Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK

DISSERTATION

DATA ANALYTICS IN HEALTHCARE

JOSÉ DIOGO SEQUEIRA BERTÃO

Lisbon School of Economics & Management
Universidade de Lisboa

MARCH - 2024

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK
DISSERTATION

DATA ANALYTICS IN HEALTHCARE

JOSÉ DIOGO SEQUEIRA BERTÃO

SUPERVISION:
MÁRIO CALDEIRA

MARCH - 2024

ABBREVIATIONS

AUC – Area Under the Receiver Operating Characteristic Curve.

JEL – Journal of Economic Literature.

KNN –K-Nearest Neighbors.

ML – Machine Learning.

RL – Reinforcement Learning.

ABSTRACT

With an emphasis on the creation of machine learning models for the early diagnosis of diabetes, this thesis investigates the potential of data analytics in the healthcare industry. This study attempts to address the growing global prevalence of diabetes and the urgent need for accurate and widely available early detection techniques at a low cost. Even though they are useful, traditional diagnostic techniques frequently detect diseases at a later stage and may not be accessible to everyone in need, which raises the possibility of harsh consequences. This study uses a dataset from Kaggle that includes several features relevant to the diagnosis of diabetes to create prediction models that try to detect the disease early on.

The methodology employed involves the application of several machine learning techniques, including Logistic Regression, Decision Tree, Random Forest, Extreme Gradient Boosting (XGBoost), Gradient Boosting, Naive Bayes, K-Nearest-Neighbors and Neural Networks (Multi-layer Perceptron), implemented in Python. These models were evaluated based on their accuracy and precision metrics for diabetes detection. Furthermore, this thesis also delves into the importance of feature selection to enhance the predictive performance of the models.

The primary findings of this study highlight how data analytics can transform healthcare, especially in managing chronic diseases. The machine learning models that were created showed good levels of accuracy, suggesting that data-driven procedures can greatly enhance conventional diagnostic techniques. In addition to supporting current initiatives to prevent diabetes by early identification, this work sheds light on the wider health implications of data analytics and offers directions for future investigation into the use of technology to enhance medical outcomes.

KEYWORDS: Data analytics; Machine learning; Health; Healthcare; Diabetes.

JEL CODES: C38; C45; C52; I10.

TABLE OF CONTENTS

TABLE OF FIGURES

ACKNOWLEDGMENTS

First, I wish to thank Professor Mário Caldeira for his guidance.

I am also grateful to my great friends, Carolina, Inês and Mafalda – they gave me the encouragement and strength necessary for the completion of this work.

I am also thankful to my family for their support from the first day of school until the last step of my student's life.

Finally, I am also thankful to my colleague and friend Catarina for the great insights and discussion through all the stages of this thesis.

DATA ANALYTICS IN HEALTHCARE

By José Diogo Bertão

1. INTRODUCTION

In an era marked by rapid advancements in technology, the domain of healthcare has seen transformative changes, particularly through the lens of data analytics. This thesis ventures into this dynamic landscape with a focus on leveraging machine learning (ML) algorithms for the early diagnosis of diabetes, a condition that has emerged as a global health epidemic.

At the core of this research is a meticulous methodology that employs a variety of ML models, including Logistic Regression, Decision Trees, Random Forest, Extreme Gradient Boosting, Naive Bayes, K-Nearest Neighbors, and Neural Networks. Through a detailed analysis of algorithms this thesis compares the models used as well as the impact of using a balanced dataset over an imbalanced dataset, with the objective of not just enhancing the accuracy of diabetes predictions but also exploring the nuances of each algorithm's performance in healthcare settings. The choice of Python as the programming language for implementation underscores the accessibility and robustness of tools available for data analytics in health research. The datasets were retrieved from Kaggle [1].

The significance of this thesis lies in its potential to contribute to the early detection of diabetes, a crucial factor in managing the disease and mitigating its long-term impacts. By examining the predictive power of various ML models, this study aims to identify the most effective techniques for diagnosing diabetes at its nascent stages, thereby opening doors to timely intervention and treatment. Furthermore, the research delves into the importance of feature selection and model optimization, highlighting the intricate balance between model complexity and practical utility in healthcare applications.

This exploration is framed within the broader context of the challenges and opportunities presented by data analytics in healthcare. It acknowledges the pressing need for innovative approaches to disease diagnosis and management in the face of rising healthcare demands and the increasing prevalence of chronic conditions like diabetes. Through its findings, the thesis aims to contribute to the dialogue on how data analytics

1) https://www.kaggle.com/datasets/julnazz/diabetes-health-indicators-dataset?select=diabetes_binary_health_indicators_BRFSS2021.csv

can serve as a catalyst for healthcare transformation, offering insights into the development of more personalized, efficient, and accessible medical care.

In conclusion, this thesis not only showcases the application of machine learning in tackling one of the most pressing health challenges of our time but also sets the foundation for future research in the field. It is a testament to the potential of data analytics to redefine healthcare paradigms, offering a glimpse into a future where technology and healthcare converge to enhance patient outcomes and quality of life. As it navigates through the intricacies of data analytics in health, this thesis contributes a vital discourse on the role of technology in advancing healthcare, proposing a future where data-driven insights inform a more proactive, personalized, and preventive approach to health management.

To ensure the clarity and accuracy of my writing, I use Grammarly for English language corrections throughout this thesis.

2. LITERATURE REVIEW

The transformative potential of data analytics in healthcare is widely acknowledged, offering unprecedented opportunities for enhancing patient care, optimizing treatment protocols, and managing resources efficiently. The application of machine learning models, particularly in the diagnosis and management of chronic diseases like diabetes, represents a significant area of research and development within this domain. This literature review synthesizes findings from recent studies, focusing on the use of reinforcement learning, big data analytics, predictive analytics, and data-driven decision-making in health, with a particular emphasis on diabetes management.

2.1. Reinforcement Learning and Treatment Policy Optimization

The incorporation of reinforcement learning (RL) into healthcare, particularly in critical care settings, represents a significant advancement in developing personalized treatment policies. This approach is particularly effective in addressing the challenges posed by limited data availability for underrepresented patient populations. By utilizing novel methodologies that leverage variational inference, such as Noisy Bayesian Policy Updates, RL can select high-performing treatment policies and accurately predict their performance for patients with non-typical clinical characteristics. This showcases the

potential of data-driven personalization in critical care, where tailored treatment strategies can significantly impact patient outcomes (Baucum et al., 2022).

2.2. Big Data Analytics in Healthcare Transformation

Big data analytics is playing a transformative role in healthcare, illustrating the profound impact of integrating analytics capabilities with IT-enabled transformation practices. This synergy enhances organizational practices and patient care, driving improvements across healthcare systems. The strategic importance of big data analytics lies in its ability to process vast amounts of health data, enabling systematic improvements from operational efficiency to personalized patient interventions and care (Wang et al., 2018).

2.3. Predictive Analytics for Resource Optimization

Predictive analytics is crucial for optimizing hospital resources and improving patient care, especially in managing chronic diseases. By accurately predicting the length of stay for patients, hospitals can manage their resources more efficiently, ensuring that patient care is both effective and sustainable. This study also highlights the importance of choosing the correct features and the importance of historical data for models' performance. (Zolbanin et al., 2022)

2.4. Machine Learning for Diabetes Classification

An emerging method for processing and analyzing vast datasets, specifically for diabetes classification, involves the utilization of advanced neural network architectures. These networks, particularly capsule networks, are adept at recognizing complex patterns and spatial hierarchies within medical data. To efficiently manage the computational demands of large datasets, the MapReduce programming model is employed. This model distributes data processing tasks across multiple computing nodes, thereby enhancing scalability and performance. This innovative approach signifies a step forward in utilizing big data analytics for improved diagnostic accuracy and management in the medical field, especially for chronic conditions like diabetes. (Arun & Marimuthu, 2024).

