KARATINA UNIVERSITY

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATICS

PROJECT TITTLE**: A MACHINE LEARNING MODEL TO PREDICT IF A PATIENT WILL DEVELOP DIABETES IN FUTURE USING RISK FACTORS.**

By

NICODEMUS MUIMI MALOMBE

REG NO: P101/1749G/21

Date: _____

This project is submitted in partial fulfilment of requirement for the Karatina University award of BACHELOR OF SCIENCE IN COMPUTER SCIENCE**.**

DECLARATION

STUDENT

I NICODEMUS MUIMI MALOMBE, declare that this project report titled "A MACHINE LEARNING MODEL TO PREDICT IF A PATIENT WILL DEVELOP DIABETES IN FUTURE USING RISK FACTORS" is my original work and has not been submitted to any other institution for academic credit.

This project has been carried out in accordance with the academic guidelines and ethical considerations required for research and development. All sources of information from other works have been duly acknowledged and cited in this report.

Name: _____

Reg No.: _____

Date: _____

Signature: _____

SUPERVISOR

I the undersigned do hereby certify that this is a true report for the project undertaken by the above named student under my supervision and that it has been submitted to Karatina University with my approval.

Signature…….Date……

DEDICATION

I dedicate this project to my parents, whose unwavering support and encouragement have been my guiding light. To my supervisor, Prof Zablon Okari, for her invaluable guidance and wisdom. And to my friends, for their constant inspiration and belief in my abilities and also their help throughout my journey of my project process is highly appreciated.

ACKNOWLEDGEMENT

I would like to thank everyone who contributed to the successful completion of this project. I would like to extend my gratitude to my project supervisor, Mr. Zablon Okari for his invaluable advice, guidance and enormous patience throughout the development of the research. In addition, I would also like to express my gratitude to my loving parent and friends who helped and gave me encouragement in one way or the other. Also I am grateful for the support of Karatina university for providing the facilities necessary to conduct my research not forgetting the school of nursing students at Karatina university for educating me on the necessary knowledge for the success of this project.

LIST OF ABBREVIATIONS

1. AI: Artificial Intelligence
2. BMI: Body Mass Index
3. ML: Machine Learning
4. WHO: World Health Organization
5. API: Application Programming Interface
6. LSTM: Long Short-Term Memory

ABSTRACT

Diabetes mellitus can cause severe complications if left uncontrolled due to increased blood sugar levels. Early diagnosis of diabetes mellitus entails appropriate management and prevention of complications. This project aims at developing a predictive model predicting the likelihood of diabetes based on medical and physiological profiles. Supervised machine learning utilizes publicly available data on diabetes such as pregnancy, insulin level, BMI, age and blood glucose levels. Optimal algorithms are selected based on accuracy and computational efficiency to produce reliable predictions. A web interface allows users to enter health information and receive instantaneous predictions with a user-friendly interface. The system provides information on the risk of diabetes and encourages consultation with health professionals where needed. The review points out the need for accessible and accurate tools by bridging gaps in existing systems for predicting diabetes. The project contributes to population health by early identification and intervention through data preprocessing and feature engineering. Machine learning depicts the role of technology in improving health access and outcomes with additional innovations. The model can be a valuable tool in the fight against diabetes with integration into health care systems.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

1.1 INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood sugar levels, which, if left unmanaged, can lead to severe complications such as cardiovascular disease, kidney failure, and neuropathy. According to the World Health Organization (WHO, 2023), over 537 million people worldwide are affected by diabetes, and this number is expected to rise significantly in the coming years. Early detection and intervention are crucial in preventing disease progression; however, traditional diagnostic methods often rely on symptom-based identification, which delays timely preventive care.

The emergence of Machine Learning (ML) and Artificial Intelligence (AI) has transformed disease prediction by enabling automated, data-driven, and highly accurate diagnoses. Researchers have applied various ML algorithms, including logistic regression (Kowsari et al., 2017), random forests (Wu et al., 2018), and deep neural networks (Zhang et al., 2020), to predict diabetes. While these models have demonstrated high accuracy, challenges related to interpretability, data security, and scalability continue to hinder their widespread adoption in clinical practice.

