C964 Java Capstone

Vincent Wong

Student ID: 011319241

11/22/2023

Diabetes Classification Application

Table of Contents

[Part A: Letter of Transmittal 3](#_Toc151507915)

[Letter of Transmittal 3](#_Toc151507916)

[Part B: Project Proposal Plan 5](#_Toc151507917)

[Project Summary 5](#_Toc151507918)

[Data Summary 5](#_Toc151507919)

[Implementation 6](#_Toc151507920)

[Timeline 7](#_Toc151507921)

[Evaluation Plan 8](#_Toc151507922)

[Resources and Costs 8](#_Toc151507923)

[Part C: Application 10](#_Toc151507924)

[Part D: Post-implementation Report 11](#_Toc151507925)

[Solution Summary 11](#_Toc151507926)

[Data Summary 11](#_Toc151507927)

[Machine Learning 12](#_Toc151507928)

[Validation 13](#_Toc151507929)

[Visualizations 14](#_Toc151507930)

[User Guide 16](#_Toc151507931)

[Reference Page 18](#_Toc151507932)

# Part A: Letter of Transmittal

## Letter of Transmittal

11/20/2023

Some Diabetes Clinic LLC  
12345 East 123rd Street  
Random Place, ABC, 12345

Subject: Proposal to Increase Patient Experience and Satisfaction

Dear Some Person,

Hello, I am writing to you to offer a solution to a problem I’ve recently observed while accompanying a female relative of mine to your clinic. My relative was referred to your clinic because her general physician suspected that she might have diabetes but was unsure and suggested she get a specialist opinion and diagnosis. However, upon arriving, we observed a packed lobby and waited for an excruciating long length of time only to be told by the attending physician that she did not have diabetes. Excellent news I say, but as a patient, the waiting experience left much to be desired.

I am a veteran machine learning engineer with a decade of experience, I believe I can utilize a machine learning model to aid your physicians in diagnosing whether a female patient has diabetes or not to expediate the process leading to a shorter wait time in the lobby for the patients. Of course, my solution is not aimed to replace the analysis and expertise of a medical professional, but merely aid them as a preliminary analysis tool, so they can utilize their time more efficiently with female patients that are more likely to have diabetes. Not only will the model help in diagnosing diabetes, but I also aim to create the model to provide data to help physicians determine what are the key features leading to a positive diagnosis of diabetes. This knowledge can potentially impact the medical field with farther studies and research to advance the topic. The faster diagnosis time would also lead to a greater patient experience instead of the excruciating long wait time my relative and I had to go through.

Of course, since the application will be using patient information to determine if they have diabetes or not, we must follow proper federal laws to ensure patient information is secured and protected (Centers for Disease Control and Prevention, 2022).

With your best interests in mind, the project to build the application will focus on reducing your spending. I aim to build the application on a web accessible notebook like Jupyter Notebook or Google Colab, both of which are free for you the client to use in your clinic if you have an internet connection. There will be no licensing fees. Hardware costs can be negated if your current technological systems already have the required specifications to run the application on the notebook. We can farther save money on not utilizing a database since we will not be saving the data and result to secure patient information. No personal or sensitive patient information is needed to use the application. The dataset used to train the model can also be acquired for free through websites like <https://www.kaggle.com/>. If you so decide to take up on my offer, an estimated initial breakdown of the project cost is as follows:

* Labor cost for building the application - $50,000.
* Hiring testers for the application - $10,000.
* Potential hardware - $3,000.

Thank you for time in considering my project proposal. If you have any questions, feel free to contact me.

Sincerely,

Vincent Wong

# Part B: Project Proposal Plan

## Project Summary

From personal experience and a plethora of negative Google reviews, patients are complaining about long wait times at the clinic. The project proposal aims to reduce patient wait time and increase turn around time to increase patient experience and satisfaction. Leveraging machine learning, the goal is to develop a predictive application that can assist in identifying female individuals at risk of developing diabetes.

Deliverables for the project include a machine learning application capable of predicting the likelihood of diabetes from input features, a user-friendly interface allowing medical professionals in the clinic to input patient data and receive real-time predictions, documentation, and user guides to assist in using and navigating the application.

The client can utilize the application to optimize and allocate resources more efficiently by focusing on high-risk individuals, thus improving patient care. The application will integrate into the existing healthcare workflow, minimizing disruption and ensuring easy adoption of the application. With the application, medical professionals can use data-driven decision to make decision faster to improve patient turnaround time. This project aims to create happier patients and increase the reputability of the client.

## Data Summary

Find the dataset for training our machine learning model from Kaggle at the following link: <https://www.kaggle.com/datasets/pentakrishnakishore/diabetes-csv/data?select=diabetes.csv>. This dataset contains a dependent variable outcome representing whether a female has diabetes with various other independent variables that attribute to the outcome, thus satisfying the needs of the project. The dataset is open source and does not contain any personal information, so we do not have to worry about any ethical or legal concerns.