2.5. Data-Driven Chronic Disease Management

Enhancing the knowledge of healthcare professionals and caregiving staff through data analytics significantly improves chronic disease management. Data-driven insights

into patient management strategies and care delivery can lead to better healthcare outcomes, demonstrating the value of analytics in supporting healthcare professionals in delivering personalized and effective care (Liu & Kauffman, 2021).

2.6. Open Health Data for Medication Management

The use of open health data to inform medication management practices offers insights into prescribing trends and supports better decision-making in healthcare. While the focus has been on specific areas such as antidepressant prescribing, the principles can be applied more broadly to chronic disease management, highlighting the potential of open health data in improving healthcare practices and outcomes (Cleland et al., 2018).

2.7. Integrating Technology and Methodology in Big Data Analytics

The integration of technology and methodological approaches in health big data analytics addresses the challenges of processing and interpreting vast amounts of health data. This integration supports decision-making in chronic disease management and emphasizes the need for innovative analytical models that can navigate the complexities of health data, providing actionable insights for healthcare providers (Gonzalez-Alonso et al., 2017).

2.8. Sequential Decision-Making and Personalized Treatment

The optimization of decision-making processes and the design of personalized treatment plans based on data analytics are critical for managing chronic diseases effectively. Tailored treatment strategies, developed through data-driven insights and sequential decision-making models, optimize patient care and health outcomes, showcasing the effectiveness of analytics in creating personalized healthcare solutions (Denton, 2018).

2.9. Enhancing Diabetes Management through Outcome-Driven Personalized Treatment Plans

The integration of data analytics into health practices, particularly for diabetes management, has led to significant advancements in personalized treatment strategies. A notable approach detailed in recent research involves a novel model that utilizes mathematical modeling and data analytics to tailor personalized treatment plans for managing diabetes, including gestational and type 2 diabetes. This model innovatively

predicts the relationship between drug dosage and its impact on blood glucose levels through non-invasive measures, using fluid dynamics, optimization techniques, and statistical analyses. By incorporating clinical constraints and personalized dose-effect knowledge into a multi-objective optimization framework, the model enables the formulation of optimized treatment plans. These plans not only achieve better glycemic control with reduced medication but also significantly lower the treatment costs, demonstrating the potential of data analytics to enhance patient care and healthcare delivery in diabetes management (Lee et al., 2018).

2.10.Strategic Innovations in Data-Driven Preventive Healthcare

In the realm of healthcare, the proactive allocation of preventive treatments stands as a pivotal strategy for managing chronic conditions such as Type II Diabetes Mellitus, offering a blueprint for significantly enhancing patient outcomes while concurrently optimizing healthcare spending. A novel decision model capitalizes on the synergy of counterfactual inference, machine learning, and optimization techniques to allocate preventive care resources judiciously. This model, evaluated using a substantial dataset comprising 89,191 prediabetic patients, underscores the utility of high-dimensional health data to inform and optimize preventive treatment decisions. By integrating a dynamic allocation framework that maximizes the expected number of prevented disease onsets within budgetary constraints, the model not only demonstrates a remarkable ability to improve disease prevention rates but also to achieve substantial cost savings. Its comparative analysis with traditional practices reveals a substantial performance leap, advocating for a transition towards risk reduction-focused allocations rather than solely risk-based strategies. This shift underscores the critical role of rigorous, data-driven decision-making frameworks in healthcare, promoting an efficient, targeted approach to preventive care that promises to reshape the management of diabetes mellitus and potentially other preventable diseases, fostering a future where healthcare resources are utilized in the most impactful manner (Kraus et al., 2023).

2.11. Conclusion

The literature reviewed underscores the critical importance of data analytics in revolutionizing health care, with a particular focus on diabetes management. The findings from these studies provide a robust foundation for the thesis' exploration of ML models

trained on Kaggle diabetes datasets. The nuanced understanding of the impact of dataset characteristics on model performance, particularly the advantages of balanced datasets in improving diabetic case identification, aligns with broader healthcare objectives. This literature review not only contextualizes the thesis within the current research landscape but also highlights the contribution of the thesis to advancing data analytics applications in health, especially in optimizing diabetes prediction and management.

3. METHODOLOGY

The main objective of the chosen datasets was to create and compare different models to predict diabetes in early stages and to compare different types of datasets. It was used two types of datasets, one with a balanced number of positive and negative cases and another with an imbalanced number of positive and negative cases, prevailing the number of negative cases. In both datasets the target variable was binary, "0" for no diabetes and "1" for prediabetes or diabetes. The balanced dataset had 67 136 cases and the imbalanced dataset had 236 378 cases. The best models for this type of prediction are models like Logistic Regression, Decision Tree, Random Forest, Extreme Gradient Boosting, Gradient Boosting, Naive Bayes, K-Nearest-Neighbors (KNN), and Neural Networks (Multi-layer Perceptron). The execution of these models in Python was possible with libraries like *pandas* and *numpy* for dataset manipulation and *sklearn*, *xgboost*, and *imblearn* for model implementation. The proposed methodology taken is represented in Figure 1.

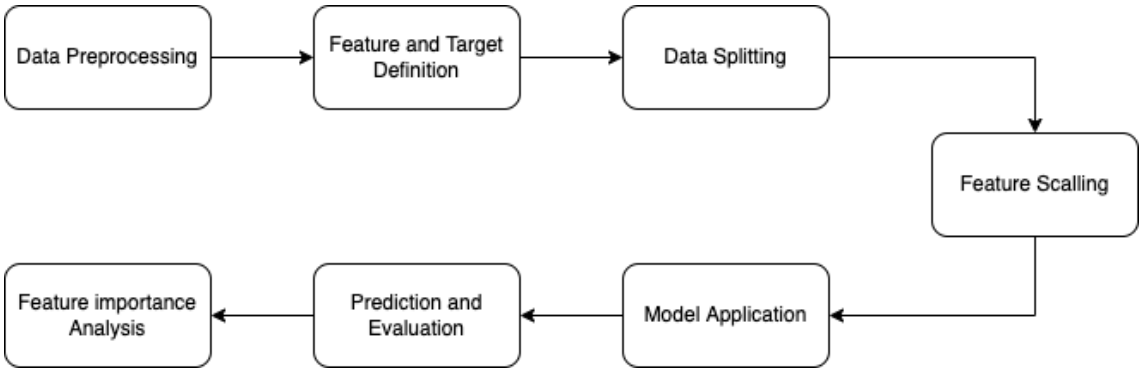

Figure 1-Proposed Methodology.

3.1. Data Preprocessing

Data preprocessing is the crucial first step in the ML pipeline, where raw data is cleaned and prepared for analysis. This stage involves handling missing values, identifying and removing outliers, and correcting inconsistencies in the data. The importance of data preprocessing lies in its capacity to enhance the quality of the data, ensuring that subsequent models are trained on accurate and representative information. This step directly impacts the reliability of the predictive models, as high-quality data is fundamental to generating meaningful and actionable insights (Ramadhan et al., 2021).

As the dataset was already clean, the only task needed was to remove duplicates to ensure that the dataset was ready for further analysis without data repetition or redundancy.

3.2. Feature and Target Definition

Defining the features and the target variable is a critical step that involves identifying which variables will be used as inputs (features, represented by X) for the models and which variable will be predicted (target, represented by Y). This is an essential step for supervised learning models like diabetes prediction (Ahsan et al., 2021).

