DIABETES PREDICTION USING ML

An Internship Project Report

Submitted to

DLITHE CONSULTANCY SERVICES PRIVATE LIMITED

by

SHREYAS P

Under the guidance of

Ms. Medini B V

Robotics & Design Engineer

1. ACKNOWLEDGEMENT

I would like to extend my heartfelt gratitude to DLithe Consultancy Services Private Limited for providing me with the invaluable opportunity to undertake my internship at their esteemed institution. Their support and guidance were instrumental in shaping my project, and I am truly grateful for the experience.

I would also like to express my deepest thanks to Ms. Medini, who served as an exceptional guide throughout this internship journey. Her expertise, encouragement, and mentorship were pivotal in making this experience rewarding and enriching.

Additionally, I am grateful to Mangalore Institute of Technology & Engineering for giving me the chance to pursue this internship at DLithe Consultancy Services Private Limited. The college's support and encouragement have been crucial in enabling me to gain practical industry experience and further my learning.

Finally, I want to acknowledge all the individuals who assisted me in various ways, providing valuable insights and contributing to my growth and knowledge during the internship period. Thank you all for your unwavering support.

2. ORGANIZATIONAL INFORMATION

DLithe is a technology-based product company that has been serving IT companies and academic institutions since the year 2018. The company is led by industry professionals with two decades of experience. For IT companies, DLithe offers services such as technology consultancy, project development, IT recruitment, staffing, competency development, and content development. On the other hand, the company serves academic institutions by providing competency development services in niche technologies like artificial intelligence, internet of things, robotics, cybersecurity, augmented reality, and more. DLithe has also developed the arm-based Cortex M3 series microcontroller and the ioCube product in the embedded and IoT domain.

During my rewarding internship in the field of Artificial Intelligence and Machine Learning, I had the privilege of being a part of an exceptional program under the guidance of this renowned organization. Throughout the internship, I gained comprehensive insights into diverse industry verticals, spanning from understanding project requirements to the final deployment phase.

DLithe's internship program provided me with a valuable opportunity to immerse myself in real-world scenarios, gaining exposure to industry best practices and learning how to implement AI and ML solutions within an agile project life cycle. The supportive environment and dedicated mentors at the organization ensured that I could explore practical use cases for AI and ML implementation, enabling me to grow and learn during insightful post-mentoring sessions.

One notable aspect of the internship was the opportunity to work on real-world project, including a diabetes prediction using aiml. DLithe provided guidance and mentoring throughout the project, allowing me to gain hands-on experience with different machine learning models and their application in solving practical problems like crop price prediction.

This internship experience has equipped me with a strong foundation in artificial intelligence and machine learning, positioning me well for a career in this dynamic and rapidly evolving field.

2. CONTENTS

4 . ABSTRACT

Diabetes is a prevalent chronic health condition that affects millions of people worldwide. Early diagnosis and effective management are crucial to mitigate its adverse effects on individuals' health. Machine Learning (ML) techniques have shown promise in predicting the risk of diabetes based on various patient attributes. This project presents a user-friendly web application built using Streamlit to provide a practical and accessible solution for diabetes prediction.

The proposed system leverages a dataset containing patient information such as age, gender, BMI, family history, and other relevant features. A machine learning model is trained using this dataset to predict the likelihood of an individual developing diabetes. Various ML algorithms, such as logistic regression, random forests, and support vector machines, are explored to identify the most accurate predictive model.

The Streamlit framework is used to create an interactive and user-friendly web application that allows users to input their health data and receive an instant prediction of their diabetes risk. The application provides not only a binary prediction (diabetic or non-diabetic) but also a probability score, which helps users understand the level of risk associated with their health status.

This project aims to empower individuals with a convenient tool to assess their diabetes risk and make informed decisions about their health. By combining the power of machine learning and the accessibility of Streamlit, this application offers a practical and efficient means of early diabetes detection, ultimately contributing to improved healthcare and the well-being of individuals.

