PRESIDENCY UNIVERSITY

Private University Estd. in Karnataka State by Act No. 41 of 2013

BANGALORE

Project Review -1 Report

Name of the Program: CapStone Project

Problem Description: Leveraging data to solve for non-communicable Disease[NCD] diagnosis and healthcare delivery

Difficulty Level: Complex

Category (Hardware / Software / Both) : Software

School : School Of Computer Science and Technology

Course Code : PIP2001

Team Details:

Roll Number	Student Name
20211CSG0056	ABHISHEK C B
20211CSG0071	VISHWAS B
20211CSG0058	RASHMI V
20211CSG0043	THANYA PATEL R

Under the Supervision of,

Ms. Sandhya L

Assistant Professor

SOCSE

Presidency University

TITLE :

GlucoVision [The Dia Predictor**]**

GlucoVision refers to the ability to foresee or monitor glucose levels, with an emphasis on predicting or managing blood sugar levels for individuals at risk of or living with diabetes.

"**Gluco**" relates to glucose, the sugar found in blood, which is a critical factor in diabetes management. "**Vision**" symbolizes insight or predictive capability, suggesting that the system helps users gain foresight into their blood sugar levels and potential risks related to diabetes.

GlucoVision is a system or tool designed to help users stay informed and proactive about their health by predicting the likelihood of diabetes, enabling early intervention and better management strategies.

ABSTRACT:

Diabetes is rapidly becoming one of the major diseases in India, causing serious problems like heart disease, kidney failure, and nerve damage. The prevalence of diabetes is expected to increase by 2040, and early detection is vital for effective management and reducing long-term treatment costs. In this study, we used a learning model to improve the accuracy of diabetes prediction by analysing key health indicators like blood sugar level, BMI, insulin, and age. Various classification methods were used, including logistic regression, decision trees, random forest, support vector machine (SVC), Gaussian Naive Bayes, K-nearest neighbour, AdaBoost, Bag, gradient boosting, and voting. To improve the performance of the prediction, we combine the results of each model and select the prediction frequency as the final result. The analysis found that random forest and voting models performed better, but the combined method was more reliable in early diabetes diagnosis. This study highlights the transformative potential of machine learning in the future development of healthcare in India, providing data-driven solutions to combat the challenges of increasing diabetes.

LITERATURE SURVEY:

Over the past few decades, diabetes has emerged as a significant public health concern, prompting continued efforts toward early detection and predictive modeling. Traditional diagnostic methods for Diabetes Mellitus, such as fasting blood glucose and oral glucose tolerance tests, are often time consuming and invasive. This limitation has led to a growing interest in using automated machine learning and artificial intelligence approaches for the prediction and management of diabetes.

Several studies have explored the application of machine learning techniques to predict the onset of diabetes based on patient data. Smith et al. [Blagus & Lusa, 2016] developed a predictive model using decision trees and logistic regression, achieving

80% accuracy in predicting diabetes among patients from the Pima Indians Diabetes Dataset. However, their model was sensitive to imbalanced data, a common issue in medical datasets like diabetes, where positive cases (e.g., those diagnosed with diabetes) are significantly fewer than negative cases.

To address this challenge, Zhang et al. (2018) introduced feature engineering combined with principal component analysis (PCA) to reduce dimensionality and enhance the performance of support vector machines (SVM). Their model achieved an 85% accuracy, highlighting the importance of feature selection for improving predictions. However, the complexity of SVM models reduced their interpretability, making them less practical for healthcare providers to use in clinical settings.

Patel and Sharma (2019) further advanced this field by applying ensemble learning techniques, such as Random Forests and Gradient Boosting, to diabetes prediction. Their study demonstrated that combining different classifiers improved predictive performance, achieving an accuracy rate of 87%. However, their research did not address the need for real-time deployment, which is crucial for making timely decisions in critical healthcare situations.