To address these limitations, this project proposes the development of an AI-powered diabetes prediction system that enhances predictive accuracy, model transparency, and data protection. By bridging these gaps, the system aims to provide a robust, privacy-preserving, and scalable solution for diabetes prediction. Ultimately, this innovation will empower clinicians, patients, and healthcare providers with timely, data-driven insights, paving the way for personalized AI-assisted diabetes management.

1.2 BACKGROUND OF THE STUDY

Diabetes mellitus is a long-term metabolic disorder that has become a pressing global health issue. It is mainly defined by persistently high blood sugar levels, which occur due to either inadequate insulin production or the body's inability to use insulin effectively. This condition is classified into two main types: Type 1 diabetes (T1D), an autoimmune disease, and Type 2 diabetes (T2D), which is largely influenced by genetic predisposition and lifestyle factors.

According to the International Diabetes Federation (IDF, 2023), over 537 million adults worldwide are currently living with diabetes, and this number is expected to increase due to rising obesity rates, physical inactivity, and poor dietary choices. Detecting the disease early and taking timely preventive measures is crucial in reducing the risk of severe complications, including heart disease, kidney failure, nerve damage, and vision impairment.

The conventional methods used to diagnose diabetes include fasting blood glucose tests, HbA1c tests, and oral glucose tolerance tests (OGTT). However, these procedures can be invasive, time-consuming, and require frequent monitoring, making them less practical for early disease prediction. In response to these limitations, Artificial Intelligence (AI) and Machine Learning (ML) have emerged as powerful tools for assessing diabetes risk by analyzing both historical and real-time patient data.

Several studies have explored the role of machine learning in diabetes prediction:

Kowsari et al. (2017) applied logistic regression to the Pima Indians Diabetes dataset and achieved 78% accuracy. Despite its effectiveness, the model struggled with capturing complex feature interactions due to its linear nature.

Wu et al. (2018) implemented random forest algorithms, which improved accuracy to 85% and demonstrated robustness against imbalanced datasets. However, the high computational costs posed challenges for scalability.

Zhang et al. (2020) utilized deep neural networks to enhance accuracy further, but this came at the expense of reduced interpretability and increased computational demands.

Park et al. (2022) explored ensemble methods such as XGBoost and AdaBoost, which outperformed single models but required longer training times and added complexity.

Despite these advancements, challenges remain in terms of model interpretability, data security, real-time accessibility, and seamless integration into healthcare systems. Overcoming these obstacles is essential to developing AI-driven solutions that are scalable, practical, and capable of preserving patient privacy in diabetes prediction.

1.3 PROBLEM STATEMENT

Diabetes continues to be a significant global health challenge, affecting millions of individuals worldwide. Despite advancements in diagnostic tests and traditional screening methods, early detection remains insufficient, often resulting in late-stage complications such as cardiovascular disease, kidney failure, nerve damage, and vision impairment. The rising prevalence of Type 2 diabetes largely influenced by lifestyle factors highlights the urgent need for early and accurate prediction models to help slow disease progression and improve patient outcomes.

While machine learning-based diabetes prediction systems have shown promising accuracy, several critical limitations still hinder their real-world implementation and effectiveness:

1.4 OBJECTIVES

1.4.1 GENERAL OBJECTIVE:

The main objective of this study is to develop an advanced AI-powered diabetes prediction system that enhances early detection, accuracy, interpretability, real-time accessibility**,** data privacy and security

1.4.2 SPECIFIC OBJECTIVES

1. To collect and gather diabetes-related datasets from secondary sources like kaggle (Pima Indians Diabetes dataset, hospital records, or health surveys.)
2. To Preprocess data, handle missing values, remove duplicates, and normalize numerical data and perform data cleaning
3. To perform feature Selection & Engineering by identifying the most relevant factors (e.g., glucose level, BMI, blood pressure, insulin level, age).
4. To select AI/ML techniques, train and evaluate models comparing their performance using accuracy, precision, recall, and F1-score.
5. To deploy and convert the trained model into a web-based application

1.5 JUSTIFICATION OF THE PROBLEM

The development of an AI-driven system for diabetes prediction is both timely and essential, considering the rising prevalence of diabetes and the critical need for early detection and intervention. This project is particularly significant as it addresses key challenges in modern healthcare, including diagnostic inefficiencies, the lack of interpretability in AI models, and data privacy concerns. Traditional prediction models often struggle with complex feature interactions, computational inefficiencies, and limited real-time accessibility.