For this case the features consisted in high blood pressure (HighBP), general health (GenHlth, on a scale 1-5 where 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor), cholesterol check in last 5 years (CholCheck, “0” for no, “1” for yes), high cholesterol (HighChol, “0” for no, “1” for yes), heavy alcohol consumption (HvyAlcoholConsump, “0” is adult men having less or 14 drinks per week and adult women having less or 7 drinks per week, “1” is adult men having more than14 drinks per week and adult women having more than 7 drinks per week), age (Age, on a scale 1 to 13 with values starting in 18 years old to 99 years old divided into intervals of 4 years, expect the first and last, the first being from 18 to 24 years old and the last being from 80 to 99 years old), heart disease or attack (HeartDiseaseorAttack, “0” is no coronary heart disease or myocardial infarction, “1” for coronary heart disease or myocardial infarction), difficulty in walking or climbing stairs (DiffWalk, “0” for no, “1” for yes), Body Mass Index (BMI), physical activity in past 30 days (PhysActivity, “0” for no, “1” for yes), sex (Sex, “0” for female, “1” for male), if ever had a stroke (Stroke, “0” for no, “1” for yes), annual family income

(Income, from a scale from 1 to 11, “1” being less than $10 000 and “11” being more than $200 000), smoked at least 100 cigarettes in the individuals life (Smoker, “0” for no, “1” for yes), number of days many days during the past 30 days was the individuals mental health not good, including stress, depression and problems with emotions (MentHlth), if there was a time in the past year when the individual needed a doctor but couldn’t because of cost (NoDocbcCost, “0” for no, “1” for yes), education level (Education, on a scale from 1 to 6, where “1” is never attended school or only kindergarten, “2” is elementary, “3” is some high school, “4” is high school graduate, “5” is some college or technical school and “6” is college graduate), 1 or more vegetables consumed per day (Veggies, “0” for no, “1” for yes), if the individual has any kind of health care coverage (AnyHealthcare, “0” for no, “1” for yes), the number of days during the past 30 days the physical health was not good, including illness and injury (PhysHlth) and 1 or more fruits consumed per day (Fruits, “0” for no, “1” for yes).

3.3. Data Splitting

Splitting the data into training and testing sets is an essential methodological step to evaluate the performance of ML models objectively. The training set is used to train the models, while the testing set, which consists of data not seen by the models during training, is used to assess their predictive performance. This separation is vital for preventing overfitting, where a model might perform well on the training data but poorly on new, unseen data. By using a separate testing set, the study ensures that the model's performance metrics accurately reflect its ability to generalize to new cases.

In this case, the data splitting consisted in random splitting the data with 70% to training and the remaining 30% to testing (Kaveripakam et al., 2024).

3.4. Feature Scaling

Feature scaling is a technique used to normalize the range of features in the data. Many ML algorithms, particularly those that rely on distance calculations like KNN, require data to be scaled to perform optimally. Scaling ensures that all features contribute equally to the model's predictions, preventing variables with larger scales from dominating those with smaller scales. This step is crucial for maintaining the balance and fairness of the model's consideration of input features, directly impacting the accuracy and fairness of the predictions (Ahsan et al., 2021).

For this case, it was used the functions from the *sklearn* library *fit_transform* for the features training dataframe and the function *transform* for the features testing dataframe.

3.5. Model Application

Applying different ML models to the prepared data is a core aspect of exploring various approaches to diabetes prediction. This step involves selecting, configuring, and training multiple models, each with its strengths and limitations. The diversity of models applied allows the study to compare performance across different algorithms, identifying which models are most effective for diabetes prediction given the specific characteristics of the dataset. This comparative analysis is fundamental to selecting the most appropriate model for deployment in a real-world healthcare setting.

The ML models applied include Logistic Regression, Decision Tree, Random Forest, KNN, Naive Bayes, Multi-layer Perceptron (Neural Network), Gradient Boosting, and XGBoost.

3.6. Prediction and Evaluation

Once the models are trained, making predictions on the testing set and evaluating their performance are crucial for understanding the models' effectiveness. This step involves using a variety of metrics, such as accuracy, precision, recall (or True Positive Rate, TPR), F1-score, and the Area Under the Receiver Operating Characteristic curve (AUC), to assess each model's performance. These metrics provide a comprehensive overview of the models' strengths and weaknesses, guiding the selection of the most suitable model for predicting diabetes. The evaluation process is critical for ensuring that the chosen model meets the desired standards of reliability and accuracy for clinical applications (Kaveripakam et al., 2024).

3.7. Feature Importance Analysis

Analyzing feature importance is the process of identifying which features have the most significant impact on the model's predictions. This analysis provides insights into the underlying relationships between the features and the target variable, offering a deeper understanding of the factors that contribute to diabetes. Feature importance analysis is invaluable for interpreting the model's predictions, informing clinical decision-making, and guiding future data collection and research. It highlights the variables that should be

prioritized in preventive healthcare strategies and patient education to mitigate the risk of diabetes (Amin et al., 2019).

This analysis was conducted for models that provide insights into the importance of features (Logistic Regression, Decision Tree, Random Forest, and XGBoost).

3.8. Conclusion

Together, these methodological steps form a comprehensive framework for developing, evaluating, and interpreting machine learning models for diabetes prediction. This structured approach ensures the study's findings are robust, reliable, and relevant to healthcare professionals seeking to leverage machine learning to improve diabetes diagnosis and management.

4. RESULTS

In evaluating the performance of various machine learning models for diabetes prediction, the analysis spans across two distinct datasets: the original, imbalanced dataset and a balanced dataset. This comparison illuminates the profound impact dataset composition has on the predictive accuracy, fairness, and overall utility of each model within a healthcare context. The detailed results can be found in the *Appendices* section, from figures 2 to 25.

4.1. Impact of Dataset Composition

The imbalanced dataset presents a common challenge in medical diagnostics: the prevalence of negative (non-diabetic) instances outweighs positive (diabetic) cases. Models trained on this dataset, including Logistic Regression, Decision Trees, Random Forest, KNN, Naive Bayes, Multi-layer Perceptron, Gradient Boosting, and XGBoost, generally exhibited high overall accuracy (around 83%). This high accuracy, however, often masked deficiencies in predicting the less represented diabetic class (diabetic). Specifically, the precision and recall for diabetic predictions were notably lower than those for non-diabetic predictions across most models. For instance, while Logistic Regression achieved commendable accuracy, its ability to correctly identify diabetic instances was limited, reflecting a systemic bias towards the majority class inherent in the dataset. This phenomenon underscores a critical challenge in using imbalanced datasets

for training predictive models in healthcare: the risk of overlooking the very outcomes most critical to detect.

Conversely, training on the balanced dataset markedly improved the recall and precision for diabetic cases across all models, indicating a heightened sensitivity to identifying diabetes. This improvement was not without trade-offs, while the sensitivity to diabetic cases increased, the overall accuracy of the models slightly decreased in some cases (around 72%). This decrease in accuracy reflects the more challenging nature of prediction when both classes are equally represented, requiring the model to discern more subtle patterns distinguishing between diabetic and non-diabetic instances without relying on class prevalence as a heuristic.

4.2. Model-Specific Observations

The nuanced performance of specific models further reveals the complex interplay between algorithm characteristics and dataset composition. For example, Naive Bayes, known for its simplicity and probabilistic approach, showed an intriguing trade-off on the imbalanced dataset, with a relatively high recall for diabetic cases (57%) but at the expense of lower overall accuracy (76%). This suggests that Naive Bayes, while prone to higher false positive rates, has an inherent capacity to detect diabetic instances more effectively than some more complex models. This capacity didn't have the same impact in the balanced dataset scenario, as the other models showed improvement in recall and precision for diabetic cases.

