March 21, 2025

C964: Computer Science Capstone Project

Hao Chen

010771133

Machine Learning Diabetes Prediction Application

Part A: Letter of Transmittal

May 21st, 2025

Joe Biden

Madagascar Family Hospital Corporation

101 Washington Street, Washington, WS 20909

Dear Mr. Biden,

It has come to my attention as of late that our hospitals are facing an unprecedented number of patients with diabetes. Even with the recent hirings, our staff and doctors are all severely overworked and can not keep up with the mass number of patients who have arrived expressing symptoms of diabetes. In essence, the problem boils down to the mass delayed diagnosis of diabetes, which in turn leads to more costly treatments and worse symptoms for our patients as they go longer with untreated and undiagnosed diabetes.

The proposed application uses available patient data (glucose levels, BMI, age, etc.) and a trained Random Forest machine learning model to assess the likelihood of diabetes, based on the Pima Indian Diabetes dataset. This application was developed using Python and hosted in Streamlit, making it robust and lightweight. Furthermore, it features an interactive dashboard and a simplistic user interface making it perfect for any medical staff to pick up and use.

This solution benefits Madagascar Family Hospitals by supporting earlier interventions, reducing long-term treatment costs, and offering a simple screening aid where lab work is not readily available. The entire solution is cost-effective and deployable on a local machine or through a free cloud service (Streamlit Cloud), with minimal technical setup.

The total development was completed over a four-week period. The only dataset used was publicly available, and no personal or identifiable health data was involved. There are no ethical concerns present, as the application does not provide a diagnosis or store any data — it purely offers a predictive tool based on historical trends. The entire model is reproducible and maintainable, with monitoring capabilities included via Python logging and version control on GitHub. The solution is low cost, with its primary costs being just hosting the application on a cloud service provider as volume increases.

Sincerely,

Hao Chen

Hao Chen, Chief Executive Officer

Part B: Project Proposal Plan

Project Summary

The project will address the critical issue of delayed and often overlooked diabetes diagnosis, which contributes to poor patient outcomes and increased healthcare costs. The client, Madagascar Family Hospital, requires a solution that improves early detection of diabetes risk while integrating seamlessly into existing workflows. To meet this need, a web-based application will be developed to predict diabetes risk based on key health metrics including glucose levels, BMI, and age. This application will incorporate a trained Random Forest Classifier, present predictions clearly, and include supportive visualizations such as a confusion matrix and comparative histograms to facilitate interpretation.

The final deliverables will consist of the complete application, a trained machine learning model, the supporting dataset used for model training, and a user guide outlining installation, usage, and interpretation of outputs. An optional cloud deployment will be provided for easy access via a secure web link. The proposed solution will enable the client to reduce reliance on manual analysis, increase prediction accuracy, and improve patient outcomes by supporting timely interventions. The project will align with the client's goal of utilizing data science to enhance operational efficiency and patient care.

Data Summary

The data for this project will be sourced from the publicly available Pima Indian Diabetes dataset, which includes anonymized health data of female patients aged 21 and older. This dataset will be collected in CSV format and stored locally within the project directory to facilitate processing and model training. During the design and development phases, the data will be cleaned to handle missing values and outliers, such as replacing or dropping records with invalid measurements and normalizing the feature distributions. The development process will incorporate clear versioning of both the raw and processed data to ensure reproducibility and data integrity. The data meets the needs of the project because it contains all relevant health metrics used to train and validate the machine learning model, and its features align with real-world indicators of diabetes risk. No sensitive or personally identifiable information is present in the dataset, and no new data will be collected from users, ensuring there are no ethical or legal concerns related to privacy or data handling. The dataset will be retained solely for project reproducibility and demonstration purposes.

Implementation

The project will adopt an Agile methodology, incorporating iterative development and continuous feedback to refine the application. The implementation will begin with the design of the data preprocessing and model training pipeline using the Pima Indian Diabetes dataset. This will include feature selection, data cleaning, and model training using a Random Forest Classifier. Once the model demonstrates satisfactory performance, a web-based user interface will be developed with Streamlit, integrating interactive forms, prediction outputs, and visualizations such as histograms, feature importance charts, and a confusion matrix. The final phase will focus on deployment and testing, including error logging and optional cloud hosting via Streamlit Cloud. The entire project will be version-controlled using GitHub to ensure traceability and reproducibility.

