**C964 Computer Science Capstone**

**Renae Villarreal**

Table of Contents

[Part A: Letter of Transmittal 2](#_Toc151567672)

[Letter of Transmittal 2](#_Toc151567673)

[Part B: Project Proposal Plan 4](#_Toc151567674)

[Project Summary 4](#_Toc151567675)

[Data Summary 5](#_Toc151567676)

[Implementation 5](#_Toc151567677)

[Timeline 6](#_Toc151567678)

[Evaluation Plan 7](#_Toc151567679)

[Resources and Costs 7](#_Toc151567680)

[Part C: Application 9](#_Toc151567681)

[Part D: Post-implementation Report 10](#_Toc151567682)

[Solution Summary 10](#_Toc151567683)

[Data Summary 10](#_Toc151567684)

[Machine Learning 11](#_Toc151567685)

[Validation 12](#_Toc151567686)

[Visualizations 13](#_Toc151567687)

[User Guide 17](#_Toc151567688)

# Part A: Letter of Transmittal

## Letter of Transmittal

11/22/2023

Justin Ray, CEO

Nae’s Healthcare Organization

19510 Papago Dr. Tomball, TX, 77377

Dear Mr. Ray,

Here at Nae’s Healthcare, we aim to provide the best possible care for our employees and patients. With the rise in breast cancer diagnoses, I am here to present a proposal for the development of a breast cancer prediction application. Patients’ lives depend on our company for a correct and timely diagnosis. Recognizing these critical needs aligns with our shared commitment to advancing healthcare outcomes.

Our current methods for the process of diagnosing a breast cancer tumor are out of data and lack effectiveness and timeliness. Not having a reliable tool can affect our company in every department. Having an application can help our healthcare professionals ability to make informed decisions quickly, helping the patients determine which treatment route to take. Our solution to these issues can be solved with the development of a breast cancer prediction application. The application will use a supervised machine learning algorithm in our specific case a support vector machine applied with a linear kernel. Paired with our dataset on already diagnosed tumors the application will be able to train and essentially learn on its own. Using this application will be simple for our healthcare professionals simply enter the patient’s tumor attributes and click a button. The application will be able to quickly predict whether a tumor is malignant or benign based on training data.

The several benefits this tool brings to our company are astonishing. To improve our accuracy, the rate our application will aim for is 90-97%. Another one is resource optimization, streamlining our process for this will allow the company to reallocate resources. Rearranging our resources can help our company keep up with the high rate of patients and further develop other departments. Timely and accurate predictions are the first step in finding individual treatment plans for each patient. Providing a solution to not only our problems but the patients’ lives. Finally, we can benefit through market competitiveness, adopting this application can position Nae’s Healthcare well beyond health industry competitors.

Our estimated cost for this project will initially be $14,925 to complete and $1,500 per month to keep the application up and running. The details for these numbers will be attached in the proposal found under the section “Resources and Cost”. Once initiated the estimated timeline for this project will begin on November 24, 2023, and the completion date is December 16, 2023. This time will be broken down into our six-stage implementation plan. The data gathered for this machine learning project is composed of tumor attributes is publicly available and does not break any laws or regulations to alter and obtain. We will be using the Wisconsin breast cancer dataset to train and develop our model. I will be the lead developer assigned to this project. My background includes 4 years of building machine learning applications and I hold a bachelor’s in computer science. I’ve worked for top companies in the industry and will continue to expand my resume. I believe in this application for Nae’s Healthcare and will meet the deadlines and budgets set.

Thank you for considering my application, and I look forward to the future this company holds in the healthcare industry.

Sincerely,

Renae Villarreal

Renae Villarreal, Director of Data Science

# Part B: Project Proposal Plan

## Project Summary

* **Describe the problem.**

Determining whether a tumor is malignant or benign can be a challenge not only to determine but also to determine with high accuracy. The method being used lacks effectiveness and quick response time. Using these conventional diagnosis methods can hinder the delivery of critical information, patient treatment, and lives. Time and correct diagnosis can potentially save more patients and allow the company to be able to expand and further develop in the field of tumor diagnosis.

