clusters Based on Age(yrs) and Blood Pressure(mmHg) with FAIR Data



(b) OPD Uganda Patient Data K-Means Clustering Plots, By 3 Clusters, Based on Age(yrs) and Blood Sugar(mg/dL) Using FAIR data



(c) OPD Uganda Patient Data K-Means Clustering Plots, By 3 Clusters, Based on Blood Pressure(mmHg) and Blood Sugar(mg/dL) Using FAIR data

Figure 8.2: K-Means clustering of OPD Uganda patient data with three different parameter pairings (where K=3, n=168)

ing process grouped patients according to similarities in their clinical measurements, enabling the identification of distinct patient profiles within the dataset. Prior to clustering, the data was standardized to ensure equal contribution of each feature during model training. After clustering, the cluster centers were inversely transformed back to the original scale to facilitate meaningful interpretation of the resulting patient groups.

Figure 8.2(a) presents the clustering distribution based on age and blood pressure measurements. The visualization demonstrates the presence of distinct patient groups characterized by varying blood pressure levels across different age ranges. Figure 8.2(b) illustrates the clustering pattern for age and blood sugar values, highlighting variations in glucose levels among patients belonging to different age categories. Similarly, Figure 8.3(c) shows the clustering re-

sults based on blood pressure and blood sugar, providing insights into the relationship between hypertension-related indicators and glucose measurements within the patient population.

The cluster centers provide a quantitative summary of the average characteristics of each identified group. Cluster 1 represents patients with moderate age and relatively high physiological measurements, while Cluster 2 corresponds to comparatively younger patients with lower average measurements. In contrast, Cluster 3 represents older patients with elevated physiological indicators, suggesting potentially higher health risks. These clustering results demonstrate the ability of unsupervised learning techniques to identify meaningful structures and hidden patterns within healthcare datasets, thus supporting patient stratification and data-driven clinical analysis.

From a FAIR data perspective, the clustering analysis demonstrates the value of using data that are Findable, Accessible, Interoperable, and Reusable for secondary analytical purposes. The FAIR-compliant OPD Uganda dataset enabled the integration and standardized processing of clinical variables such as age, blood pressure, and blood sugar, facilitating reliable unsupervised learning analysis. The interoperability of the dataset ensured that measurements were consistently represented and machine-readable, while its reusability supported the discovery of clinically meaningful patient groups without requiring extensive data restructuring. These findings illustrate how FAIR data principles can enhance healthcare analytics by enabling transparent, reproducible, and scalable data-driven investigations.

Table 8.1: Cluster descriptions for patient data based on different parameter pairings

Parameter Pairing	**Cluster**	**Patients**	**Description**
Age vs Blood Pressure	Cluster 1	41	Elderly individuals with high blood pressure, indicating potential risks of hypertension and diabetes.
	Cluster 2	71	Younger individuals with normal blood pressure, indicating lower risks of hypertension and diabetes.
	Cluster 3	56	Elderly individuals with normal blood pressure, indicating low risks of hypertension and diabetes.
	Total	**168**	
Age vs Blood Sugar	Cluster 1	34	Elderly individuals with normal blood sugar, indicating low risks of hypertension and diabetes.
	Cluster 2	76	Younger individuals with normal blood sugar, indicating lower risks of hypertension and diabetes.

Continued on next page

Table 8.1 continued from previous page

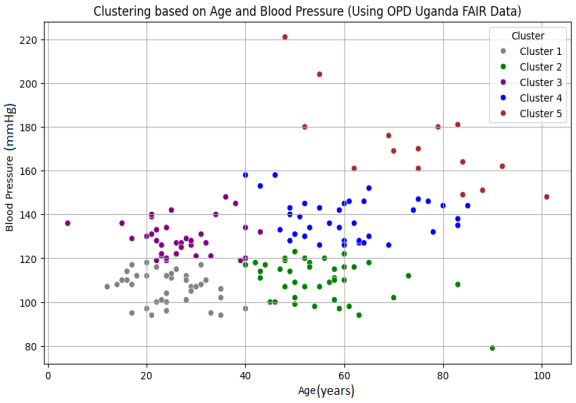
Parameter Pairing	Cluster	Patients	Description
	Cluster 3	58	Elderly individuals with normal blood sugar, indicating low risks of hypertension and diabetes.
	Total	**168**	
Blood Pressure vs Blood Sugar	Cluster 1	68	High blood pressure with normal blood sugar, indicating medium risks of hypertension.
	Cluster 2	97	Normal blood pressure with low blood sugar, indicating lower risks of hypertension and diabetes.
	Cluster 3	3	High blood pressure with high blood sugar, indicating high risks of hypertension and diabetes.
	Total	**168**	

Cluster Centers and Clustering Plots using FAIR OPD Uganda data (5 Clusters):

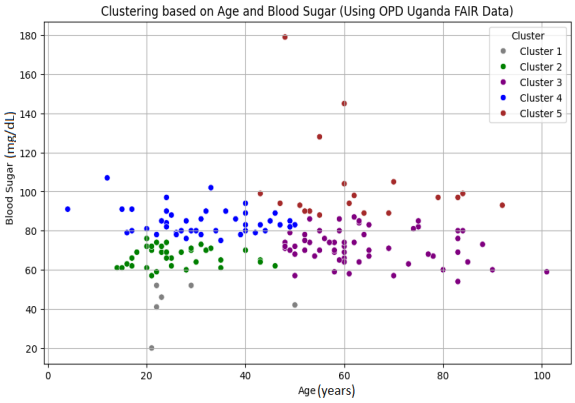
K-Means Clustering Plots (no. of clusters = 5)

Figure 8.3 presents the K-Means clustering results for the OPD Uganda patient dataset using FAIR data principles. The clustering analysis was performed using three different combinations of clinical parameters, namely age versus blood pressure, age versus blood sugar, and blood pressure versus blood sugar. These visualizations demonstrate how patient records can be grouped into distinct clusters based on similarities in physiological characteristics and health indicators. The clustering approach enables the identification of meaningful patterns within the dataset, supporting exploratory analysis and population stratification for healthcare decision-making.

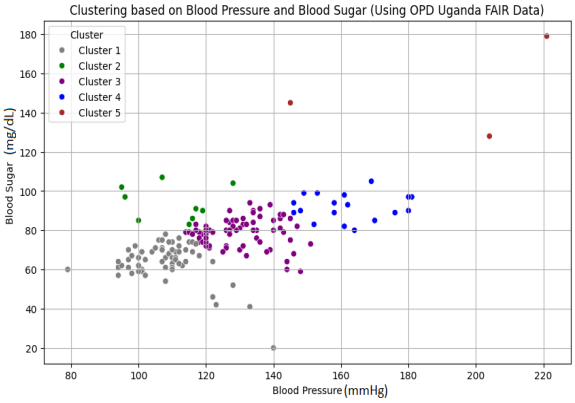
In Figure 8.3(a), patients are clustered according to age and blood pressure values, revealing distinct groups associated with varying blood pressure trends across different age ranges. The visualization indicates that certain patient groups exhibit elevated blood pressure levels at younger ages, while others maintain relatively normal blood pressure distributions. Figure 8.3(b) illustrates clustering based on age and blood sugar levels, highlighting variations in blood glucose patterns among different age groups. Similarly, Figure 8.3(c) shows clustering based on blood pressure and blood sugar measurements, demonstrating relationships between hypertension-related indicators and glucose levels. Collectively, these clustering results provide insights into patient health profiles and may support the early identification of high-risk patient groups for targeted healthcare interventions.



(a) OPD Uganda patient data K-Means clustering based on age(yrs) and blood pressure(mmHg) using FAIR data



(b) OPD Uganda patient data K-Means clustering based on age(yrs) and blood sugar(mg/dL) using FAIR data



(c) OPD Uganda patient data K-Means clustering based on blood pressure(mmHg) and blood sugar(mg/dL) using FAIR data

Figure 8.3: K-Means clustering plots for OPD Uganda patient data using FAIR data ($k = 5$, $n = 168$).

Table 8.2: Cluster descriptions for patient data based on different parameter pairings

Parameter Pairing	Cluster	Patients	Description
Age vs Blood Pressure	Cluster 1	38	Young with normal blood pressure, indicating low risks of hypertension.
	Cluster 2	41	Elderly with normal blood pressure, indicating lower risks of hypertension and diabetes.
	Cluster 3	36	Young with high blood pressure, indicating high risks of hypertension and diabetes.
	Cluster 4	38	Elderly with high blood pressure, indicating high risks of hypertension and diabetes.
	Cluster 5	15	Young with high blood pressure, indicating high risks of hypertension and diabetes.
Age vs Blood Sugar	Cluster 1	6	Young with low blood sugar, indicating low risks of hypertension and diabetes.
	Cluster 2	38	Young with low blood sugar (range 26/67), indicating lower risks of hypertension and diabetes.
	Cluster 3	61	Elderly with normal blood sugar (range 64/72), indicating low risks of hypertension and diabetes.
	Cluster 4	44	Young with normal blood sugar, indicating low risks of hypertension and diabetes.
	Cluster 5	19	Elderly with normal blood sugar, indicating low risks of hypertension and diabetes.
Blood Pressure vs Blood Sugar	Cluster 1	65	Normal blood pressure with low blood sugar, indicating low risks of hypertension.