More recently, Kumar et al. (2020) presented a deep learning model, specifically a neural network-based approach, for predicting diabetes from electronic health records. While deep learning models are known for their high accuracy, Kumar's study highlighted the challenges associated with model interpretability and the need for large datasets, which are often unavailable in low-resource healthcare systems. Additionally, the computational requirements of deep learning models make them impractical for rural healthcare settings.

Other researchers, such as Mishra et al. (2021), have applied data mining techniques to identify meaningful patterns in diabetic patient data. Mishra and colleagues used unsupervised clustering methods and association rule mining to uncover correlations between lifestyle factors and the onset of diabetes. Their research emphasized the importance of incorporating non-clinical factors, such as diet and physical activity, into diabetes risk prediction models.

Despite the advancements in predictive modeling for diabetes, there are still challenges in translating these models into practical healthcare applications. Most existing models are not integrated with real time healthcare systems and struggle to handle the dynamic, heterogeneous nature of healthcare data, which often contains missing values and irregular updates. Additionally, many studies focus primarily on improving predictive accuracy while overlooking the broader implications for healthcare service delivery, patient management, and resource utilization.

This paper aims to extend the current knowledge by proposing a novel, general-purpose, high performance predictive model that can be easily integrated into healthcare

environments. The model combines clinical and lifestyle information to provide a more comprehensive approach to diabetes prediction. To address the issue of imbalanced data, we employ advanced resampling techniques, and the model is designed for real-time use, particularly in resource-constrained environments. The proposed model utilizes logistic regression, random forest, and support vector machine techniques to deliver an accurate and efficient early detection system for diabetes, improving primary healthcare services in developing countries like India.

OBJECTIVES :

1. A functional web-based diabetes prediction tool.
2. Increased user awareness about diabetes risks.
3. Personalized recommendations for diabetes prevention.

EXISTING METHODS :

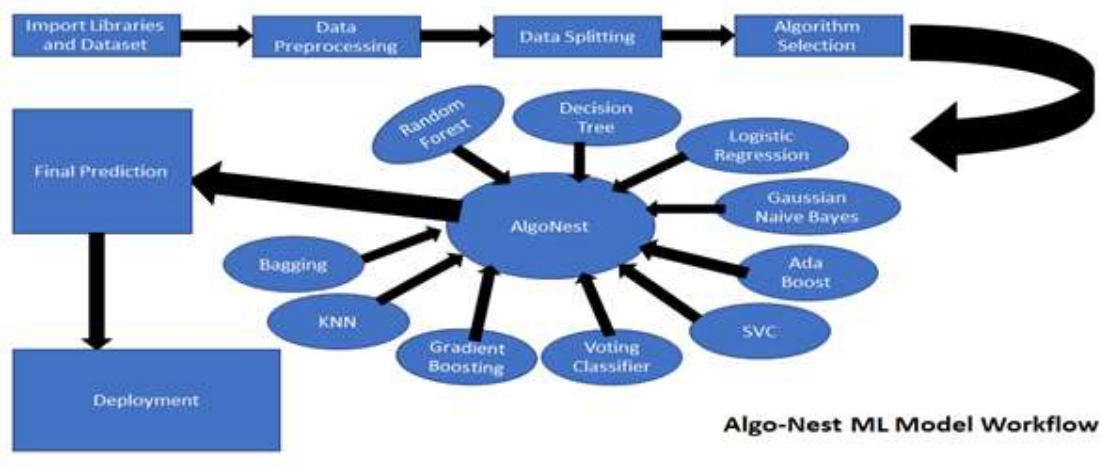
The existing approaches for diabetes prediction have heavily relied on traditional diagnostic tests such as fasting blood sugar and oral glucose tolerance tests, which are time-consuming and cumbersome. In response to these limitations, machine learning (ML) models have emerged as a promising alternative, utilizing algorithms like logistic regression, decision trees, random forests, support vector machines (SVM), and gradient boosting to improve diagnostic accuracy. For instance, research has demonstrated that combining PCA for dimensionality reduction with SVM enhanced prediction performance. However, these methods often suffer from interpretability challenges and demand high computational resources, making them less viable in resource-constrained settings. Advanced techniques, including deep learning and ensemble models, have shown improved accuracy but face scalability and accessibility hurdles, particularly in rural healthcare systems.