5. INTERNSHIP OBJECTIVES

The primary objectives of the AI and ML internship was designed to equip us with comprehensive skill set and practical knowledge in various areas of Artificial Intelligence and Machine Learning. The key objectives included:

- **Learning Python Basics:** The internship aimed to provide a strong foundation in Python programming, as it is one of the most widely used languages in AI and ML. Participants were introduced to Python syntax, data structures, and essential libraries used in AI and ML development.
- **Gain Practical Experience:** The primary goal of this internship was to gain practical, hands-on experience in the field of artificial intelligence and machine learning. This involved working on real-world projects and applying AI and ML concepts to solve practical problems.
- **Understanding ML Algorithms:** The internship focused on making us understand fundamental ML algorithms such as Linear Regression, Binary Classification, and Decision Trees. These algorithms form the building blocks for more advanced techniques and are crucial for understanding the basics of supervised learning.
- **Exploring Neural Networks:** We delved into the world of Neural Networks, understanding their architecture and how they mimic the human brain's functioning. Topics covered included Activation Functions and Forward Propagation, which are essential concepts for building and training neural networks.
- **Mentorship and Feedback:** Receive guidance and mentorship from industry experts to enhance skills and knowledge in AI and ML. Use feedback to continuously improve and refine project work.
- **Emphasizing GitHub and LinkedIn Profile Maintenance:** The internship recognized the importance of a strong online presence for aspiring AI and ML professionals. We were encouraged to maintain an active GitHub repository showcasing our projects and contributions, as well as a well-curated LinkedIn profile to showcase our skills and accomplishments.
- **Master AI/ML Tools and Platforms:** To become proficient in using AI and ML tools and platforms widely used in the industry. This includes working with frameworks like TensorFlow, scikit-learn, and exploring cloud-based AI services.

6. WEEKLY OVERVIEW OF INTERNSHIP ACTIVITY

Week 1: Python Fundamentals for AI & ML
Objective: Understand Python Fundamentals for AI & ML. Activities:
Covered Python syntax and data structures.
Explored essential libraries used in AI and ML.
Worked on basic Python programming exercises and projects.

Week 2: Exploring Machine Learning Algorithms in Python Objective:
Study and Implement ML Algorithms.
Activities:
Logistic Regression: Learned and implemented binary classification using logistic regression in Python, with real-world applications.
Support Vector Machines (SVM): Explored SVM for classification and regression, including different kernels.
Naive Bayes: Introduced probabilistic classification with Naive Bayes and its applications.
Decision Tree: Explored decision tree algorithms, implemented classifiers in Python, and addressed overfitting.
Neural Networks: Introduced neural networks and implemented simple feedforward networks using libraries like TensorFlow

Key Learnings:

Gained knowledge and hands-on experience with logistic regression, SVM, Naive Bayes, decision trees, and neural networks in Python.
Understood the strengths and weaknesses of each algorithm for various use case.

7. CHALLENGES AND LEARNING OUTCOMES

CHALLENGES

Developing a diabetes prediction application using machine learning and Streamlit can be a promising endeavor, but it also comes with various challenges. Some of the key challenges you might face during this project include:

1. Data Quality and Quantity: Obtaining a high-quality and sufficiently large dataset for training your machine learning model can be challenging. Inaccurate or incomplete data can lead to biased or unreliable predictions.

2. Feature Selection: Identifying the most relevant features from the dataset to build an accurate predictive model is crucial. Feature engineering and selection can be challenging, especially when dealing with medical data.

3. Model Selection: Choosing the right machine learning algorithm for diabetes prediction can be complex. Different algorithms have different strengths and weaknesses, and selecting the most suitable one may require experimentation.

4. Model Evaluation: Assessing the performance of the model accurately is essential. You may need to employ appropriate evaluation metrics, such as accuracy, precision, recall, and F1 score, to ensure the model's reliability.

5. Imbalanced Data: Medical datasets often suffer from class imbalance, where there are significantly more instances of non-diabetic individuals compared to diabetic individuals. Handling this imbalance while training the model is a challenge.

6. Data Privacy and Security: Health data is highly sensitive, and ensuring the privacy and security of the data used in the application is critical. Compliance with data protection regulations like HIPAA can be demanding.

LEARNING OUTCOMES

1. Understanding Diabetes: Gain a deeper understanding of diabetes, including its types, risk factors, symptoms, and complications. This knowledge is essential for framing the problem correctly.

2. Data Collection and Preprocessing: Learn how to collect and preprocess data relevant to diabetes prediction. This involves cleaning, normalizing, and transforming data to make it suitable for ML algorithms.

