PREDICTIVE_MODEL_DIABETES_PINAKA_LATEST

by Drea Agoncillo

General metrics

76,849	10,886	820	43 min 32 sec	1 hr 23 min
characters	words	sentences	reading time	speaking time

Score

This text scores better than 98% of all texts checked by Grammarly

Writing Issues

127	13	114
Issues left	Critical	Advanced

Plagiarism

This text hasn't been checked for plagiarism

Writing Issues

34	**Correctness**	
9	Text inconsistencies	
2	Improper formatting	
1	Citation style options	
10	Unknown words	
5	Punctuation in compound/complex sentences	
3	Misplaced words or phrases	
1	Redundant words	
2	Determiner use (a/an/the/this, etc.)	
1	Incorrect phrasing	
90	**Clarity**	
70	Passive voice misuse	
12	Intricate text	
6	Wordy sentences	
2	Unclear sentences	
3	**Engagement**	
3	Word choice	

Unique Words

Measures vocabulary diversity by calculating the percentage of words used only once in your document

18%

unique words

Rare Words

Measures depth of vocabulary by identifying words that are not among the 5,000 most common English words.

44%

rare words

Word Length

Measures average word length

5.6

characters per word

Sentence Length

Measures average sentence length

13.3

words per sentence

PREDICTIVE_MODEL_DIABETES_PINAKA_LATEST

3

Prediction of Diabetes Using Logistic Regression with Elastic Net Regularization

Crisan Josh S. Buendicho

Cris Deine L. Pomperada

Notre Dame of Marbel University

Bachelor of Science in Computer Science

Brenda M. Balala, MIT

December 2024

Approval Sheet

This Thesis[1] entitled Prediction of Diabetes Using Logistic Regression with Elastic Net Regularization, prepared and submitted by Crisan Josh S. Buendicho and Cris Deine L. Pomperada, in partial fulfillment of the requirement for the degree Bachelor of Science in Computer Science, is hereby accepted.

Brenda M. Balala, MIT

Thesis Adviser

Merch Jay P. Dollaga, MIT

Panel Chair

Hajah T. Sueno, DIT(Mama)

Engr. Victorino R. Tobias, Jr., MEP-ECE

Panel Member

Panel Member

Accepted and approved for the conferral of the degree

Bachelor of Science in Computer Science

Victorino R. Tobias, Jr., MEP-ECE

Dean, College of Engineering, Architecture and Computing

6

Abstract

The essential problem of diabetes[2] requires continued early preventive interventions due to its global health significance. The[3] researchers develop a diabetes prediction model by implementing Logistic Regression with Elastic Net Regularization due to its recognized easy-to-understand properties and interpretability. The[3] model[4] analyzes blood pressure, BMI, age, cholesterol measurements, and other factors to establish a tool for early diabetes prevention.

Research data stems from the Behavioral Risk Factor Surveillance System (BRFSS) 2015, which the researchers obtained preprocessed[5] from the Kaggle website to maintain data quality. The[3] Logistic Regression with Elastic Net Regularization model reaches its evaluation outcome through standard classification metrics using accuracy, precision, recall, and F1-score. The[3]

Receiver Operating Characteristic (ROC) curve revealed the strong performance of the model through its AUC values, which showed excellence in recognizing diabetic from non-diabetic cases. Some predictive performance levels were lower in selected classes, indicating potential enhancement areas.

Users can access real-time diabetes predictions through the deployed Flask web app, which inputs medical data. The analysis demonstrates that logistic regression is a dependable, economical method for early diabetes detection. Research in this area should focus on adding new modeling techniques and increasing the dataset size to enhance performance quality and diversity.

Acknowledgments

We thank all those raised by the arms and those contributors who helped make our studies successful. We are immensely grateful to our beloved thesis adviser, Brenda M. Balala, MIT, for her generous assistance in our study. She guided and encouraged us throughout our research journey. Her patience and steady support made a big difference in shaping this thesis.

We also want to thank our thesis panel members, Merch Jay P. Dollaga, MIT (Panel Chair), Hajah T. Sueno, DIT, Al Amin-Mama, and Engr. Victorino R. Tobias, Jr., MEP-ECE. Their brilliant feedback, helpful critiques, and ideas boosted the quality of our study. A special shout-out goes to Al Amin-Mama. He gave us spot-on insights that helped our study take on real meaning. We want to thank Victorino R. Tobias, Jr., MEP-ECE, Dean of the College of Engineering, Architecture, and Computing, for giving us the green light and backing us up to finish this thesis.

We also want to thank our families for always having our backs, cheering us on, and getting where we're coming from during this process. Their faith in us has kept us going the whole time.

Additionally, we want to acknowledge how our talents and hard work allowed us to perform our thesis. It has been the core behind overcoming challenges and achieving this milestone, complementing it with our perseverance knack. And lastly, we would like to thank all our friends for their encouragement and support during this study. We want to thank everyone for your support and your trust in us.

Table of Contents

List of Tables

List of Figures

6

Introduction

Background of the Study

Diabetes Mellitus (DM), commonly known as diabetes, is a chronic disease characterized by high blood sugar levels. More cases continue to emerge, emphasizing the need for improved methods of diagnosis and predicting diabetes Costi et al. (2024). According to the World Health Organization(2023),

type 2 diabetes, linked to insulin resistance, is rising worldwide, while type 1 diabetes results from little or no insulin production. There are around 830 million people out there who are dealing with diabetes, among whom are more than half living in the developing world and of whom fewer than half are on treatment. Low-cost treatment is needed, and a global target has been set to reduce diabetes and obesity by 2025.

Predictive modeling using machine learning has emerged as a promising method for identifying individuals at risk of diabetes. Logistic regression, known for its simplicity and interpretability, has been widely used in diabetes prediction tasks. Recent studies, such as that of Mundargi et al. (2024), have demonstrated the effectiveness of logistic regression in accurately predicting diabetes risk based on various health parameters. Researchers have combined it with techniques such as regularization to improve its accuracy and generalization. A study by Rajendra and Latifi (Rajendra & Latifi, 2021) found that Logistic Regression is highly efficient when paired with feature selection, data pre-processing, and cross-validation, significantly boosting prediction performance. Similarly, Joshi and Dhaka (2021) demonstrated the effectiveness of logistic regression in predicting diabetes, achieving a prediction accuracy of 78.26% while emphasizing the role of cross-validation in improving model reliability and reducing errors. These findings highlight the importance of optimizing predictive algorithms for better diabetes management.

This study focuses on developing a prediction model for diabetes using Logistic Regression with Elastic Net Regularization. Elastic Net Regularization combines the strengths of L1 (Lasso) and L2 (Ridge) penalties. This is particularly beneficial when dealing with multi-class datasets, such as health-related data, where some features may be correlated or redundant. Using Elastic Net Regularization, the study aims to enhance the predictive capability

and generalization of Logistic Regression, making it more effective for classifying individuals into three risk categories: healthy, pre-diabetes[13], and diabetes[2].

Objectives of the Study

The researchers aim to develop a model for predicting diabetes[2] using Logistic Regression with Elastic Net Regularization.

The specific objectives of this study are as follows:

Develop a model that predicts diabetes[2] by:

Utilizing Logistic Regression with Elastic Net Regularization

Optimizing the model's[4] performance using Grid Search with 7-Fold Cross-Validation.

Classify individuals into one of three categories:

Diabetes: High predicted risk score.

Pre-diabetes[13]: Moderate predicted risk score.

Healthy: Low predicted risk score.

Measure and compare the efficacy of Logistic Regression with Elastic Net Regularization against Logistic Regression with the following criteria:

Predictive Accuracy

Precision

Recall

F1 Score

Multi-Class[8] ROC and AUC Curve

Integrate the Logistic Regression with the Elastic Net Regularization Model into a Web Application.

Review of Related Literature

This chapter offers an overview of relevant research and literature from academic databases, journals, and e-books, among other online sources, that

are crucial to understanding the subject under study.

Diabetes Mellitus

Diabetes mellitus is one of the most significant and continually emerging health threats to the population worldwide. Whereas the well-recognized end-organ complications in diabetes, including cardiovascular diseases, stroke, and renal failure, have been established, new complications arising from increased longevity of patients with diabetes have been identified. These complications include cancer, liver disorder, dementia, and infection, respectively. Often with no early diagnosis, cancer, including hepatocellular and pancreatic cancer, has emerged as one of the major killer diseases in diabetic patients. These changes, like the complications in diabetes, mean that diabetes today requires comprehensive care and frequent screenings for these new compounding factors that need to be incorporated with traditional care. Additionally, Glycemia has been found to cause the risk of both cognitive disabilities as well as infections, making the outcomes of the patients worse (Tomic et al., 2022). Figure 1. The Traditional and Emerging Complications of Diabetes Mellitus.