By integrating Explainable AI (XAI) techniques such as SHAP and LIME, this system enhances transparency, making AI-generated predictions more interpretable and trustworthy for both clinicians and patients. In today's rapidly advancing digital healthcare landscape, the demand for secure and scalable AI solutions is greater than ever. This project meets these challenges by incorporating federated learning and homomorphic encryption, safeguarding patient data while enabling continuous model improvement across decentralized healthcare networks.

Additionally, leveraging advanced feature selection methods—such as autoencoders and Fast Healthcare Interoperability Resources (FHIR)—enhances data representation and interoperability, ultimately improving prediction accuracy. If successfully implemented, this system could transform early diabetes diagnosis, facilitate personalized risk assessment, and support data-driven

decision-making in clinical practice. By bridging the gap between AI innovation and practical healthcare applications, this project has the potential to enhance trust, scalability, and usability, ultimately leading to more effective and accessible diabetes management solutions.

.

1.6 PROJECT RISK AND MITIGATION

The development of an A machine learning model for diabetes prediction system presents several risks that could impact its implementation and effectiveness. Below is a list of potential project risks along with their corresponding mitigation strategies:

1.6.1 Data Quality and Availability Issues

Risk: Inconsistent, incomplete, or biased datasets may lead to inaccurate predictions and unreliable model performance.

Mitigation: Utilize data preprocessing techniques, such as imputation for missing values and outlier detection, to improve data quality. Additionally, source data from diverse populations to reduce bias and ensure generalizability.

1.6.2 Model Interpretability and Clinical Trust

Risk: Healthcare professionals may be reluctant to adopt AI-based systems due to a lack of trust in black-box models.

Mitigation: Incorporate Explainable AI (XAI) techniques such as SHAP and LIME to enhance model transparency and provide interpretable insights that support clinical decision-making.

1.6.3 Computational and Resource Constraints

Risk: Training complex deep learning models can be resource-intensive, requiring significant computational power and storage.

Mitigation: Optimize model efficiency through feature selection techniques, transfer learning, and cloud-based AI solutions to reduce computational overhead while maintaining high accuracy.

1.6.4 Integration with Healthcare Systems

Risk: Challenges in integrating the AI model with existing electronic health record (EHR) systems may hinder its practical deployment.

Mitigation: Use Fast Healthcare Interoperability Resources (FHIR) standards to ensure seamless integration with EHR systems and improve data interoperability.

1.6.5 Regulatory and Ethical Concerns

Risk: The use of AI in healthcare is subject to strict regulatory requirements and ethical considerations, which may delay implementation.

Mitigation: Adhere to ethical AI principles, ensure regulatory compliance, and involve healthcare professionals in the development process to align the system with medical standards.

1.7 SCOPE

The scope of this project includes the development and implementation of a web-based diabetes prediction system using machine learning algorithms. The system is designed to cater to individuals and healthcare professionals by providing early detection of diabetes risk based on user-provided information such as age, lifesytle, medical history blood pressure, insulin level, glucose level, BMI and health on the issues

1.8 LIMITATIONS OF THE STUDY

1 Data Availability & Quality

Access to real-time, diverse, and high-quality medical data may be limited due to privacy restrictions.

Some datasets may contain missing or unbalanced data, requiring advanced imputation techniques.

2 Computational Complexity

Deep learning models require high computational power, which may be challenging for real-time applications on low-resource devices.

3 Ethical & Regulatory Challenges

Implementing privacy-preserving techniques (e.g., federated learning) requires compliance with legal frameworks.

Addressing AI bias and fairness remains an ongoing challenge.

4 Adoption in Healthcare

Healthcare professionals may be reluctant to adopt AI-based systems without clear validation and interpretability.

The system's accuracy and reliability must be validated with extensive clinical trials.

