Task 2 parts A, B, C and D

[Part A: Letter of Transmittal 2](#_Toc98598250)

[Letter of Transmittal Requirements 2](#_Toc1738085866)

[Summary of the Problem 2](#_Toc1738085866)

[Benefits of the Data Product 2](#_Toc1738085866)

[Outline of the Data Product 2](#_Toc1738085866)

[Description of the Data 2](#_Toc1738085866)

Objectives and Hypothesis 3

[Project Methodology 3](#_Toc1738085866)

[Funding Requirements 3](#_Toc1738085866)

[Impact on Stakeholders 3](#_Toc1738085866)

[Ethical and Legal Considerations 3](#_Toc1738085866)

Expertise 4

Part B: Project Proposal Plan 5

[Decision Support Problem](#_Toc1738085866) 5

[Description of Customer and their Needs](#_Toc1738085866) 5

[Decision and Support Problem](#_Toc1738085866) 5

[Existing Gaps in Data Products](#_Toc1738085866) 5

[Available Data](#_Toc1738085866) 5-6

[Deliverables 6](#_Toc904507251)

[Implementation Plan 6](#_Toc904507251)

[Validation and Verification Methods 6](#_Toc904507251)

[Programming and Environment Costs 6](#_Toc904507251)-7

Projected Timeline 7

[Part D: Post-implementation Report 8](#_Toc651895932)

[Business Vision](#_Toc201059345) 8

[Visualizations 9-11](#_Toc201059345)

[User Guide 12](#_Toc1365484010)

# Part A: Letter of Transmittal

**Letter of Transmittal**

**To:** Senior Management

**From:** Eduardo Ramirez

**Date:** 9/28/2024

**Company: Di-a-bee Analysis**

**Subject:** Proposal for Implementation of a Predictive Data Model for Diabetes Diagnosis.

**Project Proposal for Predictive Diabetes Diagnosis Model**

1. **Summary of the Problem**

Diabetes is a chronic health condition that affects millions globally, and early diagnosis is critical for effective management and prevention of complications. Given clinical data about a patient, this project aims to predict whether they are likely to have diabetes. Current diagnostic methods often rely on a range of medical tests, which can be costly and time-consuming. With advancements in machine learning, we can streamline this process by using predictive analytics to support quicker, data-driven decisions.

1. **Benefits of the Data Product**

The proposed predictive data model will benefit healthcare professionals by providing:

* **Improved Efficiency:** Automated predictions based on clinical parameters allow doctors to quickly assess the likelihood of diabetes, leading to faster decision-making.
* **Cost Savings:** By reducing the number of unnecessary tests, healthcare providers can save costs while focusing on high-risk patients.
* **Data-Driven Decisions:** It supports decision-making by providing accurate predictions based on patient data.

1. **Outline of the Data Product:** The data product is a machine learning model that predicts the likelihood of a patient having diabetes based on clinical parameters such as glucose levels, blood pressure, BMI, etc.… The model is trained using supervised learning techniques, including logistic regression, support vector machines (SVM), and K-nearest neighbors (KNN). It will be integrated into healthcare systems to assist medical professionals in diagnosing diabetes.
2. **Description of the Data:** The dataset used in this project is sourced from the National Institute of Diabetes and Digestive and Kidney Diseases and is available on Kaggle: <https://www.kaggle.com/datasets/mathchi/diabetes-data-set/data>. The dataset includes the following features: Pregnancies, glucose concentration, insulin levels, body mass index (BMI), and a binary outcome indicating whether the patient has diabetes or not.
3. **Objectives and Hypotheses:**

* **Objective:** To develop a machine learning model that accurately predicts whether a patient has diabetes based on clinical data.
* **Hypothesis:** A predictive model using features such as glucose levels, BMI, and insulin levels, and number of pregnancies will have an accuracy of 75% in predicting diabetes.