According to (Tomic et al., 2022), in Figure 1, the traditional diabetic complications are coronary heart disease and heart failure, diabetic kidney disease, stroke, retinopathy, peripheral neuropathy, and peripheral vascular disease. The emerging complications are cancer, affective disorder, liver disease, infections, cognitive disability and functional disability, etc, became an issue of concern. Clinical management and prediction models must consider these emergent risks to better control diabetes. As the complexity of managing diabetes continues to increase, integrating predictive models into healthcare becomes essential. This is where machine learning techniques can play a crucial role.

Machine Learning

According to Mahesh (2020), Data science, which allows computers to learn independently from data without being specifically programmed, is in great demand. It greatly enhances the capability of machines in the handling & analysis of data, wherein data can be vast and convoluted and not easily understandable by human intervention. Research has also shown that machine learning algorithms effectively predict diabetes mellitus based on patient information and laboratory details. (Lai et al., 2019) developed the prediction models out of the 13,309-patient dataset from Canada through Gradient Boosting Machine (GBM) and Logistic regression algorithms. These models were evaluated on how they identify patients at risk of developing diabetes and achieved an area under the receiver operating characteristic curve (AROC) value of 84.7% for the GBM model of 84.0% for the Logistic Regression model for early prospective identification of the development of diabetes. Sensitivity rates were 71.6% and 73.4%, respectively, higher than other machine learning techniques like Random Forest and Decision Tree. This study found that the four most essential predictors of diabetes were fasting blood glucose, body mass index (BMI), high-density lipoprotein (HDL), and triglycerides. These models provide valuable frameworks for incorporating machine learning into clinical practice through early identification of high-risk patients and the ability to intervene (Lai et al., 2019). This research also contributes to a growing body of literature on the application of machine learning in healthcare, specifically in chronic conditions such as diabetes, which requires accurate, speedy diagnosis for proper preventative action to be taken.

Lasso Regression

LASSO was derived from an acronym, Least Absolute Shrinkage and Selection Operator. The model is essential when developing a balance between an

accurate and one that is easy to understand. It[3] is possible to apply a penalty, commonly called the l1 penalty; this type has the unique property of shrinking some regression coefficients while setting coefficients to an exact zero. This[3] process[7] of 'shrinkage' assesses variance more accurately and currently performs variable selection in a manner that will minimize overfitting. The[3] objective of Lasso is slightly different; instead of reducing the residual sum of squares, it determines the model parameters subject to the sum of the absolute of the regression coefficients being less than or equal to a fixed value. Therefore[3], Lasso can expand tolerance for multicollinearity in datasets and produce more convenient models than methods such as subset selection and ridge regression. Due[3] to the extreme selection of the variables, Lasso helps decipher associations within the data more effectively and utilizes them in practical scenarios. In[3] large data sets such as medical or financial data, it is often desirable to choose models with few parameters to minimize the potential overfitting of the model[4] and, at the same time, increase predictive power; Lasso accomplishes both of these goals. Moreover[3], it combines the attractive features of subset selection and ridge regression into Lasso to ensure the last outcome is simple to understand without producing overfitting and noisy patterns, making it a balanced predictive tool (Kumar, 2019)[11].

Ridge Regression

Ridge regression is a method that is applied to process multidimensional data or when characteristics measured are significantly larger than the samples (p > n). This[3,21] is quite common in medical research, whereby data matrices become many-sided, with the sides representing characteristics and the rows representing the samples. This[3,22] is because usually, when the number of variables is significantly larger than the number of samples, there is a problem of collinearity, and this, therefore, complicates traditional forms of regression[6]

because one or more of the variables are often very much linearly related. This collinearity makes the differentiation of the contribution of each variable to the results quite challenging. This problem is handled in ridge regression by adding a penalty to the regression model, thus controlling collinearity and offering stable estimates of the regression parameters with improved reliability in prediction. Therefore, ridge regression can address the problem associated with a high dimension, which may lead to the failure of standard regression (Wieringen, 2019).

L1 and L2 regularization

According to Alejandro (Alejandro Ito Aramendia, 2020), regularization techniques are essential and necessary for regression models when the data set contains many predictors. Overfitting is eliminated using the penalty of the absolute value of coefficients, and L1 or Lasso produces a model with few coefficients. This is done using L1 regularization, which employs the method of least absolute shrinkage to set the coefficients of less significant features equal to zero, thus decreasing the model's complexity while making it easier to interpret. L2 Regularization (Ridge) applies a penalty based on the squared characteristic of the coefficients. It does not remove features but shrinks correlated predictors' coefficients towards zero to avoid space occupation. Both these methods help build accurate predictive models, especially in domains such as healthcare. The selection of the penalty strength (λ) is a critical factor in model performance and is usually optimized using techniques such as k-fold cross-validation. Understanding these regularization methods is useful when implementing an Elastic Net Regression model with L1 and L2 advantages.

Elastic Net Regularization

Elastic Net is a notable regularization method that merges the L1 (Lasso) and L2 (Ridge) penalties, which makes it ideal for datasets characterized by many or correlated features. The glmnet R package is widely used to create Elastic Net models. It is capable of processing diverse Generalized Linear Model (GLM) families, supplying the flexibility to refine regression models through the adjustment of hyperparameters such as lambda (λ) and gamma (γ) (Tay et al., 2023). A key benefit of using the 'glmnet' package is its ability to handle multiple performance indications during model assessment. Researchers can immediately validate model functions with cross-validation by using functions like "cv.glmnet" and calculating relevant evaluation metrics, such as Area Under the Curve (AUC) or model-family-based deviance. The package includes the assessment. glmnet function permits additional evaluation of test data to determine performance across many metrics, eliminating the need to rerun cross-validation. For binary outcomes, glmnet has access to specific resources, including "roc.glmnet" and "confusion.glmnet," which empowers users to create Receiver Operating Characteristic (ROC) curves and confusion matrices. These capabilities must be leveraged in areas like healthcare, where accurately predicting diabetes risk is essential. Because of its talent at balancing feature selection and regularization, Elastic Net is commonly chosen for predictive modeling, particularly for exploring and clarifying complex data interconnections (Tay et al., 2023).

Logistic Regression

The notion of the odds of a binary outcome predicates logistic regression. Odds are defined as the number of favorable outcomes to the number of unfavorable outcomes, otherwise known as the ratio probability. More simply, if one of these outcomes is taken as the event of interest, the odds answer a question like how likely this event is compared to not happening. Chances are used in situations

such as betting; for instance, even odds or odds of 1 mean the event happens 50/50 times, as when rolling a standard die, getting an even number. When the odds of a given event, such as rolling a 5, are 2, it means it is occurrence is twice as likely to happen as compared to an unfavorable event (LaValley, 2018). Medical research extensively uses logistic regression prediction models because they calculate probabilities and establish connections between dependent and multiple independent variables. Logistic regression operates effectively as a classification model in diabetes prediction when it incorporates data preprocessing with feature selection and ensemble methods. Logistic regression demonstrates excellent accuracy in structured datasets such as the PIMA Indians Diabetes dataset because variables such as glucose levels, BMI, and age directly affect diabetes risk measurements, according to research. Logistic regression enhances medical diagnostic performance when it operates jointly with ensemble techniques Max Voting and Stacking, according to Rajendra and Latifi (2021). Medical establishments use logistic regression to develop predictive models that aid healthcare providers in making better decisions. The reliability of disease classification depends on using standardized predictor variables and internal and external calibration techniques in logistic regression models, which serve as clinical decision-support systems, according to Shipe et al. (2019).

fold Cross Validation

Machine learning model performance assessment frequently uses k-fold cross-validation to divide the data into k-equal parts. Tests are performed with one of the k folds while training occurs through all other folds. During k iterations, the average performance calculation is incorporated. The identification procedure carries out k repetitions followed by performance averaging across all execution round results. Cross-validation procedures

commonly use k=10 for splitting data since this setting best manages bias and variance. Still[3], researchers can select other k values, such as k=5 or k=7, depending on the specific task (Nti et al., 2021)[11].

The authors analyzed various k values from 3 to 5 with 7 and 10 and up to 15 and 20 using different machine learning models, which included Gradient Boosting Machine (GBM), Logistic Regression, Decision Tree, and K-Nearest Neighbors according to Nti (Nti et al., 2021)[11]. The[3] analysis demonstrated k=7 as a better choice than k=10 because it enhanced validation performance and decreased complex calculations in typical machine learning systems. The[3] research indicates that k=7 might be an ideal choice because it provides high performance and reduces computational costs depending on the situation.