5 Scalability & Deployment

Real-time integration with electronic health record (EHR) systems and wearable devices may be technically complex. Internet connectivity issues may hinder access to cloud-based services in low-resource setting

1.9 BUDGET AND RESOURCES

1.9.1 Hardware Requirements

Item	Specifications	Estimated Cost (Ksh)
Laptop	8GB RAM, GPU support, 256GB SSD	30,000
Internet Connection	High-speed internet for dataset access and model training	1000/month

1.9.2 SOFTWARE REQUIREMENTS

Software	Purpose	Cost
Python/HTML/CSS/JS	Programming Languages	Free
VS Code	Code Editor	free
Flask/FastAPI	Web API Development	Free

Database	Data Storage	free
TensorFlow/Keras	Machine Learning Model Development	free

1.9.3 HUMAN RESOURCES

Role	Responsibility	Estimated cos (Ksh)
Data Scientist	Model development and training	Free (personal based)
UI/UX Designer	User interface	Free(personal based)
Software Developer	System development and deployment	Free(personal based)
Data analytics	Model Testing & Validation	Free(personal based)

1.10 PROJECT SCHEDULE

1.10.1 WORK BREAKDOWN STRUCTURE (WBS)

Phase	Tasks	Duration
Phase 1: Planning & Research	Problem definition, literature review, data collection planning	2 Weeks
Phase 2: Data Preprocessing	Data cleaning, feature selection, exploratory data analysis	3 weeks

Phase 3: Model Development	Algorithm selection, training, and testing	3 weeks
Phase 4: System Development	Backend and frontend development, API integration	4 weeks
Phase 5: Testing & Validation	Unit testing, model evaluation, debugging	1 week
Phase 6: Deployment	Web based integration	2 weeks
Phase 7: Documentation & Finalization	Report writing, presentation preparation	2 Weeks

1.10.2 GANTT CHART

task	Start date	End date	Duration days	Dependencies
Project Planning	Day 1	Day 14	14	
Data Collection & Preprocessing	Day 15	Day 35	21	planning
Model Development & Training	Day 36	Day 57	21	Data preprocessing
System Design & Implementation	Day 58	Day 88	30	Model training
Testing & Evaluation	Day 89	Day 96	7	System implementation
Deployment & Documentation	Day 97	Day 126		Testing