3. Feature Engineering: Understand the process of feature selection and engineering to identify which data attributes are most informative for predicting diabetes.

4. Machine Learning Algorithms: Gain expertise in various ML algorithms like logistic regression, decision trees, random forests, support vector machines, and deep learning models. Understand when and how to use each one for diabetes prediction.

5. Model Training and Evaluation: Learn how to train ML models, split data into training and testing sets, and evaluate model performance using metrics like accuracy, precision, recall, F1-score, and ROC AUC.

6. Hyperparameter Tuning: Understand the importance of hyperparameter tuning to optimize model performance.

7. Cross-Validation: Learn about k-fold cross-validation to assess the model's generalization performance.

8. Feature Importance: Interpret and analyze the importance of different features in the prediction model.

8. PROJECT DETAILS

CHAPTER 1

8.1 INTRODUCTION

Diabetes is a chronic medical condition of growing concern worldwide, characterized by elevated blood sugar levels that can have severe health implications. Early detection and management of diabetes are essential to mitigate its complications and improve the quality of life for those affected. Machine Learning (ML) techniques have demonstrated their potential in predicting diabetes risk, offering a powerful tool for early diagnosis and prevention. To make this predictive capability accessible to a broader audience, we propose the development of a user-friendly web application utilizing Streamlit, a modern web application framework for Python.

This project seeks to bridge the gap between the sophisticated world of machine learning and the practical needs of individuals concerned about their health. By combining ML with the simplicity and interactivity of Streamlit, our application aims to provide an intuitive and efficient means for users to assess their risk of developing diabetes. The system leverages a dataset containing various health-related attributes, including age, gender, body mass index (BMI), family history, and other factors, to train a predictive model. Multiple ML algorithms are explored to identify the most accurate model for diabetes risk prediction.

Through this web application, users can input their personal health data, and within moments, receive a prediction of their likelihood of developing diabetes. The application not only offers a binary classification (diabetic or non-diabetic) but also provides a probability score, enabling users to gauge the level of risk associated with their health status.

This project carries the potential to empower individuals with a valuable tool for informed decision-making regarding their health. By democratizing the use of machine learning for diabetes prediction, we aim to contribute to the early detection of this prevalent condition, ultimately fostering better healthcare outcomes and enhancing the well-being of individuals. In the following sections, we will delve into the technical aspects of the project, including data collection and preprocessing, machine learning model development, and the user interface design using Streamlit.

CHAPTER 2

8.2 LITERATURE SURVEY

1. Diabetes Prediction with Machine Learning:

- Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., & Vlahavas, I. (2017). "Machine Learning and Data Mining Methods in Diabetes Research." Computational and Structural Biotechnology Journal, 15, 104-116.
- Sharma, A., & Bali, T. (2020). "A Survey of Machine Learning in Big Data Analytics for Healthcare." Journal of King Saud University - Computer and Information Sciences.

2. Diabetes Risk Factors and Features:

- Meigs, J. B., Shrader, P., & Sullivan, L. M. (2008). "Practice Tools: Genotype Score in Addition to Common Risk Factors for Prediction of Type 2 Diabetes." New England Journal of Medicine, 359(21), 2208-2219.
- Tabak, A. G., Herder, C., Rathmann, W., Brunner, E. J., & Kivimäki, M. (2012). "Prediabetes: A High-Risk State for Diabetes Development." The Lancet, 379(9833), 2279-2290.

3. Machine Learning Algorithms for Diabetes Prediction:

- Zhang, X. Y., Dong, H. H., & Lin, J. Z. (2019). "Performance Comparison of Machine Learning Algorithms in Predicting Type 2 Diabetes." Technology and Health Care, 27(S1), 13-22.
- Liu, Y., Zheng, X., Yu, Z., Zhang, Z., & Ma, X. (2020). "A Comparative Study of Machine Learning Algorithms in Predicting Type 2 Diabetes Mellitus." Computers in Biology and Medicine, 117, 103624.

4. Streamlit for Web Application Development:

- Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., ... & Wickham, H. (2022). "R Markdown: The Definitive Guide." CRC Press.
- Abhinav Sagar. (2020). "Building Interactive Web Applications in Python." Towards Data Science.

