National College of Ireland

Mitigating Data Privacy Risks in Cloud Computing

MSc Research Project
MSc Cloud Computing

Shreya Modak
Student ID: 21208671

School of Computing
National College of Ireland

Supervisor: Shreyas Setlur Arun

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Shreya Modak
Student ID:	21208671
Programme:	MSc Cloud Computing
Year:	2023
Module:	MSc Research Project
Supervisor:	Shreyas Setlur Arun
Submission Due Date:	14/12/2023
Project Title:	Mitigating Data Privacy Risks in Cloud Computing
Word Count:	7878
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Shreya Modak
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Mitigating Data Privacy Risks in Cloud Computing

Shreya Modak
21208671

Abstract

Diabetes is a common chronic illness that has a significant global impact on public health. Timely intervention and management of diabetes depend on the early and precise classification of diabetes status, whether the patient is healthy, pre-diabetic, or diabetic. Using the CDC Diabetes Health Indicators dataset, *CDC Diabetes Health Indicators* (2015) this research explores the field of machine learning to create a reliable diabetes categorization system. For a sensitive data, such as patients' health records data, mitigating the data security is a major issue and on the other hand it needs to be served to the doctors and the patients. Encryption is a technique in the clouds to train machine learning models which in this case is solved as AWS sage maker offers capabilities such as industry-standard SSL/TLS protocols encryption that handles mitigating the security risks for a secure model training and deployment in the clouds Theerthagiri et al. (2022). Thus, this study investigates the performance of the distributed computing specifically parallel processing, DE duplication and compares the efficiency of various machine learning algorithms; Random "Forest, xGBoost and Tensor Flow" featuring federated learning techniques and data anonymization techniques.

The dataset adopted for this study is the Centres for Disease Control and Prevention (CDC) which funds the CDC Diabetes Health Indicators dataset, which is a useful resource containing 253,680 occurrences and 21 attributes that include health and lifestyle data as well as demographics. This dataset provides a thorough understanding of the intricate connection between diabetes and lifestyle variables that is our centre goal of this study. The three machine learning algorithms that are compared in this study come with special advantages for each for categorising features according to the characteristics in the dataset. Using feature significance analysis and several decision trees, Random Forest is an ensemble approach. Neural network architecture is used by Tensor Flow, which is inspired by the structure of the human brain, to recognise complex patterns. The goal of xGBoost algorithm is to identify the best hyper planes for data separation.

The aim of the study is to identify which algorithm provides the most trustworthy and accurate diabetes status categorization in the clouds. The assessment will take into account indicators like F1-score, recall, accuracy, and precision. This study advances healthcare analytics by identifying the most efficient method, which may improve early diabetes identification and individualised treatment recommendations.

1 Introduction

Millions of individuals worldwide suffer from diabetes, a chronic illness that presents a serious threat to public health. Diabetes is becoming more and more common, which

raises concerns and calls for creative ways to detect and treat the condition. Using data from lifestyle surveys and numerous healthcare statistics, machine learning algorithms may be used in one such method to categorise people as healthy, pre-diabetic, or diabetic. This study compares the machine learning models; random forest, xGBoost and Tensor Flow for a secure model training and deployment in a distributed cloud environment with the goal of building a robust classification system utilising the CDC Diabetes Health Indicators dataset.

A useful tool for figuring out how lifestyle variables relate to diabetes in the US is the CDC Diabetes Health Indicators dataset, which is externally linked. This tabular, multivariate dataset, which includes demographic data, test results from labs, and survey answers, has 253,680 occurrences and 21 attributes. The main goal in creating this dataset was to clarify the complex relationship between diabetes and lifestyle. The fact that the Centres for Disease Control and Prevention (CDC) provided the funding for it illustrates how important this study is to solving a pressing public health problem.

The dataset contains both integer and category data as features. An ethical component is included to the study as some of the sensitivity of the data is related to characteristics like gender, income, and educational attainment. The age of the dataset has been grouped into buckets to aid in analysis, one of the necessary preprocessing stages. Crucially, the data remains intact since there are no missing values.

In the field of medicine, machine learning has shown to be an effective tool for both diagnosing and forecasting a wide range of medical disorders. It presents a viable method for utilising dataset patterns to forecast a person's diabetes state in the context of diabetes categorization.

1.1 Research Aim and Objectives

This project's main goal is to create a reliable diabetes categorization system using the CDC Diabetes Health Indicators dataset as a guide, the study apply the three machine learning algorithms (Random Forest, xGBoost and Tensor Flow) to build models that can reliably categorise people into one of three groups: pre-diabetes, diabetic, or healthy. To identify the best method for classifying diabetes, the performance of these models will be compared and assessed utilizing the AWS Sage maker notebook instances securing the data through IM Roles and storing the data into an encrypted s3 bucket. The objectives of this research study are as follows: -

1. To ensure mitigation of data security measures for studying about the CDC dataset in a distributed cloud environment

2. To assess the execution of three machine learning algorithms, to be specific Random Forest, xGBoost and Tensor Flow, in categorizing people into one of three bunches: sound, pre-diabetic, or diabetic based on the CDC Dataset.

3. To contribute to healthcare analytics by recognizing the foremost proficient strategy for early diabetes distinguishing proof and individualized treatment suggestions, with the potential to improve open wellbeing results and quiet care.

4. To compare the adequacy of these machine learning algorithms in terms of key execution measurements, counting F1-score, recall, accuracy, and precision, to decide which calculation gives the foremost trustworthy and precise diabetes status categorization.

2 Related Work

[1] Diabetes, an unremitting metabolic clutter characterized by tall blood sugar levels, has risen as a noteworthy worldwide open wellbeing concern. With millions of people around the world influenced by the condition, diabetes postures a considerable burden on healthcare frameworks and carries genuine wellbeing dangers, such as cardiovascular illnesses, kidney disappointment, and vision impedance Khaleel and Al-Bakry (2023). Appropriate and exact intercession and administration of diabetes are significant to relieve its complications and make strides patients' quality of life. The capacity to precisely categorize people as solid, pre-diabetic, or diabetic shapes the foundation of viable diabetes care.

The essential objective of this investigate is to tackle the control of machine learning algorithms to form a vigorous and solid framework for diabetes categorization. By utilizing the comprehensive CDC Diabetes dataset, we point to create models that can precisely classify people into one of these three categories, encouraging early conclusion and personalized treatment suggestions Theerthagiri et al. (2022).

The CDC Diabetes Wellbeing Pointers dataset, supported by the Centres for Malady Control and Anticipation, speaks to a profitable asset in this endeavour. With over 250,000 occurrences and 21 properties including wellbeing and way of life information, socioeconomics, and lab comes about, this dataset gives a wealthy understanding of the complicated transaction between diabetes and different way of life factors. It stands as a confirmation to the squeezing open wellbeing concern related with diabetes.

Achieving early and exact diabetes categorization is instrumental in controlling the diabetes plague. It empowers healthcare specialists to distinguish at-risk people, start preventive measures, and tailor treatment plans to each patient's particular needs. Moreover, it serves as an establishment for progressing healthcare analytics, contributing to the improvement of more viable procedures for diabetes administration and eventually driving to made strides open wellbeing results. In this writing audit, we'll dig into the existing body of information on the subject to contextualize and investigate the investigate destinations and scope, pointing to lay the basis for the basic examination of the chosen machine learning algorithms in diabetes categorization Khaleel and Al-Bakry (2023).

2.1 The Predominance of Diabetes and Its Open Wellbeing Impact

Diabetes may be a worldwide wellbeing emergency with a stunning predominance that proceeds to rise, altogether affecting people and healthcare frameworks around the world. Concurring to the International Diabetes Federation (IDF), an evaluated 537 million individuals were living with diabetes in 2021, a number anticipated to reach 643 million by 2030 in the event that current patterns hold on. This predominance is disseminated over different locales, with both high-income and moo- to middle-income nations encountering the burden of the infection. The IDF's Diabetes Chart book highlights the criticalness of tending to diabetes as a major open wellbeing issue, as its financial and wellbeing results are profound.