Continued on next page

Table 8.2 continued from previous page

Parameter Pairing	Cluster	Patients	Description
	Cluster 2	9	Normal blood pressure with normal blood sugar, indicating lower risks of hypertension and diabetes.
	Cluster 3	73	High blood pressure with normal blood sugar, indicating high risks of hypertension.
	Cluster 4	18	High blood pressure with normal blood sugar, indicating high risks of hypertension.
	Cluster 5	3	High blood pressure with high blood sugar, indicating high risks of hypertension and diabetes.

The differences between the 3 and 5 cluster analysis using the FAIR OPD Uganda data are:

Granularity of Age Groups: The 3-cluster approach produced broad age categories (younger 37, middle-aged 44, older 74), while the 5-cluster approach offered more nuanced age groupings (young 26, younger 29, middle-aged 51-58, older 81).

Health Risk Stratification: The 3-cluster approach provided general health categories (healthy, at-risk, high-risk), while the 5-group approach produced more detailed risk profiles, including a distinct group for very high-risk individuals.

Blood Pressure Patterns: The 3-cluster approach involved simple categorizations (normal, high), while the 5-cluster approach revealed more varied levels (healthy, slightly elevated, moderately high, very high).

Blood Sugar Patterns: The 3-cluster approach defined basic categories (normal, elevated, moderately elevated), while the 5-cluster approach identified finer distinctions in blood sugar levels. This includes three clusters for normal, high, and very high blood sugar levels to provide a better understanding of population glycemic control.

Insights into Young Adult Health: The 3-cluster approach identified one young group, while the 5-cluster approach provided more detailed insights by dividing the younger demographic into two groups. This differentiation enabled the identification of individuals with normal and those with slightly higher values, providing a better understanding of the variations in the health of young adults.

Insights into the Middle-Aged Population: The 3-cluster approach involved categorizing all middle-aged people into one cluster. However, the 5-cluster approach offered a more refined analysis by separating the middle-aged population into two distinct groups.

They are those who fall into the middle-risk category and those that fall into the high-risk category. This finer stratification makes it easier to conduct health risk assessments for the middle-aged population.

The 5-cluster approach provides a more detailed view of the health patterns of the population, potentially allowing for more targeted interventions and risk assessments. However, the 3-cluster approach offers a simpler, more general overview that might be easier to interpret and act upon in some contexts.

Normal Blood Pressure and Blood Sugar Levels by Age Group

Blood Pressure:

Normal blood pressure is typically considered to be around 120/80 mmHg.

The ideal range can vary by age. Older adults may have slightly higher normal levels (e.g., up to 140/90).

Blood Sugar Levels:

Fasting blood glucose levels for most people should be between 70-100 mg/dL.

Postprandial (after-meal) blood sugar levels are typically expected to stay below 140 mg/dL.

These ranges serve as general guidelines; individual health status or conditions (like diabetes) can influence what is considered "normal."

The following table 8.3 presents the normal ranges for blood pressure and blood sugar levels, categorized by age group. These values serve as a baseline to detect any anomalies or abnormalities in an individual's health data [318].

Table 8.3: Normal Blood Pressure and Blood Sugar Levels by Age Group

Age Group	**Normal Blood Pressure (Systolic/Diastolic) (mmHg)**	**Normal Blood Sugar Level (Fasting) (mg/dL)**	**Normal Blood Sugar Level (Postprandial) (mg/dL)**
0-9 years	90/60 to 110/75	70-100	90-140
10-19 years	90/60 to 120/80	70-100	90-140
20-29 years	90/60 to 120/80	70-100	90-140
30-39 years	90/60 to 120/80	70-100	90-140
40-49 years	90/60 to 130/85	70-100	90-140
50-59 years	90/60 to 130/85	70-100	90-140
60-69 years	90/60 to 140/90	70-100	90-140
70+ years	90/60 to 140/90	70-100	90-140

Table 8.3 presents the normal reference ranges for blood pressure and blood glucose levels across different age groups. These clinical thresholds serve as baseline indicators for identi-

fying physiological abnormalities within the ANC dataset. Blood glucose measurements are particularly important for detecting diabetes-related conditions, including gestational diabetes mellitus, while blood pressure measurements assist in identifying hypertensive disorders such as gestational hypertension and preeclampsia. By comparing patient observations against these established reference ranges, potential health anomalies can be identified and incorporated into subsequent machine learning analyses, including anomaly detection and patient clustering, thereby supporting early risk identification and improved maternal healthcare outcomes.

8.3.5 Use Case 2: FAIR Integrated Antenatal Care (ANC) Uganda Data Anomaly Detection and Patient Data Clustering Modelling

This section focuses on the process followed to train and apply the Isolation Forest and K-Means Clustering algorithms to analyze the Integrated Antenatal Care Data from Uganda.

FAIR ANC Uganda Acquisition and Preparation

The data used for training machine learning models to analyze patterns among pregnant women according to Integrated Antenatal Care Uganda were acquired from the repositories of the Uganda Allegrograph databases that store data collected from hospital visits. The ANC Uganda dataset contains information on pregnant women, with features: the woman's age, gestational age, date of patient visit, weight and height measurements, and blood pressure. The following query was run on the mentioned repositories to retrieve the required data:

Query 5 on the ANC Uganda repository

SPARQL Query 5

```
PREFIX UgANC: <https://dashboard.vodana.health/>
SELECT ?s?Gestational_age ?Age ?Date_of_Patient_Visit
WHERE {
?s UgANC:Gestational_age ?Gestational_age .
OPTIONAL { ?s UgANC:Age ?Age }
OPTIONAL { ?s UgANC:Date_of_Patient_Visit ?Date_of_Patient_Visit }
}
```

FAIR ANC Uganda Dataset Preprocessing

After data cleaning through handling missing values and removing outliers, the cleaned ANC Uganda dataset had 794 instances and 7 features. A function was created to identify and encode categorical and numeric columns. One-hot encoding was used to convert categorical variables into a format that can be provided to ML algorithms to perform better. The numeric columns were concatenated with the categorical columns encoded into a single data frame. This was followed by dropping rows with missing values to ensure a complete dataset for verification. The "BinaryLabelDataset" from the "AIF360" library created a dataset with the specified

labels ("target") and protected attributes [320]. In this case, the protected attribute is a specific age group. The privileged and unprivileged groups were then defined on the basis of the protected attribute. In this case, people aged 20-30 were considered privileged, while others were unprivileged. The "Reweighing algorithm" was used to adjust the weights of the instances in the data set to mitigate bias. It ensures that the learning algorithms trained on the data are less likely to inherit biases in the unaltered data. The transformed data set was converted back to a Pandas DataFrame. The one-hot encoded columns were then reversed to their original categorical form.

FAIR ANC Uganda Anomaly Detection and Clustering among Pregnant Women

The Isolation Forest algorithm was implemented to detect anomalies in the data, focusing on the features "gestationalAge", "age", and "visit_year" . The dataset contains first time visit instances of the pregnant women in Uganda. This algorithm is an unsupervised learning method for the detection of anomalies. It works by isolating anomalies in the data through random partitioning, exploiting the fact that anomalies are typically fewer and different from normal instances. This approach is particularly effective for high-dimensional data sets and is computationally efficient compared to many other anomaly detection techniques [197, 198, 199]. In practice, the model identifies approximately 10% of the data as anomalies and assigns an anomaly score to each data point. The results are visualized in a scatter plot showing maternal versus gestational age. The anomalies in the scatter plot were highlighted in different colors, allowing a clear identification of unusual patterns in the data set.

Recent advancements in anomaly detection, such as the MS2OD method [199], have demonstrated improved accuracy by combining minimum spanning trees with medoid-based selection strategies. These methods are particularly promising for complex medical datasets where traditional models may miss subtle irregularities. While Isolation Forest offers speed and generalizability, integrating ideas from structure-aware algorithms like MS2OD could enhance the detection of context-specific outliers in future work.

Findings from Anomaly Detection in ANC Uganda using Isolation Forest Algorithms

The plot 8.4 shows the relationship between the age of the women (x-axis) and the gestational age (y-axis). The ages of women range from about 15 to 45 ywars, while gestational ages range from approximately 5 to 40 weeks. This data distribution provides an overview of the range of ages and gestational periods observed in the study. Most data points are classified as normal (red dots). Normal data points indicate typical cases and help to identify general trends. The anomalies (blue dots) are scattered throughout the plot, but are more clearly concentrated at the edges of the data distribution.

The detected anomalies are primarily located at the boundaries of the maternal age and gestational age distributions, particulary among very young and older women, as well as cases with extremely low or high gestational ages. These observations deviate from the dense central cluster of normal cases identinfied by the isolation forest model. The highest density of normal data points appears to be for women aged 20-35 years with gestational ages between 15-35 weeks, which is the sample population's most common age and gestational age range. This finding may be useful for targeting healthcare resources and interventions to the demographic that is most in need of care, such as pregnant women seeking clinic services for the first time.

