

Predictive Modeling for Cancer Diagnosis Using Machine Learning

**CAPSTONE PROJECT**

**Submitted to**

**THE COLLEGE OF INFORMATION AND TECHNOLOGY**

**BACHELOR of ARTS in COMPUTER SCIENCE**

**Western Governors University**

**Malek W. Gaghman**

**June 19th, 2024**

Task 2 parts A, B, C and D

[Part A: Letter of Transmittal 3](#_Toc169723782)

[Part B: Project Proposal Plan 5](#_Toc169723783)

[Project Summary 5](#_Toc169723784)

[Data Summary 6](#_Toc169723785)

[Implementation 6](#_Toc169723786)

[Timeline 7](#_Toc169723787)

[Evaluation Plan 7](#_Toc169723788)

[Resources and Costs 8](#_Toc169723789)

[Part C: Application 10](#_Toc169723790)

[Part D: Post-implementation Report 11](#_Toc169723791)

[Solution Summary 11](#_Toc169723792)

[Data Summary 11](#_Toc169723793)

[Machine Learning 12](#_Toc169723794)

[Validation 12](#_Toc169723795)

[Visualizations 13](#_Toc169723796)

[User Guide 15](#_Toc169723797)

# Part A: Letter of Transmittal

June 18, 2024

Dr. Emily Thompson  
HealthFirst Medical Center  
123 Wellness Avenue  
MediCity, MC 56789

Dear Dr. Thompson,

I hope this letter finds you well. I am writing to propose an innovative project aimed at leveraging machine learning to predict cancer diagnoses, which could significantly enhance HealthFirst Medical Center's ability to identify high-risk patients early. This initiative promises to improve patient outcomes, optimize resource allocation, and solidify our position as a leader in cutting-edge healthcare solutions.

Cancer remains one of the most pressing health challenges globally, with early detection being paramount to successful treatment and improved survival rates. Despite advancements in medical technology, our current methods for identifying patients at high risk of cancer can be time-consuming, costly, and less precise. This gap in early diagnosis often results in delayed treatments and increased healthcare costs. To address this critical issue, I propose developing a predictive model using a comprehensive ‘Cancer Prediction’ dataset. This model will analyze various factors that contribute to cancer risk and, through machine learning, accurately predict which patients are most likely to receive a cancer diagnosis.

The implementation of this predictive model will offer numerous benefits to HealthFirst Medical Center. Firstly, it will enable earlier diagnosis and intervention, thereby significantly improving patient prognosis. By identifying high-risk patients at an early stage, we can prioritize their treatment and allocate our medical resources more effectively. This targeted approach will not only enhance patient care but also reduce overall healthcare costs by minimizing unnecessary tests and treatments. Furthermore, the success of this project will demonstrate our commitment to adopting innovative healthcare solutions, enhancing our reputation and attracting more patients who seek the best medical care available.

The implementation plan for this project is meticulously designed to ensure success. The first step involves gathering and preprocessing the cancer prediction data to ensure it is clean and suitable for analysis. Following this, we will develop and train the machine learning model using Python, utilizing libraries such as NumPy for efficient numerical computations, Matplotlib for creating insightful visualizations, and Scikit-learn for robust machine learning algorithms. Once the model is developed, we will rigorously test and validate its performance to ensure its accuracy and reliability. Additionally, we will create comprehensive documentation and visualization tools to facilitate understanding and usage of the model.

The project will incur some costs, primarily associated with data acquisition, software tools, and personnel time. However, the long-term benefits, such as improved patient outcomes and reduced healthcare costs, far outweigh these initial investments. Detailed cost estimates will be provided upon approval. The project timeline is set for six months from the start date, with regular updates and checkpoints to ensure steady progress. Ethical considerations are also paramount; we will handle all data in strict compliance with privacy regulations and ethical guidelines, ensuring patient confidentiality and data security throughout the project.

With extensive experience in machine learning and data analysis, I am confident in my ability to lead this project to success. My background includes numerous successful projects in predictive modeling and healthcare analytics, which have equipped me with the necessary skills and knowledge to deliver a robust solution. At HealthFirst Medical Center, I have been involved in several key initiatives that have enhanced our data-driven decision-making processes, and I am eager to bring this expertise to the cancer prediction project.