Upon preliminary inspection, zero values are present in the dataset:

A table of numbers and a few black text

Description automatically generated with medium confidence

The following assumptions will be made for the zero values to clean the data:

* It is possible to have zero pregnancies.
* Values of zero in the outcome column represent negative result.
* Other values of zero are missing data, thus remove the rows with missing data from the following columns: glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function and age.

After cleaning the dataset, split the dataset to a training set and a test set for training and testing purposes. Then, apply standardization to the dataset to ensure that every column representing an independent variable will have equal contribution to the dependent variable.

## Implementation

Utilize the SEMMA methodology (Hotz, 2021) to implement the project per the following criteria:

* Sample: From the dataset, the outcome column will be the dependent variable and all other columns will be the independent variables.
* Explore: Assess the dataset. Analyze and predict which independent variables will have the greatest correlations with the dependent variable.
* Modify: Clean the dataset for missing data. Standardize the data for training and testing with the machine learning models.
* Model: Utilize multiple classification models to determine which model provide the best results.
* Assess: Evaluate the results from the different classification models and choose the model with the best mean accuracy for our project.

After choosing the best model, create a cell in the notebook utilizing the best model for users to interact with.

## Timeline

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone or Deliverable** | **Duration** | **Projected Start Date** | **Anticipated End Date** |
| Data Preprocessing | 1 Day | 12/01/23 | 12/02/23 |
| Data Exploration and Analysis | 2 Days | 12/02/23 | 12/04/23 |
| Feature Engineering | 1 Day | 12/04/23 | 12/05/23 |
| Models Training | 3 Days | 12/05/23 | 12/08/23 |
| Cross-Validation | 1 Day | 12/08/23 | 12/09/23 |
| Models Evaluation | 2 Days | 12/09/23 | 12/11/23 |
| Model Selection | 1 Day | 12/12/23 | 12/13/23 |
| Results Visualizations | 2 Days | 12/13/23 | 12/15/23 |
| Create User Interface | 2 Days | 12/15/23 | 12/17/23 |
| Documentation | 1 Day | 12/17/23 | 12/18/23 |
| Deployment and Testing | 1 Week | 12/18/23 | 12/25/23 |

## Evaluation Plan

Verification is crucial for the development of the project. While developing the project, look to verify the completeness of the preprocessing steps by ensuring all zero values are filter correctly, split, and standardize accordingly. Unit tests will check for the presence of zero values, length of training and test sets and new value ranges after the standardization of the dataset. Then, use cross validation to ensure consistency of the different classification models to choose the best model. At the last step, use unit tests again to check for the completeness of the user interface interaction.

After deploying the application, validate the model impact on the business, involve end-users to validate the model’s usability and practically. Then per users’ feedback, improve the application accordingly. Third party testers with background in machine learning will assess the model’s performance, bugs, errors, and check if it meets the project requirements.

## Resources and Costs

See table below for detailed information on cost of project:

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| Hardware Upgrades | Upgrading computers in the clinic to run the application. | ~$3,000 |
| Software Upgrades | Ensuring computers have valid operating system and software to run application. | ~$1,000 |
| Coding Environment | Jupyter Notebook | Free |
| Database Software (Optional) | MySQL – If want to save data for future reference. | Free |
| Data Security Measures (Optional) | Cost to implement security measures to protect data in database. | ~5,000 |
| Third Party Testers | Hiring testers to test overall application towards the end of the project. | ~$10,000 per testing |
| Labor | Cost to build and deploy the application. | ~$50,000 per month |
| Maintenance | Expenses for updating application, bug fixes and patching vulnerabilities. | $500 per month |

The total cost will be around $64,000 plus monthly $500 maintenance fees.

# Part C: Application

See provided c964.ipynb file in zip archive.

# Part D: Post-implementation Report

## Solution Summary

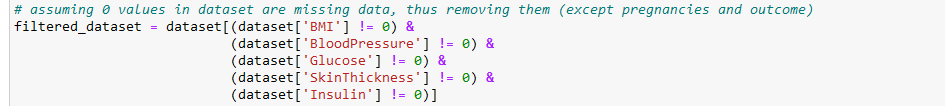
First and foremost, the application is to be used as a helping tool to aid physicians in evaluating if a female has diabetes or not. It is not meant to be a deterministic application used to replace a physician’s expertise judgements. By using the application, medical professionals at Some Diabetes Clinic LLC can do preliminary analysis based on provided medical records to see whether a female is more susceptible to having diabetes. The final decision should be determined by lab works and a thorough analysis of the female patient’s history.