Models like Gradient Boosting and XGBoost, which leverage ensemble methods to iteratively correct errors, also demonstrated notable adaptability to the balanced dataset. These models, already proficient in handling imbalances through their inherent mechanisms, exhibited significant gains in both precision and recall for diabetic cases when trained on the balanced dataset. This improvement underscores the effectiveness of ensemble and boosting techniques in mitigating the adverse effects of class imbalance on model sensitivity and specificity.

4.3. Feature Importance Analysis

The feature importance analysis conducted as part of this thesis offers critical insights into the underlying factors that contribute most significantly to the prediction of diabetes across the evaluated machine learning models. This analysis not only illuminates the

relative importance of various predictors but also underscores the models' interpretability, an essential aspect of applying machine learning in healthcare. For instance, features such as general health, age, BMI, and high blood pressure emerged as top predictors across several models, reflecting well-established clinical understandings of diabetes risk factors. The prominence of these features aligns with epidemiological evidence linking lifestyle, physiological, and demographic factors with diabetes prevalence. Notably, the analysis revealed differences in feature importance rankings between models, illustrating the unique ways in which each algorithm processes and prioritizes information. Such insights are invaluable for healthcare professionals, as they provide a data-driven basis for targeted interventions and patient education. Moreover, understanding which features significantly influence diabetes predictions enhances the transparency of ML applications in clinical settings, fostering trust and facilitating the integration of these technologies into patient care. Ultimately, the feature importance analysis not only contributes to the predictive performance of the models but also enriches our understanding of diabetes, offering a bridge between machine learning innovation and clinical practice.

4.4. Conclusion

The transition from the imbalanced to the balanced dataset underscores a pivotal insight: balancing the representation of outcomes in training data is crucial for developing predictive models that are not only accurate but also fair and clinically useful. While the slight reduction in overall accuracy on the balanced dataset may initially seem counterintuitive, the substantial improvement in correctly identifying diabetic cases (as evidenced by increased Recall) represents a meaningful advancement in model utility for healthcare applications. This improvement aligns with the primary goal of medical diagnostics: to accurately identify conditions for timely intervention. The analysis emphasizes the importance of considering dataset composition in the model development process, advocating for a balanced approach that prioritizes equitable sensitivity across outcomes.

This comparative analysis of model performances across two datasets highlights the critical role of dataset composition in predictive modeling for healthcare. It illustrates the necessity of a nuanced approach to model selection and evaluation, one that balances

overall accuracy with the ethical and clinical imperatives of sensitivity and fairness in patient care.

5. CONCLUSION

This thesis has embarked on a comprehensive journey through the landscape of data analytics within the healthcare sector, highlighting its transformative potential from theoretical underpinnings to practical applications, particularly in the early detection of diseases such as diabetes. Through the meticulous analysis of various machine learning algorithms applied to health data, it has uncovered the profound impact these technologies can have on enhancing diagnostic accuracy, optimizing treatment protocols, and ultimately improving patient outcomes.

The exploration has not only showcased the capabilities of algorithms like Logistic Regression, Decision Trees, and Neural Networks in processing and analyzing complex datasets but has also underscored the critical importance of feature selection and model optimization and analysis that accompany the deployment of these technologies in a healthcare context. The findings underscore a pivotal shift towards a more data-driven approach in healthcare, promising to usher in an era of precision medicine characterized by more personalized, predictive, and preventive care strategies.

The journey through this thesis has underscored the significant challenges and opportunities that lie ahead in integrating data analytics into healthcare. It has navigated through the complexities of model selection, the nuances of data preprocessing, and the intricacies of evaluation metrics, emerging with a deeper understanding of how data-driven approaches can improve diagnostic processes and patient outcomes. This work contributes to a growing body of evidence that supports the adoption of ML in healthcare, advocating for a future where data analytics serves as a cornerstone of disease prevention and management.

As we look to the future, the role of data analytics in health is poised for exponential growth, driven by advancements in technology, the increasing availability of health data, and the continuous push towards integrated care models. Collaboration among data scientists, healthcare professionals, and patients is essential to develop robust, ethical, and

sustainable analytics solutions that can adapt to the evolving healthcare needs of our global population.

The integration of data analytics into healthcare represents a beacon of hope for the future of medical diagnostics and treatment. With careful consideration of the ethical, social, and technical challenges, data analytics has the potential to significantly enhance healthcare delivery and patient care. It is incumbent upon us to navigate this complex landscape with foresight, diligence, and a commitment to equity, ensuring that the benefits of data analytics in health are realized for all members of society.

REFERENCES

Arun, G., & Marimuthu, C. N. (2024). Diabetes classification using MapReduce-based capsule network. *Automatika*, *65*(1), 73–81. https://doi.org/10.1080/00051144.2023.2284031

Denton, B. T. (2018). *Optimization of Sequential Decision Making for Chronic Diseases: From Data to Decisions. In INFORMS TutORials in Operations Research* (Issue October). https://doi.org/10.1287/educ.2018.0184

Baucum, M., Khojandi, A., Vasudevan, R., & Davis, R. (2022). Adapting Reinforcement Learning Treatment Policies Using Limited Data to Personalize Critical Care. *INFORMS Journal on Data Science*, *1*(1), 27–49. https://doi.org/10.1287/ijds.2022.0015

Wang, Y., Kung, L. A., Wang, W. Y. C., & Cegielski, C. G. (2018). An integrated big data analytics-enabled transformation model: Application to health care. *Information and Management*, *55*(1), 64–79. https://doi.org/10.1016/j.im.2017.04.001

Liu, N., & Kauffman, R. J. (2021). Enhancing healthcare professional and caregiving staff informedness with data analytics for chronic disease management. *Information and Management*, *58*(2), 103315. https://doi.org/10.1016/j.im.2020.103315

Zolbanin, H. M., Davazdahemami, B., Delen, D., & Zadeh, A. H. (2022). Data analytics for the sustainable use of resources in hospitals: Predicting the length of stay for patients with chronic diseases. *Information and Management*, *59*(5), 103282. https://doi.org/10.1016/j.im.2020.103282

Cleland, B., Wallace, J., Bond, R., Black, M., Mulvenna, M., Rankin, D., & Tanney, A. (2018). Insights into Antidepressant Prescribing Using Open Health Data. *Big Data Research*, *12*, 41–48. https://doi.org/10.1016/j.bdr.2018.02.002

Gonzalez-Alonso, P., Vilar, R., & Lupianez-Villanueva, F. (2017). Meeting Technology and Methodology into Health Big Data Analytics Scenarios. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, *2017-June*, 284–285. https://doi.org/10.1109/CBMS.2017.71

Lee, E. K., Wei, X., Baker-Witt, F., Wright, M. D., & Quarshie, A. (2018). Outcome-driven personalized treatment design for managing diabetes. *Interfaces*, *48*(5), 422–435. https://doi.org/10.1287/inte.2018.0964

Kraus, M., Feuerriegel, S., & Saar-Tsechansky, M. (2023). Data-Driven Allocation of Preventive Care with Application to Diabetes Mellitus Type II. *Manufacturing & Service Operations Management*, *February*. https://doi.org/10.1287/msom.2021.0251

Kaveripakam, D., & Ravichandran, J. (2024). Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction. *International Journal of Scientific and Research Publications (IJSRP)*, *11*(1), 339–346. https://doi.org/10.29322/ijsrp.11.01.2021.p10936

Ramadhan, N. G., Adiwijaya, & Romadhony, A. (2021). Preprocessing Handling to Enhance Detection of Type 2 Diabetes Mellitus based on Random Forest. *International Journal of Advanced Computer Science and Applications*, *12*(7), 223–228. https://doi.org/10.14569/IJACSA.2021.0120726

Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies*, *9*(3), 5–9. https://doi.org/10.3390/technologies9030052

Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*, *36*(August 2018), 82–93. https://doi.org/10.1016/j.tele.2018.11.007

APPENDICES

Logistic Regression

Imbalanced Dataset

Confusion Matrix

		Predicted values	
		0	1
Actual values	0	55 725	1 263
	1	8 526	1 459

Classification Report

	Precision	Recall	F1-score	Support
0	0.87	0.98	0.92	56 988
1	0.54	0.15	0.23	9 985
Accuracy			0.85	66 973
Macro avg	0.71	0.57	0.58	66 973
Weighted avg	0.82	0.85	0.82	66 973

AUC	0.806

Figure 2-Logistic Regression Imbalanced Dataset key metrics.