MinMaxScaler

The data normalization method MinMaxScaler converts feature values to a specific range, which typically extends from 0 to 1. The[3] transformation method ensures each feature plays an equivalent role in machine-learning models by stopping attributes with wide numeric spans from controlling the learning process. The[3] machine learning discipline frequently uses MinMaxScaler and StandardScaler as feature scaling tools. MinMaxScaler[3] proves most helpful when data distributions differ from Gaussian patterns and when variable value relationships need to[40] be preserved[41]. The[3] scaling range of MinMaxScaler gets significantly distorted by outliers because this normalization technique is sensitive to outlier points. StandardScaler[3] proves superior for handling outliers because it transforms data to center around zero means with unit variance. StandardScaler[3] provides better results when data follows a normal distribution and when variables possess highly disparate measurement ranges according[42] to Banerjee (2023).

The application of MinMaxScaler on cardiotocogram (CTG) data for fetal health classification enhanced the detection capabilities of XGBoost, Decision Trees, and Multi-SVM machine learning classifiers. Feature normalization enables models to extract delicate patterns from data, which boosts their accuracy levels and stability during classification operations (Alkurdi & Abdulazeez, 2024). The combination of MinMaxScaler missing data imputation and Tomek Links enabled bank telemarketing neural networks to optimize their binary classification performance. The research by Halim et al. (2020) proved that MinMaxScaler produced superior results to MaxAbsScaler and StandardScaler because it enhanced classification accuracy, specifically when datasets were balanced. The selection between MinMaxScaler and StandardScaler relies on the dataset's characteristics and the machine-learning operation's requirements.

Research findings demonstrate that MinMaxScaler is a highly effective tool to enhance the predictive capabilities of machine learning models for air pollution predictions. The research about differing feature scale methods for particulate matter concentration forecasts assessed MinMaxScaler against StandardScaler, RobustScaler, and Normalizer. MinMaxScaler demonstrated its most significant advantages for Random Forest Regressor and XGBoost Regressor because these models need optimized numerical inputs to achieve accurate predictions. MinMaxScaler maintains original data distribution during scaling operations, which proves beneficial for datasets that do not adopt a typical distribution pattern. StandardScaler becomes the preferred option when dealing with outliers because it handles extreme values effectively without altering the data range significantly (Thakker & Buch, 2024). The selection between MinMaxScaler and StandardScaler depends on the unique

characteristics of the dataset together with the particular requirements of the applied machine learning system.

Multi-class[8] Area Under the ROC curve (MAUC)

MAUC extends traditional AUC by assessing classification models that work with multiple classes. The[3] calculation of MAUC functions to extend binary AUC methodology for various courses by averaging pairwise AUC scores. This[3] evaluation technique produces reliable results for classifier assessment of multi-class[8] situations, particularly in datasets with class imbalance problems because standard accuracy metrics become unreliable. Researchers[3] have recently invested efforts to develop efficient MAUC calculation methods because of their complex computational demands. The[3] researchers developed Prequential Multi-Class[8] AUC (PMAUC) as a system that integrates sliding window methodology and red-black tree data architecture to calculate real-time MAUC values. The[3] methodology provides both increased efficiency during computation and dependable evaluation methods for online learning models dealing with imbalanced multi-class[8] data streams (Wang & Minku, 2020).[11]

The optimization of MAUC requires a solution for the distribution imbalance because the metric tends to favor majority classes. The[3] M metric represents one of the proposed optimization frameworks that seeks fair class pair ranking across all groups. The[3] M metric prevents minority class pair discrimination by distributing their weight equally toward the total MAUC score calculation. Research[3] teams developed efficient empirical surrogate risk minimization techniques for optimizing MAUC with reduced computational expenses and theoretical proof. The[3] studied approaches have enhanced classification accuracy in situations with abundant class imbalance, such as medical diagnosis, fraud detection, and recommender systems (Yang et al., 2021).[11] The[3]

superiority of MAUC emerges through recent developments that prove its ability to outperform standard metrics for multi-class classification evaluation. Research on multi-class imbalanced classification boosting methods has evaluated MAUC as a performance measure against metrics such as G-mean and Matthews Correlation Coefficient (MCC). MAUC offers significant insights about classifier performance across different classes, but several researchers indicate that MCC delivers superior balanced evaluation when dealing with heavily imbalanced cases. Experimental testing of different boosting algorithms demonstrated that MAUC maintains its reliability for model performance assessment when used with ensemble learning methods alongside AdaBoost, CatBoost, and LogitBoost. The study indicates that MAUC provides optimal results when quality ranking precedes the raw prediction accuracy of class labels (Tanha et al., 2020). The continuous optimization of MAUC techniques and their intersection with advanced machine learning methods confirms its essential role in multi-class classification applications.

Grid Search Method

The grid search method selects parameters for the model to provide the best chance of guessing the following data parameters that still need to be labeled. According to Hsu, Chang, and Lin (Chih-Wei Hsu, Chih-Chung Chang, 2020), when using cross-validation, the grid search method is recommended for identifying the best values of two parameters: C and y. It checks variations of (C, y) and selects the pair with the best cross-validation accuracy. The findings highlight two key reasons for choosing the grid search approach: 1) the potential problem of using the approximation heuristics that often avoid parameter search throughout the range of possible values, and 2) the time taken to search for proper parameters using the grid search approach is only slightly more than the time required for other refined approaches when the

number of parameters remains small. This study applied Five classification algorithms to the voice frequency dataset: KNN, Logistic Regression, Random Forest, Decision Tree, and Neural Network. The accuracy measures of these classifiers are shown in table 1, following the outcomes of the algorithms.

Table 1. Accuracy of Classification Algorithm on the Dataset Before and After Transformation.

The classifiers' accuracy was estimated using a 7-fold cross-validation technique. According to the findings, Random Forest yields the best performance of the tested algorithms by having the highest accuracy of 0.9675 on the dataset before and after data transformation. Data pre-processing was done by scaling the main variables into the range [0, 1] (RAMADHAN et al., 2017).

Figure 2. Literature Map

Related Studies

A Comprehensive Review of Machine Learning Techniques on Diabetes Detection

Sharma & Shah (Sharma & Shah, 2021) reviewed different machine-learning methods used in Diabetes Mellitus (DM) detection. Their study points to the developing concern about diabetes, especially in response to its rising frequency and related health issues. The review categorizes and evaluates numerous algorithms, including Support Vector Machines (SVM), Decision Trees (DT), and Logistic Regression (LR), while also exploring deep learning methods such as Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN).

Sharma & Shah points out the benefits of machine learning in healthcare, particularly in automating diagnostics, to improve human accuracy. Sharma[3] and Shah reviewed the difficulties encountered in the industry, with a particular focus on data inadequacy, which may affect model performance and reliability. In[3] addition, the authors noted the essential requirement for the effective deployment of models in real-world clinical contexts. They[3] concluded that machine learning progress, notably through deep learning applications, shows great promise for improving methods of diabetes detection. These[3] models can potentially improve timely interventions and, ultimately, patients' outcomes and strategies for diabetes management through their automation and refining of diagnostic processes.

Predicting Diabetes in Adults: Identifying Important Features in Unbalanced Data Over a 5-Year Cohort Study Using Machine Learning Algorithm.

According to Moghaddam (Moghaddam, 2021)[11], Type 2 diabetes mellitus is a chronic, noncommunicable disease affecting insulin demand or supply, attributed to a considerable disease burden worldwide. Diabetes[3] is linked[45] to these educated health risks such as cardiovascular diseases, stroke, kidney failures, blindness as well as the amputation of limbs. It[3] has also been linked to worsened dementia, hearing issues, and certain types of cancer – an inevitable reason for early mortality. Interval[3] datasets are very problematic in predictive modeling as they have skewed distributions, which increase the probability of the automated learning model, making unfair outcomes unreliable. In[3] such scenarios, models tend to give way to the majority class while completely disregarding the minority class, which, in value-oriented applications, as seen in the diabetes prediction scenario, leads to inferior results. This[3,46] is a critical problem to tackle if we build practical and dependable models to reduce inferential errors.

Moghaddam[47] is confined to reducing the problem of data skewness in predicting Type 2 diabetes using machine learning approaches. Information[3] obtained through the Fasa Adult Cohort Study (FACS), where 10,000 participants underwent follow-up over five years, was used to develop and calibrate prediction algorithms. The[3] large-scale dataset provided a rich learning environment for understanding the impact of data imbalance on the performance of a model and assessing techniques on how to mitigate the effects of data imbalance. Their[3] goal is to improve the actual predictive performance of the machine learning models for high-risk patients developing Type 2 diabetes with inherent problems of an imbalanced data set.