* **Summarize the client and their needs as related to the problem.**

Nae’s Healthcare Organization is a healthcare provider that takes exams from various medical institutions. With the rising need and pressure to continue the enhancement of accurate diagnosis of breast cancer tumors, the organization is always looking to improve. With the nature of the issue stemming from time sensitivity the adoption of a predictive model becomes the beneficial object saving time and improving patient outcomes. Seeking a solution that integrates seamlessly into the existing workflow we aim to provide a reliable tool for healthcare professionals. That not only improves speed but can also contribute to the overall efficiency of the healthcare system. By adopting the machine learning model Nae’s Healthcare will be able to stay at the top of the healthcare industry and continue to innovate and be aligned with the latest advancement of technology. The implementation of the predictive breast cancer model serves as a strategic move towards a greater future for delivering fast responses and high-accuracy diagnoses for every patient in need.

* **Provide descriptions of all deliverables. For example, the finished application and a user guide.**

Breast Cancer Prediction Model

1. The model uses SVM (Support Vector Machine) that is trained off a dataset of previously diagnosed tumors. The user will see the different parameters and kernels for linear and rbf and compare the accuracy. For this application, we will choose the best kernel due to the higher accuracy rate in prediction. The application will utilize visuals and samples for the users to understand the metrics behind the data and model. Finally, the user will be able to input attributes related to the tumor and the model predicts with a click of a button.

User Guide

1. The guide is for our users to be able to successfully run the application with a detailed walk-through for each step on how to use and obtain the application. This makes it easy to share and explain throughout the company and employees.

* **Provide a summary justifying how the application will benefit the client.**

The application will allow healthcare professionals to diagnose a tumor accurately and quickly. Having an application with a high accuracy rate allows the company to streamline the task and reallocate resources that were once needed to complete this task. It allows the application to be a catalyst for transformative change within the organization. Which can lead to timely, personalized treatment plans and the improvement of patient outcomes. Nae’s Healthcare Organization will also benefit from cost reduction, market competitiveness, and most importantly save lives with the adoption of this application.

## Data Summary

* **Provide the source of the raw data, how the data will be collected, or how it will be simulated.**

The dataset will be sourced from kaggle.com. This Wisconsin breast cancer dataset contains thirty-two attributes about a tumor and its diagnosis. I will be using the CSV files to upload the data to my GitHub allowing the application to read the data using pandas and create a data frame for the model to use.

* **Describe how data will be processed and managed throughout the application development life cycle: design, development, maintenance, or others.**

During the process, the data will go through multiple steps to benefit the accuracy of the model. We will check the sample data for any null columns removing any if needed. We are moving on to deleting any columns our model will not use such as the id column. The machine learning algorithm we will be utilizing does not need this information to be trained. Once the dataset is fit for training and cleaned, we will move on to splitting the data into smaller portions for the SVM model to be trained with. This will allow the model to have sufficient data to learn with and avoid any repetition.

* **Justify why the data meets the needs of the project. If relevant, describe how data anomalies, e.g., outliers, incomplete data, etc., will be handled.**

The Wisconsin breast cancer dataset offers us comprehensive and detailed data related to tumors. This will give the model every feature it needs to “learn” and return a correct diagnosis. Any incomplete data will be removed from the dataset as it is not fit for our model. Outliers will be accounted for within our algorithm.

* **Address any ethical or legal concerns regarding the data. If there are no concerns, explain why.**

The dataset is available to further the healthcare industry. I do not intend to issue any harm with my application and is solely made to benefit the community it involves. There are no ethical or legal concerns regarding the chosen dataset. The dataset is publicly available and adheres to privacy regulations and laws.

## Implementation

* **Describe an industry-standard methodology to be used.**

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a framework that we will be implementing for the breast cancer prediction application. This framework will ensure a well-structured and systematic development process for the application. It consists of six main segments business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This method will be implemented in an agile manner. Each segment can build upon the other and can always be built on later during evaluation.