Logistic Regression

Balanced Dataset

Confusion Matrix

		Predicted values	
		0	1
Actual values	0	7 050	2 788
	1	2 423	7 659

Classification Report

	Precision	Recall	F1-score	Support
0	0.74	0.72	0.73	9 838
1	0.73	0.76	0.75	10 082
Accuracy			0.74	19 920
Macro avg	0.735	0.74	0.74	19 920
Weighted avg	0.74	0.74	0.74	19 920

AUC	0.814

Figure 3-Logistic Regression Balanced Dataset key metrics.

Decision Tree

Imbalanced Dataset

Confusion Matrix

		Predicted values	
		0	1
Actual values	0	48 742	8 246
	1	6 815	3 170

Classification Report

	Precision	Recall	F1-score	Support
0	0.88	0.86	0.87	56 988
1	0.28	0.32	0.3	9 985
Accuracy			0.78	66 973
Macro avg	0.58	0.59	0.59	66 973
Weighted avg	0.79	0.78	0.78	66 973

AUC	0.587

Figure 4-Decision Tree Imbalanced Dataset key metrics.

Decision Tree

Balanced Dataset

Confusion Matrix

		Predicted values	
		0	1
Actual values	0	6 277	3 561
	1	3 621	6 461

Classification Report

	Precision	Recall	F1-score	Support
0	0.63	0.64	0.64	9 838
1	0.64	0.64	0.63	10 082
Accuracy			0.64	19 920
Macro avg	0.64	0.64	0.64	19 920
Weighted avg	0.64	0.64	0.64	19 920

AUC	0.640

Figure 5-Decision Tree Balanced Dataset key metrics.

Random Forest

Imbalanced Dataset

Confusion Matrix

		Predicted values	
		0	1
Actual values	0	55 205	1 783
	1	8 378	1 607

Classification Report

	Precision	Recall	F1-score	Support
0	0.87	0.97	0.92	56 988
1	0.47	0.16	0.24	9 985
Accuracy			0.85	66 973
Macro avg	0.67	0.57	0.58	66 973
Weighted avg	0.81	0.85	0.82	66 973

AUC	0.776

Figure 6-Random Forest Imbalanced Dataset key metrics.

Random Forest

Balanced Dataset

Confusion Matrix

		Predicted values	
		0	1
Actual values	0	6 818	3 020
	1	2 406	7 676

Classification Report

	Precision	Recall	F1-score	Support
0	0.74	0.69	0.72	9 838
1	0.72	0.76	0.74	10 082
Accuracy			0.73	19 920
Macro avg	0.73	0.73	0.73	19 920
Weighted avg	0.73	0.73	0.73	19 920

AUC	0.796

Figure 7-Random Forest Balanced Dataset key metrics.

K-Nearest Neighbors

Imbalanced Dataset

Confusion Matrix

		Predicted values	
		0	1
Actual values	0	53 949	3 039
	1	7 972	2 013

Classification Report

	Precision	Recall	F1-score	Support
0	0.87	0.95	0.91	56 988
1	0.4	0.2	0.27	9 985
Accuracy			0.84	66 973
Macro avg	0.64	0.58	0.59	66 973
Weighted avg	0.8	0.84	0.81	66 973

AUC	0.705

Figure 8-KNN Imbalanced Dataset key metrics.

K-Nearest Neighbors

Balanced Dataset

Confusion Matrix

		Predicted values	
		0	1
Actual values	0	6 636	3 202
	1	2 714	7 368

Classification Report

	Precision	Recall	F1-score	Support
0	0.71	0.67	0.69	9 838
1	0.7	0.73	0.71	10 082
Accuracy			0.7	19 920
Macro avg	0.71	0.70	0.70	19 920
Weighted avg	0.7	0.7	0.7	19 920

AUC	0.757

Figure 9-KNN Balanced Dataset key metrics.

Naive Bayes

Imbalanced Dataset

Confusion Matrix

		Predicted values	
		0	1
Actual values	0	45 223	11 765
	1	4 263	5 722

Classification Report

	Precision	Recall	F1-score	Support
0	0.91	0.79	0.85	56 988
1	0.33	0.57	0.42	9 985
Accuracy			0.76	66 973
Macro avg	0.62	0.68	0.64	66 973
Weighted avg	0.83	0.76	0.78	66 973

AUC	0.770

Figure 10-Naive Bayes Imbalanced Dataset key metrics.

Naive Bayes

Balanced Dataset

Confusion Matrix

		Predicted values	
		0	1
Actual values	0	7 054	2 784
	1	2 833	7 249

Classification Report

	Precision	Recall	F1-score	Support
0	0.71	0.72	0.72	9 838
1	0.72	0.72	0.72	10 082
Accuracy			0.72	19 920
Macro avg	0.72	0.72	0.72	19 920
Weighted avg	0.72	0.72	0.72	19 920

AUC	0.779

Figure 11-Naive Bayes Balanced Dataset key metrics.

Neural Network, Multi-layer Perceptron

Imbalanced Dataset

Confusion Matrix

		Predicted values	
		0	1
Actual values	0	55 786	1 202
	1	8 633	1 352

Classification Report

	Precision	Recall	F1-score	Support
0	0.87	0.98	0.92	56 988
1	0.53	0.14	0.22	9 985
Accuracy			0.85	66 973
Macro avg	0.70	0.56	0.57	66 973
Weighted avg	0.82	0.85	0.81	66 973

AUC	0.804

Figure 12-Multi-Layer Perceptron Imbalanced Dataset key metrics.

Neural Network, Multi-layer Perceptron

Balanced Dataset

Confusion Matrix

		Predicted values	
		0	1
Actual values	0	7 010	2 828
	1	2 534	7 548

Classification Report

	Precision	Recall	F1-score	Support
0	0.73	0.71	0.72	9 838
1	0.73	0.75	0.74	10 082
Accuracy			0.73	19 920
Macro avg	0.73	0.73	0.73	19 920
Weighted avg	0.73	0.73	0.73	19 920

AUC	0.804

Figure 13-Multi-Layer Perceptron Balanced Dataset key metrics.

Gradient Boosting

Imbalanced Dataset

Confusion Matrix

		Predicted values	
		0	1
Actual values	0	55 802	1 186
	1	8 453	1 532

Classification Report

	Precision	Recall	F1-score	Support
0	0.87	0.98	0.92	56 988
1	0.56	0.15	0.24	9 985
Accuracy			0.86	66 973
Macro avg	0.72	0.57	0.58	66 973
Weighted avg	0.82	0.86	0.82	66 973

AUC	0.811

Figure 14-Gradient Boosting Imbalanced Dataset key metrics.

Gradient Boosting

Balanced Dataset

Confusion Matrix

		Predicted values	
		0	1
Actual values	0	6 890	2 948
	1	2 176	7 906

Classification Report

	Precision	Recall	F1-score	Support
0	0.76	0.7	0.73	9 838
1	0.73	0.78	0.76	10 082
Accuracy			0.74	19 920
Macro avg	0.75	0.74	0.75	19 920
Weighted avg	0.74	0.74	0.74	19 920

AUC	0.819

Figure 15-Gradient Boosting Balanced Dataset key metrics.