Machine Learning-Based Diabetes Classification and Prediction for Healthcare Applications

Butt et al.'s study, titled "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications," seeks to utilize machine learning techniques to detect and predict diabetes mellitus, a deadly disease with a global health impact. Using[3] the PIMA Indian Diabetes dataset, three classification models (Random Forest, Logistic Regression, and Multilayer Perceptron (MLP)), as well as three predictive models (Long Short Term Memory (LSTM), Moving Averages and Linear Regression), are evaluated. The[3] results show that MLP achieved the highest classification accuracy (86.08%), and LSTM performed better than other predictive models (87.26%) in handling healthcare data, proving the robustness of these algorithms with such data. In[3] addition to machine learning approaches, a hypothetical Internet of Things (IoT)-based diabetes monitoring system is proposed. The[3] system combines Bluetooth Low Energy (BLE) sensors and smartphones to collect and process real-time health data, including blood glucose levels, blood pressure, and weight. Data[3] streaming is made[48] possible using tools like Apache Kafka and

storage using MongoDB to enable the transmission and analysis of data to get real-time health recommendations and early diagnosis. The design improves the practicability of monitoring and controlling diabetes by providing actionable insights into patients' health. The proposed system enhances diabetes management by integrating prediction analytics with IoT. Using fine-tuned algorithms such as MLP for classification and LSTM for prediction improves diagnostic accuracy and provides a basis for real-time user-centric healthcare. We find that the study highlights the need for additional research in the deployment of these models in practical healthcare applications and also suggests the inclusion of other technologies, such as genetic algorithms, for monitoring the disease (Butt et al., 2021).

Synthesis

The importance of using machine learning to predict and diagnose Diabetes Mellitus (DM) has increased significantly over the years, and the reviewed studies have shown this. Machine learning helps improve how diabetes is detected, but there is still work to do, such as handling poor-quality data and making the models more straightforward to understand, they say. As per Sharma & Shah (Sharma & Shah, 2021), A common idea in these studies is that different machine learning algorithms may enhance the accuracy in detecting diabetes. Sharma & Shah have reviewed different ML techniques, such as Support Vector Machines (SVM), Decision Trees (DT), and deep learning techniques, such as Artificial Neural Networks (ANN). However, their study proves that machine learning can help the process of diagnosing diabetes to be more accurate and faster. (Graph Neural Networks, GNN, as suggested by Costi (Costi et al., 2024), can more accurately predict diabetes.) Costi also employed techniques to make the machine learning models more understandable by doctors and patients, who must trust the results. (Moghaddam, 2021) One of

the significant challenges faced in these studies is the problem of imbalanced data. When the data used to train the models has a small number of one group (people with diabetes) vs. another group (people without diabetes), this occurs. It focuses on how the imbalanced data affects the performance of ML models. Monghaddam used a large dataset from the Fasa Adult Cohort Study to investigate how to solve this problem and improve the reliability of the models. Sharma & Shah also pointed out that imbalanced data makes applying these models in the real world difficult. Butt et al. (Butt et al., 2021) also perform the study on datasets like PIMA Indian Diabetes and classify diabetes amongst different machine learning models like random forest, logistic regression, and MLP. They conducted their study together to establish the ability of MLP to provide high classification accuracy and propose new predictive models such as Long Short-Term Memory (LSTM) to analyze resilient healthcare data. They also suggest an IoT system for people with diabetes that captures real-time data and proposes action plans that will be helpful in their prevention and control of diabetes. However, all the studies conclude that machine learning has great potential to improve how diabetes is detected and managed. As machine learning methods mature, they can make more accurate predictions, Sharma & Shah, Moghaddam, Costi, and Butt show. The work of Costi on making the models more understandable, together with Monghaddam's focus on fixing data issues and Sharma and Shah's various techniques, demonstrate how machine learning can help in healthcare. Overall, these studies provide valuable insight into how machine learning can more effectively and earlier identify diabetes. Each study is a unique challenge, whether improving data quality accuracy or making the model more interpretable. These results influence the proposed study's design, and a focus on imbalanced datasets is made.

Concept of the Study

Figure 3. Conceptual[3] Framework of the Study

Data Acquisition

The dataset employed in this research is from the Behavioral Risk Factor Surveillance System (BRFSS) 2015, acquired by the researchers from the Kaggle website. The[3] BRFSS data is an annual health telephone survey conducted by the Centers for Disease Control and Prevention (CDC) and includes 253,680 responses and 21 related feature variables: Diabetes_012, HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education and Income. The[3] target variable, Diabetes_012, categorizes individuals into three classes: 0, meaning Healthy, a low predicted risk score; 1, meaning Pre-diabetes,[13] a moderate predicted risk score; and 2, meaning Diabetes,[2] a high-risk

score. Being the dominant class, Healthy has more samples than the diabetes and pre-diabetes class, which shows that the class is imbalanced.

Data Preparation

Data preparation will focus on ensuring data quality and addressing the class imbalance in the target variable (Diabetes_012), where the dataset contains a higher proportion of healthy individuals than those with diabetes and prediabetes. Missing values will be identified to ensure the dataset is clean before processing, and appropriate imputation or removal strategies will be applied based on the results. Additionally, data types will be verified to ensure numerical and categorical variables are correctly formatted. If necessary, incorrect data types will be converted for consistency in preprocessing. To address the class imbalance, the Synthetic Minority Over-Sampling Technique with Tomek Links (SMOTETOMEK) will be applied. This technique helps create a more balanced dataset by synthesizing new minority class samples while removing noisy or borderline samples using Tomek Links. Applying SMOTETOMEK during preprocessing will give the dataset an equal class distribution, enhancing the model's ability to generalize and improve classification performance.

Model Development: Logistic Regression with Elastic Net Regression

In this part, the whole dataset is divided into a training set (70 percent) and a test set (30 percent) for the training and evaluation of the model. The Logistic Regression model is wrapped with a OneVsRestClassifier to implement the one-vs-rest strategy for multi-class prediction. In this approach, each binary classifier is trained for one class against all the other courses, allowing the logistic regression model to handle multiple classes effectively. Logistic Regression uses the sigmoid function on the linear combination of input features to estimate the probability of a class, making it suitable for multi-

class prediction tasks. The model also incorporates Elastic Net Regularization, which combines L1 and L2 penalties to address issues with dependent features and improve feature selection.

Model Validation

To ensure consistent and unbiased model performance, 7-fold Cross-Validation will be used in the Grid Search Method to select the best combination of hyperparameters (alpha and lambda) in the model.

Model Evaluation

Predictive Accuracy, Precision, Recall, F1 Score, Multi-Class ROC, and AUC Curve will evaluate model performance. Logistic Regression with Elastic Net Regularization performance metrics will be tested against standard Logistic Regression. The overall correctness of the model will be measured by predictive accuracy. At the same time, precision and recall will judge its capacity to find true positives and thus avoid false negatives. The researchers of this study will have a balanced view of how the model performs by using the F1 score, which is the harmonic mean of precision and recall. It compares approaches to see which is better at predictive capability and generalization to unseen data.

Integration of the model into the web application

After evaluating the proposed diabetes prediction model, it will be implemented in a web application through Flask API. The model will be saved in a local server so the web application can input the users' health data. The model will then analyze the same data and predict, categorizing the user as healthy, pre-diabetic, or diabetic. This integration makes passing information back and forth between the front-end user interface and the trained model

easier, making predictions possible in real-time and offering users an easy, manageable means of health risk evaluation.

Significance of the Study

This study aims to develop a model that predicts diabetes[2] using Logistic Regression with Elastic Net Regularization. The[3] significance of this study will be beneficial to the following:

Patients and Individuals at Risk. By[3] providing accurate risk assessments, this study will help individuals make informed choices about their health, potentially lowering their chances of developing diabetes[2] and its complications.

Healthcare Providers. The[3] findings of this study will give healthcare professionals a reliable model to tailor patient care, enabling targeted actions that improve patient outcomes and better use of healthcare resources.

Public Health Authorities. This[3] study will support public health efforts by identifying common risk factors in communities, which will help create specific prevention programs to reduce the occurrence of Diabetes[2].

Researchers in the Machine Learning Field. This[3] study will add to the understanding how machine learning can be used in health research, encouraging further exploration of advanced techniques for predicting chronic diseases.

Future Researchers. This[3] study will provide a helpful guide for future studies on diabetes prediction, emphasizing the need to consider various lifestyle and genetic factors in assessing risk. As[3] the complexity of managing diabetes[2] continues to increase, integrating predictive models into healthcare becomes essential. This[3,62] is where machine learning techniques can play a crucial role.

Scope and Limitations

The study's main aim is to apply a logistic regression with elastic net regularization on the BRFSS2015 dataset in constructing a diabetes prediction model. This study is centered on using the strengths of Elastic Net Regularization to handle the multicollinearity of health indicators and to allow for an efficient variable selection by combining the Lasso and Ridge penalties. The study is fundamental in establishing whether the model can perform better or is equal to the logistic regression model. Data collection in the present research is derived from a health-related survey called BRFSS2015, available in the public domain. It contains variables considered crucial for the risk of developing diabetes prediction. However, there are limitations involved in the present study: Fasting blood glucose as a clinical parameter capable of improving the predictions has yet to be included.

Operational Definition of Terms

The following terms are defined according to how they are used in the study.

Predictive Model

A computational framework that uses historical data to forecast future outcomes.