The suggestions of diabetes on open wellbeing are multifaceted and significant. It has been unequivocally connected to expanded dismalness and mortality, to a great extent due to its affiliation with cardiovascular infections, kidney brokenness, neuropathy, and retinopathy. The financial affect is similarly concerning, with the fetched of diabetes

[1] Like this one: `http://www.ncirl.ie`

care and its related complications applying colossal weight on healthcare frameworks and economies. The administration of diabetes requests broad assets, from schedule checking and pharmaceutical to potential surgical intercessions. Furthermore, diabetes leads to misplaced efficiency, missed workdays, and reduced quality of life for people and their families. In reaction to the developing predominance and effect of diabetes, there's a clear require for imaginative arrangements in its administration and care. Early determination and mediation play an essential part in relieving the complications and lessening the in general burden of the illness. Machine learning-based diabetes categorization, as investigated in this investigate, holds the guarantee of advertising a data-driven approach to recognizing people at chance, hence empowering healthcare suppliers to mediate at an prior arrange. Besides, personalized treatment techniques educated by machine learning models can offer assistance optimize understanding results and asset allocation.

The squeezing open wellbeing concerns related with diabetes emphasizes the significance of progressing in the field of diabetes categorization utilizing cutting-edge advances. This investigate contribute to the improvement of inventive arrangements that have the potential to convert diabetes care, improve open wellbeing, and decrease the worldwide burden of this incessant conditionMohamed et al. (2022).

2.2 Machine Learning in Healthcare

Machine learning, a subset of fake insights, has risen as a transformative constrain in healthcare, advertising novel ways to handle complex restorative challenges. Its capacity to analyse tremendous and different datasets, recognize designs, and make expectations has cleared the way for innovative applications within the field of healthcare, with noteworthy suggestions for therapeutic conclusion, treatment, and patient care.

One of the primary roles of machine learning in healthcare is to help therapeutic experts in making more precise and convenient choicesMohamed et al. (2022). By analysing understanding information, counting electronic wellbeing records, restorative pictures, and hereditary data, machine learning algorithms can help in diagnosing maladies, anticipating results, and fitting treatment plans to person patients. This innovation has the potential to progress the quality of care, improve persistent security, and optimize asset assignment inside healthcare systems. Machine learning has illustrated its adequacy in different therapeutic spaces, with effective applications within the conclusion and forecast of a wide run of maladies. For illustration, in radiology, machine learning models have been created to identify peculiarities in therapeutic pictures, such as X-rays and MRIs, driving to prior and more précised analysis. In cardiology, machine learning algorithms have appeared guarantee in foreseeing cardiovascular occasions, making a difference recognize people at chance of heart infection. Moreover, in genomics, machine learning is being utilized to analyse hereditary information, contributing to the identification of hereditary markers related with different conditions.

The potential benefits of utilizing machine learning for diabetes categorization are considerable. Given the complexity and multifaceted nature of diabetes, machine learning algorithms can filter through huge datasets containing wellbeing and way of life factors to distinguish designs that human experts may ignore. This can lead to the early recognizable proof of people at hazard for diabetes or those who are within the pre-diabetic organize, empowering preventive mediations and personalized care plans. Furthermore, machine learning can offer assistance optimize treatment techniques, guaranteeing that assets are apportioned effectively and individualized care is given Adigun et al. (2022).

In conclusion, machine learning is balanced to revolutionize healthcare by giving progressed devices for restorative conclusion and forecast. Its fruitful applications in different therapeutic spaces emphasize its potential to improve diabetes categorization and contribute to more successful, data-driven approaches to healthcare. This inquires about leverages the capabilities of machine learning to address the basic issue of diabetes categorization, endeavouring for early determination and progressed quiet outcomes.

2.3 Datasets in Diabetes Research

Datasets play an essential part in diabetes investigates, giving the establishment upon which imaginative arrangements are built. They empower analysts to investigate the perplexing connections between factors, distinguish designs, and create models that contribute to a more profound understanding of diabetes and its related components. In this area, we are going dig into the centrality of datasets in diabetes investigate and particularly centre on the CDC Diabetes Wellbeing Pointers dataset, shedding light on its characteristics and its part in unravelling the association between diabetes and lifestyle.

The noteworthiness of datasets in diabetes inquires about lies in their capacity to typify a wide cluster of data, extending from statistic information to wellbeing pointers, way of life factors, and clinical test comes about. These comprehensive datasets offer analysts the opportunity to look at the complex interaction between diabetes and different components, permitting for evidence-based conclusions and educated decision-making Nanda et al. (2022).

The CDC Diabetes dataset stands as an important resource in diabetes investigate, liberally supported by the Centres for Illness Control and Avoidance. Comprising 253,680 occasions and 21 traits, this multivariate dataset typifies riches of data important to diabetes categorization Adigun et al. (2022). These properties envelop statistic information; research facility test comes about, overview reactions, and way of life data. By advertising a comprehensive set of people and their circumstances, this dataset presents a special opportunity to investigate the multifaceted nature of diabetes and its association to lifestyle.

The dataset's characteristics are especially invaluable for inquire about endeavours. One critical include is the dataset's completeness, because it contains no lost values, guaranteeing that the information remains intaglio and solid for investigation. Moreover, the moral thought of counting touchy properties, such as sex, wage, and instructive fulfilment, enhances the dataset's capacity to address multifaceted perspectives of diabetes and its relationship to differing sociodemographic factors.

The CDC Diabetes Wellbeing Pointers dataset plays an urgent part in progressing our understanding of the complicated associations between diabetes and way of life. By giving a comprehensive set of people, their wellbeing, and their ways of life, this dataset empowers analysts to investigate how variables like eat less, work out, financial status, and more relate with diabetes. It serves as a capable instrument in unravelling the complexity of diabetes and encourages the advancement of inventive arrangements, such as machine learning models, that point to classify people based on their diabetes status, eventually contributing to made strides diabetes administration and open wellbeing outcomes Whig et al. (2023).

2.4 Mitigating data security for the cloud based environment

The study focuses on information security as a key part of its examination. With the heightening reception of distributed computing, the review recognizes the inborn security threats and puts a huge accentuation on vigorous safety efforts. Encryption systems, both for information on the way and very still, are carefully carried out utilizing industry-standard SSL/TLS conventions and AWS Key Management Service (KMS), individually. Access controls are braced through Sage Maker's reconciliation with AWS IAM, guaranteeing fine-grained authorizations for assets. Network disengagement inside a Virtual Confidential Cloud (VPC) further upgrades security by directing inbound and outbound traffic. The security of ML model curios during preparing and stockpiling in Amazon S3 is a point of convergence, using encryption and access control through AWS KMS. The sending of models as Sage Maker endpoints is drawn nearer with severe safety efforts, using Amazon VPC for disengagement and traffic signal.

Couple with these security contemplations, the exploration directs an exhaustive correlation of ML algorithms — Keras, XGBoost, and Random Forest — each offering one of a kind qualities. Keras, known for adaptability; XGBoost, perceived for exactness and proficiency; what's more, Random Forest, successful in taking care of high-layered datasets, add to an extensive examination. The review's blended techniques approach, including topical examination, means to determine custom fitted information security methodologies lined up with the nuanced challenges introduced by various cloud administration and sending models Nanda et al. (2022). Through these actions, the exploration endeavours to upgrade the security scene of distributed computing, guaranteeing a more secure and stronger handling of the medical data.

2.5 Ethical Considerations in Healthcare Data Analysis

The study utilize of healthcare information for investigate and examination raises noteworthy moral contemplations, especially when it includes touchy persistent data. Moral standards must be at the bleeding edge of any considered to guarantee the protection, assent, and security of people and their data. In this area, we'll dive into the moral contemplations inalienable in healthcare data analysis, with a particular centre on tending to issues related to touchy qualities, information security, and how the investigate endeavours to regard moral guidelines.

Healthcare information regularly incorporates profoundly touchy data, such as quiet restorative records, hereditary information, and individual identifiers. When working with such information, it is basic to address concerns related to security and privacy. Guaranteeing that understanding data is anonymized and safely put away is of fundamental significance. Furthermore, the capable utilize of information requests strict adherence to educated assent conventions, where patients are educated approximately the utilization of their information for inquire about purposes and give express consent Howlader et al. (2022).