* **An outline of the project’s implementation plan. This outline can focus on the project’s development as a whole, or it may focus on only the implementation of the machine learning solution.**

1. **Business Understanding**

Finding the goal/objective of the application. What problem will we be solving? In our case diagnosis accuracy, streamlining company processes, and providing patients with a better outcome.

1. **Data Understanding**

Data Analysis breaks down our data to further understand the structure, features, and challenges we are faced with. The dataset will need to be evaluated and preprocessed before use. Understanding these attributes beforehand will save us time and avoid future issues when making our model.

1. **Data Preparation**

Data cleaning addresses issues within the dataset like null values, outliers, and any inconsistent data. This stage allows for data quality. Features can be added and modified to further benefit the model.

1. **Modeling**

The SVM model is developing and utilizing the support vector model. Focusing on the parameters and kernels to potentially optimize the algorithm. Modifying the algorithm to reach our accuracy goal.

1. **Evaluation**

The application performance is evaluated at this stage. Testing throughout different stages of the model, the performance is tracked. We will be using the independent variables to cross-validate the algorithm's predictions.

1. **Deployment**

User Training provides support to healthcare professionals to ensure proper use of the application and seamless integration into Nae’s Healthcare Organization. Making a smooth transition to the new application.

## Timeline

* **Provide a projected timeline, including projected start dates and end dates for each milestone (a table is not required but encouraged).**

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone or deliverable | Duration  (hours or days) | Projected start date | Anticipated end date |
| Business Understanding | 20 hours | 11/24/2023 | 11/26/2023 |
| Data Understanding | 25 hours | 11/27/2023 | 11/30/2023 |
| Data Preparation | 30 hours | 12/1/2023 | 12/4/2023 |
| Modeling | 40 hours | 12/5/2023 | 12/10/2023 |
| Evaluation | 25 hours | 12/11/2023 | 12/13/2023 |
| Deployment | 20 hours | 12/14/2023 | 12/16/2023 |

## Evaluation Plan

* **Describe the verification method(s) to be used at each stage of development.**

1. **Business Understanding**

Constant meetings with the stakeholders to validate that the objectives are in line with the company's views.

1. **Data Understanding**

Use stakeholder feedback to ensure the features are chosen properly and align with the project.

1. **Data Preparation**

Performing regular data cleaning checks to make sure the data meets the preferred quality.

1. **Modeling**

Using cross-validation scores to determine the model's performance allows us to be specific during the modifications and tuning of the algorithm.

1. **Evaluation**

Testing metrics in the algorithm to try and predict any future scenarios. To avoid any potential problems.

1. **Deployment**

Gathering feedback from the healthcare professionals while being trained on how to use the application and during the integration of the model into the company workflow.

* **Describe the validation method to be used upon completion of the project.**

To ensure the application meets the healthcare industry standard. Upon completion, the application will undergo test trials of various independent data of the patient’s tumor data. The data will be evaluated with the original process and compared to this new process. This allows us to ensure that the algorithm is indeed producing the correct outcome and can be trusted. Once validated the company will be allowed to produce feedback on how the new application fits in the company and any future issues will be accessed during this time to make the application more user-friendly.

## Resources and Costs

* **Itemize hardware and software costs.**

|  |  |  |
| --- | --- | --- |
| **Item:** | **Type:** | **Cost:** |
| Hardware | Operating System /  High Standard Computing | $1,500 |
| Hardware | AWS Cloud Services | $2,000 |
| Software | Python Libraries | Free |
| Software | Tools and Frameworks | Free |

* **Itemize estimated labor time and costs.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Process:** | **Role:** | **Estimated Hours:** | **Cost:** |
| Data Understanding / Data Preparation | Data Scientists | 25 Hours (50 per hour) | $1,250 |
| Modeling | Machine Learning Engineers | 30 Hours (45 per hour) | $1,350 |
| Testing / Validation | QA Engineers | 15 Hours (35 per hour) | $525 |
| Deployment | DevOps Engineers | 20 Hours (35 per hour) | $1,050 |
| Documentation | Data Scientists / Machine Learning Engineers | 10 Hours (45 per hour) | $450 |