Diabetes

Chronic metabolic disorder creates elevated blood sugar because the body lacks proper insulin production or fails to use insulin effectively.

Dataset

A structured data collection is used to train and evaluate the predictive model.

SMOTETOMEK (Synthetic Minority Over-Sampling Technique with Tomek Links)

A data preprocessing technique used to handle class imbalance. It combines SMOTE, which generates synthetic samples for minority classes, with Tomek Links, which removes overlapping majority class instances. In this study,

SMOTETOMEK is applied[70,71] to balance the dataset, ensuring the model[4] can effectively classify healthy, pre-diabetic, and diabetic cases.

MinMaxScaler

A feature scaling technique that transforms numerical values into a fixed range (typically [0,1]) to ensure uniformity across different input features. In[3] this study, MinMaxScaler is applied[72] to standardize 21 related feature variables (e.g., Diabetes_012, HighBP) before training the logistic regression model, preventing certain features from dominating the predictions.

Logistic Regression with Elastic Net Regularization

A machine learning classification algorithm that integrates L1 (Lasso) and L2 (Ridge) regularization to enhance predictive performance and prevent overfitting. In[3] this study, logistic regression[6] with elastic net regularization is employed[73] to classify individuals into three categories: Healthy, Pre-Diabetic, and Diabetic.

Grid Search with 7-Fold Cross-Validation

A hyperparameter tuning method that systematically evaluates different model parameter combinations to identify the best coefficients of regularization for logistic[74] regression model's regularization strength and L1/L2 ratio, ensuring the best predictive accuracy.

Flask Web Application

A lightweight web framework for Python used to deploy machine learning models.

Method

Materials

Hardware

The hardware in this study consists of a high-performance laptop with an AMD Ryzen 7 6800H HS Mobile Processor (8-core/16-thread, up to 4.7 GHz) and 16GB DDR5-4800 RAM for multitasking and data processing. Storage[3] is made of 512GB M.2 NVMe PCIe 4.0 SSD, allowing quick access to information. Data[3] can then[75] be presented accurately through an IPS-level display with a 144Hz refresh rate. Also[3], the discrete Nvidia GeForce RTX 3050 GPU (4 GB GDDR6) integrated into the device offers improved machine learning operations performance.

Software

For running ML operations, the Google Collab will be used to[76,77] execute Python scripts to analyze data and train and evaluate models. Google[3] Collab has an interactive environment where it accompanies GPU for fast computation, especially when executing large computations, and it supports group work. It[3] can accommodate a range of prominent machine learning libraries, including TensorFlow, sci-kit-learn, and Pandas, respectively; thus, it is perfect for performing the Logistic Regression with Elastic Net Regularization model and every other associated analysis. This[3,78] guarantees that this study can accommodate big data and execute heavy computations as required.

Data

The dataset used for this study is from the BRFSS2015 acquired from the Kaggle website.

Procedures

Data Gathering

The dataset used in this study is from the Behavioral Risk Factor Surveillance System (BRFSS) 2015, obtained via the Kaggle website. The[3] BRFSS is an annual health survey conducted by telephone under the Centers for Disease Control and Prevention (CDC). It[3] consists of 253,680 responses and 21 associated

feature variables, including Diabetes_012, HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education, and Income. The target variable, Diabetes_012, classifies individuals into three categories: 0 for Healthy (low predicted risk), 1 for Pre-diabetes (moderate predicted risk), and 2 for Diabetes (high predicted risk).

Pre-processing

The BRFSS2015 dataset will undergo pre-processing before model development to ensure it is ready for analysis. Although the dataset has undergone initial cleaning, it will be further examined to verify its cleanliness by checking for missing values and validating data types. Class imbalance and data formatting will also be addressed to ensure the dataset is properly prepared for training and testing.

Figure 4. Load the BRFSS2015 diabetes dataset.

To manage, analyze, and compute data, the study used pandas, seaborn, matplotlib, and numpy libraries to enhance the effectiveness of the analysis. Furthermore, the os library was used for path handling to make path operations and file handling problem-free. The data set was obtained from the CDC's BRFSS2015, which was then read into a DataFrame pandas.read_csv() assuming the data is in the local directory.

Figure 5. Ensuring data quality

The code checks for missing values in the dataset using df.isnull().sum() and verifies the data types of each column using df.dtypes to ensure proper formatting.

Figure 6. Data Separation as X and Y

The data was preprocessed before being analyzed and fed into the model in the next step. This process started by splitting the dataset into feature sets (X) and dependent variables (Y). In particular, the variable of interest Y was operationalized as an indicator of diabetes presence in the dataset, and other health markers were considered as features X.

Figure 7. Splitting the dataset

In data splitting the train_test_split function from sklearn.model_selection was used to split the data into training and testing data sets. The dataset was divided at a 70:30 ratio to evaluate the proposed model, where 70% was employed for training the model and 30% for model evaluation. The parameter random_state was set to 42 to make the result reproducible.

Figure 8. Applying SMOTETOMEK to the training data

In the following processing stage, SMOTETomek handles class imbalance by combining the Synthetic Minority Over-sampling Technique (SMOTE) and Tomek Links. The initial strategy produces synthetic samples to expand the minority groups of pre-diabetes and diabetes before achieving data set equilibrium. The next operation employs SMOTETomek to eliminate majority class instances that overlap with each other while removing noise and enhancing class discriminability. The approach improves the accuracy of categorized data while keeping a pure decision boundary.

Figure 9. Apply MinMaxScaler to the data.

The code in Figure 9 implements MinMaxScaler from sklearn, preprocessing for normalizing features by scaling into the specified range, which typically spans from zero to one. The fit_transform() method is executed on x_train_resampled to rescale every feature value between the training-set-based minimum and maximum values. To prevent leakage during testing the transform() method uses the same scaling parameters from training which were learned from x_train_resampled. The normalization procedure makes all features equally important, resulting in stable model performance.

Building the Model (application of the algorithm)/ Simulation Process

The Logistic Regression model with Elastic Net Regularization algorithms is used to combine the strengths of L1 (Lasso) and L2 (Ridge) regularization techniques for both feature selection and shrinkage. The Elastic Net Regularization approach helps manage multicollinearity and improves model generalization.

Figure 10. Training the Model using Logistic Regression with Elastic Net

The code in Figure 10 defines a Logistic Regression model with Elastic Net regularization, combining L1 and L2 penalties for feature selection. The solver='saga' is used for efficient optimization with large datasets, and max_iter=10000 ensures sufficient iterations for convergence, especially with complex data.

Figure 11. Wrapping the model in OneVsRestClassifier

Wrapping the LogisticRegression model with a OneVsRestClassifier performs the one-vs-rest strategy on a given feature. Each binary classifier is trained for

one class against all other classes; therefore, the logistic regression model can handle multiple classes.

Figure 12. Hyperparameter Tuning with Grid Search

Grid Search executes the model hyperparameter tuning process, optimizing its functional capabilities. The tuning procedure used two fundamental controls as its main elements:

Estimator__C: A wide scale of inverse regularization values was tested to maintain equilibrium between bias and variance.

Estimator__l1_ratio: Elastic Net controls the L1 (Lasso) versus L2 (Ridge) regularization through its mixing parameter to achieve an ideal balance between essential features and predictive capabilities.

Figure 13. Performing Grid Search with 7-Fold Cross-Validation

The code performs hyperparameter tuning using GridSearchCV on the OneVsRestClassifier with the LogisticRegression model wrapped inside it. The GridSearchCV searches over a range of hyperparameters, defined in param_dist, to identify the best combination for optimal model performance. It uses 7-fold cross-validation (cv=7) and parallelizes the process with n_jobs=-1 to speed up computation. The verbose=True option provides detailed output during the search process.

Evaluating the Model

In the evaluation of the performance of the trained model for diabetes prediction, specific performance measures are employed, hence the accuracy of the model. The measures are predictive accuracy, precision, recall, and F1 score. These assessment factors give the overall perspective of how the model

operates while offering the essential requirements to assess risk and applicable in the real environment.

Predictive Accuracy

Accuracy of prediction is typically considered[109] the primary metric of success in Machine Learning (ML) applications and how well a model is likely to perform during prediction. Hence,[3] it is essential to attain high predictive accuracy because the model[4] has to work in the natural environment with optimal efficiency (Accuracy, n.d.).