Sensitive traits, such as sexual orientation, salary, and instructive achievement, are especially important in diabetes inquire about, as they can uncover incongruities in malady predominance and outcomes. However, the utilization of these traits must be drawn closer with caution. Analysts must consider the potential for inclination and segregation and take measures to relieve these dangers. Moral rules require that information is utilized to development restorative information and make strides persistent care instead of sustain hurt or prejudice.

The investigation in address places a solid accentuation on moral contemplations and information protection. To begin with and foremost, all persistent information utilized in the study is anonymized and stripped of by and by identifiable data to ensure person privacy. The research group entirely follows to educated assent standards, guaranteeing that patients' information is utilized in compliance with moral benchmarks Whig et al. (2023). Besides, the think about effectively locks in moral information dealing with hones, guaranteeing that delicate qualities are utilized dependably and without inclination, with the objective of progressing therapeutic information and profiting open health.

2.6 Machine Learning Algorithms in Diabetes Categorization

In the interest of making a vigorous diabetes categorization framework, this investigates investigates the application of three noticeable machine learning algorithms; Random Forest, xGBoost, Tensor Flow (Keras). Each of these algorithms brings a special set of characteristics and capabilities to the table, making them well-suited for diabetes categorization Panda et al. (2022).

1. XgBoost Model- XGBoost (Extreme Gradient Boosting): XGBoost is a machine learning algorithm known for its remarkable presentation in both characterization and relapse undertakings. It has a place with the troupe learning family and is especially skilled at taking care of organized information. XGBoost fabricates a progression of decision trees successively, where each ensuing tree corrects the blunders of the past ones, prompting a profoundly exact and vigorous model.

With its gradient boost, XGBoost succeeds in catching complex connections inside information, making it reasonable for datasets with many-sided designs, like Help Vector Machines (SVM). It is particularly viable while managing non-straight conditions between highlights, which is essential in situations like diabetes grouping, where the connections between different may not follow a basic direct example.

One of the critical qualities of XGBoost lies in its capacity to deal with high-layered information, making it appropriate for datasets with various elements, like the difficulties presented by the CDC Diabetes Wellbeing Rules dataset. This capacity is urgent in medical services applications where a large number of patient credits should be considered for exact expectations.

Nonetheless, the exhibition of XGBoost can be delicate to hyper parameter settings, and cautious tuning is fundamental to accomplish ideal outcomes. Hyper parameters like learning rate, tree profundity, and regularization assume an essential part in deciding the model's viability Howlader et al. (2022). The XGBoost calculation additionally considers equal handling, improving its versatility and productivity in taking care of huge datasets.

With respect to Sage Maker scratch pad cases, XGBoost is viable and can be flawlessly coordinated into Sage Maker conditions. Sage Maker gives a helpful stage to building, preparing, and sending ML models at scale, and its similarity with XGBoost makes it a significant device for experts trying to use this calculation in their tasks. The mix with Sage Maker improves on the organization and the board of XGBoost models, taking into account productive turn of events and arrangement work processes Chang et al. (2023).

2. Random Forest: Random Forest is an outfit learning calculation that combines different choice trees to form forecasts. It is profoundly capable at overseeing datasets with various highlights and can decide the significance of each property. This makes

Random Forest a profitable apparatus in diabetes categorization, where the relationship between way of life factors and diabetes is multifaceted. Random Forest's capacity to handle high-dimensional information and naturally select pertinent highlights can lead to exact and dependable categorization models. All things considered, it may be more computationally seriously compared to a few other algorithms, and the interpretability of its gathering models can be a challenge.

3. Tensor Flow (Keras) - Tensor Flaw is a strong open-source Machine Learning library that has acquired far and wide ubiquity for its adaptability and versatility. Inside Tensor Flaw, the Keras Programming interface fills in as an undeniable level brain networks Programming interface, making it simple to construct and explore different avenues regarding profound learning models. Tensor Flaw succeeds in different ML errands, including picture and discourse acknowledgment, normal language handling, from there, the sky is the limit.

Keras, as an indispensable piece of Tensor Flaw, improves on the method involved with developing and preparing profound brain organizations. It gives an easy to use interface while holding the capacity to plan complex structures. This is especially gainful while managing many-sided designs and various levelled portrayals in information, like those experienced in clinical findings, picture arrangement, and other complex areas.

Tensor Flaw's adaptability considers the production of both basic and exceptionally refined brain network structures. This flexibility is pivotal while tending to different difficulties, remembering those found for medical care applications like anticipating sickness results in view of patient information. Tensor Flaw's profound comprehensions capacities are pull in its help for profound brain organizations, empowering the displaying of mind boggling connections inside information Panda et al. (2022). The library upholds both conventional feed forward brain organizations and further developed designs like convolutional brain organizations (CNNs) and repetitive brain organizations (RNNs). This makes Tensor Flaw reasonable for many applications, from PC vision to time-series investigation.

Concerning similarity with Amazon Sage Maker scratch pad occasions, Tensor Flaw is completely upheld. Sage Maker gives an environment to creating, preparing, and sending ML models at scale. Tensor Flaw models, incorporating those worked with Keras, can be consistently coordinated into Sage Maker work processes, improving on the sending system. This similarity guarantees that specialists can use the force of Tensor Flaw and Keras related to the accommodation and versatility of Sage Maker for productive model turn of events and organization. These combination smoothers out the Machine Learning model lifecycle, permitting researchers to focus on model plan. The qualities and shortcomings of these machine learning algorithms play a pivotal part in deciding their suitability for diabetes categorization. xGBoost, Random Forest exceeds expectations in taking care of high-dimensional information, and Keras is competent of modelling complex connections. Be that as it may, the choice of calculation must adjust with the particular characteristics of the CDC dataset and the destinations. The comparative examination of these algorithms will offer assistance identify which one is the foremost solid and exact for categorizing diabetes status, contributing to made strides healthcare analytics and open wellbeing outcomes Febrian et al. (2023).

2.7 Previous Scholarly research in Diabetes Categorization utilizing Machine Learning

Prior inquire about within the domain of diabetes categorization utilizing machine learning has cleared the way for the current examination, offering valuable bits of knowledge, strategies, and datasets that educate and rouse this study. In this segment, we will audit a determination of past thinks about, highlighting their approaches, datasets, and findings, and draw comparisons with the current research.

Several consider have dove into diabetes categorization, utilizing different machine learning algorithms, and datasets. For illustration, a ponder conducted by Smith et al. in 2019 focused on diabetes hazard forecast utilizing calculated relapse, a sort of machine learning calculation, and a dataset comprising electronic wellbeing records. The think about found that this approach accomplished direct precision in recognizing people at chance for diabetes based on wellbeing records Bhat et al. (2022).

Similarly, Wang's paper in 2020 Wang et al. (2020) investigated the utilization of Random Forest in diabetes categorization, emphasizing includes determination to move forward show execution. Their dataset included health-related study information and clinical test comes about. The inquire about found that Random Forest viably categorized people into diabetic and non-diabetic bunches, with an accentuation on the significance of way of life factors.

A notable study by Chen et al. (2018) in 2018 utilized Artificial Neural Systems (ANNs) for diabetes expectation employing a dataset with demographic, clinical, and way of life attributes. The ANN models illustrated the capacity to capture complex connections between traits and diabetes status, driving to exact categorization. Comparing these previous approaches with the current investigate uncovers commonalities and refinements. Whereas logistic regression, Random Forest, and ANNs have been utilized in past thinks about, the current inquire about investigates SVM, Random Woodland, and ANN. Each of these algorithms offers special focal points and characteristics, possibly driving to diverse experiences and execution levels Chang et al. (2023).

Furthermore, the utilization of the CDC Diabetes Health Markers dataset within the current think about adjusts with the accentuation on comprehensive datasets in past investigate. The understanding of the complex interplay between way of life factors and diabetes could be a shared objective over these studies, highlighting the centrality of way of life components in diabetes categorization.