1. **Project Methodology:**

**The methodology follows a standard machine learning workflow:**

* **Data Collection and Preprocessing:** 
  + Gather data from the diabetes dataset.
  + Clean and preprocess the data for analysis.
* **Exploratory Data Analysis (EDA):**
  + Perform EDA using Pandas, Seaborn, and Matplotlib to uncover patterns and correlations in the data.
  + Visualize the data using bar graphs, scatter plots, correlation matrices, and other types of charts.
* **Model Development:** 
  + Train machine learning algorithms such as Logistic Regression, SVC, and KNN to build predictive models.
* **Model Optimization:** 
  + Using GridSearchCV and RandomizedSearchCV to fine-tune the model’s hyperparameters.
* **Evaluation:** 
  + Evaluate the model using accuracy, precision, recall, and F1-score metrics to ensure reliable predictions.

1. **Funding Requirements:**

* **Development Tools:** There will be no cost for development tools. I will be using Jupyter Notebook which is open source.
* **Google Cloud Platform (GCP):** There will be a cost in using the GCP services. Estimated costs are approximately $5,000.

1. **Impact on Stakeholders:**

* **Patients:** Early diagnosis leads to better health outcomes, with preventive measures taken before the disease progresses.
* **Healthcare Providers:** This tool will aid doctors in making faster and more accurate diagnoses, allowing them to prioritize patients at risk.
* **Management:** Reduced operational costs due to fewer unnecessary tests, leading to overall healthcare savings.

1. **Ethical and Legal Considerations:** Given the sensitivity of the data, the project will adhere to strict data privacy regulations, including adhering to ethical guidelines and legal regulations such as HIPAA. Data will be anonymized to protect patient privacy, and informed consent will be obtained for data usage. Transparency will be provided to clearly explain how model predictions work to avoid any bias in decision-making.
2. **Expertise:** I have extensive experience with Python, data visualization, and machine learning. My expertise includes developing predictive models, data analysis, and integrating machine learning solutions.

By implementing this predictive data model, we can improve diagnostic accuracy, reduce operational costs, and promote proactive healthcare measures. I hope this proposal provides a comprehensive overview of the project and its potential benefits.

Sincerely,

Eduardo Ramirez

# Part B: Project Proposal Plan

**1. Decision Support Problem:** The primary decision support problem we are addressing is the need for early detection of diabetes in patients based on clinical parameters. Healthcare professionals need efficient tools to predict diabetes risk based on patient clinical data. Early diagnosis is crucial for effective management and prevention of sever complications associated with diabetes.

**2.** **Description of Customers and Their Needs:**

* **Healthcare Providers:** They require tools that can assist in diagnosing patients efficiently, enabling early intervention in diabetes care. (doctors, nurses, and specialists)
* **Patients:** Early and accurate diagnosis enables timely intervention and improving patient outcomes.
* **Healthcare Administrators:** Need tools to reduce the cost and time associated with extensive testing while improving diagnostic accuracy.

**3. Existing Gaps in Data Products:** Currently, the diagnostic process for diabetes involves multiple clinical tests that can be expensive and time-consuming. There being a lack of efficient and accurate predictive models for diabetes diagnosis in clinical settings. The key gaps are:

* Lack of automation in early diagnosis
* Absence of tools that integrate predictive analytics to assess diabetes risk using existing clinical data.

This product addresses these gaps by using machine learning to predict diabetes risk more efficiently and accurately.

**4. Available Data:** The data used for this project is sourced from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset I’m using was downloaded from this source in Kaggle: <https://www.kaggle.com/datasets/mathchi/diabetes-data-set/data>. The data set includes features such as pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, age, and target (outcome).

**5.** **Methodology:**

* **Data Collection:** Gather data from the specified source.
* **Data Preprocessing:** Clean and preprocess the data for analysis.
* **Exploratory Data Analysis (EDA):** Visualizing the relationships and patterns in the dataset to guide model selection.
* **Model Development**: Develop and train machine learning models using the dataset. Machine learning models being used are Logistic Regression, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN).
* **Model Hyperparameter Tuning**: Using RandomizedSearchCV and GridSearchCV to optimize the model’s hyperparameters.
* **Model Evaluation**: Evaluate the model’s performance using metrics such as accuracy, precision, recall, and F1-score using cross validation.