Accuracy=(TP + TN)(TP +FP+TN+FN)

Precision

Precision derives from dividing correctly[110] predicted positive cases against all predicted instances, which should be positive. The[3] ratio shows how precisely the model[4] identifies positive outcomes in situations where evaluation centers on accurate positive recognition. According[3] to the model,[4] high precision indicates that the predicted positive outcome will probably be correct (Murphy & Moore, 2019).[11]

Precision= TPTP+TN

Recall

The analysis technique Recall detects actual positive cases through its sensitivity measurement. The[3] model[4] demonstrates its ability to identify all genuine positive results within the dataset. Severe[3] negative consequences emerge from no positive cases, so a high recall value helps the model[4] detect the most actual positive instances (Murphy & Moore, 2019).[11]

Recall= TPTP+FN

F1 Score

The F1-Score is a two-sided measure that incorporates precision and recall with a provision for their balance. It[3] is an average of precision and recall, which

is why it weighs both values – achieving a high value of both metrics is critical. The F1-Score takes its highest at the precision-recall point of (0.5, 0.5), which depicts that the proposed model has the best nearness of topped performance with the least false positive and false negative (Murphy & Moore, 2019).

F1-Score= 11recall+1precision

Figure 14. Predictive Accuracy

The code determines model accuracy through the accuracy_score function, which compares the predicted labels y_pred against true y_test. The program displays the accuracy measurement that shows the ratio of successful predictions made in the test dataset. The measurement method allows the model to be evaluated for performance.

Figure 15. Precision

The code to decide the model precision is achieved by using the method called precision_score, where y_pred is the predicted label, and y_test is the actual label. The average='macro' option gives precision for each class individually and then provides the evaluation with the average. Precision defines the accuracy of positive classification of outcomes to distinguish those correctly predicted from those not, thus evaluating the number of actual positives among all the optimistic predictions made by the model. The result is then printed.

Figure 16. Recall

The code then computes the model's recall using the recall_score function with y pred and y_test being used to check for the correctness of the predictions. The average= 'macro' option calculates each class's recall and then takes its mean. Accuracy measures the ratio of corresponding correct predictions to the

total sum of actual cases, showing which percentage of optimistic scenarios the model predicts. The result is then printed.

Figure 17. F1 Score

The code calculates the F1 score, which balances precision and recall, using the f1_score function with average='macro' to compute the score for each class and then average it. The result is printed.

Figure 18. Compute Multi-Class ROC and AUC Curve

Performance evaluation of the multi-class Logistic Regression Model with Elastic Net regularization requires the calculation of Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) measurements for all separate classes. The One-vs-Rest method served the process because the dataset has three classes: 0 for no diabetes, 1 for pre-diabetes, and 2 for diabetes. Each category was evaluated against every other class using multi-class categorizations during the computations of ROC curves. The Receiver Operating Characteristic (ROC) curve depicts how True Positive Rate (TPR) relates to False Positive Rate (FPR) as the threshold changes to reveal model discrimination for positive class and negative class data points. The AUC value computation for each class depended on totaling the ROC curve area measurement. Model performance directly rises with increasing AUC measurements from zero through one. Model performance detecting positive cases shows its ability to correct recognition by approaching an AUC value of 1, but random guessing is indicated by AUC values approaching 0.5.

Integration of the model into the web application

Figure 18. Integration of the model into the web application

The Flask API web application uses 21 feature variables to predict diabetes. The Flask API application uses model.pkl and scaler.pkl to load trained preprocessing components for the 21 feature variables. The index.html web page contains an interface that lets users input their health data, which consists of age, education level, income, general health rating, history of heart disease, sex, difficulty of walking, physical health, no visit to the doctor because of cost, mental health, BMI, smoking history, cholesterol check, cholesterol, eating fruits or vegetables, high blood pressure, physical activity, blood pressure, cholesterol levels or stroke. The backend system receives this data from the JavaScript script (script.js). After scaling features on the backend, the model gets them and generates a prediction amongst three possibilities (0, 1, or 2). The application converts numeric outputs to more straightforward readable categories where zero stands for "Healthy," 1 represents "Pre-Diabetes," and 2 indicates "Diabetes" to make results easily understandable for users.

The Flask /predict API endpoint processes health data sent through JSON by converting it to a pandas DataFrame, leading to feature scaling and model prediction using trained models. The outcome from the model returns a human-friendly response containing diagnostic labels matching "Healthy," "Pre-Diabetes," or "Diabetes." JavaScript APIs automatically update the webpage

with prediction results for user display. The system utilizes Flask APIs, machine learning, and user-friendly interfaces to transfer health-related user input data into a practical prediction program.

Results and Discussions

The model performance evaluation consists of precision, recall, and F1-score and accuracy, together with multi-class ROC and AUC curves. The study utilized 253,680 BRFSS 2015 survey participants who received evaluation as either healthy (0) pre-diabetes (1) or diabetes (2)—the splitting of data involved training the data at 70%, with testing data of 30%. The Logistic Regression model with Elastic Net regularization completed its training stage until it assessed performance values.

Table 2. Results of the Logistic Regression Model Performance Evaluation.

Predictive Accuracy	Precision	Recall	F1-Score
0.6613	0.4393	0.5152	0.4274

The baseline Logistic Regression model performance measuring criteria are reported in Table 2, which shows the results for predictive accuracy, precision and recall, and F1-score values. The Logistic Regression model's results correctly predicted 66.13% of the time, thus achieving a 0.6613 predictive

accuracy rate. The model showed a precision of 0.4393 and recall of 0.5152 with an F1-score of 0.4274, demonstrating proper positive instance classification and precision-recall trade-off ability.

Table 3. Results of Logistic Regression Model with Elastic Net Regularization Performance Evaluation.

Predictive Accuracy	Precision	Recall	F1-Score
0.6614	0.8464	0.6614	0.7286

The performance measurements of the Elastic Net Regularized Logistic Regression model appear in Table 3. The predictive results, including accuracy and precision, together with recall and F1-score values, are displayed in the table. The updated model achieved improved accuracy compared to the initial model at 0.6613, with its performance reaching a new mark of 0.6614. The model achieves improved precision at 0.8464 because it detects more authentic positive cases than the baseline model. The recall measures have grown to 0.6614 because the model detects more actual positive instances. The F1-score improved significantly to 0.7286 after regularization implementation, confirming the method enables better generalization capability from the model . Elastic Net Regularization enhanced the model classification by successfully improving evaluation metrics compared to logistic regression results.

Figure 19. Multi-Class ROC and AUC Curve Plot

A multi-class classification model demonstrates its performance balance between actual positive rate (recall) and false positive rate through the Receiver Operating Characteristic (ROC) curve. The model's performance to differentiate particular classes from others can be seen in separate curves. The Area Under the Curve (AUC) measurement helps assess model discrimination ability across all groups, yet its higher value reflects superior classification recognition. Relative to random guessing, the predictive power of Class 2 (AUC = 0.82) matches Class 0 (AUC = 0.81) in strength since their curves approach the upper left sector. A lower AUC score (0.68) signifies that Class 1 poses considerable challenges to the model for accurately identifying instances belonging to that class. All curves demonstrate higher precision than random guessing based on the diagonal reference line because they exceed its value of AUC = 0.5.

Table 4. Results of Multi-Class ROC and AUC Curve.

Comparison
Logistic Regression
Logistic Regression with Elastic Net Regularization
Class 0 (No Diabetes)
AUC = 0.81
AUC = 0.81
Class 1 (Pre-Diabetes)
AUC = 0.68

AUC = 0.68

Class 2 (Diabetes)

AUC = 0.82

AUC = 0.82

In Table 4, The base model (Logistic Regression) achieved an AUC of 0.81 for Healthy individuals, 0.68 for Pre-Diabetes, and 0.82 for Diabetes. These values indicate that the model performs well in identifying Healthy and Diabetic individuals but struggles with Pre-Diabetes, which has the lowest AUC score. This suggests that the decision boundary between Healthy and Pre-Diabetes is not well-defined, likely due to overlapping feature distributions in the dataset. The Logistic Regression with Elastic Net Regularization yielded identical AUC values (0.81, 0.68, and 0.82). While Elastic Net Regularization is expected to improve generalization and prevent overfitting by penalizing unnecessary features, its impact on AUC scores was minimal. This could indicate that the dataset was not highly prone to overfitting or that the regularization strength did not significantly alter feature selection to enhance class discrimination.

Figure 20. Diabetes Prediction Web Application.

Figure 20 presents the web application implementing the Logistic Regression with Elastic Net Regularization. This application features twenty-one input boxes where users can input a specific number for each of 21 related feature variables. Upon clicking the Predict button, the model will provide a real-time output of Healthy, Pre-Diabetes, or Diabetes.

Figure 21. Actual Data vs. Predicted Data

User

Actual Data	Predicted Data
1	
0	
2	
2	
0	
0	
3	
1	
1	
4	
2	
2	
5	
2	
2	
6	
0	
1	
7	
0	
0	
8	
0	
1	
9	

0

1

10

0

0

The table displays results from the diabetes classification predictions against the actual data through which the model was examined. Each user receives model predictions in the Predicted Data table through evaluations of high blood pressure (HighBP), cholesterol levels (High Chol), BMI, smoking status, and additional health-related features. The Actual Data table shows the classification of diabetes (0 means healthy, 1 means pre-diabetes, and 2 means diabetes). A model accuracy assessment becomes possible by aligning the predicted data with the actual diabetes risk classifications obtained from the Actual Data table.