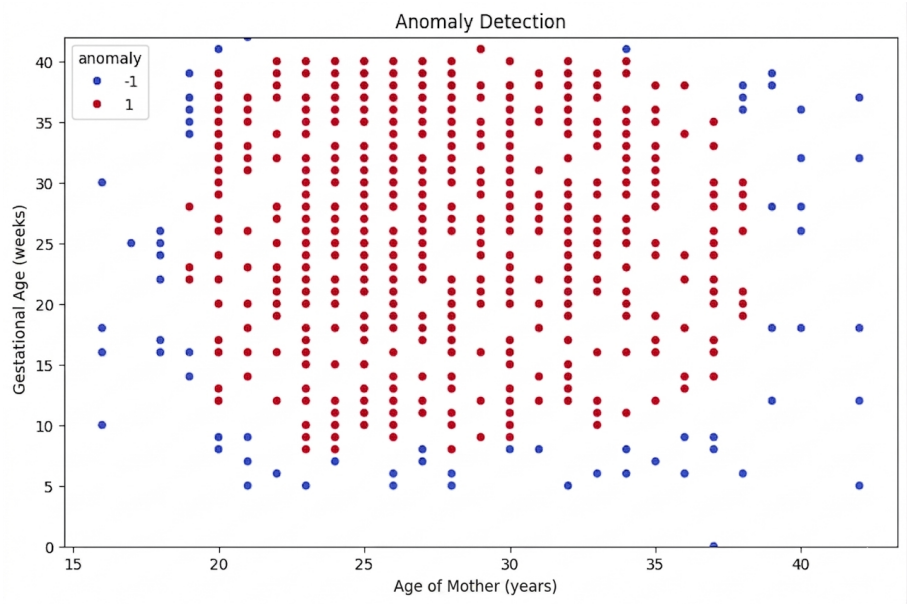


Figure 8.4: ANC Uganda Anomaly Detection Using Isolation Forest Algorithm by Age(yrs) and Gestational age(wks)

The anomalies observed at both extremes of the maternal age spectrum suggest that younger women may present at relatively higher gestational ages compared to other age groups.These age-related patterns suggest that specialized care and attention may be necessary for these groups due to their increased risk factors.

The wide range of gestational ages in all maternal ages indicates that women seek antenatal care at various stages of their pregnancy, irrespective of their age. This variation in gestational age highlights the diverse timing of antenatal visits and the need for flexible care strategies. Anomalies with very high or very low gestational ages can indicate high-risk pregnancies that require special monitoring and intervention within the Ugandan healthcare system. Identifying these anomalies can help prioritize and manage high-risk cases more effectively.

The data distribution suggests that most women are recorded in the dataset during mid-pregnancy. This pattern could be attributed to standard antenatal care schedules in Uganda or local healthcare seeker behaviors. Understanding these data collection patterns is crucial to improving the timing and effectiveness of antenatal care services.

The plot depicted in Figure 8.4 shows the outcome of the anomaly detection process. Using the Isolation Forest algorithm, patient records were classified as normal or anomalous, with anomalies representing observations that exhibit unusual characteristics of maternal age and gestational age relative to the overall population distribution.

8.3.6 Use Case 3: FAIR ANC Uganda Data Pregnancy Risk Factors and Fetal States Determinants' Clustering Analysis

FAIR ANC Uganda Data K-Means Clustering Analysis by Pregnancy Risk Factors

The data set consists of 909 records and 70 attributes. The numerical features that include "Weight(kg)", "Height_(cm)", and "Blood_pressure" were converted to appropriate data types, with non-numeric values being systematically replaced with NaN values. A Body Mass Index (BMI) feature was created using the following formula:

$$\text{BMI} = \frac{\text{Weight (kg)}}{\left(\frac{\text{Height (cm)}}{100}\right)^2} \tag{8.1}$$

Pregnant risk factors were determined using a function that classified patients based on multiple clinical parameters: blood pressure measurements and BMI thresholds. The classification included four risk categories: Hypertension, Obesity, Malnutrition, "no risk factor", which was the default category for cases that did not meet any risk criteria. The two cluster analyzes for the age-risk and gestational age-risk relationships were performed simultaneously. Categorical risk factors were encoded using LabelEncoder and numerical features were scaled using RobustScaler to ensure comparable feature ranges. To address class imbalance issues, SMOTE was applied to generate synthetic samples for underrepresented classes. For gestational age analysis, domain-specific constraints were implemented to filter the data to clinically relevant ranges (8-42 weeks). The K-Means algorithm was configured to partition the data into three distinct clusters (n_clusters= 3) for both analyzes. The quality of the clustering was evaluated using silhouette scores, which measure the cohesion and separation of the resulting clusters [321]. To evaluate the quality and effectiveness of the K-Means clustering results, the silhouette score was used as an internal validation metric. The silhouette score measures how similar a data point is to its own cluster compared to other clusters. Provides an indication of the degree of separation and cohesion among the clusters, helping to determine whether the selected number of clusters is appropriate for the data set.

The silhouette score ranges from -1 to 1. A score close to 1 indicates that the data points are well matched to their assigned clusters and clearly separated from neighboring clusters. A score around 0 suggests overlapping clusters or weak separation between groups, while negative values indicate that some data points may have been assigned to incorrect clusters.

In general, silhouette scores above 0.70 are considered strong and indicate highly distinct clusters. Scores between 0.50 and 0.70 represent reasonably good clustering performance with acceptable separation between groups. However, silhouette scores below 0.50 are typically considered weak or poor, suggesting that clusters are not clearly distinguishable and that there is significant overlap between data groups. Such low scores may indicate that the selected number of clusters is not optimal or that the dataset does not naturally form well-separated clusters.

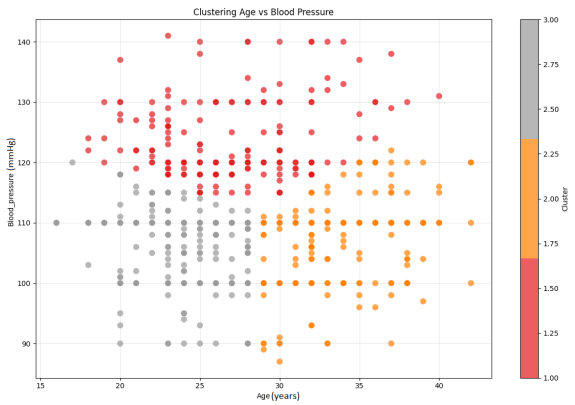
Therefore, in this study, the silhouette score was used as an important criterion to evaluate the reliability of the cluster structure and to determine the suitability of the chosen value of K for patient grouping and exploratory healthcare analysis.

Visualization was accomplished through scatter plots using matplotlib, with the original (non-scaled) age and gestational age values plotted against risk factors. BMI and blood pressure variables were useful for the cluster analysis, but excluded from predictive modeling to avoid inflated classification performance. The plots were enhanced with appropriate labels, color-coding by cluster assignment, and gridlines to facilitate interpretation. The resulting visualizations demonstrate the distribution patterns of risk factors between different age groups and gestational periods in the Ugandan ANC population.

Model Performance Analysis: FAIR ANC Uganda Data K-Means Clustering by Pregnancy Risk Factors

The K-Means cluster analysis produced silhouette scores of 0.41 for age and blood pressure clustering and 0.48 for gestational age and blood pressure clustering. These scores indicate moderate cluster quality, suggesting that while meaningful grouping patterns exist within the FAIR ANC Uganda dataset, the clusters are not strongly separated. This is evident in Figure 8.5, where substantial overlap exists between cluster assignments, as shown by the intermixing of colors across the plots. The gestational age-based clustering demonstrates slightly better separation than age-based clustering, reflected by its higher silhouette score, suggesting that gestational age may be a more informative variable for identifying pregnancy risk-related patient groups. Nevertheless, the absence of clearly distinct cluster boundaries indicates that the underlying data contain gradual transitions rather than naturally well-separated groups.

Findings from FAIR ANC Uganda Data K-Means Clustering by factors Contributing to Risks in Pregnancy



(a) FAIR ANC Uganda Patient Data K-Means Clustering by Age(Yrs) and Blood Pressure(mmHg)

(b) FAIR ANC Uganda Patient Data K-Means Clustering by Gestational Age(wks) and Blood Pressure(mmHg)

Figure 8.5: K-Means clustering plots for FAIR ANC Uganda patient data.

The K-Means clustering analysis revealed distinct patient groupings based on age and blood pressure patterns in plot 8.5(a) within the FAIR ANC Uganda dataset. The observed clusters suggest the presence of varying maternal health profiles and potential pregnancy risk categories associated with elevated blood pressure and demographic characteristics. Although clustering

itself does not measure FAIRness, the FAIR-compliant data infrastructure enabled efficient preprocessing, integration, and reuse of semantically structured clinical variables for advanced analytical tasks. The interoperability and reusability of the FAIR data framework facilitated machine learning-driven exploration of maternal risk patterns while preserving data consistency and accessibility across analytical workflows.