* **Itemize estimated environment costs of the application, e.g., deployment, hosting, maintenance, etc.**

|  |  |  |
| --- | --- | --- |
| **Process:** | **Type:** | **Cost:** |
| Deployment | Cloud Platform | $1,500 |
| Hosting | Hosting Fee Per Month | $500 |
| Maintenance | Per Month | $1,000 |
| User Support | Training IT Helpdesk | $1,800 |
| Security | Cybersecurity | $2,000 |

**Total Cost for Development: $14,925**

**Monthly Cost: $1,500**

# Part C: Application

**Part C is your submitted application. This part of the document can be left blank or used to include a list of any submitted files or links.**

**Google Colab Notebook Link** - <https://colab.research.google.com/drive/1e3JiYG2YTzudrYrmYlfk1-jVaIJ137JG?usp=sharing>

**Source/Raw file**- (.ipynb please double click icon)

# Part D: Post-implementation Report

## Solution Summary

* **Summarize the problem and solution.**

The main issue addressed in this project was to provide a solution for Nae’s Healthcare Organization. The issues stemmed from slow patient results to an abundant number of resources spent on the diagnosis phase of the tumor. We have a goal set for providing medical institutions with a reliable tool to help patients in the detection of whether their breast cancer tumor is malignant or benign. Our application implements the use of a supervised machine learning algorithm specifically the support vector machine with a linear kernel. This type of machine learning allows us to provide a solution accurately and easily to these issues we face.

* **Describe how the application provides a solution to the problem from parts A and B.**

Our application provides a solution to these issues by allowing the user to enter specific attributes related to the patient. Once entered our trained algorithm can make a prediction based on the validated dataset. We are producing a result with an accuracy of up to 97%. That can higher the probability of early detection for every case entered. This allows the medical institution to effectively plan a treatment for everyone. This allows for time-saved, resourced relocation and provides the ability to potentially save lives.

## Data Summary

* **Provide the source of the raw data, how the data was collected, or how it was simulated.**

The raw dataset (Wisconsin Breast Cancer Dataset) was retrieved from Kaggle.com. This data is sourced from the UCI Machine Learning Repository, which holds very comprehensive data on attributes related to a malignant tumor or benign. The data was downloaded from a CSV file and uploaded to my personal GitHub. The application then reads in the data using pandas to load all the data into the data frame created.

(The screenshot shows the first 5 and last 5 rows in our data frame)A screenshot of a computer screen

Description automatically generated

* **Describe how data was processed and managed throughout the application development life cycle: design, development, maintenance, or others.**

Once loading in the data to our data frame we first started by checking if any of the features were left null.

print(df.isnull().sum())

After checking that no rows were left empty, we proceeded to move on to cleaning the data. As shown the diagnosis column type is an object. For our application to run efficiently, I changed the column type to an integer. Giving us the preferred execution rate for the SVM model.

#label encoder to convert 'M' and 'B' to 1 and 0

nt = LabelEncoder()

df['diagnosis'] = nt.fit\_transform(df['diagnosis']

(The screenshot below shows the type of change from object to integer)

A screenshot of a computer code

Description automatically generated

The id column plays no role in our application working correctly. Removing this column allowed us to clean our data and avoid potential bugs in the future. Finally, our data is now ready to be implemented into the model.

#drop id column data not needed

df = df.drop('id',axis=1)

(The screenshot below shows the ID column removed)

A black text on a white background

Description automatically generated

## Machine Learning

**For each employed method (at least one is required) provide the following:**

* **Identify the method and what it does (the “what”).**

The model used is the SVM (Support Vector Machine). It is a supervised machine-learning algorithm that is used for classification. The main goal of an SVM is to find a decision boundary known as a hyperplane. This allows for the data to be separated into “categories” in our case malignant or benign tumor attributes. I used a linear kernel which helps the model issue the decision boundary linearly. Essentially it makes a straight line in a higher dimensional space.