Conclusions

The following conclusions are drawn based on the results of the study.

The study created an effective predictive model by applying Elastic Net Regularization to Logistic Regression. The study reached performance optimization through Grid Search with 7-fold Cross-Validation.

The model successfully divided people into groups of Healthy, Pre-diabetes, and Diabetes status. Through its evaluation, the model demonstrated excellent diagnostic capabilities for Class 0 (No Diabetes) and Class 2 (Diabetes) by

achieving AUC scores of 0.81 and 0.82. The predictive ability of Class 1 (Pre-diabetes) measured with an AUC value of 0.68 requires additional enhancements to balance class management effectively.

Standard precision, recall score, and F1-score improved significantly in the Logistic Regression with Elastic Net Regularization model compared to the baseline model. The improved detection abilities of positive cases between multiple classes become evident through these recent model enhancements. The predictive accuracy stayed stable as the performance improvements in precision, recall, and F1-score validated the significance of regularization techniques in multiclass classification problems. Model performance metrics indicated that the AUC scores validated its successful discrimination of different classes.

The model implementation appeared within an interactive web interface that allowed real-time diabetes risk assessment. The model found widespread practical application through integration, which proved its value as an identification tool for early risks and a support system for clinical choices within healthcare settings.

The examined study proves that machine learning methods demonstrate substantial promise for diabetes prediction. A model based on logistic regression with elastic net regularization enables the research to develop a solution that enhances predictive accuracy and scalability for real-time risk evaluation. The results establish fundamental standards for diabetes prediction improvements, which will help future research in this domain. The model succeeds in reliable prediction-making because it improved all key performance metrics, including accuracy, precision, recall, and F1-score, alongside practical web application integration for real-world healthcare usage.

Recommendations

The researchers present potential recommendations for future investigators to progress the study from the examination findings and assessment results:

Researchers can explore alternative fine-tuning approaches for better performance of Logistic Regression with Elastic Net Regularization in diabetes prediction by optimizing regularization parameters and implementing feature engineering methods.

The model needs more considerable, more varied health datasets to enhance scalability to deliver effective performance across different population groups and health risk characteristics.

Measure how well the proposed model identifies health risks from chronic diseases and its effectiveness for healthcare predictions outside the scope of diabetes detection and classification.

Healthcare experts or medical professionals should evaluate the model to establish its practical application under hospital conditions. Such an assessment helps determine if the model suits hospital workflows for generating compelling healthcare predictions.

An evaluation of Random Forest and XGBoost machine learning models through performance comparison with Logistic Regression models will identify the most appropriate solution for real-time healthcare predictions.

References

Accuracy, B. (n.d.). Machine learning:

Alejandro Ito Aramendia. (n.d.). L1 and L2 Regularization (Part 1): A Complete Guide. L1 and L2 Regularization (Part 1): A Complete Guide. https://medium.com/@alejandro.itoaramendia/l1-and-l2-regularization-part-1-a-complete-guide-51cf45bb4ade#:~:text=What exactly is L1 and, term to the loss function

Banerjee, C. (2023). MinMaxScaler. MinMaxScaler.

Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. R. (2021). Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications. Journal of Healthcare Engineering, 2021. https://doi.org/10.1155/2021/9930985

Chih-Wei Hsu, Chih-Chung Chang, and C.-J. L. (2020). A sociological analysis of the satanic verses affair. Theory, Culture and Society, 17(1), 39–61. https://doi.org/10.1177/02632760022050997

Hassan, M. (2024). What is a kaggle? https://www.kaggle.com/discussions/general/328265

Joshi, R. D., & Dhakal, C. K. (2021). Predicting type 2 diabetes using logistic regression and machine learning approaches. International Journal of

Environmental Research and Public Health, 18(14). https://doi.org/10.3390/ijerph18147346

Kumar, D. (2019). Ridge Regression and Lasso Estimators for Data Analysis.

Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. BMC Endocrine Disorders, 19(1), 1–9. https://doi.org/10.1186/s12902-019-0436-6

LaValley, M. P. (2018). Logistic regression. Circulation, 117(18), 2395–2399. https://doi.org/10.1161/CIRCULATIONAHA.106.682658

Mahesh, B. (2020). Machine Learning Algorithms - A Review. International Journal of Science and Research (IJSR), 9(1), 381–386. https://doi.org/10.21275/art20203995

Moghaddam, M. T. (2021). Predicting Diabetes in Adults : Identifying Important Features in Unbalanced Data Over a 5-Year Cohort Study Using Machine Learning Algorithm.

Murphy, A., & Moore, C. (2019). Confusion matrix. Radiopaedia.Org, October. https://doi.org/10.53347/rid-68081

Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation. International Journal of Information Technology and Computer Science, 13(6), 61–71. https://doi.org/10.5815/ijitcs.2021.06.05

Rajendra, P., & Latifi, S. (2021). Prediction of diabetes using logistic regression and ensemble techniques. Computer Methods and Programs in Biomedicine Update, 1, 100032. https://doi.org/10.1016/j.cmpbup.2021.100032

RAMADHAN, M. M., SITANGGANG, I. S., NASUTION, F. R., & GHIFARI, A. (2017). Parameter Tuning in Random Forest Based on Grid Search Method for Gender Classification Based on Voice Frequency. DEStech Transactions on Computer

Science and Engineering, Cece. https://doi.org/10.12783/dtcse/cece2017/14611

Sharma, T., & Shah, M. (2021). A comprehensive review of machine learning techniques on diabetes detection. Visual Computing for Industry, Biomedicine, and Art, 4(1). https://doi.org/10.1186/s42492-021-00097-7

Tang, K., Wang, R., & Chen, T. (2011). Towards Maximizing the Area Under the ROC Curve for Multi-Class Classification Problems. Proceedings of the 25th AAAI Conference on Artificial Intelligence, AAAI 2011, Elkan 2001, 483–488. https://doi.org/10.1609/aaai.v25i1.7901

Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic Net Regularization Paths for All Generalized Linear Models. Journal of Statistical Software, 106(1). https://doi.org/10.18637/jss.v106.i01

Tomic, D., Shaw, J. E., & Magliano, D. J. (2022). The burden and risks of emerging complications of diabetes mellitus. Nature Reviews Endocrinology, 18(9), 525–539. https://doi.org/10.1038/s41574-022-00690-7

Wieringen, W. N. Van. (2019). Lecture notes on ridge regression. 0–30.

World Health Organization. (2023). Diabetes.

Alkurdi, A. A. H., & Abdulazeez, A. M. (2024). Comprehensive classification of fetal health using cardiotocogram data based on machine learning. Indonesian Journal of Computer Science, 13(1), 277-290.

Halim, K. N. A., Fadzil, A. F. A., & Jaya, A. S. M. (2020). Data pre-processing algorithm for neural network binary classification model in bank telemarketing. International Journal of Innovative Technology and Exploring Engineering, 9(3), 272-284.

Thakker, Z. L., & Buch, S. H. (2024). Effect of feature scaling pre-processing techniques on machine learning algorithms to predict particulate matter

concentration for Gandhinagar, Gujarat, India. International Journal of Scientific Research in Science and Technology, 11(1), 410-419.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: An experimental review. Journal of Big Data, 7(70). https://doi.org/10.1186/s40537-020-00349-y

Wang, S., & Minku, L. L. (2020). AUC estimation and concept drift detection for imbalanced data streams with multiple classes. In Proceedings of the IEEE Conference on Machine Learning and Applications (ICMLA) (pp. 1-10). IEEE.