Table 8.4: Cluster Analysis for Age and Blood Pressure (BP)

Cluster	**Sample Size**	**Age (Mean, Range)**	**Blood Pressure (Mean, Range)**	**Dominant Features**
Young & High BP	203 (27.2%)	27.0 (18.0–40.0)	123.3 (11.5–141.0)	Fetal state: Non-reassuring (53.7%); Age group: 26–30 (39.9%)
Mid-age & Normal BP	232 (31.1%)	33.3 (22.9–42.0)	106.7 (87.0–122.0)	Fetal state: Reassuring (53.9%); Age group: 31–35 (48.3%)
Young & Normal BP	312 (41.8%)	23.8 (16.0–28.0)	106.6 (90.0–120.0)	Fetal state: Reassuring (43.6%); Age group: 20–25 (51.6%)
Total	**747 (100%)**	–	–	–

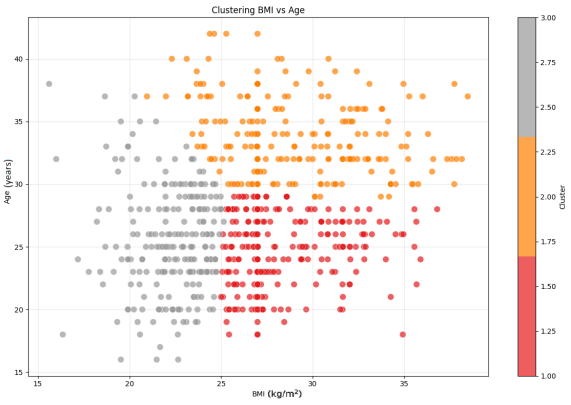
The K-Means cluster analysis revealed clear patient groups based on gestational age and blood pressure trends in plot 8.5(b) within the FAIR ANC Uganda dataset. These groups indicate different maternal health profiles and possible pregnancy risk categories related to variations in blood pressure at various stages of gestation. Although cluster analysis itself does not directly assess FAIRness, the FAIR-compliant data infrastructure supported the integration, preprocessing, and reuse of semantically organized clinical data for advanced analytical applications. The interoperability and reusability provided by the FAIR data framework enabled machine learning-based investigation of maternal risk patterns while promoting standardized and accessible healthcare data analysis.

A comparison of the four plots reveals notable differences in the clustering structure and the strength of the separation between physiological and demographic variables. The Age vs. blood pressure clustering plot shows moderate overlap between clusters, suggesting that although blood pressure varies between age groups, age alone may not strongly differentiate maternal risk categories. However, clusters with elevated blood pressure values appear to be concentrated between certain reproductive age ranges, indicating that blood pressure remains a clinically important indicator of maternal risk.

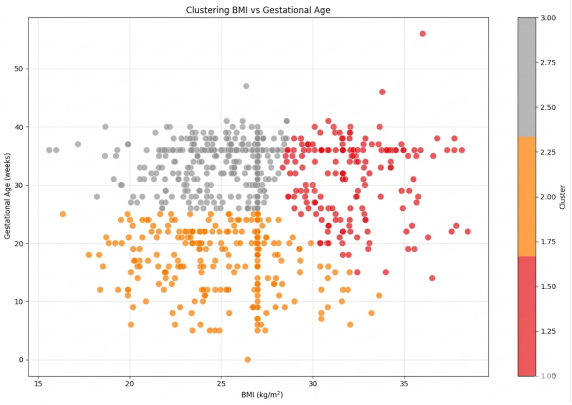
The Gestational Age vs Blood Pressure clustering plot demonstrates stronger and more structured clustering patterns than the Age vs Blood Pressure analysis. Different groups emerge in the gestational stages, particularly among patients who have elevated blood pressure during later periods of pregnancy. This suggests that gestational age may have a stronger relationship

Table 8.5: Cluster Analysis for Gestational Age and Blood Pressure (BP)

Cluster	**Sample Size**	**Gestational Age (Mean, Range)**	**Blood Pressure (Mean, Range)**	**Dominant Features**
Mid Term & Normal BP	339 (45.4%)	33.3 (21.0–56.0)	105.5 (87.0–114.0)	Fetal state: Reassuring (41.6%); Age group: 26–30 (33.0%)
Early Term & Normal BP	236 (31.6%)	16.6 (0.0–25.0)	110.5 (89.0–133.0)	Fetal state: Reassuring (55.5%); Age group: 20–25 (30.9%)
Early Term & High BP	172 (23.0%)	31.6 (15.0–46.0)	123.1 (11.5–141.0)	Fetal state: Non-reassuring (54.7%); Age group: 26–30 (33.1%)
Total	**747 (100%)**			



(a) FAIR ANC Uganda Patient Data K-Means Clustering by Age (Yrs) and BMI (kg/m^2)



(b) FAIR ANC Uganda Patient Data K-Means Clustering by Gestational Age(Wks) and BMI (kg/m^2)

Figure 8.6: K-Means clustering plots for FAIR ANC Uganda patient data.

with variability in blood pressure and progression of pregnancy-related risk than chronological age alone. Clearer separation of clusters implies that gestational progression plays a critical role in maternal cardiovascular changes during pregnancy.

The Age vs BMI clustering plot reveals a broader dispersion and partial overlap among clusters, indicating that BMI varies considerably across age groups. Nevertheless, several clusters containing moderate-to-high BMI values appear to be concentrated within the middle reproductive age ranges. This suggests the presence of heterogeneous nutritional and metabolic maternal profiles within the population. Compared to blood pressure clustering, BMI-based clusters appear less sharply separated, indicating weaker discrimination between maternal groups when age is used as the primary demographic variable.

In contrast, the Gestational Age vs BMI clustering plot demonstrates more distinct clustering structures than the Age vs BMI analysis. Patients with elevated BMI values appear to be concentrated within specific gestational stages, suggesting that gestational progression may influence maternal weight-related physiological patterns more strongly than age alone. The clearer cluster separation observed in this plot indicates that gestational age contributes substantially to differentiating maternal health profiles associated with BMI and pregnancy risk.

In general, a comparison of the four clustering analyzes suggests that gestational age produces stronger and more distinguishable clustering patterns than chronological age when combined with physiological indicators such as blood pressure and BMI. This finding implies that gestational progression may be a more influential factor in identifying maternal risk profiles within the FAIR ANC Uganda dataset. Furthermore, blood pressure-based clustering demonstrated stronger separation of potential risk categories compared to BMI-based clustering, highlighting blood pressure as a potentially more sensitive indicator of maternal health variation in pregnancy.

From a FAIR data perspective, these clustering analyzes demonstrate the practical analytical benefits of semantically structured and interoperable healthcare data. Although clustering itself does not directly measure FAIRness, the FAIR-compliant infrastructure enabled the integration, preprocessing, and reuse of standardized maternal health variables for advanced machine learning analysis. The interoperability and reusability of the FAIR ANC Uganda dataset facilitated the exploratory identification of maternal risk patterns while supporting accessible, machine-readable, and reusable healthcare analytics suitable for future federated learning and predictive modeling applications.

Tables 8.6 and 8.7 provide quantitative summaries of the clusters visualised in Figure 8.6. While Figure 8.6 illustrates the spatial distribution and separation of the K-Means clusters, the tables describe their demographic and clinical characteristics, including mean BMI, mean age or gestational age, cluster size, and dominant fetal state. Together, the visual and tabular analyses indicate that gestational age combined with BMI produces more distinct and clinically meaningful patient groupings than chronological age combined with BMI, as evidenced by the clearer cluster separation observed in Figure 8.6(b) and the well-defined cluster profiles reported in Table 8.7.

Table 8.6: Cluster Analysis for Age and BMI

Cluster	**Sample Size**	**BMI (Mean, Range)**	**Age (Mean, Range)**	**Dominant Features**
High BMI & Young	283 (37.9%)	28.2 (25.0–36.8)	24.6 (18.0–29.0)	Fetal state: Reassuring (47.7%); Age group: 20–25 (47.7%)
High BMI & Mid-age	227 (30.4%)	29.1 (21.0–38.5)	33.7 (29.0–42.0)	Fetal state: Reassuring (52.4%); Age group: 31–35 (54.6%)
Normal BMI & Young	237 (31.7%)	22.4 (15.6–25.1)	25.4 (16.0–38.0)	Fetal state: Distressed (38.4%); Age group: 20–25 (40.9%)
Total	**747 (100%)**			

Table 8.7: Cluster Analysis for Gestational Age and BMI

Cluster	**Sample Size**	**BMI (Mean, Range)**	**Gestational Age (Mean, Range)**	**Dominant Features**
High BMI & Early Term	176 (23.6%)	32.2 (28.4–38.5)	31.2 (14.0–56.0)	Fetal state: Non-reassuring (60.2%); Age group: 31–35 (33.0%)
Normal BMI & Early Term	245 (32.8%)	24.9 (16.4–33.6)	17.0 (0.0–25.0)	Fetal state: Reassuring (53.9%); Age group: 20–25 (32.7%)
Normal BMI & Mid Term	326 (43.6%)	24.9 (15.6–28.6)	33.8 (26.0–47.0)	Fetal state: Reassuring (36.2%); Age group: 20–25 (35.6%)
Total	**747 (100%)**			

8.3.7 Use Case 4: FAIR ANC Uganda Data Fetal State Prediction Tool Modelling

Sub-use Case 4.1: FAIR ANC Uganda Data Fetal State Prediction Tool Modelling

This involved data preprocessing, feature engineering, and fetal state determination. Multiple models were trained to predict the fetal state using ANC Uganda data. They include MLP, KNN, and Gradient Boosting models.

ANC Uganda Data Pre-processing
A search was conducted across all the data in the triplestore. The data set was then pivoted to convert the SPO structure into a more conventional DataFrame format, with subjects as rows and predicates as columns. This pivot operation significantly altered the shape of the dataset, organizing it into a suitable structure for machine learning tasks.

Feature Engineering and Transformation
In this step, numeric columns such as "Weight(kg)", "Height_(cm)", and "Blood_pressure" were converted to numeric data types to simplify data handling. Non-numeric values in the other columns were replaced with NaN to make sure that the data were consistent. This was followed by the calculation of the Body Mass Index (BMI) using the standard formula given in Equation 8.1. BMI was added as a new feature in the data set.