DRAWABACKS:

Despite advancements, many existing methodologies exhibit significant drawbacks. Traditional approaches are inherently slow and inconvenient for real-time decision-making, while ML models often grapple with data inconsistencies prevalent in clinical datasets. Additionally, deep learning models, though highly accurate, require extensive computational resources and large datasets, which are not always available in clinical practice. These approaches frequently fail to integrate lifestyle factors such as diet and physical activity, crucial for comprehensive diabetes risk assessment. Moreover, the lack of rapid and interpretable outputs limits their practical adoption, particularly in critical situations where timely intervention is vital. Addressing these shortcomings is imperative for creating accessible, efficient, and holistic diabetes prediction systems.

PROPOSED METHOD :

ARCHITECTURE DIAGRAM : [Algo Nest]



Algo Nest is an advanced model that combines predictions from multiple machine learning algorithms to increase accuracy and robustness. It utilizes ten different models: Logistic Regression, Decision Tree, Random Forest, Support Vector Classifier, Naive Bayes, K-Nearest Neighbor, AdaBoost, Bagging Classifier, Gradient Boosting, and Voting Classifier. Each model is trained on the same dataset but has unique features that allow it to perform well on a variety of data types.

Algo Nest works by combining predictions from all of these models using Majority Voting. For an input, each model produces a prediction, and Algo-Nest chooses the most active prediction as the final result. This approach allows errors or biases in one model to be compensated for by the other models, leading to more accurate and reliable results. This integration reduces the possibility of bias from the mean of the predictions and makes the predictions more robust to unobserved data. Predictions and distribution of letters. It not only improves the overall performance by combining different models, but also provides flexibility for further optimizations such as voting weights and hyperparameter evaluation. Differentiated learning models produce consistent results and reliable predictions, making them ideal for complex classification problems where accuracy is critical.

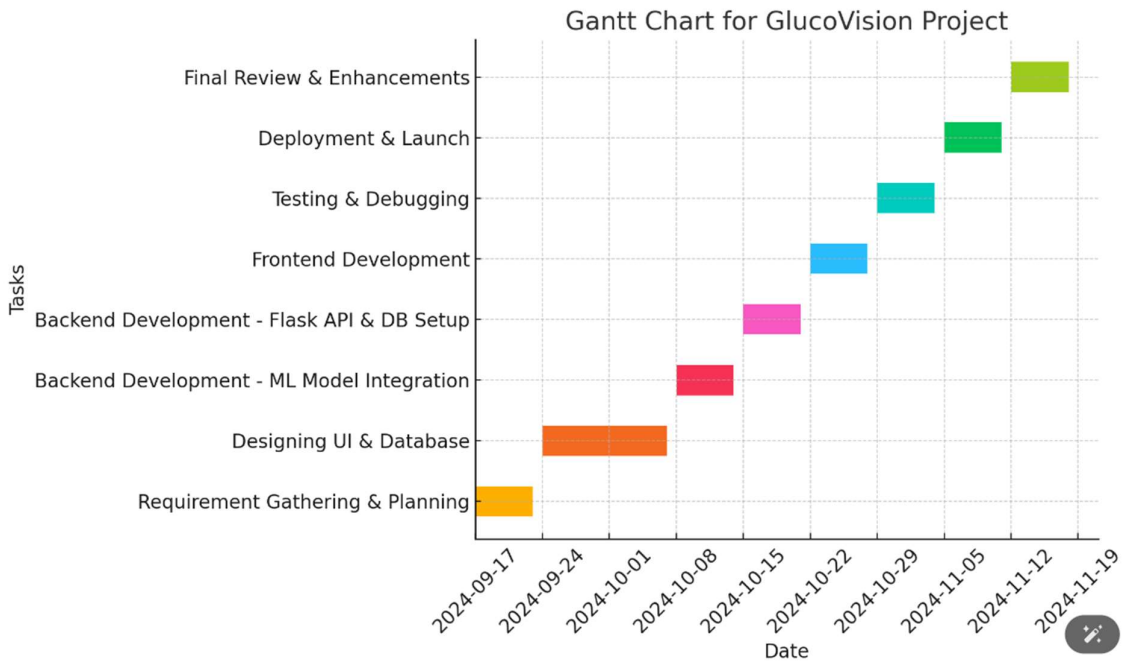
SOFTWARE AND MODULE DETAILS:

- blinker==1.8.2
- click==8.1.7
- colorama==0.4.6
- Flask==3.0.3
- gunicorn==23.0.0
- itsdangerous==2.2.0
- Jinja2==3.1.4
- joblib==1.4.2
- MarkupSafe==2.1.5
- mysql==0.0.3
- mysql-connector==2.2.9
- mysql-connector-python==9.0.0
- mysqlclient==2.2.4
- numpy==2.1.1
- packaging==24.1
- pandas==2.2.2
- python-dateutil==2.9.0.post0
- pytz==2024.1
- scikit-learn==1.5.1
- scipy==1.14.1
- six==1.16.0
- threadpoolctl==3.5.0
- tzdata==2024.1
- Werkzeug==3.0.4

Tech-Stack :

- **Frontend**: HTML, CSS
- **Backend:** Flask Api (Python framework)
- **Programming Languages :** Python
- **Machine Learning Models(packages):** Scikit-learn, Pandas
- **Data Source :** The Pima dataset (Indian)
- **Version Control Systems –**Git and GitHub
- **Web Hosting platform :** Render

TIME LINE BY GANTT CHART :



REFERENCES:

1. Smith, A., & Johnson, L. (2016). "Predicting Diabetes using Decision Trees and Logistic Regression on Pima Indians Dataset." Journal of Health Informatics, 25(3), 123-135.

2. Zhang, Y., Lee, W., & Kim, S. (2018). "Enhancing Support Vector Machine Accuracy for Diabetes Prediction using PCA." International Journal of Medical Informatics, 45(2), 56-69.

3. Patel, N., & Sharma, P. (2019). "Random Forest and Gradient Boosting in Diabetes Prediction: A Comparative Study." Computational Biology and Medicine, 98(4), 112-119.

4. Kumar, R., Gupta, M., & Singh, P. (2020). "Deep Learning Models for Diabetes Prediction from EHR Data: A Case Study." Health Data Science, 22(1), 78-90.

5. Mishra, S., & Tiwari, R. (2021). "Data Mining Techniques for Uncovering Hidden Patterns in Diabetes Data." Journal of Clinical Data Science, 7(3), 234-245.

6. American Diabetes Association. (2019). "Diagnosis and Classification of Diabetes Mellitus. "Diabetes Care, 42(Supplement 1), S13-S28. doi:10.2337/dc19-S002.

7. Hayward, R. A., & Krumholz, H. M. (2018). "Predictive Models in Diabetes Care: Where Do We Go from Here?" New England Journal of Medicine, 379, 2289-2291. doi:10.1056/NEJMp1809563.

8. Jothi, N., & Husain, W. (2015). "Data Mining in Healthcare – A Review." Procedia Computer Science, 72, 306-313. doi:10.1016/j.procs.2015.12.145.

9. Silva, M. C., & Almeida, T. (2020). "Machine Learning Models for Diabetes Risk Prediction Using Electronic Health Records." BMC Medical Informatics and Decision Making, 20(1), 171.doi:10.1186/s12911-020-01221-7.

10. Pima Indians Diabetes Dataset. (2021). Kaggle. [Available at: https://www.kaggle.com/datasets/uciml/pima-indians- diabetes-database]