XGBoost

Imbalanced Dataset

Confusion Matrix

		Predicted values	
		0	1
Actual values	0	55 635	1 353
	1	8 380	1 605

Classification Report

	Precision	Recall	F1-score	Support
0	0.87	0.98	0.92	56 988
1	0.54	0.16	0.25	9 985
Accuracy			0.85	66 973
Macro avg	0.71	0.57	0.59	66 973
Weighted avg	0.82	0.85	0.82	66 973

AUC	0.808

Figure 16-XGBoost Imbalanced Dataset key metrics.

XGBoost

Balanced Dataset

Confusion Matrix

		Predicted values	
		0	1
Actual values	0	6 855	2 983
	1	2 199	7 883

Classification Report

	Precision	Recall	F1-score	Support
0	0.76	0.7	0.73	9 838
1	0.73	0.78	0.75	10 082
Accuracy			0.74	19 920
Macro avg	0.75	0.74	0.74	19 920
Weighted avg	0.74	0.74	0.74	19 920

AUC	0.811

Figure 17-XGBoost Balanced Dataset key metrics.

Logistic Regression

Imbalanced Dataset

Features importance

Feature	Importance
GenHlth	0.48925
Age	0.395103
BMI	0.371649
HighBP	0.359398
CholCheck	0.276575
HighChol	0.27124
HvyAlcoholConsump	0.177272
Sex	0.137639
Income	0.112458
HeartDiseaseorAttack	0.077148
PhysActivity	0.06847
PhysHlth	0.043269
Smoker	0.04297
MentHlth	0.042204
DiffWalk	0.040817
Stroke	0.030118
AnyHealthcare	0.02846
Education	0.027127
Fruits	0.013596
NoDocbcCost	0.005123
Veggies	0.00081

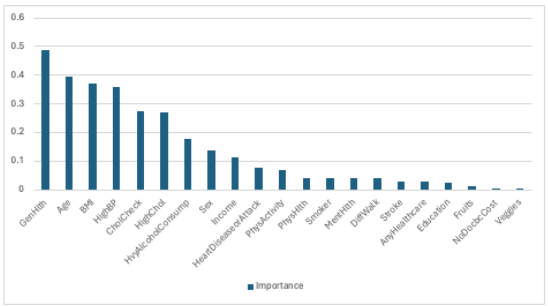

Figure 18-Logistic Regression Imbalanced Dataset features importance.

Logistic Regression

Balanced Dataset

Features importance

Feature	Importance
GenHlth	0.567347
BMI	0.451843
Age	0.416271
HighBP	0.360976
HighChol	0.290979
CholCheck	0.225672
HvyAlcoholConsump	0.155367
Sex	0.15492
Income	0.136705
HeartDiseaseorAttack	0.110781
PhysActivity	0.087341
PhysHlth	0.078539
Stroke	0.063151
MentHlth	0.04497
DiffWalk	0.044087
Education	0.032091
AnyHealthcare	0.019843
Smoker	0.018835
Fruits	0.012108
Veggies	0.010856
NoDocbcCost	0.007382

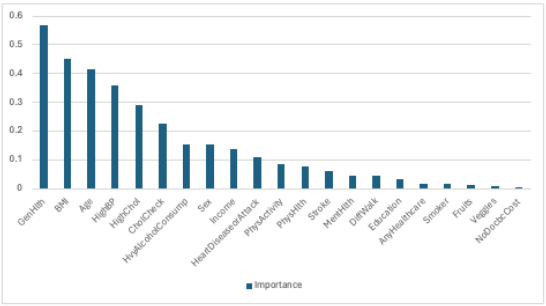

Figure 19-Logistic Regression Balanced Dataset features importance.

Decision Tree

Imbalanced Dataset

Features importance

Feature	Importance
BMI	0.148728
Income	0.12641
Age	0.104969
PhysHlth	0.081863
Education	0.076834
MentHlth	0.073904
HighBP	0.067677
GenHlth	0.060213
Fruits	0.03778
Smoker	0.037388
Sex	0.028844
PhysActivity	0.028256
Veggies	0.027932
DiffWalk	0.019919
HighChol	0.015923
HeartDiseaseorAttack	0.015915
Stroke	0.015080
NoDocbcCost	0.013461
HvyAlcoholConsump	0.009293
AnyHealthcare	0.006112
CholCheck	0.003498

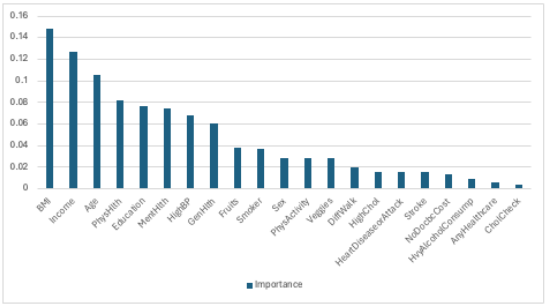

Figure 20-Decision Tree Balanced Dataset features importance.

Decision Tree

Balanced Dataset

Features importance

Feature	Importance
BMI	0.139314
HighBP	0.138337
Age	0.112003
Income	0.101226
GenHlth	0.089008
PhysHlth	0.063692
Education	0.059971
MentHlth	0.057519
Smoker	0.032029
Fruits	0.031903
Sex	0.027347
PhysActivity	0.027049
Veggies	0.024811
HighChol	0.022093
DiffWalk	0.016342
HeartDiseaseorAttack	0.013743
Stroke	0.011430
HvyAlcoholConsump	0.010215
NoDocbcCost	0.009138
CholCheck	0.006539
AnyHealthcare	0.00629

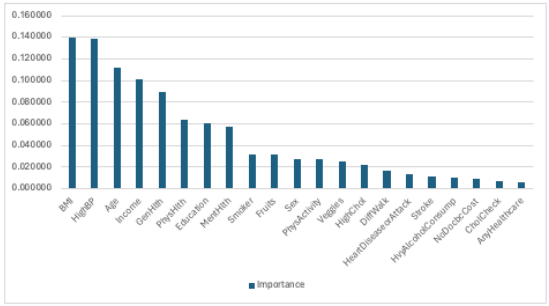

Figure 21-Decision Tree Balanced Dataset features importance.

Random Forest

Imbalanced Dataset

Features importance

Feature	Importance
BMI	0.177541
Age	0.128377
Income	0.121235
PhysHlth	0.077132
MentHlth	0.069239
Education	0.067546
GenHlth	0.066331
HighBP	0.041939
Fruits	0.033923
Smoker	0.033304
Sex	0.027863
HighChol	0.025995
Veggies	0.025766
DiffWalk	0.022084
PhysActivity	0.021538
HeartDiseaseorAttack	0.017711
Stroke	0.012815
NoDocbcCost	0.011729
HvyAlcoholConsump	0.008572
AnyHealthcare	0.005812
CholCheck	0.003548

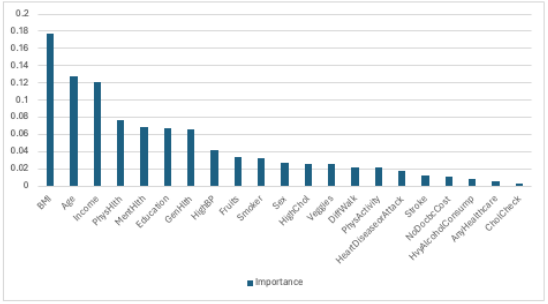

Figure 22-Random Forest Imbalanced Dataset features importance.