5. Privacy and Data Security in Healthcare Applications:

- El Emam, K., Jonker, E., Moher, E., Walker, J., & Neisa, A. (2011). "A Systematic Review of Re-Identification Attacks on Health Data." PLOS ONE, 6(12), e28071.
- Ohno-Machado, L., & Loukides, G. (2014). "Sharing Clinical Trial Data: A Proposal from the Institute of Medicine." PLOS Medicine, 11(2), e1001602.

CHAPTER 3

8.3 PROBLEM STATEMENT

Developing a diabetes prediction model using machine learning poses a set of challenges. Firstly, obtaining a diverse and high-quality dataset for model training is a critical concern, as data quality significantly impacts prediction accuracy. Second, the selection and evaluation of the most suitable machine learning algorithm for diabetes prediction require careful consideration, as various algorithms offer different trade-offs between accuracy and interpretability. Additionally, handling imbalanced data, a common issue in medical datasets, is essential to prevent bias in the predictive model. Ensuring user privacy and data security, especially in a healthcare context, demands robust measures to comply with regulatory standards such as HIPAA. Designing an intuitive, user-friendly interface for the application to cater to users with varying levels of technical expertise is another crucial challenge. Finally, promoting ethical use of the application by providing responsible and empowering information to users while avoiding unnecessary anxiety adds another layer of complexity to the problem. Addressing these challenges is essential to create a reliable and user-centric diabetes prediction system.

CHAPTER 4

8.4 PROJECT OBJECTIVE

For a diabetes prediction application using machine learning and Streamlit, the objectives can be framed as follows:

1. Data Collection and Preprocessing:

- Gather a diverse and reliable dataset, ensuring data quality and relevance to diabetes prediction.
- Implement efficient data preprocessing techniques to handle missing values and outliers, and standardize the data for model training.

2. Model Development and Evaluation:

- Train and evaluate multiple machine learning algorithms to identify the most accurate and interpretable model for diabetes prediction.
- Employ cross-validation techniques to assess model performance, considering metrics such as accuracy, precision, recall, and F1 score.

3. Imbalanced Data Handling:

- Address class imbalance in the dataset by exploring techniques like oversampling, undersampling, and synthetic data generation.
- Ensure that the predictive model is trained to be robust and unbiased.

4. User-Friendly Interface Design:

- Design an intuitive and visually appealing user interface using Streamlit that caters to users with varying levels of technical expertise.
- Create an interactive platform for users to input their health data and receive predictions effortlessly.

5. Model Explainability and Transparency:

- Implement model explainability techniques to make predictions more transparent and understandable to users.
- Ensure users can access information about the factors influencing their risk of diabetes.

6. Privacy and Security Compliance:

- Develop and enforce stringent privacy and security measures to protect user data and comply with healthcare data privacy regulations, such as HIPAA.
- Ensure that sensitive health information is handled with the utmost care and that user confidentiality is maintained.

CHAPTER 5

8.5 METHODOLOGIES

The methodology for developing a diabetes prediction application using machine learning and Streamlit involves a series of steps, from data collection to application deployment. Here is a concise outline of the methodology:

1. Data Collection and Preprocessing:

- Collect a diverse and reliable dataset containing health-related features, such as age, gender, BMI, family history, and glucose levels.
- Clean the data by handling missing values, outliers, and inconsistencies.
- Normalize or standardize the data to ensure consistency in feature scales.

2. Exploratory Data Analysis (EDA):

- Perform EDA to gain insights into the dataset, identify correlations, and understand the distribution of features.
- Visualize key statistics and relationships between variables using plots and charts.

3. Feature Engineering:

- Select the most relevant features for diabetes prediction through feature selection techniques.
- Create new features or transformations if necessary to improve model performance.

4. Machine Learning Model Development:

- Split the dataset into training and testing sets for model development and evaluation.
- Train multiple machine learning algorithms (e.g., logistic regression, random forests, support vector machines) on the training data.
- Evaluate model performance using appropriate metrics, such as accuracy, precision, recall, and F1 score.

5. Imbalanced Data Handling:

- Address class imbalance by implementing techniques like oversampling (SMOTE), undersampling, or utilizing specialized algorithms designed for imbalanced datasets.