**6. Deliverables**

* **Predictive Data Model:** A machine learning model capable of predicting the likelihood of diabetes based on patient clinical parameters.
* **Data Visualizations and Exploratory Data Analysis (EDA):** Visual insights into the key features influencing diabetes risk.
* **Model Performance Report:** A detailed analysis of the model’s performance, including accuracy, precision, recall and F1-score metrics.
* **Documentation:** A detailed report on the model’s performance and potential improvements.

**7. Implementation Plan**

* **Phase 1: Data Collection and Preprocessing (10 hours):** Gathering and cleaning the dataset.
* **Phase 2: Exploratory Data Analysis (EDA) (20 hours):** Visualizing the data to identify key trends and patterns.
* **Phase 3: Model Development ( 40 hours):** Training various machine learning models on the dataset.
* **Phase 4: Hyperparameter Tuning (20 hours):** Optimizing the model using GridSearchCV and RandomizedSearchCV.
* **Phase 5: Model Validation (20 hours):** Testing the model on thetest set and evaluating its performance metrics.
* **Phase 6: Deployment (30 hours):** Deployment of flask application using Google Cloud Platform (GCP).

**8. Validation and Verification Methods**

To ensure that the developed data product meets the requirements and the needs of the customers, we will employ the following methods:

* **Cross-Validation:** Accuracy, precision, recall, and F1-scores will be cross-validated to ensure reliable model performance.
* **Confusion Matrix:** Will be used to evaluate the models true positives and true negatives, ensuring minimal errors in predictions.

**9. Programming Environments and Costs**

* **Programming Language:** Python: for data analysis and model development
* **Development Tools:** Jupyter Notebook: For interactive data analysis and visualization
* **Scikit-learn:** Machine learning library for model development
* **Google Cloud Platform:** Cloud Provider that will host the Diabetes App predicting application.
* **Pandas and NumPy:** Libraries for data manipulation and analysis
* **Matplotlib and Seaborn:** Libraries for data visualization

**Costs:**

* **Cloud Computing Resources:** $5,000
* **Data Scientist:** $8,000
* **ML Engineer:** $5,000
* **IT:** $2,000

Total: $20,000

**10. Projected Timeline:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Milestone | Start Date | End Date | Duration (Hours) | Dependencies | Resources Assigned |
| Data Collection and Preprocessing | **09/27/2024** | **09/27/2024** | **10** | **None** | **Data Scientist** |
| Exploratory Data Analysis | **09/28/2024** | **09/28/2024** | **20** | **Data Collection** | **Data Scientist** |
| Model Development | **09/30/2024** | **10/01/2024** | **40** | **Data Preprocessing** | **Data Scientist, ML Engineer** |
| Hyperparameter Tuning | **10/02/2024** | **10/03/2024** | **20** | **Model Development** | **Data Scientist** |
| Model Validation | **10/04/2024** | **10/05/2024** | **20** | **Hyperparameter Tuning** | **Data Scientist** |
| Deployment and Testing | **10/07/2024** | **10/10/2024** | **24** | **None** | **IT, Data Scientist** |
| Total Hours: 144 |  |  |  |  |  |

# Part D: Post-implementation Report

**D. Documentation for the Predictive Diabetes Diagnosis Model**

**1. Business Vision**

The vision of this project is to develop a machine learning-based data product that predicts the likelihood of diabetes in patients based on clinical parameters. By providing an automated and efficient diagnostic tool, healthcare institutions will save time, reduce operational costs, and enhance the accuracy of diabetes diagnosis. Ultimately improving patient outcomes and reducing healthcare costs.

**3. Code for Data Analysis (descriptive and predictive)**

**Correlation Matrix**

**A white background with a couple of people

Description automatically generated with medium confidence**

**Bar Graph:** Diabetes per Number of Pregnancies

A computer code with colorful text

Description automatically generated with medium confidence

**Histogram:**

A computer screen shot of a computer code

Description automatically generated

**Line Graph**

A computer screen shot of a code

Description automatically generated

**Box, Count, and Distribution plot**

**A white rectangular object with a black border

Description automatically generated**

**Predictive Code:**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A white background with blue and black text

Description automatically generated**

**A close-up of a computer screen

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A computer screen shot of a computer code

Description automatically generated**

****

**A screenshot of a computer program

Description automatically generated**

**4. Assessment of Hypotheses for Acceptance or Rejection**

**Hypothesis:** A predictive model using clinical features such a glucose, insulin, BMI, and blood pressure can reliably predict diabetes.