Random Forest

Balanced Dataset

Features importance

Feature	Importance
BMI	0.161684
Age	0.137272
Income	0.104282
GenHlth	0.093247
HighBP	0.075221
PhysHlth	0.062688
MentHlth	0.059270
Education	0.056257
HighChol	0.038905
Smoker	0.028874
Fruits	0.028401
Sex	0.025916
DiffWalk	0.024378
PhysActivity	0.022691
Veggies	0.021507
HeartDiseaseorAttack	0.017808
HvyAlcoholConsump	0.010283
Stroke	0.009724
NoDocbcCost	0.009394
CholCheck	0.006432
AnyHealthcare	0.005766

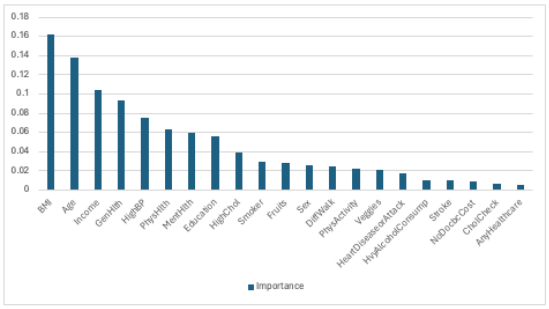

Figure 23-Random Forest Balanced Dataset features importance.

XGBoost

Imbalanced Dataset

Features importance

Feature	Importance
HighBP	0.483359
GenHlth	0.113883
CholCheck	0.056139
HighChol	0.053393
HvyAlcoholConsump	0.048699
Age	0.036693
HeartDiseaseorAttack	0.032489
DiffWalk	0.022609
BMI	0.021763
PhysActivity	0.018817
Sex	0.018092
Stroke	0.011961
Income	0.010898
Smoker	0.010688
MentHlth	0.009118
NoDocbcCost	0.009014
Education	0.008898
Veggies	0.008836
AnyHealthcare	0.008446
PhysHlth	0.008231
Fruits	0.007974

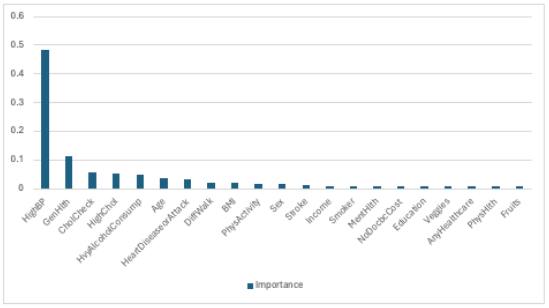

Figure 24-XGBoost Imbalanced Dataset features importance.

XGBoost

Balanced Dataset

Features importance

Feature	Importance
HighBP	0.476762
GenHlth	0.110122
HighChol	0.058382
CholCheck	0.053199
HvyAlcoholConsump	0.03838
Age	0.034832
HeartDiseaseorAttack	0.032246
DiffWalk	0.022655
BMI	0.021991
Sex	0.021149
PhysActivity	0.017538
Stroke	0.015408
Income	0.013164
NoDocbcCost	0.01136
MentHlth	0.010875
Education	0.010836
PhysHlth	0.010687
Veggies	0.010403
Smoker	0.010122
AnyHealthcare	0.010025
Fruits	0.009865

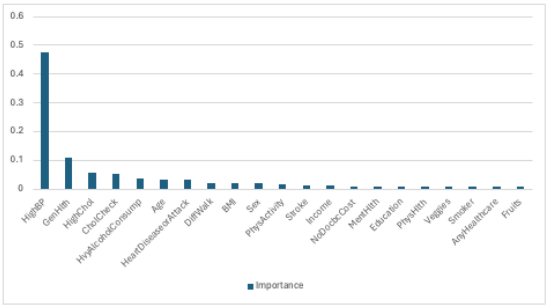

Figure 25-XGBoost Balanced Dataset features importance.

```python
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.neural_network import MLPClassifier
from sklearn.ensemble import GradientBoostingClassifier
from xgboost import XGBClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from sklearn.feature_selection import RFE
from sklearn.metrics import roc_auc_score
```

Python

Importing the dataset for the analysis

```python
#importing the dataset
df = pd.read_csv('/Users/diogobertao/Desktop/Data Analytics for Business/MFW/Dataset/diabetes_binary_5050split_health_indicators_BRFSS2021.csv')
```

Python

```python
#checking for duplicates
df.duplicated().sum()
```

Python

Figure 26-Code implementation 1

```python
#removing duplicates
df.drop_duplicates(inplace=True)
```

Python

```python
#checking for duplicates
df.duplicated().sum()
```

Python

Defining the features vector and the and the target variable

```python
#setting the X (features) and Y (target) dataframes
X = df.drop('Diabetes_binary', axis=1)
Y = df['Diabetes_binary']
```

Python

Spliting the data into two categories, train data and test data

```python
#counting the number of positives and negatives of the dataframe
Y.value_counts()
```

Python

Figure 27-Code implementation 2

```python
#spliting the data into training data (70%) and testing data (30%)
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3, random_state=42)
```

Python

Scaling the features data to bring all features to a similar scaling

+ Code + Markdown

```python
#standardizing the data
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
#ros = RandomOverSampler(random_state=42)
#X_train, Y_train = ros.fit_resample(X_train, Y_train)
```

Python

Applying differnt models to our training data

```python
#Logistic Regression
LogReg = LogisticRegression(random_state=42)
LogReg.fit(X_train, Y_train)
```

Python

Figure 28-Code implementation 3

```python
#Decision Tree
DecTree = DecisionTreeClassifier(random_state=42)
DecTree.fit(X_train, Y_train)
```

Python

```python
#Random Forest
RandForest = RandomForestClassifier(random_state=42)
RandForest.fit(X_train, Y_train)
```

Python

```python
#K-Nearest Neighbors
KNN = KNeighborsClassifier()
KNN.fit(X_train, Y_train)
```

Python

```python
#Naive Bayes
NaiveBayes = GaussianNB()
NaiveBayes.fit(X_train, Y_train)
```

Python

```python
#Neural Network, Multi-layer Perceptron
NeuralNet = MLPClassifier(random_state=42)
NeuralNet.fit(X_train, Y_train)
```

Python

Figure 29-Code implementation 4

```python
#Gradient Boosting
GradBoost = GradientBoostingClassifier(random_state=42)
GradBoost.fit(X_train, Y_train)
```

Python

```python
#XGBoost
XGBoost = XGBClassifier(random_state=42)
XGBoost.fit(X_train, Y_train)
```

Python

Creating the prediction of our Logistic Regression model based on our test data

```python
#Logistic Regression
LogReg_pred = LogReg.predict(X_test)
```

Python

```python
#Decision Tree
DecTree_pred = DecTree.predict(X_test)
```

Python

Figure 30-Code implementation 5

```python
#Random Forest
RandForest_pred = RandForest.predict(X_test)
```

Python

```python
#K-Nearest Neighbors
KNN_pred = KNN.predict(X_test)
```

Python

```python
#Naive Bayes
NaiveBayes_pred = NaiveBayes.predict(X_test)
```

Python

```python
#Neural Network, Multi-layer Perceptron
NeuralNet_pred = NeuralNet.predict(X_test)
```

Python

```python
#Gradient Boosting
GradBoost_pred = GradBoost.predict(X_test)
```

Python

```python
#XGBoost
XGBoost_pred = XGBoost.predict(X_test)
```

Python

Figure 31-Code implementation 6

Evaluation metrics

```python
#Logistic Regression
print('**************************')
print('** Logistic Regression **')
print('************************** \n')

#Accuracy
print('Accuracy:', accuracy_score(Y_test, LogReg_pred),'\n')

#Area Under curve
auc_lr = roc_auc_score(Y_test, LogReg.predict_proba(X_test)[:, 1])
print('AUC:', auc_lr,'\n')

#Confusion matrix
print('Confusion Matrix:\n', confusion_matrix(Y_test, LogReg_pred),'\n')

#Classification report
print('Classification Report:\n', classification_report(Y_test, LogReg_pred))
```