Overall, by using the application that has a mean accuracy of 78.9% in determining whether a female patient has diabetes, Some Diabetes Clinic LLC can expediate the screening process to improve patient experience and satisfaction. The application allows the physicians to focus on which individual is more susceptible to diabetes as to provide a speedy response time instead of wasting valuable time interacting with a patient that might not have diabetes.

## Data Summary

The dataset that was used to create the training and testing sets for the application was obtained from the following link: <https://www.kaggle.com/datasets/pentakrishnakishore/diabetes-csv/data?select=diabetes.csv>.

The first step taken to preprocess the dataset prior to training was to remove all the zero values in the dataset which were presumed to be missing data with the following code:



The pregnancies and outcome columns were excluded because a female can have zero pregnancies and the zero in the outcome column represents a negative result. A print function was used to verify that the dataset was filtered correctly:

A screenshot of a computer

Description automatically generated

The dataset was then separated based on the position of the independent and dependent variables:

A white rectangular object with black text

Description automatically generated

Feature scaling was then applied to the training and test sets to ensure that the different independent variables contribute equally to the learning process:

A screenshot of a computer

Description automatically generated

## Machine Learning

The chosen machine learning algorithm for the application was the Support Vector Machine model. It is a supervised learning model developed by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1964 (Wikipedia Contributors, 2019). The primary objective of the model is to find a hyperplane with the largest margin or separation that best separates the data points into different classes with the help of mathematical functions known as kernels that map the original input features into a higher-dimensional space (*Support Vector Machine (SVM) Algorithm*, 2021).

To ensure that the application would be able to produce the best results to aid physicians in evaluating whether patients have diabetes or not, the Support Vector Machine model was chosen from among multiple other classification machine learning models including:

* Random Forest Classification
* Naïve Bayes
* Logistic Regression
* K-NN
* Kernel Support Vector Machine
* Decision Tree Classification
* XGBoost

Every model was cross validated with a test set of size ten to ensure that luck wasn’t a factor for one model having the best accuracy with a single test run. The Support Vector Machine model was the model with the best performance with a mean accuracy of 78.9% and was thus chosen to be the machine learning model to be used with the application.

A screenshot of a computer code

Description automatically generated

## Validation

Validating the Support Vector Machine model involved assessing its performance and classification ability on data that it has not seen during training. After filtering the dataset, two subsets, a training set and testing set were created based on a 75-25 split. Feature scaling was done to ensure all features contributed equally to the model. The model was then trained using the training set to find the optimal hyperplane that separates the classes. The model was used to evaluate the testing set and the results were used to examine the importance of features to see which features contribute the most to the model’s predictions. If necessary, the model can be fine-tuned by adjusting the hyperparameters to see if the model can provide an even better mean accuracy score.

## Visualizations

Provided below are visualizations of choosing the machine learning model for the application and the results derived from said model:

* How the Support Vector Machine model was chosen among other classification models:

A graph of different colored rectangular objects

Description automatically generated with medium confidence

* How the selected model performed versus the test set:

A diagram of a heatmap

Description automatically generated

* Which features contributed the most to the model’s predictions based on the coefficients returned for each feature from running the Support Vector Machine model:

A green bar graph with black text

Description automatically generated

## User Guide

Please follow the following instructions to run and interact with the application:

1. Install Jupyter Notebook via Anaconda or Windows terminal.

* Download and install Anaconda via <https://www.anaconda.com/>.
* To install and run Jupyter Notebook via Windows terminal, run the following commands -pip install jupyter and -jupyter notebook. If pip is not installed on the Windows system, please run python -m pip install –upgrade pip to install and upgrade pip on to the system.

1. Extract the c964.ipynb file from the provided zip archive.
2. Extract diabetes.csv from provided zip archive.
3. Upload diabetes.csv to Jupyter Notebook.
4. Open the c964.ipynb file in Jupyter Notebook.
5. Run all cells:

A screenshot of a computer

Description automatically generated

1. Scroll to the last cell.
2. Input values to the input fields and click on the “Run Prediction” button to see if a female has diabetes or not based on the Support Vector Machine model:

A screenshot of a computer

Description automatically generated

# Reference Page

* Wikipedia Contributors. (2019, April 20). *Support-vector machine*. Wikipedia; Wikimedia Foundation. <https://en.wikipedia.org/wiki/Support_vector_machine\>
* *Support Vector Machine (SVM) Algorithm*. (2021, January 20). GeeksforGeeks. [https://www.geeksforgeeks.org/support-vector-machine-algorithm/#](https://www.geeksforgeeks.org/support-vector-machine-algorithm/)
* Centers for Disease Control and Prevention. (2022, June 27). *Health insurance portability and accountability act of 1996 (HIPAA)*. Centers for Disease Control and Prevention. <https://www.cdc.gov/phlp/publications/topic/hipaa.html>
* Hotz, N. (2021, May 14). *SEMMA*. Data Science Process Alliance. https://www.datascience-pm.com/semma/