Yang, Z., Xu, Q., Bao, S., Cao, X., & Huang, Q. (2021). Learning with multiclass AUC: Theory and algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2021.3101125

Rajendra, P., & Latifi, S. (2021). Prediction of diabetes using logistic regression and ensemble techniques. Computer Methods and Programs in Biomedicine Update, 1, 100032. https://doi.org/10.1016/j.cmpbup.2021.100032

Shipe, M. E., Deppen, S. A., Farjah, F., & Grogan, E. L. (2019). Developing prediction models for clinical use using logistic regression: An overview. Journal of Thoracic Disease, 11(Suppl 4), S574-S584. https://doi.org/10.21037/jtd.2019.01.25

Mundargi, Z., Dabade, M., Chindhe, Y., Bondre, S., Chaudhary, A. (2024). Diabetes Prediction Using Logistic Regression. In: Gundebommu, S.L., Sadasivuni, L., Malladi, L.S. (eds) Renewable Energy, Green Computing, and Sustainable Development. REGS 2023. Communications in Computer and Information Science, vol 2081. Springer, Cham. https://doi.org/10.1007/978-3-031-58607-1_4

1.	*Thesis; thesis*	Text inconsistencies	Correctness
2.	*diabetes; Diabetes*	Text inconsistencies	Correctness
3.	*. The; . Some; . Research; . We; . She; . Her; . A; . He; . Their; . It; . And; . Compute; . Accuracy; . Results; . Results; . Literature-Map; . Conceptual; . Load; . Ensuring; . Data; . Splitting; . Applying; . Apply; . Training; . Wrapping; . Hyperparameter; . Performing; . Predictive; . Precisi...*	Text inconsistencies	Correctness
4.	*model; Model; model's*	Text inconsistencies	Correctness
5.	*preprocessed; Pre-processing; pre-processing; preprocessing*	Text inconsistencies	Correctness
6.	*regression; Regression*	Text inconsistencies	Correctness
7.	*process; Process*	Text inconsistencies	Correctness
8.	*Multi-class; Multi-Class; multi-class; multiclass*	Text inconsistencies	Correctness
9.	1 .	Improper formatting	Correctness
10.	*been set*	Passive voice misuse	Clarity
11.	*(Rajendra & Latifi, 2021); (Tomic et al., 2022); (Lai et al., 2019); (Kumar, 2019); (Wieringen, 2019); (Tay et al., 2023); (LaValley, 2018); (Nti et al., 2021); (Alkurdi & Abdulazeez, 2024); (Thakker & Buch, 2024); (Wang & Minku, 2020); (Yang et al., 2021); (Tanha et al., 2020); (RAMADHAN et al., 2...*	Citation style options	Correctness
12.	*This*	Intricate text	Clarity
13.	*pre-diabetes; Pre-diabetes; prediabetes; Pre-Diabetes*	Text inconsistencies	Correctness

14.	*been identified*	Passive voice misuse	Clarity
15.	*be incorporated*	Passive voice misuse	Clarity
16.	*been found*	Passive voice misuse	Clarity
17.	*This*	Intricate text	Clarity
18.	*These models were evaluated*	Passive voice misuse	Clarity
19.	*be taken*	Passive voice misuse	Clarity
20.	*was derived*	Passive voice misuse	Clarity
21.	*This*	Intricate text	Clarity
22.	*This*	Intricate text	Clarity
23.	*is handled*	Passive voice misuse	Clarity
24.	*is eliminated*	Passive voice misuse	Clarity
25.	*This*	Intricate text	Clarity
26.	*is done*	Passive voice misuse	Clarity
27.	*glmnet*	Unknown words	Correctness
28.	*is widely used*	Passive voice misuse	Clarity
29.	*glmnet*	Unknown words	Correctness
30.	*glmnet*	Unknown words	Correctness
31.	*These capabilities must be leveraged*	Passive voice misuse	Clarity
32.	*is commonly chosen*	Passive voice misuse	Clarity
33.	*are defined*	Passive voice misuse	Clarity

34.	*is taken*	Passive voice misuse	Clarity
35.	*is compared*	Passive voice misuse	Clarity
36.	*are used*	Passive voice misuse	Clarity
37.	~~fold~~ → Fold	Improper formatting	Correctness
38.	*Machine learning model performance assessment*	Intricate text	Clarity
39.	*is incorporated*	Passive voice misuse	Clarity
40.	~~need to~~ → must	Wordy sentences	Clarity
41.	*be preserved*	Passive voice misuse	Clarity
42.	, according	Punctuation in compound/complex sentences	Correctness
43.	*are shown*	Passive voice misuse	Clarity
44.	*was done*	Passive voice misuse	Clarity
45.	*is linked*	Passive voice misuse	Clarity
46.	*This*	Intricate text	Clarity
47.	*is confined*	Passive voice misuse	Clarity
48.	*is made*	Passive voice misuse	Clarity
49.	~~One of the~~ → ¶ One of the	Intricate text	Clarity
50.	*is made*	Passive voice misuse	Clarity
51.	*Missing values will be identified*	Passive voice misuse	Clarity
52.	*appropriate imputation or removal strategies will be applied*	Passive voice misuse	Clarity

53.	*data types will be verified*	Passive voice misuse	Clarity
54.	*incorrect data types will be converted*	Passive voice misuse	Clarity
55.	*To address the class imbalance*	Misplaced words or phrases	Correctness
56.	*the Synthetic Minority Over-Sampling Technique with Tomek Links (SMOTETOMEK) will be applied*	Passive voice misuse	Clarity
57.	*is divided*	Passive voice misuse	Clarity
58.	*Logistic Regression with Elastic Net Regularization performance metrics will be tested*	Passive voice misuse	Clarity
59.	*be measured*	Passive voice misuse	Clarity
60.	*compares approaches*	Redundant words	Correctness
61.	*it will be implemented*	Passive voice misuse	Clarity
62.	*This*	Intricate text	Clarity
63.	~~study's main aim is~~ → study aims	Wordy sentences	Clarity
64.	*is centered*	Passive voice misuse	Clarity
65.	*is derived*	Passive voice misuse	Clarity
66.	*be included*	Passive voice misuse	Clarity
67.	*are defined*	Passive voice misuse	Clarity
68.	*are used*	Passive voice misuse	Clarity
69.	*is used*	Passive voice misuse	Clarity

70.	*In this study, SMOTETOMEK is applied to balance the dataset, ensuring the model can effectively classify healthy, pre-diabetic, and diabetic cases.*	Unclear sentences	Clarity
71.	*is applied*	Passive voice misuse	Clarity
72.	*is applied*	Passive voice misuse	Clarity
73.	*is employed*	Passive voice misuse	Clarity
74.	the logistic	Determiner use (a/an/the/this, etc.)	Correctness
75.	~~then~~	Wordy sentences	Clarity
76.	*the Google Collab will be used*	Passive voice misuse	Clarity
77.	~~be used to~~	Wordy sentences	Clarity
78.	*This*	Intricate text	Clarity
79.	*be further examined*	Passive voice misuse	Clarity
80.	*Class imbalance and data formatting will also be addressed*	Passive voice misuse	Clarity
81.	*is properly prepared*	Passive voice misuse	Clarity
82.	adequately prepared, appropriately prepared, correctly prepared, prepared correctly	Word choice	Engagement
83.	*To manage, analyze, and compute data*	Misplaced words or phrases	Correctness
84.	*matplotlib*	Unknown words	Correctness
85.	*numpy*	Unknown words	Correctness
86.	*the os library was used*	Passive voice misuse	Clarity

87.	*The data set was obtained*	Passive voice misuse	Clarity
88.	*was then read*	Passive voice misuse	Clarity
89.	, assuming	Punctuation in compound/complex sentences	Correctness
90.	*isnull*	Unknown words	Correctness
91.	*The data was preprocessed*	Passive voice misuse	Clarity
92.	*was operationalized*	Passive voice misuse	Clarity
93.	*were considered*	Passive voice misuse	Clarity
94.	, the	Punctuation in compound/complex sentences	Correctness
95.	*sklearn*	Unknown words	Correctness
96.	~~split~~ → separate	Word choice	Engagement
97.	*was employed*	Passive voice misuse	Clarity
98.	*sklearn*	Unknown words	Correctness
99.	*is executed*	Passive voice misuse	Clarity
100.	, the	Punctuation in compound/complex sentences	Correctness
101.	, which	Punctuation in compound/complex sentences	Correctness
102.	~~which were~~	Wordy sentences	Clarity
103.	*were learned*	Passive voice misuse	Clarity
104.	~~is used to combine~~ → combines	Wordy sentences	Clarity

105.	*is trained*	Passive voice misuse	Clarity
106.	~~classes~~ → courses	Word choice	Engagement
107.	*was tested*	Passive voice misuse	Clarity
108.	*are employed*	Passive voice misuse	Clarity
109.	*is typically considered*	Passive voice misuse	Clarity
110.	*correctly*	Misplaced words or phrases	Correctness
111.	*be evaluated*	Passive voice misuse	Clarity
112.	*is achieved*	Passive voice misuse	Clarity
113.	*is then printed*	Passive voice misuse	Clarity
114.	*is then printed*	Passive voice misuse	Clarity
115.	*is printed*	Passive voice misuse	Clarity
116.	*is indicated*	Passive voice misuse	Clarity
117.	*pkl*	Unknown words	Correctness
118.	*pkl*	Unknown words	Correctness
119.	*are reported*	Passive voice misuse	Clarity
120.	*be seen*	Passive voice misuse	Clarity
121.	*This*	Intricate text	Clarity
122.	*is expected*	Passive voice misuse	Clarity
123.	*This*	Intricate text	Clarity
124.	*the model was examined*	Passive voice misuse	Clarity

125.	the Elastic	Determiner use (a/an/the/this, etc.)	Correctness
126.	~~to~~ → and	Incorrect phrasing	Correctness
127.	*The model needs more considerable, more varied health datasets to enhance scalability to deliver effective performance across different population groups and health risk characteristics.*	Unclear sentences	Clarity