The missing values were imputed by replacing them with the most frequent value in the columns "Gestational age", "Weight(kg)", "Height (cm)", "Blood pressure", and "BMI" in the data set to ensure that the data was complete and ready for modelling.

Fetal State Determination
A custom function was designed to classify the fetus into one of the four categories according to gestational age, BMI, and blood pressure. This function applied logical conditions to assign one of four fetal states as follows: "Well", "Distressed", "Non-reassuring" or "Reassuring".

'Well': gestational age ≥ 36 weeks, $23 \leq \text{BMI} \leq 28$, and blood pressure between 70 and 120 mmHg.

'Distressed': BMI < 22 or blood pressure > 140 mmHg.

'Non-reassuring': $22 \leq \text{BMI} < 23$ or $28 < \text{BMI} \leq 32$ or blood pressure between 120 and 140 mmHg.

'Reassuring': Other cases that do not fall into the above categories.

'Unknown': Cases in which data on gestational age, BMI, or blood pressure are missing.

Each row in the data set was evaluated, and the corresponding fetal state was assigned based on the criteria defined above.

Fetal State Prediction Tool Model Training and Evaluation

The processed data was split into features (X) and the target variable (y), with features including 'Age', 'Gestational_age', 'Weight(kg)', 'Height_(cm)', 'BMI', and 'Blood_pressure'. Numeric features were identified and handled separately to ensure that they were appropriately imputed and scaled using the RobustScaler to mitigate the effects of outliers. To address class imbalance in the target variable, SMOTE was applied. This resulted in a balanced dataset for model training.

Table 8.8: Comparison of model performance for fetal state based on the FAIR ANC Uganda Data

Model	Fetal State	Precision	Recall	F1-Score
K-Nearest Neighbors	Distressed	0.9023 ± 0.0267	0.8332 ± 0.0492	0.8656 ± 0.0332
	Non-reassuring	0.6691 ± 0.0389	0.6610 ± 0.0324	0.6638 ± 0.0207
	Reassuring	0.7083 ± 0.0311	0.6143 ± 0.0616	0.6561 ± 0.0394
	Well	0.8105 ± 0.0159	0.9855 ± 0.0183	0.8892 ± 0.0088
MLP	Distressed	0.9431 ± 0.0231	0.9592 ± 0.0309	0.9507 ± 0.0170
	Non-reassuring	0.7084 ± 0.0521	0.6318 ± 0.0647	0.6659 ± 0.0461
	Reassuring	0.7300 ± 0.0212	0.6143 ± 0.0474	0.6659 ± 0.0284
	Well	0.7860 ± 0.0398	0.9766 ± 0.0116	0.8705 ± 0.0260
Gradient Boosting	Distressed	**0.9971 ± 0.0058**	**0.9913 ± 0.0116**	**0.9941 ± 0.0055**
	Non-reassuring	**0.9822 ± 0.0218**	**0.9533 ± 0.0213**	**0.9674 ± 0.0184**
	Reassuring	**0.9659 ± 0.0189**	**0.9853 ± 0.0132**	**0.9754 ± 0.0108**
	Well	**0.9771 ± 0.0144**	**0.9913 ± 0.0116**	**0.9840 ± 0.0106**

Three machine learning models were trained using stratified sampling: K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), and Gradient Boosting. The selected machine learning

algorithms were chosen to represent different learning paradigms and levels of model complexity. K-Nearest Neighbors (KNN) was selected as a distance-based, instance-learning algorithm that classifies observations according to the similarity of neighbouring data points. Decision Trees (DT) were chosen because they produce human-readable decision rules, making them particularly suitable for healthcare applications where model interpretability is important. Decision Trees can capture non-linear relationships between variables and provide clear explanations of the factors contributing to a prediction. Gradient Boosting was selected as an ensemble learning technique that combines multiple weak learners to improve predictive performance. Gradient Boosting is capable of modeling complex relationships within healthcare datasets and often achieves superior classification accuracy compared to individual classifiers. Each model was optimized with hyperparameters through . This was done to ensure that the best parameters were selected based on cross-validation performance. The training methodology used a 5-fold stratified cross-validation approach to ensure robust model evaluation. This involved 80 total fits for K-Nearest Neighbors with 16 candidate hyperparameter combinations multiplied by 5 folds. For the Multi-Layer Perceptron, 100 fits were conducted with 20 candidate hyperparameter combinations multiplied by 5 folds. The Gradient Boosting model involved 100 fits with 20 candidate hyperparameter combinations multiplied by 5 folds.

ANC Uganda Data Fetal State Prediction Tool Models' Performance Comparative Analysis

The Gradient Boosting model was selected to be used for prediction based on data input by users. This selection was due to its superior performance across all categories of fetal state compared to the other trained models. It outperformed both the K-Nearest Neighbors (mean accuracy score 0.7734 ± 0.0165) and the MLP (mean accuracy score 0.7953 ± 0.0181) models by achieving an overall mean accuracy of 0.7953 ± 0.0181, making it the most reliable choice for predicting fetal states.

The resulting performance is shown in the table 8.9.

When comparing the initial results with the latest evaluations, the models showed improved mean accuracy, precision, recall, and F1 scores. The mean accuracy score range of the K-Nearest Neighbors (KNN) model improved from 0.7734 ± 0.0165 to 0.7909 ± 0.0109, and the mean accuracy score range of the MLP model improved from 0.7953 ± 0.0181 to 0.7602 ± 0.0111. These improvements suggest that data augmentation effectively improved the generalization capabilities of the models.

The Gradient Boosting model achieved an even better mean accuracy score range of 0.8622 ± 0.0161, with precision, recall, and F1 scores that demonstrate improvement in all categories of fetal state. This shows its robustness in predicting fetal states for pregnant women in Uganda. These results underscore the role of data augmentation in improving model performance and reliability, particularly in medical applications where data may be limited or imbalanced.

Table 8.9: Comparison of Model Performance per Fetal State based on FAIR ANC Uganda data After Adding Noise

Model	Fetal State	Precision	Recall	F1-Score (%)
K-Nearest Neighbors	Distressed	0.9051 ± 0.0348	0.8743 ± 0.0154	0.8888 ± 0.0102
	Non-reassuring	0.6962 ± 0.0295	0.6491 ± 0.0305	0.6716 ± 0.0272
	Reassuring	0.7467 ± 0.0261	0.6944 ± 0.0473	0.7190 ± 0.0338
	Well	0.8089 ± 0.0165	0.9460 ± 0.0176	0.8720 ± 0.0164
MLP	Distressed	0.9133 ± 0.0213	0.9269 ± 0.0369	0.9193 ± 0.0130
	Non-reassuring	0.6631 ± 0.0206	0.6212 ± 0.0392	0.6404 ± 0.0209
	Reassuring	0.7746 ± 0.0256	0.5584 ± 0.0221	0.6484 ± 0.0134
	Well	0.7072 ± 0.0336	0.9342 ± 0.0146	0.8048 ± 0.0270
Gradient Boosting	Distressed	**0.9493 ± 0.0101**	**0.9225 ± 0.0325**	**0.9352 ± 0.0128**
	Non-reassuring	**0.7716 ± 0.0283**	**0.7294 ± 0.0485**	**0.7492 ± 0.0335**
	Reassuring	**0.8251 ± 0.0282**	**0.8509 ± 0.0159**	**0.8376 ± 0.0194**
	Well	**0.9024 ± 0.0325**	**0.9459 ± 0.0120**	**0.9232 ± 0.0159**

Fetal State Prediction Tool Based on User Input

A function was implemented to provide real-time predictions of the fetal state based on user input, including age, gestational age, weight, height, and blood pressure. The purpose of this feature is to help healthcare professionals and expectant women monitor fetal well-being. Once the user inputs the required information, the system first calculates the Body Mass Index (BMI) using the weight and height provided. The features are then scaled to match the format of the training data used by the Gradient Boosting model. This model is then used to predict the fetal state. This prediction can help identify potential risks and guide appropriate interventions. The tool is designed to be used in a clinical or home-care setting, enabling real-time decision-making during prenatal check-ups or consultations. The feature provides an immediate prediction after user input, which healthcare providers can use to adjust care plans or recommend additional tests if needed.

Sub-use Case 4.2: ANC Uganda Fetal State Prediction Tool Model Training using Federated Learning with a Simulated Dataset Split

Federated learning was implemented by simulating multiple clients (e.g., hospitals) that trained machine learning models on their respective local datasets. The ANC Uganda dataset was first transformed into a patient-level dataset and records with missing fetal state labels were excluded. To ensure a realistic evaluation and avoid information leakage, the variables BMI and Blood Pressure were removed from the predictor set for all models. BMI is directly derived from weight and height, while blood pressure is strongly associated with fetal health outcomes and could artificially inflate model performance if used as a predictor.

The final feature set consisted of Age, Weight, Height, Gestational Age, Parity, Hemoglobin (Hb), Mid-Upper Arm Circumference (MUAC), and ANC visit timing. The target variable was fetal state, consisting of four classes: Distressed, Non-reassuring State, Reassuring State, and Well.

The dataset was divided into training and testing subsets before model development. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied only to the training dataset. This ensured that information from the testing dataset did not influence the training process. After oversampling, the balanced training dataset contained an equal number of samples for each fetal state class.