The potential impact of this project on our organization and our patients is immense. By embracing this advanced technology, we will not only improve the quality of care we provide but also position ourselves at the forefront of medical innovation. I am enthusiastic about the possibilities and look forward to discussing this proposal in more detail.

Thank you for considering this proposal.

Sincerely,

Malek Gaghman  
Lead Data Scientist

# Part B: Project Proposal Plan

## Project Summary

Medi al establishments face a critical challenge in accurately diagnosing cancer among its patient population. The current diagnostic process is hindered by the complexity of patient data and the variability in diagnostic criteria, leading to delays in identifying individuals at high risk of cancer. These delays not only impact patient outcomes negatively but also contribute to increased healthcare costs associated with late-stage treatments.

HealthFirst Medical Center, renowned for its commitment to delivering high-quality healthcare services, aims to enhance its diagnostic capabilities through advanced predictive analytics. By leveraging machine learning techniques, the center seeks to develop a predictive model capable of analyzing comprehensive patient data—including demographic details, medical histories, and diagnostic tests—to predict cancer diagnoses with high accuracy and efficiency.

The proposed solution involves developing a sophisticated predictive analytics system tailored specifically to meet the needs of HealthFirst Medical Center. This system will include a robust machine learning model trained on a curated dataset specifically focused on cancer prediction. The model will utilize advanced algorithms to process and analyze patient data, providing timely predictions regarding cancer likelihood based on various input factors.

Additionally, the project will deliver an intuitive command-line interface designed for healthcare professionals at HealthFirst Medical Center. This interface will facilitate seamless data input and retrieval of real-time predictions, allowing healthcare providers to make informed decisions promptly.

To ensure effective implementation and user adoption, comprehensive documentation will accompany the system. This documentation will include a detailed user guide outlining system navigation, data input requirements, interpretation of prediction outcomes, and troubleshooting tips. Technical documentation will also be provided, covering the development process of the machine learning model, data preprocessing methodologies, system architecture, and deployment procedures tailored to the specific requirements of HealthFirst Medical Center.

Implementing this predictive analytics system will provide HealthFirst Medical Center with several benefits. It will significantly enhance diagnostic accuracy, enabling healthcare professionals to identify high-risk patients early and initiate timely interventions. This proactive approach will improve patient outcomes, increase treatment efficacy, and reduce overall healthcare costs associated with late-stage treatments.

Furthermore, the system will streamline operational workflows by automating data analysis processes, thereby optimizing resource allocation and reducing the time spent on manual data interpretation. The insights generated by the predictive model will also support strategic decision-making, enabling HealthFirst Medical Center to allocate resources more effectively and improve overall organizational efficiency.

In conclusion, the development and deployment of this predictive analytics system will empower HealthFirst Medical Center to deliver superior patient care, uphold its reputation as a leader in healthcare innovation, and ensure continued excellence in diagnostic accuracy and patient outcomes.

## Data Summary

The source of data was obtained from the website, Kaggle.com, specifically from the 'Cancer Prediction Dataset,' which contains comprehensive data relevant to cancer diagnoses. This dataset includes anonymized patient information such as demographic details (age, gender), medical history (family history of cancer), lifestyle factors (smoking status, alcohol consumption), and the patients diagnosis. The dataset will be downloaded directly from Kaggle, ensuring accessibility and reliability.

Throughout the application development life cycle, a systematic approach will be taken to process and manage the data. During the design phase, rigorous data cleaning and preprocessing techniques will be applied to address any anomalies such as outliers or incomplete entries. This includes handling missing data through imputation methods and identifying outliers using statistical approaches.

During the development phase, the dataset will be split into training and testing sets to facilitate the training and validation of the machine learning model. This division ensures that the model is trained on a representative subset of the data and evaluated on unseen data to assess its generalizability and performance.