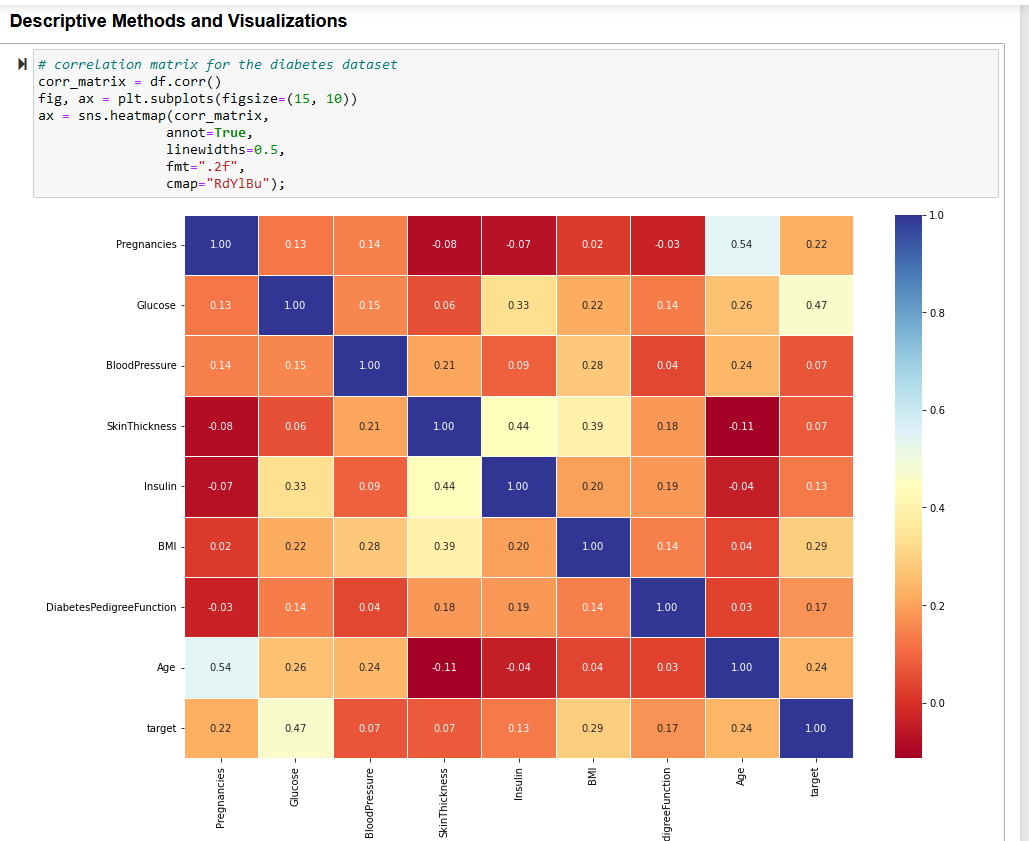
**Assessment:**

* **SVC**: showed comparable performance to Logistic Regression, but there was no significant improvement after hyperparameter tuning.
* **Logistic regression**: performed the best, with an accuracy of around 77%
* **KNearest Neighbors (KNN):** performed the lowest with an accuracy of approximately 73%.

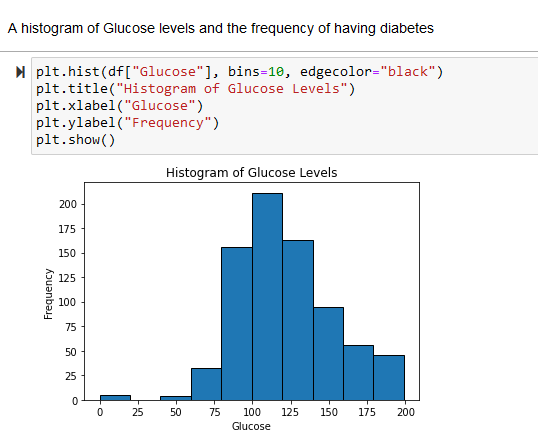
**Conclusion:** The hypothesis is accepted based on the accuracy of the models, particularly Logistic Regression, in predicting diabetes using clinical data.