Timeline

Milestone or deliverable	**Project Dependencies**	**Resources**	**Start and End Date**	**Duration**
Data Acquisition and Preprocessing	N/A	Python, Pandas	06/01/2025 - 06/03/2025	3 Days
Model Training and Validation	Data Preprocessing Completed	Scikit-learn, Joblib	06/06/2025 - 06/10/2025	4 Days
Streamlit App Development	Machine Learning Model	Streamlit, Matplotlib	06/10/2025 - 06/15/2025	5 Days

	Ready and Trained			
Deployment and Testing	Streamlit App Ready and Developed	GitHub, Streamlit Cloud	6/16/2025 - 6/17/2025	2 Days
Documentation and Maintenance Review	Streamlit App Tested and Deployed	N/A	6/18/2025 - 6/20/2025	2 Days

Evaluation Plan

Evaluation will be performed at each development stage to make sure that correctness and completeness standards are met. During the Data Acquisition and Preprocessing stage, we will extract random samples of the data and manually verify that those samples are properly preprocessed and proper handling of missing or invalid values and outliers are met. The next stage, Model Training and Development, the model will be verified using accuracy metrics including but not limited to precision, recall, F1 score, and confusion matrix generated from the validation data. The Streamlit App Development stage will have its functionality evaluated by running the app on a local computer and all its features tested thoroughly and have its logic flows examined closely. The Deployment and Testing phase will be evaluated by monitoring the application uptime with varying levels of activity and traffic. We will also be entering controlled inputs and comparing these outputs against known expectations to verify that the logic persists once hosted to the cloud. Documentation and Maintenance reviews will be thoroughly examined and checked over to make sure it adheres to the company standards and guidelines. Upon project completion, validation will involve running the final application with a set of unseen test data to confirm consistent prediction accuracy and visualizations. User acceptance testing will be conducted to ensure the interface is intuitive and delivers correct results.

Costs

The project will incur minimal hardware and software costs. Python and its libraries (e.g., Streamlit, Scikit-learn, Pandas) are open-source and free to use, and the development will be conducted on existing hardware. Estimated labor will include 40 hours of developer time at an internal rate of $50/hour, totaling $2,000. Environment costs will be minimal, as Streamlit Cloud provides free hosting for public apps, and local testing will be conducted on personal hardware. Optional expenses for cloud deployment beyond free tiers or for advanced monitoring tools are estimated at $20–$50 per month, if needed for future scaling. No additional software licenses or proprietary tools will be required.

Part C: Application

The completed application can be found and used on https://c964-machine-learning-capstone-hao-chen.streamlit.app/

Part D: Post-implementation Report

Solution Summary

The project addressed the challenge of delayed diabetes diagnosis by delivering a machine learning–based application that predicts diabetes risk. The application was built using Python and Streamlit and integrated a Random Forest Classifier trained on the Pima Indian Diabetes dataset. It accepted user inputs for key health metrics such as glucose, BMI, and age, and returned a risk prediction along with visual explanations including a confusion matrix, feature importance chart, and comparative histograms. This solution provided healthcare professionals with an accessible decision-support tool for early detection of diabetes risk, thereby improving patient outcomes and reducing long-term treatment costs.

Data Summary

The raw data used for this project was sourced from the publicly available Pima Indian Diabetes dataset, which contains anonymized health data for female patients aged 21 and older. The data was collected in CSV format and stored locally. During development, the data was cleaned and processed to handle missing values and outliers by removing or imputing invalid records. Feature distributions were normalized, and the dataset was split into training, validation, and testing subsets to support model development and evaluation. The data was maintained in version-controlled environments to ensure traceability and reproducibility. No personally identifiable information was present, and no additional data collection was performed, ensuring compliance with ethical standards.

Machine Learning

The project used a Random Forest Classifier as the machine learning model to predict diabetes risk. This model was selected for its robustness, ability to handle non-linear relationships, and strong performance on classification tasks. The model was trained using key features from the Pima dataset, including glucose levels, BMI, insulin, blood pressure, and age. Development involved splitting the dataset into training, validation, and testing sets, cleaning the data, and tuning model parameters to optimize accuracy. Performance metrics such as precision, recall, F1 score, and a confusion matrix were used to validate the model, and results demonstrated a validation accuracy of approximately 74%, with particularly high precision and recall for non-diabetic predictions.

The Random Forest Classifier was chosen because of its interpretability, feature importance analysis capabilities, and its proven effectiveness in handling complex datasets with potential outliers or missing data. Its ensemble approach, combining multiple decision trees, provided strong generalization and reduced overfitting, making it well-suited for this health-related classification task.

Validation

The model used in this project was a supervised machine learning algorithm, specifically a Random Forest Classifier, which is well-suited for classification tasks with structured data. To validate the model's performance, standard supervised learning metrics were employed, including accuracy, precision, recall, F1 score, and a confusion matrix.