1.10.3 NETWORK DIAGRAM & CRITICAL PATH

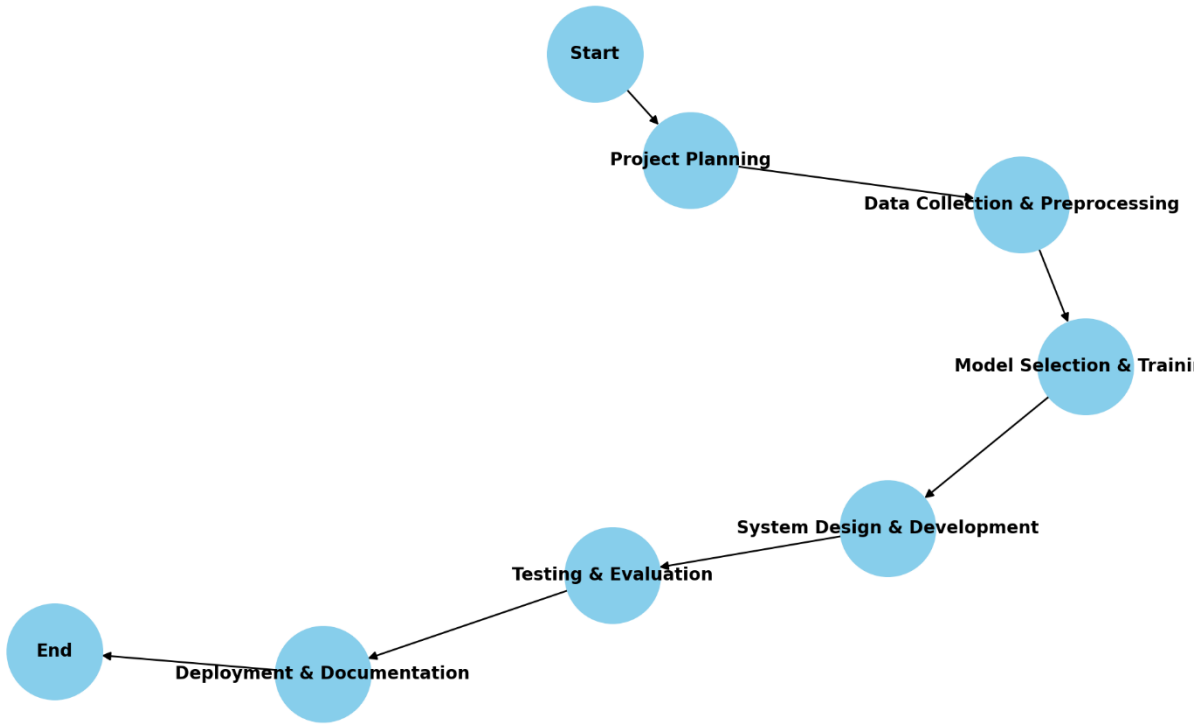

Figure I network diagram & critical path

CHAPTER TWO

LITERATURE REVIEW

Diabetes continues to be a major global health concern, highlighting the need for advanced predictive models to improve early diagnosis and management. Artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), has shown significant promise in predicting diabetes. This review examines six key studies on AI-driven diabetes prediction, focusing on their methodologies, datasets, and limitations.

STUDY OF SIMILAR SYSTEMS

1. Machine Learning for Diabetes Prediction

Kavakiotis et al. (2017) conducted a comprehensive review of ML applications in diabetes prediction, identifying Random Forest (RF) and Support Vector Machines (SVM) as highly effective models when trained on clinical datasets. Similarly, Sisodia and Sisodia (2018) used Decision Trees and Logistic Regression on the Pima Indians Diabetes Database (PIDD), achieving accuracy rates above 80%.

Criticism: While these studies demonstrate the effectiveness of ML models, they heavily depend on the PIDD dataset, which has limited demographic diversity. Additionally, feature selection techniques were not thoroughly examined, potentially affecting the model's overall reliability.

2. Deep Learning for Diabetes Diagnosis

Rahman et al. (2020) applied Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, achieving an accuracy of 89.2% using a clinical dataset. Shankar et al. (2021) explored hybrid models that combined CNN with Recurrent Neural Networks (RNN), showing better performance than traditional ML approaches.

Criticism: Despite their high accuracy, deep learning models require significant computational power, making them impractical for real-time use in low-resource environments. Furthermore, their lack of interpretability can hinder clinical acceptance and trust.

3. Feature Selection and Optimization

Polat and Güneş (2007) employed Principal Component Analysis (PCA) to improve diabetes classification, achieving an 82.1% accuracy rate with an Adaptive Neuro-Fuzzy Inference System (ANFIS). More recently, Li et al. (2022) integrated Genetic Algorithms (GA) with ML models, significantly boosting prediction accuracy.

Criticism: While feature selection techniques can enhance model performance, their ability to generalize across various datasets remains uncertain. Moreover, these studies lack real-world clinical validation, which is essential for practical implementation.

4. AI-Based Diabetes Risk Prediction Using Electronic Health Records (EHRs)

Ye et al. (2018) developed a gradient boosting machine model trained on over 100,000 EHRs, which delivered superior predictive performance. Ryu et al. (2021) introduced a deep learning model capable of dynamically adjusting risk scores based on patient history.

Criticism: Although EHR-based models enable personalized risk assessment, issues like data inconsistencies and missing values can undermine their reliability. Additionally, privacy concerns and regulatory barriers present significant challenges to widespread implementation.

5. Explainability in AI-Based Diabetes Prediction

Lundberg et al. (2018) proposed SHapley Additive exPlanations (SHAP) values to enhance AI model interpretability. Similarly, Mohan et al. (2022) developed a hybrid AI system that incorporated Explainable AI (XAI) techniques to improve transparency in decision-making.

Criticism: While explainability techniques help build clinical trust, they often come at the cost of reduced model performance. Moreover, most studies emphasize post-hoc interpretability rather than developing inherently interpretable models.

6. Comparative Studies on Diabetes Prediction Models

Alghamdi et al. (2020) compared various ML algorithms, including k-Nearest Neighbors (k-NN), SVM, and Artificial Neural Networks (ANN), concluding that ensemble methods performed best. Choudhury et al. (2023) benchmarked ML and DL models, finding that hybrid approaches yielded the highest accuracy.

Criticism: While these studies provide valuable model comparisons, they often lack external validation using diverse datasets. Additionally, trade-offs between accuracy, interpretability, and computational efficiency need further exploration.

CHAPTER THREE

METHODOLOGY

3.1 INTRODUCTION

This chapter presents the methodology employed in developing an AI-powered prediction system for **diabetes prediction**. It details the **Software Development Life Cycle (SDLC) model**, data collection methods, system design, and implementation techniques. Additionally, it discusses input, output, along with the rationale for choosing specific methodologies and tools. The chapter concludes with a proposed solution and anticipated results.