**5. Visualizations and Storytelling**

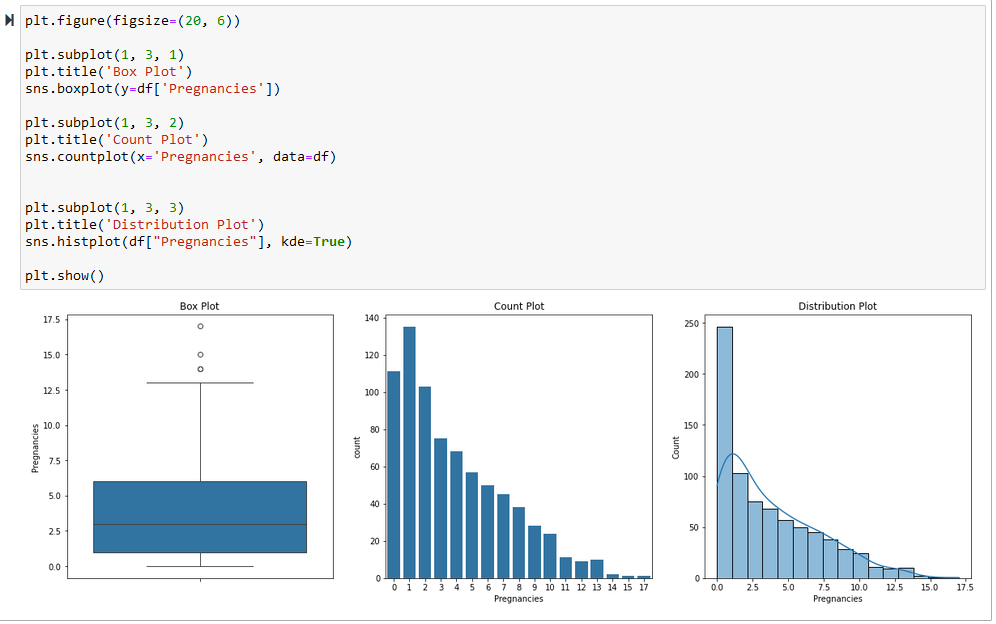
**Correlation Matrix:**



**Histogram of Glucose Levels**



**Box, Count, and Distribution plots of Pregnancies**

****

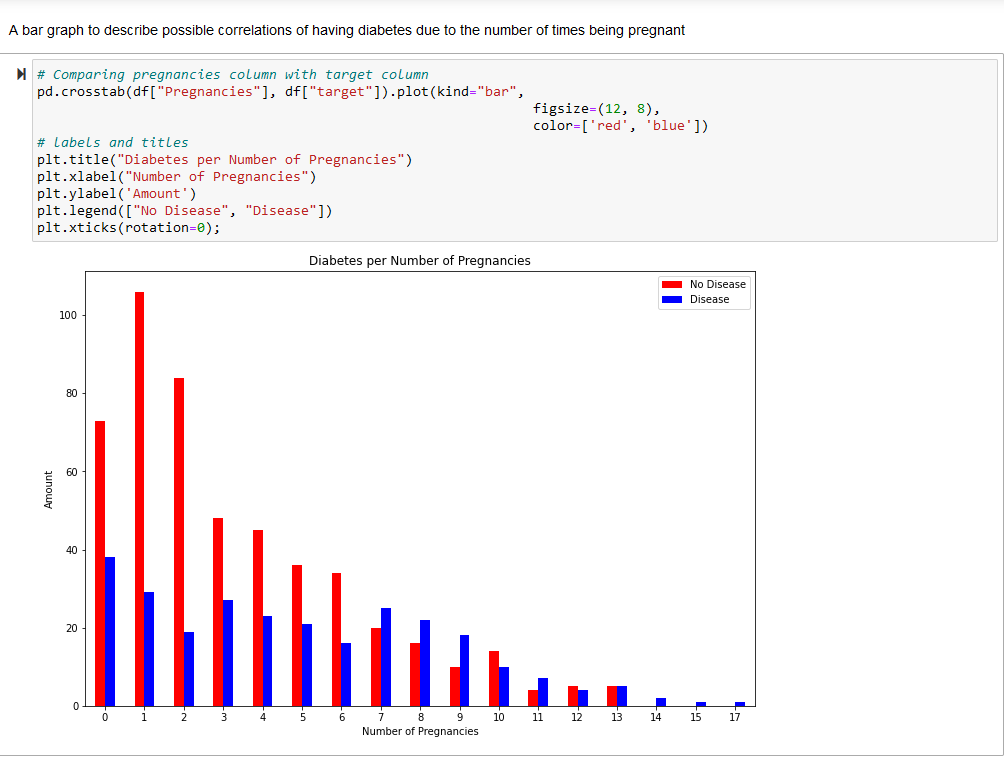
**Bar Plot for Pregnancies vs. Diabetes**