The balanced training data were subsequently partitioned among three simulated federated clients representing independent healthcare facilities. Each client trained its own local model using only its assigned data partition. Model parameters or predictions were then combined to produce a global federated model. The overall federated learning workflow is illustrated in Figure 8.7.

Federated Dataset Distribution

This setup enables the development of collaborative models without requiring the direct sharing of sensitive patient records between participating institutions, thereby preserving privacy while benefiting from distributed learning. After removing records with missing fetal state labels, 710 patient records remained for model development. A stratified 80:20 train-test split was performed, resulting in 568 training records and 142 testing records. To address class imbalance, SMOTE was applied exclusively to the training set, increasing the training data size to 868 balanced samples while preserving the independence of the test set.

The balanced training dataset was subsequently divided among three federated clients. Client 1 contained 292 samples, while Clients 2 and 3 contained 288 samples each. The class distributions were approximately equal for all clients after SMOTE balancing. Each client independently trained its local model without sharing patient-level records, thereby preserving data privacy.

The performance of the local model was also evaluated in each federated node prior to aggregation. Across the ten experimental iterations, local test accuracies generally followed the same trend as the global model results. Gradient Boosting achieved the highest local accuracies, followed by Decision Tree and SVM models, while MLP produced the lowest local accuracies. Local accuracies for the Decision Tree and Gradient Boosting models typically ranged between 0.65 and 0.75 across the three simulated clients. The consistency of local performance across nodes indicates that the balanced data partitioning strategy produced comparable learning conditions for each participating client.

To evaluate local learning performance, each client model was assessed on its corresponding local validation subset. Across the ten experimental iterations, local accuracies ranged from

approximately 64.8% to 68.3% for Decision Tree models, 66.9% to 70.4% for Gradient Boosting models, 52.1% to 62.7% for SVM models, and 28.9% to 41.5% for MLP models. These results indicate that Gradient Boosting consistently achieved the strongest local performance among the evaluated algorithms.

During federated aggregation, each client model contributed to the global prediction according to the size of its local training set. The class probability outputs produced by each local model were aggregated using weighted averaging, where the contribution of each client was proportional to the number of samples used for local training. Consequently, Client 1 contributed a weight of 292 samples, while Clients 2 and 3 each contributed a weight of 288 samples. This approach more closely reflects real-world federated learning deployments, where healthcare facilities often possess different volumes of data. Model aggregation followed the Federated Averaging (FedAvg) principle for the MLP model. Instead of averaging predicted class labels, class probability outputs from local models were aggregated using weighted averaging based on the number of training samples available at each client. Let n_k denote the number of training samples at client k, and P_k denote the predicted probability distribution produced by that client model. The global probability estimate was computed as:

$$P_{global} = \frac{\sum_{k=1}^{K} n_k P_k}{\sum_{k=1}^{K} n_k} \tag{8.2}$$

where $K = 3$ represents the total number of clients. The final predicted class was obtained by selecting the class with the highest aggregated probability. This weighted aggregation strategy ensures that clients with larger datasets contribute proportionally more to the global model while maintaining complete data privacy.

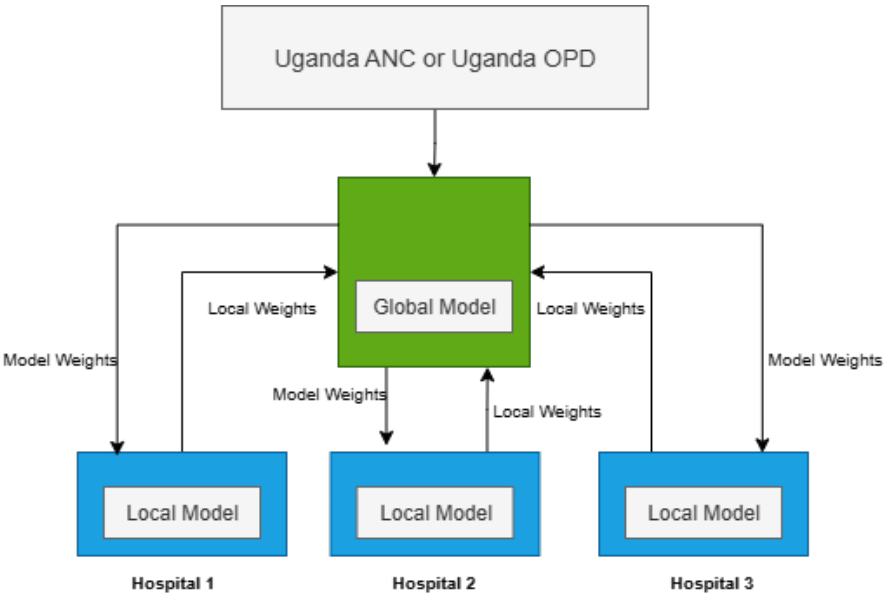


Figure 8.7: Simulated federated learning illustration

The fetal state prediction task was implemented using four machine learning models: a Multi-Layer Perceptron (MLP), Decision Tree, Support Vector Machine (SVM), and Gradient

Boosting model. These models were evaluated within the same federated learning environment to compare their effectiveness in predicting the fetal state.

Multi-Layer Perceptron Neural Network

A Multi-Layer Perceptron (MLP) classifier was implemented using Scikit-learn's MLPClassifier. The network consisted of two hidden layers containing 64 and 32 neurons, respectively. Data preprocessing included median imputation and robust scaling, while early stopping was employed to reduce overfitting and improve generalization performance. The model was trained within the federated learning framework and evaluated alongside the Decision Tree, Support Vector Machine (SVM), and Gradient Boosting models for fetal state prediction.

The MLP architecture consisted of:

An input layer receiving the selected maternal health features.

A first hidden layer containing 64 neurons.

A second hidden layer containing 32 neurons.

An output layer producing probabilities for the four fetal state classes.

Early stopping to prevent overfitting during training.

Decision Tree Model

A Decision Tree classifier was implemented within the federated learning framework to predict the outcomes of the fetal state. Decision Trees are supervised machine learning algorithms that recursively partition data into homogeneous groups by selecting features that maximize class separation. The model was configured with a maximum tree depth of five levels to reduce overfitting and improve generalization, while balanced class weighting was applied to address class imbalance within the dataset. Each participating client trained a local Decision Tree model using its local data, and predictions were aggregated through the federated learning process.

The Decision Tree model consisted of:

An input layer receiving the selected maternal health features.

Recursive binary splits based on feature values that maximize class separation.

A maximum tree depth of five levels.

An output layer producing probabilities for the four fetal state classes.

Leaf nodes representing the predicted fetal state classes.

Balanced class weighting to mitigate class imbalance effects.

Support Vector Machine Model

The Federated Support Vector Machine employed a Radial Basis Function (RBF) kernel, making it a non-linear classification model. Unlike a linear SVM, which separates classes using a single hyperplane, the RBF kernel maps the data into a higher-dimensional feature space, enabling the model to capture complex, non-linear relationships among maternal health variables and fetal state outcomes. This makes the model particularly suitable for clinical datasets

where risk factors interact in a non-linear manner. The SVM algorithm identifies the optimal decision boundaries by maximizing the margin between classes in a transformed feature space. Balanced class weighting was applied to mitigate class imbalance, while probability estimation was enabled to support federated aggregation of local model outputs. Each participating client trained a local SVM model using its own data, and the resulting class probabilities were combined through weighted federated averaging to produce the global prediction.

The SVM model consisted of:

An input layer receiving the selected maternal health features.

Transformation of the feature space using an RBF kernel.

Construction of optimal separating hyperplanes that maximize class margins.

A maximum tree depth of five levels.

Probability estimation for multi-class fetal state prediction.

Leaf nodes representing the predicted classes of fetal state.

Balanced class weighting to address class imbalance.

Gradient boosting Model

A Gradient Boosting classifier was implemented within the federated learning framework to predict the outcomes of the fetal state. Gradient Boosting is an ensemble learning method that constructs a sequence of weak decision tree learners, where each subsequent tree is trained to correct the prediction errors of the previous trees. The model combines multiple decision trees to improve predictive accuracy and reduce bias. In this study, the model was configured with 150 estimators, a learning rate of 0.05, a maximum tree depth of three, and a sub-sampling rate of 0.9. Local models were trained independently at each federated client, and prediction probabilities were aggregated using weighted federated averaging to produce the final global prediction.

The Gradient Boosting model consisted of:

An input layer receiving the selected maternal health features.

A sequence of decision trees trained iteratively.

Error correction through gradient-based optimization, where each new tree focuses on reducing residual errors from previous trees.

Aggregation of predictions from all trees to generate the final fetal state classification.

Federated aggregation of local model probabilities using weighted averaging across participating clients.

Federated Learning Process

The federated learning process was simulated using three clients as illustrated in Figure 8.7. The implementation involved the following steps: The balanced training dataset was partitioned among three simulated clients. Each client independently trained a local model using only its assigned data. Local model information was communicated to a central aggregation server. The server combined client outputs to generate a global federated model. The global model was evaluated using a separate testing dataset that was not involved in training.

Federated MLP Neural Network Training

For the neural network implementation, federated learning followed the Federated Averaging (FedAvg) algorithm proposed by McMahan et al. [226]. During each federated round: The current global model was distributed to each client. Each client trained the model locally using its private dataset partition. The updated model weights were transmitted to the server. The server aggregated client weights by averaging them to produce an updated global model.