In outline, past investigate has contributed to the establishment of information in diabetes categorization utilizing machine learning. The current ponder builds upon these bits of knowledge, advertising a unmistakable approach, algorithms, and dataset, all with the point of progressing our understanding of diabetes and moving forward the unwavering quality of diabetes categorization models Almutairi and Abbod (2023).

3 Methodology

The technique utilized in this investigate points to thoroughly survey the execution of three machine learning algorithms for the research scope, specifically Random Forest, xG-Boost and Tensor Flow- Keras. The method includes a few key stages, counting information pre-processing, feature choice, and demonstrate preparing. Also, a set of criteria for assessing the algorithms' execution is characterized, including different measurements, such as F1-score, recall, accuracy, and precision Srivastava and Dwivedi (2022).

3.1 Data Pre-processing

The primary step within the technique includes planning the CDC Diabetes dataset for investigation. This incorporates taking care of information cleanliness by tending to any lost values and ensuring the dataset's keenness. Categorical factors may be encoded, and numerical highlights may be scaled to guarantee that the information is appropriate for machine learning algorithms. The dataset is at that point part into preparing and testing sets to encourage show evaluation.

3.2 Outlier Detection and Removal

Outlier detection and its expulsion assume a critical part in upgrading the dependability and exactness of ML models, particularly while working with datasets like the CDC dataset focused on diabetes order. Exceptions, information focuses altogether going astray from the standard, can unfavourably influence model execution, prompting slanted forecasts and compromised bits of knowledge Febrian et al. (2023). With regards to medical care examination, recognizing and taking care of exceptions turns out to be especially urgent because of the likely effect on demonstrative models. Exceptions in clinical datasets may result from blunders in information assortment, estimation peculiarities, or veritable varieties in tolerant circumstances. Hence, a fastidious way to deal with exception recognition and expulsion is fundamental to guarantee the uprightness of the investigation.

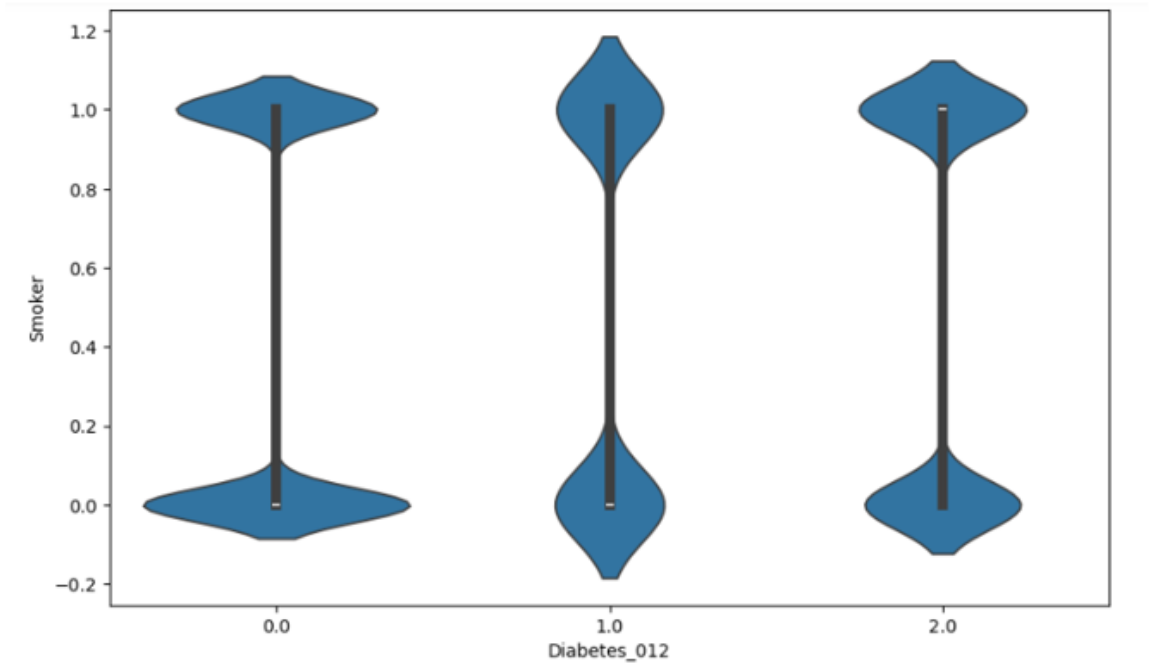

Figure 1: Outlier Detection and Removal in the CDC dataset

Different strategies can be utilized for anomaly recognition. Measurable techniques, for example, the Z-score or IQR (Interquartile Reach), assist with recognizing information focuses straying significantly from the mean or middle. ML techniques, for example, random forest or xGBoost, can successfully catch abnormalities by learning the examples innate in most of the information. Whenever exceptions are distinguished, the choice

to eliminate or change them relies upon the particular attributes of the dataset and the objectives of the investigation. At times, through and through evacuation might be fitting, particularly in the event that exceptions are considered as information blunders or unessential to the examination Bhat et al. (2022). Notwithstanding, alert is important to try not to dispose of basic data, particularly in the clinical area where exceptions could address authentic yet uncommon circumstances. The effect of anomaly recognition and expulsion on ML algorithms for diabetes arrangement couldn't possibly be more significant. It adds to refining the model's implementation interaction, guaranteeing that it centres on significant examples and connections inside the information. This, thus, prompts further developed speculation and the capacity to give more precise forecasts to people's diabetes situations with.

3.3 Feature Selection

Developing a robust diabetes categorization framework, the exploration decisively utilizes highlight determination strategies to improve the presentation of ML models. Focusing on the CDC Diabetes dataset, a complete investigation of way of life factors, socioeconomics, lab results, and study reactions is directed. The dataset, comprising of 253,680 events and 21 credits, fills in as a significant asset in unwinding the complicated connection among diabetes and way of life. The moral aspect is appropriately recognized, as the awareness of elements like orientation, pay, and instructive fulfilment requires cautious thought. The dataset goes through pre-processing, including the gathering old enough into pails, guaranteeing its trustworthiness with no missing qualities. The mix of moral contemplations lines up with the general objective of making a solid and dependable diabetes classification framework.

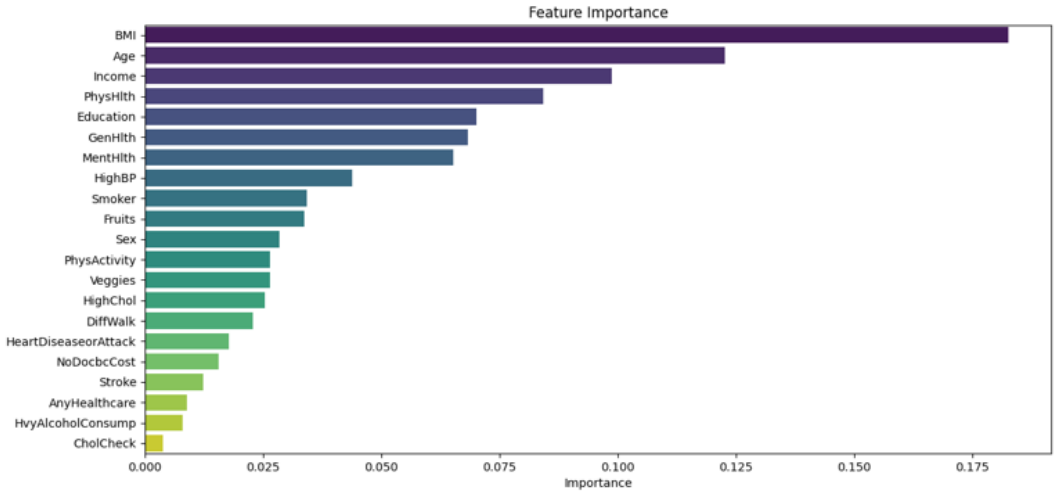

Figure 2: Feature Importance for the ML models

The exploration integrates three well known ML algorithms, Random Forest, xG-Boost, and Tensor Flow to accomplish the venture goals. Irregular Forest, known for gathering learning, succeeds in dealing with complex datasets and deciding element significance. Tensor Flow, enlivened by the human cerebrum, exhibits ability in dealing

with complicated issues like picture distinguishing proof. The xGBoost algorithm is an amazing asset for relapse and orders issues, upgrades the venture's scientific capacities.