The selected dataset meets the project's needs by providing comprehensive and diverse data points essential for building a robust predictive model for cancer diagnoses. It includes critical variables known to influence cancer risk and diagnostic outcomes, thus supporting accurate predictions and informed decision-making in healthcare settings.

Ethically and legally, there are no concerns associated with this dataset, as it is publicly available and anonymized to protect patient confidentiality. Its use aligns with established guidelines for handling healthcare-related data, ensuring compliance with privacy regulations such as GDPR and HIPAA.

* All data fields include
  + Age, Gender, BMI, Smoking, Genetic Risk, Alcohol Intake, Cancer History, Diagnosis.

## Implementation

This project was implemented following a waterfall approach. The Waterfall methodology is a linear and sequential approach to project development. Each phase must be completed before the next phase begins, ensuring that all aspects of the project are thoroughly addressed in a step-by-step manner.

* Define Project Scope and Objective.
* Gather Requirements
  + Includes meeting with healthcare professionals to gather detailed requirements.
* Develop a Project Plan
  + Includes a detailed description of tasks, timelines, and resource allocation.
* Data Acquisition
  + Obtaining the necessary dataset to be used.
* Data Exploration and Cleaning
  + Apply data cleaning techniques to handle missing values, outliers or inconsistencies.
* Data Preprocessing
* Model Selection
* Model Training
* Model Evaluation
* CLI Design
* Documentation
* Testing
* Deployment
* Continuous Improvement

## Timeline

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone or deliverable** | **Duration (hours/ days)** | **Projected start date** | **Anticipated end date** |
| Project Planning and Requirements Gathering | 30 hours | June 15th, 2024 | June 20th, 2024 |
| Data Acquisition and Preparation | 20 hours | June 19th, 2024 | June 22nd, 2024 |
| Model Development Iterations | 30 hours | June 23rd, 2024 | June 27th, 2024 |
| Continuous Integration and Testing | 15 hours | June 28th, 2024 | June 30th, 2024 |
| Deployment and User Interaction | 10 hours | July 1st, 2024 | July 2nd, 2024 |
| Final Testing and Evaluation | 10 hours | July 2nd, 2024 | July 3rd, 2024 |
| Documentation and Final Deliverables | 5 hours | July 3rd, 2024 | July 4th, 2024 |
| Review and Deployment Sign-Off | 5 hours | July 5th, 2024 | July 5th, 2024 |

## Evaluation Plan

The verification methods at each stage of the development process for the cancer diagnostic system are meticulously designed to ensure accuracy and alignment with project goals. Initially, the project scope and objectives will be reviewed and signed off by stakeholders to confirm clarity and agreement. During the requirements gathering phase, detailed sessions with healthcare professionals will ensure that all needs are accurately captured and validated through use cases. Each subsequent phase, from project planning to data acquisition and exploration, will involve thorough internal reviews, peer assessments, and stakeholder approvals to ensure that every task is executed correctly and efficiently.

As the development progresses, verification will continue with rigorous checks during data preprocessing, model selection, training, and evaluation. Comparative analysis and peer reviews will be conducted to select the most suitable machine learning model, ensuring that it meets the project's performance requirements. The command-line interface (CLI) design will undergo usability reviews, and extensive testing will be performed to ensure that it handles data input and predictions accurately. Comprehensive documentation will be reviewed to ensure it provides clear and sufficient guidance for users and technical staff.

Upon completion of the project, a robust validation process will be implemented to ensure the system's overall functionality and performance. End-to-end system testing will simulate real-world usage to validate the seamless operation of all components. User acceptance testing (UAT) with healthcare professionals will ensure the system is user-friendly and meets their needs. Performance metrics will be validated against predefined benchmarks, and compliance checks will ensure adherence to regulations such as HIPAA and GDPR. The comprehensive documentation will be reviewed for clarity and completeness, and a final stakeholder review and sign-off will confirm that the project meets all defined objectives and requirements, ensuring that HealthFirst Medical Center can deliver superior patient care and maintain its reputation for healthcare innovation.