Python

```python
#Get the coefficients for each feature
LogReg_coefficients = LogReg.coef_[0]
feature_importance_LogReg = pd.Series(LogReg_coefficients, index=X.columns)
feature_importance_LogReg = feature_importance_LogReg.abs().sort_values(ascending=False)
print('Logistic Regression Feature Importance:\n', feature_importance_LogReg)
```

Python

Figure 32-Code implementation 7

```python
#Decision Tree
print('********************')
print('** Decision Tree **')
print('******************** \n')

#Accuracy
print('Accuracy:', accuracy_score(Y_test, DecTree_pred),'\n')

#Area under curve
auc_tree = roc_auc_score(Y_test, DecTree.predict_proba(X_test)[:, 1])
print('AUC:', auc_tree,'\n')

#Confusion matrix
print('Confusion Matrix:\n', confusion_matrix(Y_test, DecTree_pred),'\n')

#Classification report
print('Classification Report:\n', classification_report(Y_test, DecTree_pred))
```

Python

```python
#Get feature importances from Decision Tree
tree_feature_importances = DecTree.feature_importances_
feature_importance_tree = pd.Series(tree_feature_importances, index=X.columns)
feature_importance_tree = feature_importance_tree.sort_values(ascending=False)
print('Decision Tree Feature Importance:\n', feature_importance_tree)
```

Python

Figure 33-Code implementation 8

```python
#Random Forest
print('********************')
print('** Random Forest **')
print('******************** \n')

#Accuracy
print('Accuracy:', accuracy_score(Y_test, RandForest_pred),'\n')

#Area under curve
auc_rf = roc_auc_score(Y_test, RandForest.predict_proba(X_test)[:, 1])
print('AUC:', auc_rf,'\n')

#Confusion matrix
print('Confusion Matrix:\n', confusion_matrix(Y_test, RandForest_pred),'\n')

#Classification report
print('Classification Report:\n', classification_report(Y_test, RandForest_pred))
```

Python

```python
#Get feature importances from Random Forest
tree_feature_importances = RandForest.feature_importances_
feature_importance_tree = pd.Series(tree_feature_importances, index=X.columns)
feature_importance_tree = feature_importance_tree.sort_values(ascending=False)
print('Decision Tree Feature Importance:\n', feature_importance_tree)
```

Python

Figure 34-Code implementation 9

```python
#K-Nearest Neighbors
print('************************')
print('** K-Nearest Neighbors **')
print('************************ \n')

#Accuracy
print('Accuracy:', accuracy_score(Y_test, KNN_pred),'\n')

#Area under curve
auc_knn = roc_auc_score(Y_test, KNN.predict_proba(X_test)[:, 1])
print('AUC:', auc_knn,'\n')

#Confusion matrix
print('Confusion Matrix:\n', confusion_matrix(Y_test, KNN_pred),'\n')

#Classification report
print('Classification Report:\n', classification_report(Y_test, KNN_pred))
```

Python

Figure 35-Code implementation 10

```python
#Naive Bayes
print('****************')
print('** Naive Bayes **')
print('**************** \n')

#Accuracy
print('Accuracy:', accuracy_score(Y_test, NaiveBayes_pred),'\n')

#Area under curve
auc_naive_bayes = roc_auc_score(Y_test, NaiveBayes.predict_proba(X_test)[:,1])
print('AUC:', auc_naive_bayes,'\n')

#Confusion matrix
print('Confusion Matrix:\n', confusion_matrix(Y_test, NaiveBayes_pred),'\n')

#Classification report
print('Classification Report:\n', classification_report(Y_test, NaiveBayes_pred))
```

Python

Figure 36-Code implementation 11

```python
#Neural Network, Multi-layer Perceptron
print('*******************************************')
print('** Neural Network, Multi-layer Perceptron **')
print('******************************************* \n')

#Accuracy
print('Accuracy:', accuracy_score(Y_test, NeuralNet_pred),'\n')

#Area under curve
auc_mlp = roc_auc_score(Y_test, NeuralNet.predict_proba(X_test)[:, 1])
print('AUC:', auc_mlp,'\n')

#Confusion matrix
print('Confusion Matrix:\n', confusion_matrix(Y_test, NeuralNet_pred),'\n')

#Classification report
print('Classification Report:\n', classification_report(Y_test, NeuralNet_pred))
```

Python

Figure 37-Code implementation 12

```python
#Gradient Boosting
print('**********************')
print('** Gradient Boosting **')
print('********************** \n')

#Accuracy
print('Accuracy:', accuracy_score(Y_test, GradBoost_pred),'\n')

#Area under curve
auc_gb = roc_auc_score(Y_test, GradBoost.predict_proba(X_test)[:, 1])
print('AUC:', auc_gb,'\n')

#Confusion matrix
print('Confusion Matrix:\n', confusion_matrix(Y_test, GradBoost_pred),'\n')

#Classification report
print('Classification Report:\n', classification_report(Y_test, GradBoost_pred))
```

Python

Figure 38-Code implementation 13

```python
#XGBoost
print('*************')
print('** XGBoost **')
print('************* \n')

#Accuracy
print('Accuracy:', accuracy_score(Y_test, XGBoost_pred),'\n')

#Area under curve
auc_xgb = roc_auc_score(Y_test, XGBoost.predict_proba(X_test)[:, 1])
print('AUC:', auc_xgb,'\n')

#Confusion matrix
print('Confusion Matrix:\n', confusion_matrix(Y_test, XGBoost_pred),'\n')

#Classification report
print('Classification Report:\n', classification_report(Y_test, XGBoost_pred))
```

Python

```python
#Get feature importance from XGBoost
xgb_feature_importances = XGBoost.feature_importances_
feature_importance_xgb = pd.Series(xgb_feature_importances, index=X.columns)
feature_importance_xgb = feature_importance_xgb.sort_values(ascending=False)
print('XGBoost Feature Importance:\n', feature_importance_xgb)
```

Python

Figure 39-Code implementation 14

	Imbalanced			Balanced		
Logistic Regression	Actual\Predicted values	0	1	Actual\Predicted values	0	1
	0	55 725	1 263	0	7 050	2 788
	1	8 526	1 459	1	2 423	7 659
Decision Tree	Actual\Predicted values	0	1	Actual\Predicted values	0	1
	0	48 742	8 246	0	6 277	3 561
	1	6 815	3 170	1	3 621	6 461
Random Forest	Actual\Predicted values	0	1	Actual\Predicted values	0	1
	0	55 205	1 783	0	6 818	3 020
	1	8 378	1 607	1	2 406	7 676
K-Nearest-Neighbors	Actual\Predicted values	0	1	Actual\Predicted values	0	1
	0	53 949	3 039	0	6 636	3 202
	1	7 972	2 013	1	2 714	7 368
Naive Bayes	Actual\Predicted values	0	1	Actual\Predicted values	0	1
	0	45 223	11 765	0	7 054	2 784
	1	4 263	5 722	1	2 833	7 249
Neural Network, Multi-layer	Actual\Predicted values	0	1	Actual\Predicted values	0	1
	0	55 786	1 202	0	7 010	2 828
	1	8 633	1 352	1	2 534	7 548
Gradient Boosting	Actual\Predicted values	0	1	Actual\Predicted values	0	1
	0	55 802	1 186	0	6 890	2 948
	1	8 453	1 532	1	2 176	7 906
XGBoost	Actual\Predicted values	0	1	Actual\Predicted values	0	1
	0	55 635	1 353	0	6 855	2 983
	1	8 380	1 605	1	2 199	7 883

Figure 40-Model's Confusion Matrix Comparison