Federated Decision Tree, Gradient Boosting, and SVM Training

Unlike neural networks, traditional machine learning models such as Decision Tree, Gradient Boosting, and SVM models do not naturally support parameter averaging. Therefore, each client trained independently trained a local model using its private data partition. Instead of aggregating model parameters, predictions from locally trained models were combined during inference to generate a global prediction. This ensemble-based federated approach allowed the evaluation of traditional machine learning algorithms within a federated learning environment while preserving data locality and privacy.

K-Nearest Neighbours (KNN) was used only in the non-federated analysis and was not included in the federated learning experiments because it is an instance-based learning algorithm that does not produce a compact set of model parameters suitable for aggregation. Unlike Decision Trees, Support Vector Machines, Gradient Boosting, and Multi-Layer Perceptrons, which learn transferable model representations, KNN relies on retaining and accessing training instances during prediction. In a federated learning environment, this would require the exchange of patient records or local datasets between participating institutions, which is inconsistent with the privacy-preserving design of federated learning. Therefore, models supporting local training and parameter aggregation were selected for federated experiments.

ANC Uganda Data Fetal State Prediction Federated Learning Models' Performance Comparative Analysis

The ANC Uganda dataset was first transformed into a patient-level table and records with missing fetal state labels were excluded. To ensure a realistic evaluation and prevent information leakage, BMI and Blood Pressure were removed from the predictor variables because BMI is derived directly from weight and height, while blood pressure is strongly associated with fetal health outcomes and could artificially inflate model performance. The final set of characteristics consisted of Age, Weight, Height, Gestational Age, Parity, Hemoglobin (Hb), Mid-Upper Arm Circumference (MUAC) and ANC visit timing.

Table 8.10: Mean ± standard deviation of federated learning model performance across ten experimental iterations for fetal state prediction using FAIR ANC Uganda data.

Federated Model	Fetal State	Precision	Recall	F1-Score
Gradient Boosting	Distressed	0.8160 ± 0.0823	0.8600 ± 0.0583	**0.8341 ± 0.0465**
	Non-reassuring	0.6882 ± 0.0487	0.5220 ± 0.0817	**0.5909 ± 0.0649**
	Reassuring	0.6579 ± 0.0667	0.8500 ± 0.1026	**0.7400 ± 0.0737**
	Well	0.7730 ± 0.0367	0.8500 ± 0.0384	0.8089 ± 0.0282
Decision Tree	Distressed	0.7516 ± 0.1120	0.8250 ± 0.0873	0.7813 ± 0.0812
	Non-reassuring	0.6974 ± 0.0775	0.4700 ± 0.0744	0.5554 ± 0.0554
	Reassuring	0.5917 ± 0.0576	0.8278 ± 0.0911	0.6851 ± 0.0428
	Well	0.7743 ± 0.0425	0.8537 ± 0.0700	**0.8090 ± 0.0270**
SVM	Distressed	0.6643 ± 0.0644	0.7700 ± 0.1030	0.7084 ± 0.0612
	Non-reassuring	0.5702 ± 0.0417	0.4680 ± 0.0515	0.5128 ± 0.0410
	Reassuring	0.5632 ± 0.0673	0.7444 ± 0.1117	0.6382 ± 0.0713
	Well	0.7138 ± 0.0375	0.7074 ± 0.0452	0.7098 ± 0.0346
MLP	Distressed	0.3068 ± 0.1574	0.4750 ± 0.3010	0.3555 ± 0.1903
	Non-reassuring	0.3837 ± 0.1416	0.2860 ± 0.2193	0.2982 ± 0.1213
	Reassuring	0.3259 ± 0.1225	0.7333 ± 0.2639	0.4456 ± 0.1555
	Well	0.5598 ± 0.0899	0.4000 ± 0.1655	0.4458 ± 0.1250

The dataset was divided into training and testing subsets prior to model development. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied only to the training set. The balanced training data set was later partitioned and distributed among three simulated federated clients representing healthcare facilities. Each client trained local machine learning models using its own partition of the data without sharing raw patient records, thereby preserving data privacy and locality throughout the federated learning process.

Four machine learning models were evaluated within the federated learning environment: Multi-Layer Perceptron (MLP), Decision Tree, Support Vector Machine (SVM) and Gradient Boosting. The federated neural network used the Federated Average (FedAvg) approach, where the weights of the local model were aggregated to update the global model. For Decision Tree, Gradient Boosting, and SVM models, local models were trained independently on each client, and their predictions were aggregated to produce a global prediction. The models were evaluated on an independent test set using precision, recall, F1-score, and overall classification accuracy. The reported precision, recall, F1-score, and accuracy values represent the mean ± standard deviation obtained across ten independent experimental iterations. Reporting both statistics provides a more reliable assessment of model stability and robustness by accounting for variation arising from random train-test partitioning and federated client allocation. Results are reported as mean accuracy ± standard deviation across ten independent experimental iterations.

Table 8.11: Global accuracy of federated learning models across ten iterations using FAIR ANC Uganda data.

Federated Model	**Global Accuracy**
Decision Tree	0.7113 ± 0.0321
Gradient Boosting	0.7359 ± 0.0332
SVM	0.6366 ± 0.0294
MLP	0.4127 ± 0.0390

The overall classification accuracies obtained by the federated models were 71.83% for Federated Gradient Boosting, 71.13% for Federated Decision Tree, 66.20% for Federated SVM, and 40.85% for Federated MLP. Among the evaluated models, Federated Gradient Boosting achieved the highest classification accuracy and demonstrated the most balanced performance across the fetal state categories.

Federated Gradient Boosting achieved strong recall for the Distressed and Reassuring State classes while maintaining competitive precision and F1-scores across all categories. The Federated Decision Tree model produced comparable results and performed particularly well for the Well class. Federated SVM achieved moderate performance, but showed reduced effectiveness in distinguishing some categories of fetal state. The Federated MLP model produced the lowest accuracy, indicating that the available data set size and feature representation were less suitable for neural-network-based learning in the federated setting.

The performance obtained in this experiment is lower than that observed in earlier experiments because potential leakage variables were removed and the evaluation was conducted using a strict train-test separation strategy. Although the resulting accuracy values are lower, they provide a more realistic estimate of model performance on unseen patient data. Consequently,

Table 8.12: Average local client performance across ten federated iterations.

Federated Model	Client	Local Accuracy (Mean)
Decision Tree	Client 1 (n=292)	∼0.65
Decision Tree	Client 2 (n=288)	∼0.68
Decision Tree	Client 3 (n=288)	∼0.65
Gradient Boosting	Client 1 (n=292)	∼0.67
Gradient Boosting	Client 2 (n=288)	∼0.70
Gradient Boosting	Client 3 (n=288)	∼0.66
SVM	Client 1 (n=292)	∼0.53
SVM	Client 2 (n=288)	∼0.52
SVM	Client 3 (n=288)	∼0.62
MLP	Client 1 (n=292)	∼0.42
MLP	Client 2 (n=288)	∼0.39
MLP	Client 3 (n=288)	∼0.29

Federated Gradient Boosting was selected as the best-performing federated learning model for fetal state prediction based on the ANC Uganda dataset.

Prediction Based on User Input Using a Federated Learning Model

A function 'predict_fetal_state' was defined to make predictions using the best model. Processed user input, calculated BMI, and handled missing values. Predictions were made, and the result was printed with a formatted message showing the user's input values.

Table 8.13: Comparison of Predictions Generated by the Traditional and Federated Gradient Boosting Models

Example	Age	Gest. Age	Weight	Height	BP	Traditional GB	Federated GB
1	25	36	70	165	110	Well	Well
2	50	40	90	165	110	Non-reassuring	Non-reassuring
3	24	28	50	170	100	Distressed	Distressed
4	30	34	70	165	150	Reassuring	Reassuring

Table 8.13: Comparison of the predictions of the fetal state produced by the Traditional Gradient Boosting and Federated Gradient Boosting models for four representative patient

cases from the Uganda dataset of the ANC. The table illustrates the consistency and differences between traditional and federated learning approaches when applied to identical patient inputs.

Performance Summary and Implications

The results obtained in this study demonstrate both the potential and current limitations of applying traditional and federated machine learning techniques to the prediction of fetal state using the FAIR ANC Uganda dataset. In all experimental settings, Gradient Boosting consistently achieved the highest performance, demonstrating superior precision, recall, F1-score, and overall classification accuracy compared to the other models evaluated. The model also exhibited relatively low standard deviations in repeated experimental runs, indicating stable and reliable performance. Although MLP produced competitive results in some categories, its overall performance remained lower than that of Gradient Boosting. Furthermore, the experiments involving noisy data showed that Gradient Boosting was comparatively robust, while other models exhibited greater sensitivity to data perturbations.

Federated learning experiments produced lower performance than traditional machine learning approaches, which was expected given the distributed nature of training and the challenges associated with client heterogeneity, model aggregation, and limited local data availability. Nevertheless, Federated Gradient Boosting achieved the highest global accuracy among the federated models and maintained balanced performance across the different fetal state categories. All federated models demonstrated a reasonable degree of stability across repeated iterations, although variations in performance were observed due to differences in data distribution and client-level training.

From a FAIR data perspective, this study demonstrates how FAIRification can transform maternal health records into machine-readable resources that support advanced analytics while enabling interoperability and data reuse. However, the available data set remains relatively small and does not fully capture the diversity of maternal healthcare conditions in Uganda. In addition, a substantial proportion of antenatal care records in Uganda are still maintained in paper-based systems, limiting the amount of digitized data available for analysis. Consequently, the results presented in this study should be viewed primarily as an indication of emerging trends rather than definitive clinical evidence.