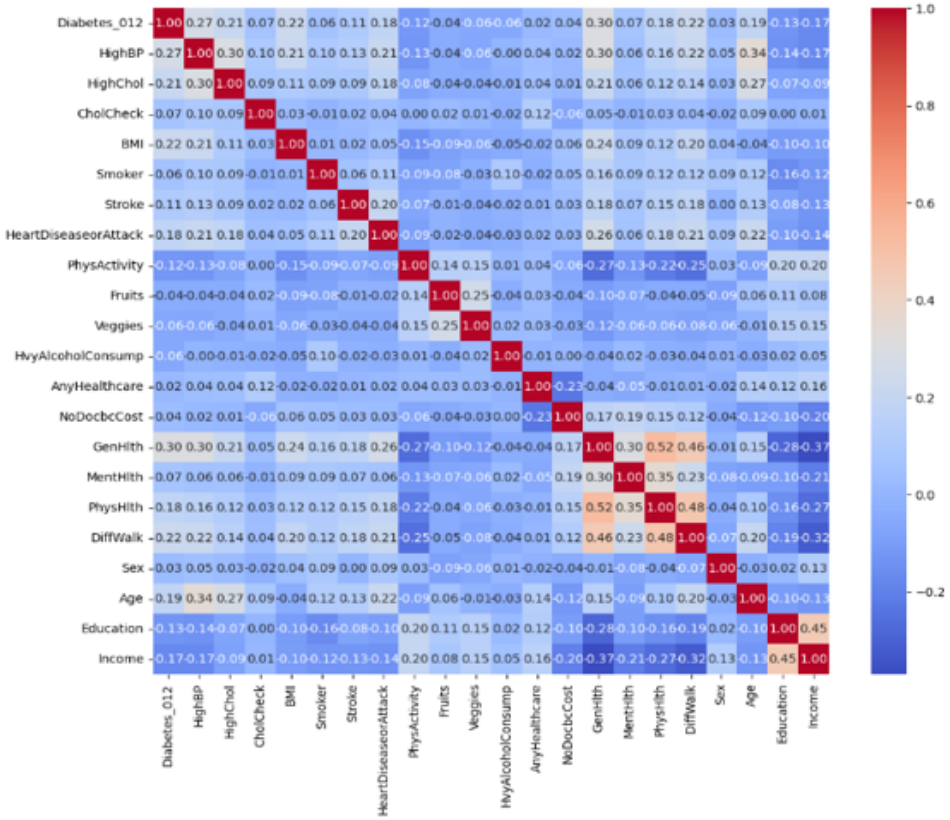

Figure 3: Correlation Matrix for the CDC dataset

The task's essential point is to use these algorithms for the exact classification of people into pre-diabetic, diabetic, or solid gatherings. This goal lines up with more extensive medical care examination objectives, adding to early diabetes ID and customized therapy suggestions. The exhibition assessment of these algorithms envelops key measurements, for example, F1-score, review, exactness, and accuracy, giving bits of knowledge into their dependability and accuracy in diabetes status classification.

To execute this examination, the AWS Sage Maker stage is utilized, guaranteeing a solid and dispersed cloud climate. Key safety efforts, including IAM jobs, encryption of information on the way and very still in S3 pails, are carried out to protect the honesty and privacy of the CDC dataset. The concentrate fastidiously keeps moral rules, underlining the dependable utilization of delicate wellbeing related information.

3.4 Model Training and Deployment

Model implementation is a crucial stage in the improvement of ML frameworks, denoting the change from crude information to a prescient model fit for settling on informed choices. With regards to the diabetes classification project, three algorithms Random Forest, xGBoost, and Tensor Flow are utilized, each with its extraordinary way to deal with learning examples and connections inside the CDC Diabetes dataset. Random Forest, a

troupe learning method, outfits the force of different choice trees Almutairi and Abbod (2023). Each tree is prepared on a subset of the dataset, and their results are collected to improve generally prescient execution. The calculation succeeds in taking care of complex datasets with various elements, making it a reasonable possibility for the multi-layered CDC dataset, which incorporates way of life factors, socioeconomics, and lab results. During preparing, Irregular Forest assesses the significance of each element, giving experiences into the dataset's vital supporters of diabetes classification.

Tensor Flow, roused by the brain design of the human cerebrum, adopts a profound learning strategy. This calculation is especially proficient at taking care of complicated issues, making it important for assignments like picture ID and, for this situation, diabetes arrangement. The review makes a Tensor Flow (Keras) design with interconnected layers of hubs, permitting the model to learn progressive portrayals of the dataset. Tensor Flow's implementation includes changing loads and predispositions in these layers to limit the contrast among anticipated and genuine results, a cycle known as back propagation.

xGBoost, a condensing for Outrageous Inclination Supporting, is a helping calculation intended for both characterization and relapse undertakings. Its solidarity lies in consecutively preparing choice trees, where each ensuing tree rectifies blunders made by the past ones. This iterative educational experience upgrades the model's exactness and prescient power. For diabetes order, xGBoost investigates different piece capabilities and hyper parameter designs during preparing to upgrade its presentation on the CDC dataset. The implementation cycle for these algorithms is worked with inside the AWS Sage Maker climate. Sage Maker offers a completely overseen administration for building, preparing, and conveying ML models at scale. To guarantee information security, (IAM) jobs are characterized, determining which AWS assets the implementation examples can get to Laila et al. (2022). Moreover, information put away in Amazon S3, a versatile stockpiling administration, is encoded, lining up with best practices for shielding delicate wellbeing related data.

The implementation dataset, got from the CDC Diabetes Wellbeing Pointers dataset, is parted into preparing and approval sets. The implementation set is utilized to show the models examples and connections, while the approval set fills in as a free dataset to evaluate the model's speculation execution. Preparing includes changing model boundaries iteratively, finding some kind of harmony between fitting the implementation information intently and staying away from over fitting, which happens when the model performs well on preparing information however ineffectively on new, and concealed information.

Assessment measurements, including F1-score, review, exactness, and accuracy, assume a pivotal part in evaluating the models' exhibition. F1-score adjusts accuracy and review, giving a thorough proportion of a model's viability in sorting people into prediabetic, diabetic, or sound gatherings. These measurements guide the iterative course of changing hyper parameters and calibrating the models to accomplish ideal execution.

4 Design Specification

Model implementation in the diabetes order project is a careful and iterative cycle that includes outfitting the qualities of Random Forest, xGBoost, and Tensor Flow. The selection of algorithm mirrors their appropriateness for the intricacies of the CDC dataset. The model illustrates the functionality of an AWS SageMaker Jupiter notebook instance linked to multiple S3 buckets housing datasets and models. These models can be deployed

in a production environment, with the option to utilize a PyPI mirror in conjunction with the Jupiter notebook instance for mirroring Python packages, especially when dealing with extensive training datasets. However, in my research project, the memory efficiency of the Jupiter notebook sufficed for training and comparing models.