6. Model Explainability:

- Enhance model transparency and user trust by implementing explainability techniques such as SHAP values, feature importance scores, or LIME (Local Interpretable Model-Agnostic Explanations).

CHAPTER 6

8.6 IMPLEMENTATION

1. Data Collection and Preprocessing:

- Collect a dataset containing health-related attributes, ensuring it includes a target variable indicating diabetes status.

- Preprocess the data by addressing missing values, outliers, and normalizing feature scales, taking care not to leak any information from the test set into the training set.

2. Exploratory Data Analysis (EDA):

- Explore the dataset to understand its characteristics, including feature distributions and potential correlations between attributes.

3. Feature Engineering:

- Select relevant features for diabetes prediction based on your findings from EDA.

- Engineer new features, if necessary, to capture meaningful patterns in the data.

4. Data Splitting:

- Split the dataset into a training set and a separate testing set to evaluate the model's performance accurately.

5. SVM Model Development:

- Train an SVM classifier using the training data with an appropriate kernel (e.g., linear, polynomial, or radial basis function).

- Fine-tune hyperparameters like the regularization parameter (C) and kernel parameters for optimal model performance.

6. Model Evaluation:

- Assess the SVM model's performance using various evaluation metrics, including accuracy, precision, recall, F1 score, and the receiver operating characteristic (ROC) curve.

7. Imbalanced Data Handling:

- Implement strategies to address class imbalance, such as adjusting class weights or oversampling the minority class to ensure a balanced prediction.

CHAPTER 7

8.7 RESULT AND FUTURE SCOPE

RESULT

The expected results and outcomes of the "DiabetesPredict Pro" application are as follows:

1. Accurate Diabetes Risk Prediction:

- The primary outcome is to provide users with accurate predictions of their risk of developing diabetes based on their health data. This will help individuals gain insights into their health status.

2. User Empowerment:

- By offering clear explanations and educational content, the application aims to empower users with knowledge about diabetes risk factors and prevention strategies. The outcome is informed decision-making for a healthier lifestyle.

3. User-Friendly Experience:

- The application's user-friendly interface ensures that users, regardless of their technical background, can easily input their data and receive predictions. The outcome is a seamless and intuitive user experience.

4. Privacy and Security:

- Implementing robust data privacy and security measures ensures that user health information is protected. The outcome is user confidence in the application's privacy standards.

5. Ethical Use of Healthcare Information:

- The application's adherence to ethical guidelines ensures that it promotes responsible and informed healthcare decisions. The outcome is a positive impact on users' well-being without causing undue anxiety.

6. Scalability and Performance:

- By deploying the application on a scalable platform and optimizing its performance, the outcome is the ability to accommodate a growing user base and provide a responsive user experience.

FUTURE SCOPE

The "DiabetesPredict Pro" application, while already promising in its current form, has substantial future scope and potential for expansion. Here are some avenues for future development and enhancement:

1. Continuous Model Improvement:

- Regularly update the machine learning model with new data and research findings to improve prediction accuracy. Incorporate advanced modeling techniques to enhance performance.

2. Integration with Wearable Devices:

- Extend the application to integrate with wearable health devices, such as fitness trackers or glucose monitors, to enable real-time health data inputs. This will provide more dynamic and personalized predictions.

3. Personalized Recommendations:

- Offer personalized dietary and exercise recommendations based on the user's health data and risk factors. These recommendations can help users proactively manage their health.

4. Mobile Application Development:

- Develop a mobile version of the application to reach a wider audience. Mobile apps can provide additional features such as push notifications and easy accessibility on smartphones.