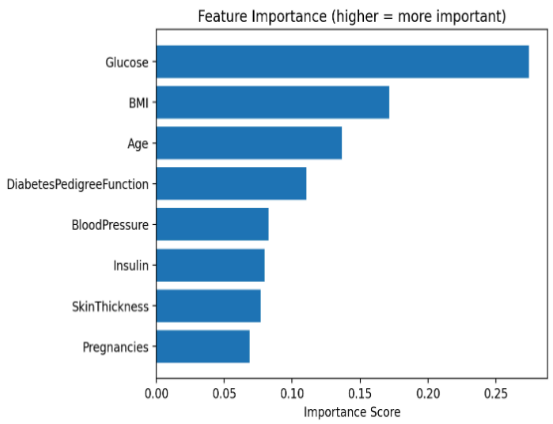
Validation involved splitting the dataset into training, validation, and testing sets. After training on the training set, the model's predictions were evaluated on the validation set. The following results were achieved: a validation accuracy of approximately 74%, a precision of 83% and recall of 80% for non-diabetic cases, and a precision of 57% and recall of 60% for diabetic cases. The F1 scores for both classes were 0.82 and 0.58, respectively. These results indicated that the model was effective at identifying non-diabetic patients while moderately successful at catching diabetic cases.

The confusion matrix visualization provided clear insights into the model's prediction strengths and weaknesses. Since the data contained clear categorical labels for outcomes (diabetic or non-diabetic), supervised learning validation techniques were appropriate, and no unsupervised methods were needed for this project.

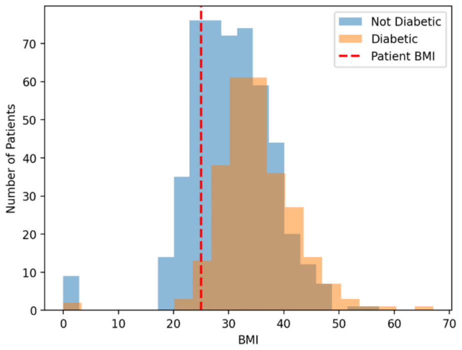
Visualizations

Upon entering the required inputs for the application, the application automatically redirects the user to the results page where it will let the user know what the machine learning model predicts whether the user is diabetic or not. The page also contains 4 unique visualizations which are the Confusion Matrix, Glucose Level Comparison Histogram, BMI Comparison Histogram, and a Feature Importance Graph.

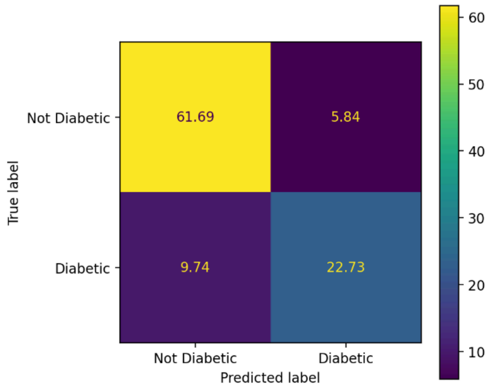
Feature Importance (Model Insights)



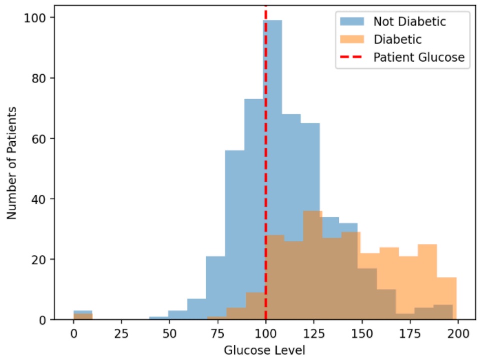
BMI Comparison



Model Evaluation: Confusion Matrix



Glucose Level Comparison



User Guide

Online Cloud Access

1. Use the link https://c964-machine-learning-capstone-hao-chen.streamlit.app/
2. Enter in patient information such as Pregnancies, BMI, Glucose Levels, Skin Thickness, Insulin, Age
3. Press Submit
4. Application automatically redirects to the results page
5. Client/User can see whether the model predicted the patient to be diabetic or not along with visualizations of the model and how the patient compared to distributions

Local Computer Access (No Internet)

1. Download and install Python 3.8 or later from the official website https://www.python.org/downloads/
2. Install required libraries from the requirements.txt file included with the project. To do so open a terminal and run the command “pip install -r requirements.txt” (remove quotation marks)
3. Run the application by running the command “streamlit run user_app.py” (remove quotation marks)
4. Application should automatically open in the default browser at http://localhost:8501
5. Enter in patient information such as Pregnancies, BMI, Glucose Levels, Skin Thickness, Insulin, Age
6. Press Submit
7. Application automatically redirects to the results page
8. Client/User can see whether the model predicted the patient to be diabetic or not along with visualizations of the model and how the patient compared to distributions