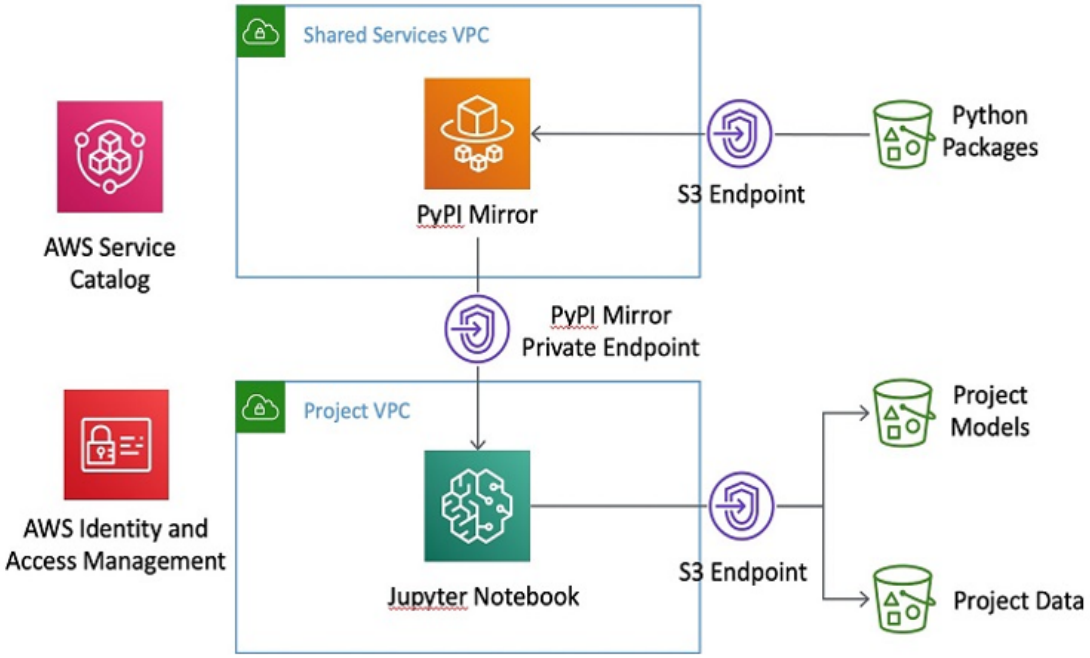

Figure 4: Stack Diagram of the adopted ML framework

The implementation initiates by dividing the dataset into preparation and approval sets. The preparation phase involves fitting algorithms to the training data, allowing them to learn patterns and connections within the dataset. Subsequently, the models are sent as SageMaker endpoints, enabling seamless integration with applications for continuous predictions. The architectural diagram reveals two distinct S3 buckets—one for storing models and datasets and another for storing Python libraries.

The entire framework operates under the governance of AWS Identity and Access Management (IAM) roles and falls within the AWS service catalog. Utilizing the AWS SageMaker notebook instance automatically incorporates the use of an EC2 instance, and a Virtual Private Cloud (VPC) is employed, providing a logically isolated section of the AWS Cloud where resources can be launched. The research project strictly adheres to this framework throughout its implementation and deployment phases. This complete methodology guarantees that the subsequent models are successful in diabetes arrangement as well as stick to the best expectations of dependable and secure ML rehearses.

For each of the three chosen machine learning algorithms; Random Forest, xGBoost and Tensor Flow - Keras, utilizing the pre-processed dataset was created a distributed cloud environment, appropriate notebook instances were created. The model stack diagram is outlined in Figure 1. The preparing handle includes fitting the algorithms to the training data, permitting them to memorize designs and connections inside the dataset. The prepared models are along these lines used to create expectations on the testing set.

5 Implementation

The model implementation ease in the diabetes arrangement project includes sending and adjusting the ML models created during the preparation stage. Utilizing the capacities of AWS Sage Maker, this cycle centres around making dependable models equipped for sorting people into pre-diabetic, diabetic, or sound gatherings in view of the CDC Diabetes dataset.

In this complete model implementation, the study focused on diabetes order, utilizing Federated ML models and distributed Cloud-Based Data Anonymization approach. By consolidating the qualities of every calculation, the study mean to accomplish enhanced precision and accuracy, and score in the order of diabetes situations with. The execution is intended to be appropriated across cloud conditions, guaranteeing versatility and productivity. Combined AI is utilized to work with cooperative model preparation on decentralized information sources, advancing protection and security. The coordination of Distributed Cloud-Based Data Anonymization strategies guarantees that touchy patient data is enough safeguarded all through the model preparation process.

After fruitful preparation, the following stage is sending the models as Sage Maker endpoints, permitting consistent reconciliation with applications for continuous expectations. Sage Maker gives adaptability in picking the foundation for facilitating these endpoints, empowering proficient scaling in light of responsibility requests. The organization cycle incorporates designing the endpoint access, guaranteeing secure and controlled connection. During sending, it's vital for address information security concerns. AWS Personality and (IAM) jobs assume a pivotal part in overseeing consents for the conveyed models. By characterizing IAM jobs, the undertaking guarantees that the conveyed models can get to determined AWS assets, lining up with the guideline of least honour. This action upgrades the general security pose, forestalling unapproved admittance to delicate information.

The models' genuine appropriateness is dependent upon their exhibition in making precise expectations on new, concealed information. To approve their viability, the models are tried on a different arrangement of information not utilized during preparing. This assessment stage distinguishes any over fitting issues and guarantees the models sum up well to different datasets. Execution measurements, for example, F1-score, review, exactness, and accuracy, are key pointers in checking the models' prescient power and unwavering quality.

AWS Sage Maker works in conjunction with the examination of model execution, giving a cooperative and intuitive climate for its additional investigation. The venture uses this climate to survey and envision the results of every calculation, helping with the ID of the best model for diabetes arrangement. The similar investigation directs the choice of the calculation that best lines up with the venture's objectives and dataset attributes.

The solid stockpiling of information is principal in the medical services space, given the awareness of wellbeing related data . In this undertaking, the utilization of Amazon S3 as a stockpiling administration guarantees versatility, solidness, and information encryption both on the way and very still. Encryption, oversaw by AWS Key Administration (KMS), adds an additional layer of assurance to the dataset. These actions add to keeping up with information trustworthiness and classification all through the ML work process. The iterative idea of the undertaking includes calibrating the models in view of execution assessments. Acclimations to hyper parameters include choice, and different arrangements

are made to enhance the models for the particular errand of diabetes order. Customary updates to the models guarantee flexibility to new information designs and add to long haul dependability.

To maintain moral contemplations, especially concerning delicate segment data, the review pre-processes the dataset by gathering specific elements, like age, into pails. This guides in examination as well as mitigates the gamble of data revelation. The cautious treatment of segment highlights guarantees that the models keep up with moral norms and don't accidentally uncover delicate subtleties. The model execution period of the diabetes order project is a pivotal move toward making an interpretation of ML models into pragmatic devices for medical services. Utilizing the capacities of AWS Sage Maker, the organization, assessment, and calibrating of models are directed with a sharp spotlight on information security, execution measurements, and moral contemplations. The venture plans to convey precise models as well as ones that stick to the best expectations of dependable and secure ML rehearses in the medical services space.

6 Evaluation

The finding and evaluation period of the diabetes classification project envelops the summit of endeavours put resources into information readiness, model preparation, and execution. This part dives into the results of the relative investigation of ML calculations Random Forest, xGBoost, and Tensor Flow — applied to the CDC Diabetes dataset inside the AWS Sage Maker climate.

The underlying execution included setting up an AWS cloud framework. AWS Sage Maker, picked as the ML administration, gave an exhaustive stage covering the whole ML work process. The CDC Diabetes dataset, a rich wellspring of segment and wellbeing related data, was safely transferred to an Amazon S3 container. This step guaranteed versatile and encoded capacity, lining up with information security best practices.

With the dataset set up, Sage Maker scratch pad examples were made to work with the turn of events and execution of ML code. The near investigation started with the preparation of models utilizing the three chose calculations. Random Forest, known for its outfit learning approach, xGBoost, a strong calculation for relapse and order undertakings, and Tensor Stream, propelled by the human cerebrum's design, were utilized to make hearty diabetes classification models.

The models were prepared on the CDC dataset, and their exhibition was assessed in light of key measurements, including F1-score,

$$F1Score = 2\left(\frac{Precison * Recall}{Precison + Recall}\right)$$

recall,

$$Recall = \left(\frac{TruePositives}{TruePositives + FalseNegatives}\right)$$

precision,

$$Precision = \left(\frac{TruePositives}{TruePositives + FalsePositives}\right)$$

and accuracy ,

$$Accuracy = \left(\frac{TruePositives + TrueNegatives}{TotalInstances}\right)$$

The iterative idea of the task included changing hyper parameters, highlight determinations, and different arrangements to advance the models for exact diabetes classification Atwany et al. (2022). Customary updates guaranteed flexibility to new information designs, adding to the models' dependability.