**6. Assessment of Product’s Accuracy**

**Logistic Regression:**

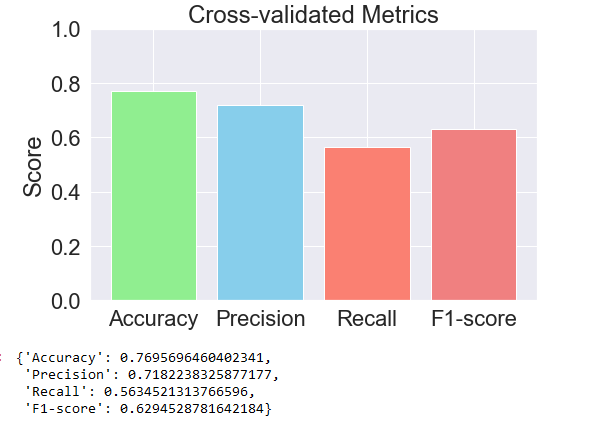
* **Accuracy: 0.77%**

**7. Data Product Testing, Revisions, and Optimization**

**Testing Results:**

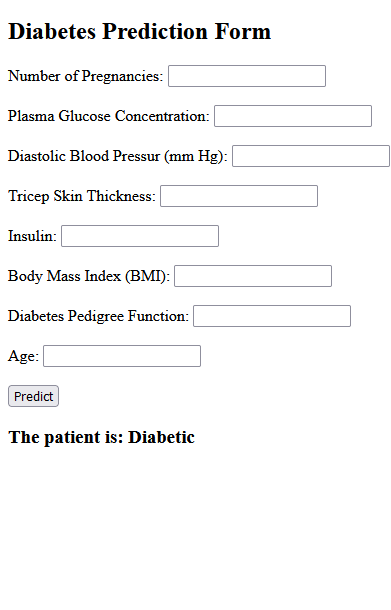
The model was tested using cross-validation with an average accuracy of 77%. Hyperparameter tuning using GridSearchCV and RandomizedSearchCV for Logistic Regression and SVC was conducted, but no significant improvement was observed for the SVC model.

**Revisions:** Based on testing, we opted to focus on Logistic Regression as the primary model for deployment due to its simplicity and strong performance

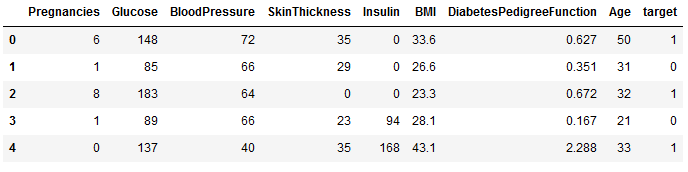


**8. A quick-start guide:**

* Visit the website: <https://flask-diabetes-predictor.uc.r.appspot.com/predict>
* Make sure to fill out all the boxes with numbers.



* + **You can use the image below as an example to know what numbers to input into the empty clinical parameter boxes in the application.**



* **Once you have filled out all the boxes you must hit the predict button and it will predict whether a patient has diabetes or not.**