* **Describe how the method was developed (the “how”).**

In this project, the SVM model was executed with the sci-kit-learn library known as sklearn. The model was then “trained” on our data that has been split up to avoid repetition. This allows the model to learn the relationships between the tumor attributes. Using this library allowed us to implement different kernel parameters. After changing a few parameters, we have settled on the linear kernel to handle teaching the model the data. Once trained the model is then moved into predicting the diagnoses column of the data it is fed. Finally, we determined the accuracy of our model based on the accuracy score and y\_pred.

(The screenshot below shows our model development)

A screen shot of a computer

Description automatically generated

* **Justify the selection and development of the method (the “why”).**

SVM paired with a linear kernel proves itself when it comes to datasets with numerous features. Our breast cancer dataset is very complex and provides any attributes related to a tumor. As with most data sets that are related to the healthcare industry. The SVM excels in classifying and navigating through our data. Having a linear kernel paired with it depended on the simplicity it brings to our machine-learning model. It provides a clear boundary for the SVM to distinguish between and determine patterns in our data. Making this our best choice for a reliable application in predicting future cases.

## Validation

**For each employed method described in the section above provide the following:**

* **An appropriate validation method.**

Cross-validation is used to validate the SVM model. This is done by separating the dataset into smaller sections for training the model. With this validation, we can see what features the model is predicting correctly and wrong narrowing down our key feature points. Allowing us to make assessments on the accuracy score leading to a reduced risk of overfitting or underfitting a set of data.

* **Results of the validation method or a plan to obtain those results.**

The results of the cross-validation show a high level of accuracy in predicting the correct type of tumor. Our model also shows metrics such as precision, recall, and F1 scores which help us have a better understanding of the outcome of our model. Helping us determine where changes need to be made or prioritized. Precision is measuring the accuracy of the model predictions, recall is the model's ability to obtain true positives, and F1 is the mean of both precision and recall. The model allows the user to change between a linear kernel and rbf. Showing the difference in accuracy, classification report, and confusion matrix for each type.

(The screenshots below show the results of the classification report and confusion matrix) A screenshot of a computer

Description automatically generated

A blue and white squares with white text

Description automatically generated

## Visualizations

**Identify the location of at least three unique visualizations. They can additionally included here.**

In the Google Colab, the visuals are under the Data Visualization heading. These visuals include a histogram, heatmap, violin plot, and scatter plot. Snippets will be provided below but for a better picture please see the notebook.

(The screenshot below shows a snippet of the histogram graph)

A group of red and white graphs

Description automatically generated

(The screenshot below shows a snippet of the heatmap)

A screenshot of a computer screen

Description automatically generated

(The screenshot below shows a violin plot)

A diagram of a diagram

Description automatically generated

(The screenshot below shows the scatter plot)

A diagram of red and blue dots

Description automatically generated

## User Guide

**Include an enumerated (steps 1, 2, 3, etc.) guide to execute and use your application.**

* **Include instructions for downloading and installing any necessary software or libraries.**

1. Sign into your Google account/make if needed (\*\*Google Account needed\*\*)
2. Click this link to my Google Colab Notebook -(<https://colab.research.google.com/drive/1e3JiYG2YTzudrYrmYlfk1-jVaIJ137JG?usp=sharing>) **(\*\*For source/raw file please see Part C\*\*)**
3. Please hit the “Runtime” option in the toolbar and then click Run All

A screenshot of a computer

Description automatically generated

1. Give the application a few seconds to execute and then move to step 5.

**(\*\*If you are going to run a single cell you will need to repeat step 3 after\*\*)**

* **Provide an example of how the client should use the application.**

1. The client should then move from top to bottom of the notebook through each tab in the table of contents on the left-hand side.
2. Reading through all text and code to understand what each section is about.
3. Once finished reading the client can move to the “1. User Kernal Scores Interactive Tool” and proceed to the drop-down box to see the different results of kernels.



1. Next move to the “User Sample Interactive Tool” proceeding to the drop box to see examples of malignant or benign tumor attributes.

A close up of a sign

Description automatically generated

1. Finally move to the “User Input Interactive Tool” where the client can input sample data of a tumor with the sliders or manually input data for the model to predict the type of tumor.

A screenshot of a computer

Description automatically generated