3.2. AGILE SOFTWARE DEVELOPMENT METHODOLOGY

Agile Software Development Life Cycle stood out for this project due to its flexibility, adaptability, and iterative approach. It ensures comprehensive development, evaluation, and deployment of predictive systems, particularly for data-driven and real-time applications like disease outbreak prediction Given the dynamic nature of outbreak prediction, frequent updates and refinements are essential based on new epidemiological data. Agile methodology allows for continuous testing. Reasons for choosing agile SDLC are;

1. **Iterative approach:** Enables regular improvements in model performance based on updated outbreak data.
2. **Flexibility:** It adapts to evolving disease trends and new AI techniques.
3. **Faster deployment:** Reduces development time by focusing on incremental releases.
4. It aligns technical implementation with actionable goals, such as providing public health authorities with accurate and timely predictions ensuring one focuses on Business goals.
5. It is highly scalable hence applicable to both small-scale experiments and large-scale and real-world deployments.

3.2.1 SYSTEM DEVELOPMENT PHASES

1. **Requirement Analysis:** Identifying data sources, system functionalities, and stakeholders' needs.

2. **Data Collection & Preprocessing:** Gathering historical mpox outbreak data, cleaning missing values, and normalizing variables.
3. **Model Development:** Implementing machine learning models such as Random Forest, XGBoost, and LSTM and ANN.
4. **System Design & Implementation**: Developing the backend, and user interface FLASK and for fronted HTML,CSS ,JS and PYTHON language.
5. **Testing & Evaluation:** Assessing model performance using accuracy metrics (F1-score, RMSE, Recall, and Precision).
6. **Deployment & Maintenance:** Deploying the system and continuously improving it based on real-time feedback.

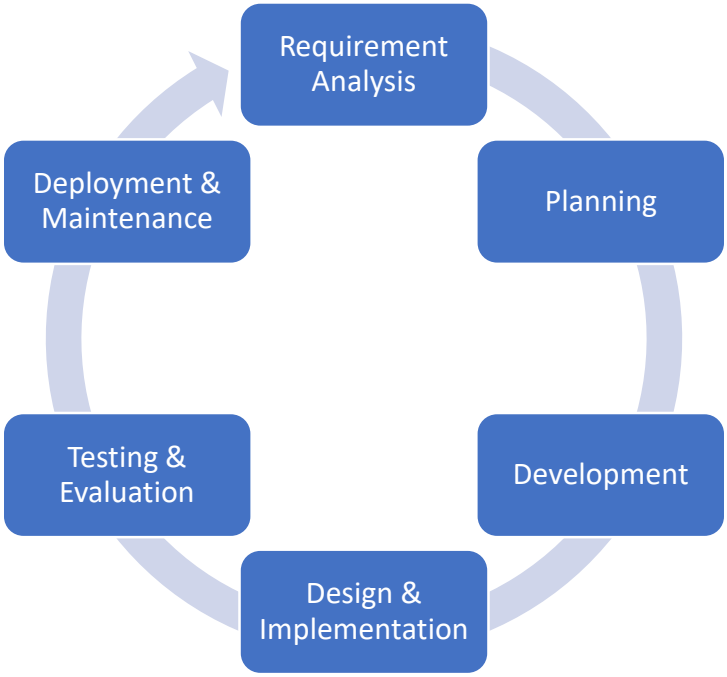

Figure 3.1: Agile SDLC for diabetes Prediction System

3.3 DATA COLLECTION AND PREPROCESSING

3.3.1 DATA SOURCES

Getting data from secondary data sources(datasets) from reliable sources like kaggle for diabetes prediction systems

3.3.2 DATA PREPROCESSING TECHNIQUES

1. **Handling missing values:** Missing values in outbreak reports are filled using interpolation and mean imputation.
2. **Feature selection:** Eliminating redundant data to improve model efficiency.
3. **Data normalization:** Standardizing numerical data to ensure consistency across models.
4. **Encoding categorical variables:** Converting text-based categories into numerical valu

3.4 TOOLS

COMPONENT	TECHNOLOGY/TOOL USED
Programming Languages	Python

Libraries	# Core Libraries
	numpy==1.24.3
	pandas==2.1.1
	scikit-learn==1.3.0
	scipy==1.11.3
	