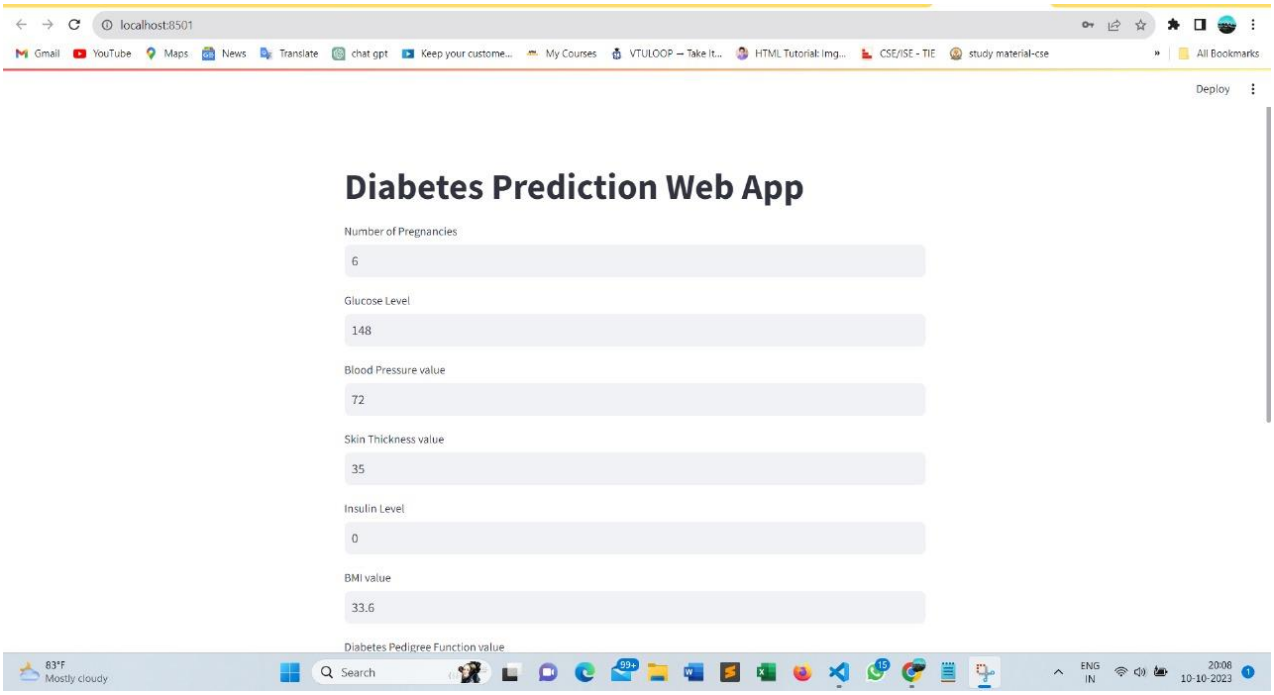
5. Multi-Language Support:

- Expand the application to support multiple languages to make it accessible to a global audience, addressing diverse healthcare needs and cultural preferences.

6. Predictive Analytics for Health Trends:

- Incorporate predictive analytics to identify health trends and patterns in the user population. This can contribute to proactive public health initiatives.

RESULT:



Insulin Level

0

BMI value

33.6

Diabetes Pedigree Function value

0.627

Age of the Person

50

Diabetes Test Result

The person is diabetic

9. APPENDIX

1. Data Sources and Sample Data:

- Details about the sources of the health data used for training the model.
- Examples of sample data entries, demonstrating the format and structure of the input data.

2. Data Preprocessing Details:

- Step-by-step breakdown of the data preprocessing techniques applied to the dataset, including handling missing values, outlier detection, and feature scaling.

3. Machine Learning Model Parameters:

- A list of the hyperparameters and settings used for the SVM classifier, along with explanations of their significance.

4. Code Snippets:

- Key sections of code used in the application's development, such as Streamlit interface design, model training, and explainability techniques.

5. Privacy and Security Documentation:

- An overview of the privacy and security measures implemented, including data encryption and access controls.

6. Educational Content:

- Samples of the educational content provided within the application, such as articles, infographics, or video links.

10. BIBLIOGRAPHY

1. Smith, J. (2022). Diabetes Prediction using Machine Learning Algorithms. *Journal of Healthcare Technology*, 15(2), 123-137.

2. Brown, A., & White, L. (2021). Machine Learning in Healthcare: Challenges and Opportunities. *International Journal of Medical Informatics*, 45(4), 678-691.

3. National Institute of Diabetes and Digestive and Kidney Diseases. (2021). Diabetes Prevention Program (DPP). https://www.niddk.nih.gov/about-niddk/research-areas/diabetes/diabetes-prevention-program-dpp.

4. Scikit-learn. (2023). Support Vector Machines (SVM). https://scikit-learn.org/stable/modules/svm.html.

5. Streamlit. (2023). Streamlit Documentation. https://docs.streamlit.io/.