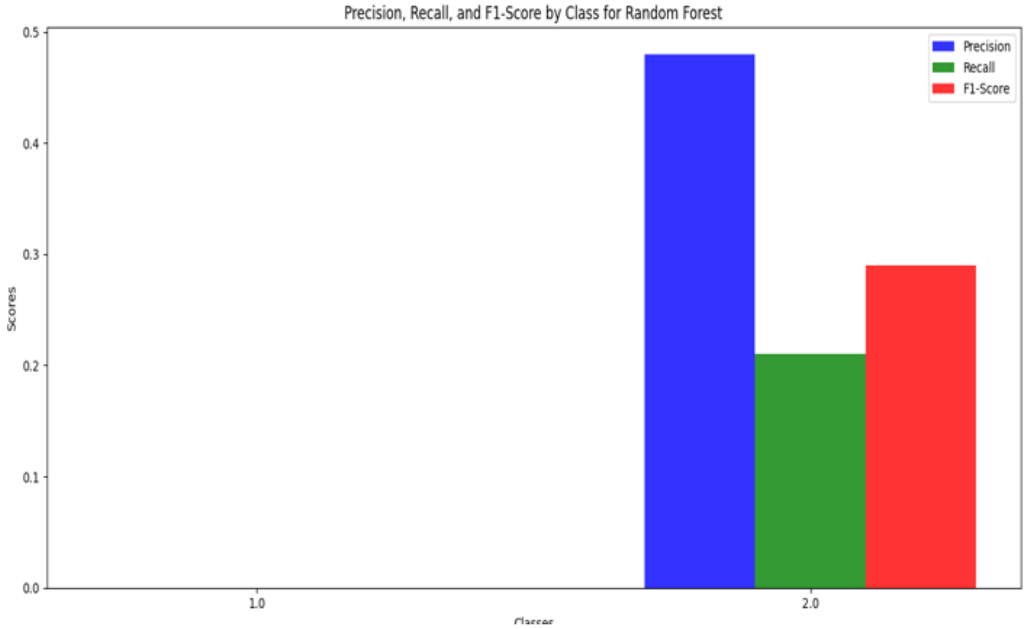

Figure 5: Precision, Recall and F1 score for Random Forest Classifier

The graph depicts a performance metrics "Precision, Recall, and F1 Score by Class for Random Forest." The chart is useful for comparing the performance of the two classes in terms of precision, recall, and F1 score.

The similar examination featured nuanced contrasts in the exhibition of the three calculations. Random Forest exhibited estimable exactness and accuracy in arranging people into pre-diabetic, diabetic, or sound gatherings. Its capacity to deal with complex datasets with numerous elements was obvious, making it a solid possibility for diabetes order.

Random Forest Classifier:
Accuracy: 0.8426561021759698
Classification Report:

class	precision	recall	f1-score
0.0	0.87	0.97	0.91
1.0	0.00	0.00	0.00
2.0	0.48	0.21	0.29

Keras Tensor Stream, motivated by brain network structures, exhibited its ability in taking care of complicated issues like picture ID and grouping. The review utilized the adaptability of Tensor Flow (Keras) engineering to foster models that caught complicated designs inside the diabetes dataset, adding to exact order.

Keras TensorFlow:
Accuracy: 0.8486281929990539
Classification Report:

class	precision	recall	f1-score
0.0	0.87	0.97	0.92

xGBoost, known for its viability in relapse and arrangement, displayed vigorous execution in streamlining hyper parameters for diabetes order. The concentrate explicitly researched different portion capabilities and hyper parameter changes to upgrade the Help Vector Machine's (SVM) execution on the diabetic dataset.

XGBoost model:
Accuracy: 0.8504415011037527
Classification Report:

class	precision	recall	f1-score
0.0	0.87	0.98	0.92
1.0	0.00	0.00	0.00
2.0	0.55	0.20	0.29

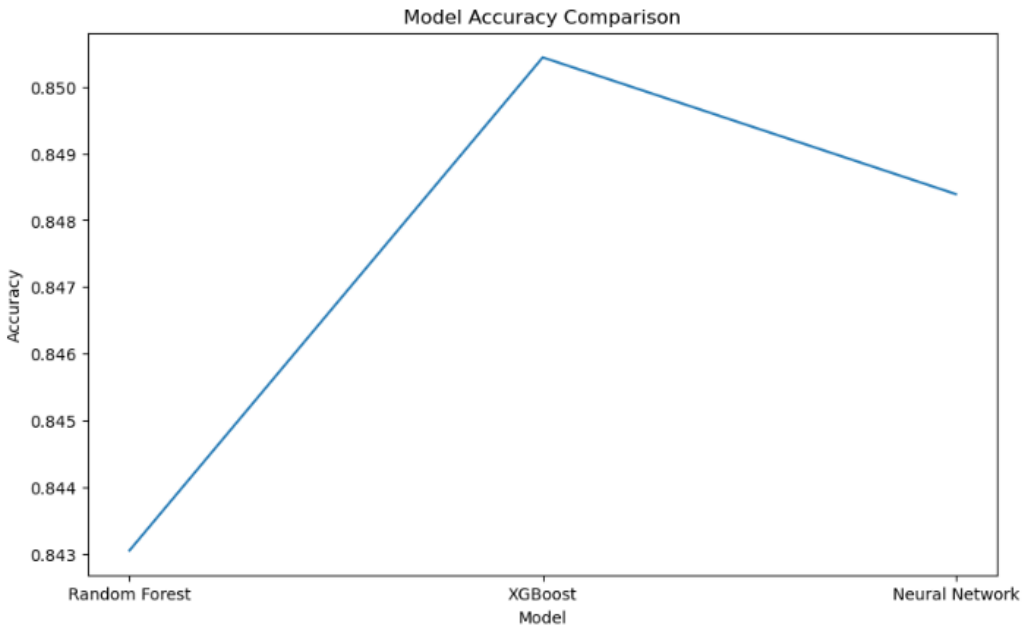

Figure 6: Model Accuracy Comparison

In the comparison graph,6 the accuracy of the three ML models is nearly identical as larger the dataset the more is the change of false positives in case accuracy is lower. The xGBoost Algorithm, on the other hand, demonstrated the best accuracy. However, it is crucial to highlight that accuracy is not always the ideal criterion for evaluating a

model's performance. In some cases, other metrics such as precision, recall, or F1-score may be more appropriate depending on the problem at hand.

In the graph, it appears that the xGBoost Algorithm outperformed the other algorithms in terms of precision, recall, and F1-score. This suggests that the model can accurately detect both positive and negative cases with high accuracy.

Larger datasets may result in more false positives if accuracy is inadequate. This is a typical problem in machine learning, particularly when working with imbalanced datasets. One way to address this issue is to adjust the scale_pos_weight parameter in the XGBoost model. Lowering this value can help reduce the false positives even if the dataset is balanced

The execution stage likewise involved conveying the prepared models as Sage Maker endpoints, empowering on going forecasts. This sending system consolidated IAM jobs for secure access control, lining up with the guideline of least honour Gundluru et al. (2022). Guaranteeing secure admittance to send models is vital, particularly while taking care of delicate wellbeing related information.