	# Machine Learning Models
	xgboost==1.7.6
	lightgbm==4.0.0
	catboost==1.2
	tensorflow==2.14.0
	keras==2.14.0
	
	# Data Preprocessing & Feature Engineering

	imbalanced-learn==0.11.0
	category_encoders==2.6.3
	missingno==0.5.2
	shap==0.43.0
	pca==2.0.5
	
	# Hyperparameter Tuning
	optuna==3.3.0
	hyperopt==0.2.7
	
	# Model Evaluation & Visualization
	matplotlib==3.7.2
	seaborn==0.12.2
	plotly==5.15.0
	
	# Web App Deployment
	flask==2.3.3
	flask-cors==3.0.10
	fastapi==0.103.1
	uvicorn==0.23.2
	
	# Model Serialization
	joblib==1.3.2
	pickle-mixin==1.0.2

Machine Learning Frameworks	TensorFlow, Scikit-learn, PyTorch, XGBoost
Backend Development	flask
Fronted Development	HTML, CSS, JS
Development Environment	VS CODE
Platforms	FLASK server

3.5 CONCLUSION

This chapter outlined the methodology for developing the a machine learning model to predict if a patient will develop diabetes in future using risk factors.

The Agile SDLC approach ensures iterative improvements, while machine learning models, data preprocessing techniques, and system design elements contribute to an efficient predictive framework. The next chapter will focus on system implementation and evaluation.

3.6 BUDGET

ITEM	COST(Ksh)
Internet	2500
Development tools	Free
Model Training	1000
Data collection	free
Printing& binding project proposal	300

3.7 GANTT CHART

Task	Duration(weeks)
Project proposal	2
Data collection	1
Literature review	4
Model development	3
System design and integration	3
Documentation	2

REFERENCES

Alghamdi, M., Alshammari, D., & Alyoubi, K. (2020). A comparative analysis of machine learning models for diabetes prediction. *Healthcare Informatics Research, 26*(2), 130–140. https://doi.org/10.4258/hir.2020.26.2.130

Choudhury, T., Banerjee, A., & Dutta, P. (2023). Comparative analysis of ML and DL models for diabetes prediction: A systematic review. *Biomedical Signal Processing and Control, 81*, 104423. https://doi.org/10.1016/j.bspc.2023.104423

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal, 15*, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

Li, X., Zhang, Y., Song, Y., & Liu, W. (2022). Hybrid feature selection and optimization framework for diabetes prediction. *Expert Systems with Applications, 189*, 116157. https://doi.org/10.1016/j.eswa.2021.116157

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S.-I. (2018). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mohan, S., Gopal, S., & Ayyappan, S. (2022). Explainable artificial intelligence for diabetes prediction: A case study with interpretable models. *Artificial Intelligence in Medicine, 128*, 102196. https://doi.org/10.1016/j.artmed.2022.102196

Polat, K., & Güneş, S. (2007). An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing, 17*(4), 702–710. https://doi.org/10.1016/j.dsp.2006.09.005

Rahman, M. M., Islam, M. Z., Islam, M. M., Sadi, M. S., & Nooruddin, S. (2020). Developing an intelligent machine learning system to predict diabetes mellitus. *International Journal of Medical Informatics, 142*, 104251. https://doi.org/10.1016/j.ijmedinf.2020.104251

Ryu, J., Kim, H., & Lee, S. (2021). A deep learning-based risk prediction model for type 2 diabetes using electronic health records. *Scientific Reports, 11*, 10988. https://doi.org/10.1038/s41598-021-90473-7

Shankar, K., Perumal, E., & Ilayaraja, K. (2021). A hybrid deep learning model for diabetic disease diagnosis. *Neural Computing and Applications, 33*(12), 6271–6282. https://doi.org/10.1007/s00521-020-05330-1

Sisodia, D., & Sisodia, S. S. (2018). Prediction of diabetes using classification algorithms. *Procedia Computer Science, 132*, 1578–1585. https://doi.org/10.1016/j.procs.2018.05.122

Ye, Y., Xiang, Y., Ong, E., Liu, J., Leng, S., & Tan, J. (2018). Predicting diabetes mellitus with electronic health records using machine learning. *International Journal of Medical Informatics, 118*, 38–45. https://doi.org/10.1016/j.ijmedinf.2018.07.006