The observed performance patterns are consistent with expectations for datasets of this size. The performance of traditional and federated models is expected to improve with access to larger, more diverse, and FAIR-compliant datasets collected from multiple healthcare facilities. Therefore, more testing, external validation, and large-scale evaluations are required to assess the generalizability and clinical applicability of the proposed approach.

The comparison between the predictions generated by the Traditional Gradient Boosting and Federated Gradient Boosting models demonstrated a high degree of agreement across representative patient cases, indicating that the federated approach was capable of capturing similar predictive patterns despite being trained in a distributed environment.

Despite the performance gap between traditional and federated learning, federated learning remains a highly promising direction for healthcare analytics. Its ability to train models across distributed datasets without sharing raw patient information provides significant privacy, security, and collaborative advantages. As healthcare institutions continue to digitize records and adopt FAIR data principles, federated learning offers a practical pathway to scalable, privacy-preserving, and data-driven maternal and fetal healthcare decision support systems. Although

further evaluation is required, the results of this study suggest that FAIR-enabled federated learning is an important and viable direction for future healthcare research and implementation.

8.4 Discussion

The results of this study demonstrate both the potential and limitations of applying federated learning and traditional machine learning techniques to maternal and fetal health prediction using the ANC Uganda dataset. While the initial centralized experiments showed strong performance, the federated learning setup produced more moderate and realistic results after applying strict preprocessing and leakage prevention measures.

A key improvement in this study was the removal of potential data leakage variables, specifically BMI and blood pressure. Although these features are clinically relevant, BMI is derived directly from weight and height, and blood pressure is strongly correlated with fetal health outcomes. Their exclusion ensured that the models were not indirectly exposed to target-related information, resulting in more reliable and generalizable performance estimates.

Furthermore, the application of SMOTE solely on the training dataset ensured that class balancing did not introduce bias into the evaluation process. This strict separation between training and testing data provides a more realistic assessment of model behavior in real-world clinical deployment scenarios, where unseen patient data must be classified without prior exposure.

The federated learning experiments showed that distributed model training is feasible for fetal state prediction; however, performance varies significantly across model types. Federated Gradient Boosting achieved the highest performance among all evaluated models, followed closely by the Federated Decision Tree model. Both models demonstrated relatively balanced precision and recall across fetal state classes. In contrast, the Federated SVM showed moderate performance, while the Federated MLP exhibited the weakest results, indicating that neural network models may require larger datasets or more optimized architectures in federated settings.

Compared to centralized machine learning approaches, federated learning produced lower accuracy values. However, this reduction is expected due to data partitioning across clients and the absence of centralized training. Despite this, federated learning provides a significant advantage in terms of privacy preservation, as sensitive patient data remains localized within each institution and is never directly shared.

The multilayer perceptron (MLP) model demonstrated the lowest performance among the evaluated algorithms. Neural network models generally require larger datasets to effectively learn complex feature representations and avoid overfitting. In this study, only 710 patient records were available after target generation and data cleaning, which limited the amount of information available for training. Tree-based models such as Gradient Boosting and Decision Trees are often more suitable for relatively small tabular healthcare datasets and therefore achieved superior performance. Future studies incorporating substantially larger datasets may allow deep learning models such as MLPs to realize their full predictive potential and potentially outperform traditional machine learning approaches.

Overall, the findings highlight a fundamental trade-off between predictive performance and data privacy. While centralized models achieve higher accuracy, federated learning offers a scalable and privacy-preserving alternative suitable for healthcare environments where data

sharing is restricted. This makes federated learning particularly valuable for multi-institutional collaboration in resource-constrained settings such as Uganda.

The machine learning and federated learning experiments conducted in this study demonstrated the potential value of FAIR-compliant health data for advanced healthcare analytics. Using FAIR ANC Uganda data, predictive models were successfully developed to classify fetal health states, with the federated Gradient Boosting model achieving the highest overall accuracy. These findings indicate that improved data management practices, standardized data collection, and FAIR data implementation can support the development of data-driven clinical decision support tools. Furthermore, the federated learning approach showed that predictive models can be trained collaboratively while preserving patient privacy, making it a promising strategy for future healthcare information systems in Uganda. The federated learning experiments demonstrated that FAIR-compliant health data can be effectively used within privacy-preserving machine learning architectures. Gradient Boosting achieved the highest overall performance, followed by Decision Tree and SVM models. The relatively lower performance of the MLP model is likely attributable to the limited size of the available dataset. Neural network-based approaches generally require substantially larger datasets to learn robust feature representations and fully exploit their modeling capacity. With larger FAIR-compliant datasets collected from multiple healthcare facilities, improved predictive performance would be expected across all evaluated models, particularly for deep learning approaches such as MLP.

Future improvements should focus on optimizing federated aggregation strategies, improving feature representation, and exploring advanced ensemble methods to further close the performance gap between centralized and federated approaches while maintaining strict privacy constraints.

8.5 Conclusion

This study investigated the integration of FAIR data principles with both traditional machine learning and federated learning approaches for fetal state prediction using the ANC Uganda dataset. The primary objective was to develop predictive models capable of classifying fetal health conditions while preserving patient privacy through distributed learning.

A key contribution of this work is the development of a leakage-free experimental pipeline. By removing derived and highly correlated variables such as BMI and blood pressure, and by applying SMOTE exclusively to the training dataset after a strict train-test split, the study ensured that evaluation was conducted under realistic and unbiased conditions. This improved the reliability and clinical relevance of the reported results.

Traditional machine learning models demonstrated strong predictive performance in earlier centralized experiments; however, the federated learning framework provided a more realistic assessment of model generalization in distributed healthcare environments. Among the federated models, Gradient Boosting achieved the highest performance with a global accuracy of 71.83%, followed closely by the Decision Tree model at 71.13%. The SVM model showed moderate performance, while the MLP model performed comparatively lower, indicating that neural networks may require larger datasets or further optimization in federated settings.

The results demonstrate that FAIR-compliant healthcare data can successfully support machine learning and federated learning workflows for maternal healthcare analytics. Among the evaluated models, the federated Gradient Boosting classifier achieved the highest overall accuracy of 73.59%, followed by the Decision Tree model with 71.13%. These findings indicate that

machine learning techniques can effectively utilize FAIR health datasets to predict fetal health states and support data-driven clinical decision making.

The findings of this chapter demonstrate that FAIR data support and strengthen the accessibility and usability of health data for machine learning applications. The transformation of health information into FAIR-compliant formats enabled the successful development of clustering, predictive analytics, and federated learning models using both the OPD Uganda and ANC Uganda datasets.

Regarding Sub-Research Question 1, FAIR data facilitated the application of machine learning algorithms for clustering and prediction by improving data interoperability, accessibility, and reusability. The clustering analyses revealed meaningful groupings of patients and diagnoses that can support healthcare planning and decision-making.

Regarding Sub-Research Question 2, the FAIR ANC Uganda dataset was sufficiently rich to support the prediction of pregnancy risks and fetal states. The predictive models successfully identified relationships between maternal characteristics and fetal outcomes, with Gradient Boosting achieving the strongest predictive performance among the evaluated models. The federated learning implementation further demonstrated that machine learning can be applied within a distributed healthcare environment while preserving patient privacy.

The federated learning experiments are particularly relevant to real-world healthcare systems because patient data remain within local institutions and only model parameters are exchanged during training. This approach reduces the need for centralized data sharing while still enabling collaborative model development across multiple healthcare facilities. The results therefore illustrate how FAIR data and federated learning can jointly support secure, scalable, and privacy-preserving health data analytics.

The federated learning framework further demonstrated that predictive models can be trained without sharing sensitive patient-level records between participating institutions. Instead, only model parameters and prediction probabilities were exchanged during aggregation, preserving patient privacy while enabling collaborative model development. This characteristic makes federated learning particularly suitable for healthcare environments where data sharing is restricted by privacy, ethical, or regulatory considerations.

Although promising results were obtained, the relatively limited dataset size constrained overall model performance, particularly for the MLP neural network model. Future work involving larger and more diverse datasets is expected to improve predictive accuracy, model generalization, and the effectiveness of federated learning systems in real-world healthcare settings. Overall, the study confirms that the combination of FAIR data principles, machine learning, and privacy-preserving federated learning provides a viable approach for developing scalable and trustworthy healthcare prediction systems.

The results highlight an important trade-off between predictive performance and data privacy. While centralized models typically achieve higher accuracy due to access to complete datasets, federated learning enables collaborative model training without sharing sensitive patient data. This makes it particularly suitable for healthcare systems where data privacy regulations and institutional constraints limit centralized data collection.

Overall, the study demonstrates that federated learning is a viable and privacy-preserving approach for fetal state prediction, especially when combined with careful preprocessing and leakage prevention strategies. Although there is a measurable reduction in performance compared to centralized learning, the approach provides a more realistic and ethically robust framework for multi-institutional healthcare analytics.

Future work should focus on improving federated optimization strategies, exploring advanced ensemble aggregation methods, and incorporating richer clinical features to further enhance predictive performance while maintaining strict privacy guarantees. Additionally, expanding the dataset across more healthcare institutions would improve model robustness and generalization across diverse populations.