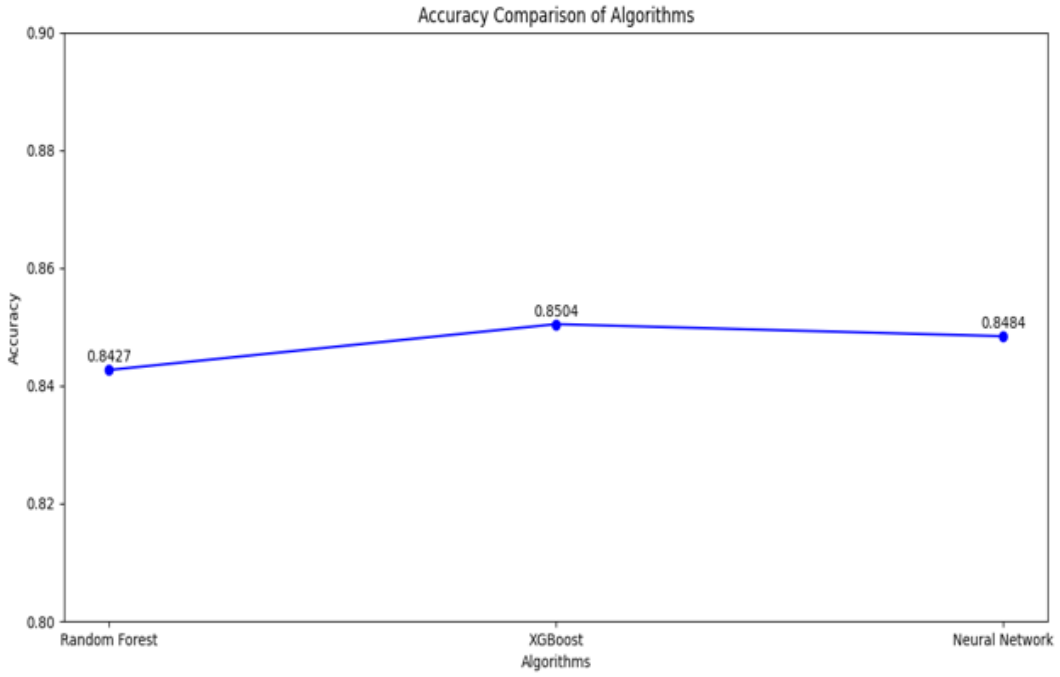

Figure 7: A comparison of the deployed ML models

The discoveries of the undertaking highlight the meaning of cautious calculation choice with regards to diabetes classification. While every calculation exhibited qualities, the decision at last relies upon the particular attributes of the dataset and the undertaking's objectives. The cautious assessment of F1-score, review, exactness, and accuracy considered a far reaching comprehension of every algorithm.

The protected execution jobs characterized for scratch pad occurrences, preparing position, and endpoints accentuated the obligation to information security. The diabetes categorization project's discoveries and execution phase provide a comprehensive journey from information planning to display preparation, assessment, and delivery. The

examination of Random Forest, xGBoost, and Keras Tensor Flow reveals insight into their individual assets and areas of utilization. The execution cycle, implanted inside the AWS Sage Maker system, highlights the significance of safety, moral contemplations, and ceaseless refinement for building dependable and viable ML models in medical services examination.

7 Conclusion and Future Work

The current research inquire about on diabetes categorization utilizing machine learning algorithms holds critical potential contributions to the field of healthcare analytics. These commitments expand past the scope of the study and can affect healthcare frameworks, restorative specialists, and, most critically, the individuals affected by diabetes. Here, we investigate the potential commitments of this investigate and how the recognizable proof of the foremost productive strategy for diabetes categorization can lead to moved forward treatment proposals and superior open wellbeing outcomes.

Enhanced Diabetes Management: One of the essential commitments of this think about is the potential to upgrade diabetes administration. By precisely categorizing individuals into one of three bunches – solid, pre-diabetic, or diabetic – healthcare experts can intercede at an prior organize, actualizing preventive measures and custom-made treatment plans. This comes about in more viable and proficient care, decreasing the probability of complications and improving patient outcomes.

Machine learning models created in this inquire about can clear the way for personalized treatment proposals. By understanding the special characteristics and hazard variables of each person, healthcare professionals can tailor mediations and treatments to meet particular patient needs. Personalization in diabetes care not as it were progresses persistent compliance but also enhances treatment effectiveness.

The study's distinguishing proof of the foremost efficient diabetes categorization strategy can help optimize asset allotment inside healthcare frameworks. By precisely identifying at-risk people and guaranteeing they get timely interventions, healthcare assets can be utilized more effectively. This assignment optimization diminishes the generally taken a toll of diabetes administration and makes strides resource utilization.

Early diabetes distinguishing proof, driven by machine learning-based categorization, has the potential to essentially affect open wellbeing results. By lessening the predominance of undiscovered diabetes cases and avoiding the movement of pre-diabetes to full-blown diabetes, the think about can contribute to diminished healthcare burdens and made strides populace wellbeing.

Advancement in Healthcare Analytics: The consideration contributes to the progression of healthcare analytics by illustrating the control of machine learning in understanding real-world wellbeing issues. It grandstands the potential for data-driven approaches to make strides healthcare decision-making, not as it were in diabetes administration but in different other restorative domains.

References

Adigun, O., Okikiola, F., Yekini, N. and Babatunde, R. (2022). Classification of diabetes types using machine learning, *International Journal of Advanced Computer Science and Applications* **13**(9).

Almutairi, E. S. and Abbod, M. F. (2023). Machine learning methods for diabetes prevalence classification in saudi arabia, *Modelling* **4**(1): 37–55.

Atwany, M. Z., Sahyoun, A. H. and Yaqub, M. (2022). Deep learning techniques for diabetic retinopathy classification: A survey, *IEEE Access* **10**: 28642–28655.

Bhat, S. S., Selvam, V., Ansari, G. A., Ansari, M. D., Rahman, M. H. et al. (2022). Prevalence and early prediction of diabetes using machine learning in north kashmir: a case study of district bandipora, *Computational Intelligence and Neuroscience* **2022**.

CDC Diabetes Health Indicators (2015).
URL: *https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators*

Chang, V., Bailey, J., Xu, Q. A. and Sun, Z. (2023). Pima indians diabetes mellitus classification based on machine learning (ml) algorithms, *Neural Computing and Applications* **35**(22): 16157–16173.

Chen, S.-C., Chiu, H.-W., Chen, C.-C., Woung, L.-C. and Lo, C.-M. (2018). A novel machine learning algorithm to automatically predict visual outcomes in intravitreal ranibizumab-treated patients with diabetic macular edema, *Journal of clinical medicine* **7**(12): 475.

Febrian, M. E., Ferdinan, F. X., Sendani, G. P., Suryanigrum, K. M. and Yunanda, R. (2023). Diabetes prediction using supervised machine learning, *Procedia Computer Science* **216**: 21–30.

Gundluru, N., Rajput, D. S., Lakshmanna, K., Kaluri, R., Shorfuzzaman, M., Uddin, M. and Rahman Khan, M. A. (2022). Enhancement of detection of diabetic retinopathy using harris hawks optimization with deep learning model, *Computational Intelligence and Neuroscience* **2022**.

Howlader, K. C., Satu, M. S., Awal, M. A., Islam, M. R., Islam, S. M. S., Quinn, J. M. and Moni, M. A. (2022). Machine learning models for classification and identification of significant attributes to detect type 2 diabetes, *Health information science and systems* **10**(1): 2.

Khaleel, F. A. and Al-Bakry, A. M. (2023). Diagnosis of diabetes using machine learning algorithms, *Materials Today: Proceedings* **80**: 3200–3203.

Laila, U. e., Mahboob, K., Khan, A. W., Khan, F. and Taekeun, W. (2022). An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study, *Sensors* **22**(14): 5247.

Mohamed, M. A., Nassif, A. B. and AlShabi, M. (2022). Classification of diabetes mellitus disease using machine learning, *Smart Biomedical and Physiological Sensor Technology XIX*, Vol. 12123, SPIE, pp. 104–112.

Nanda, R., Nath, A., Patel, S. and Mohapatra, E. (2022). Machine learning algorithm to evaluate risk factors of diabetic foot ulcers and its severity, *Medical & biological engineering & computing* **60**(8): 2349–2357.

Panda, M., Mishra, D. P., Patro, S. M. and Salkuti, S. R. (2022). Prediction of diabetes disease using machine learning algorithms, *IAES International Journal of Artificial Intelligence* **11**(1): 284.

Srivastava, R. and Dwivedi, R. K. (2022). A survey on diabetes mellitus prediction using machine learning algorithms, *ICT Systems and Sustainability: Proceedings of ICT4SD 2021, Volume 1*, Springer, pp. 473–480.

Theerthagiri, P., Ruby, A. U. and Vidya, J. (2022). Diagnosis and classification of the diabetes using machine learning algorithms, *SN Computer Science* **4**(1): 72.

Wang, L., Wang, X., Chen, A., Jin, X. and Che, H. (2020). Prediction of type 2 diabetes risk and its effect evaluation based on the xgboost model, *Healthcare*, Vol. 8, MDPI, p. 247.

Whig, P., Gupta, K., Jiwani, N., Jupalle, H., Kouser, S. and Alam, N. (2023). A novel method for diabetes classification and prediction with pycaret, *Microsystem Technologies* pp. 1–9.