## Resources and Costs

* Hardware Costs
  + The only hardware cost would be for a computer, as needed, which could range from $1,000 to $2,000 depending on the specifications chosen for optimal performance and development efficiency.
* Software Costs
  + This project will incur zero software costs as we will be utilizing a free and open-source application, Jupyter Notebook, for code development. Additionally, we will leverage free libraries for data analysis and machine learning, eliminating the need for any software-related expenses.
* Estimated labor time and costs

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| Hardware | Development computer | $1,000 - $2,000 / computer |
| Software | Jupyter Notebook, Python libraries | $0 |
| Labor time and Costs | * Project Planning & Requirements Gathering * Data Acquisition & Preparation * Model Development Iterations * Continuous Integration & Testing * Deployment & User Interaction * Final Testing & Evaluation * Documentation & Final Deliverables * Review & Deployment sign off | * 30 hours \* $50/hour = $1,500 * 20 hours \* $50/hour = $1,000 * 30 hours \* $50/hour = $1,500 * 15 hours \* $50/hour = $750 * 10 hours \* $50/hour = $500 * 10 hours \* $50/hour = $500 * 5 hours \* $50/hour = $250 * 5 hours \* $50/hour = $250 |
| Environment Cost | Azure as Cloud Hosting Service | $0 |
|  | **Total** | $7.250 - $8,250+ |

# Part C: Application

Code, as well as dataset links are uploaded via my task submission link.

# Part D: Post-implementation Report

## Solution Summary

The problem addressed in this project was the challenge faced by medical establishments in accurately diagnosing cancer among their patient populations. The current diagnostic process was hindered by the complexity of patient data and the variability in diagnostic criteria, leading to delays in identifying individuals at high risk of cancer. To overcome this challenge, a predictive analytics system using machine learning techniques was developed for HealthFirst Medical Center.

The solution provided by this application involved developing a sophisticated predictive analytics system tailored specifically to meet the needs of HealthFirst Medical Center. The system included a robust machine learning model trained on a curated dataset specifically focused on cancer prediction. This model utilized advanced algorithms to process and analyze comprehensive patient data, including demographic details, medical histories, and more, to predict cancer diagnoses with high accuracy and efficiency. Additionally, an intuitive command-line interface (CLI) was designed for healthcare professionals at HealthFirst Medical Center to facilitate seamless data input and retrieval of real-time predictions, allowing for informed decision-making.

The application addressed the problem by enhancing diagnostic accuracy, enabling healthcare professionals to identify high-risk patients early and initiate timely interventions. This proactive approach improved patient outcomes, increased treatment efficacy, and reduced overall healthcare costs associated with late-stage treatments. The system also streamlined operational workflows by automating data analysis processes, optimizing resource allocation, and reducing the time spent on manual data interpretation. Overall, the application empowered HealthFirst Medical Center to deliver superior patient care, uphold its reputation as a leader in healthcare innovation, and ensure continued excellence in diagnostic accuracy and patient outcomes.

## Data Summary

* The data used to create this prediction application was obtained from

<https://www.kaggle.com/datasets/rabieelkharoua/cancer-prediction-dataset>

* Data Management

During the application development life cycle, data processing and management played a pivotal role in ensuring the accuracy and effectiveness of the predictive analytics system. The process began with importing the dataset (‘The\_Cancer\_data\_1500\_V2.csv') using pandas, allowing for initial exploration of its structure and contents. Missing data was addressed through the SimpleImputer tool from sklearn, which replaced NaN values with the mean of the respective columns. This step was crucial for maintaining data integrity and completeness, preparing the dataset for further analysis.

Once the data was preprocessed, it was divided into input features (x) and the target variable (y) for training the machine learning model. The dataset underwent a training and testing split using train\_test\_split from sklearn.model\_selection, allocating 80% of the data for training and 20% for testing. This partition ensured that the model was trained on a representative subset while retaining unseen data for evaluating its generalization performance.

Standardization of specific columns (0, 2, 5, 6) was performed using StandardScaler from sklearn.preprocessing, ensuring that all features were on a similar scale and preventing bias in the model.

These data processing and management steps were crucial components of the application's development, ensuring that the dataset was appropriately prepared, missing values were handled effectively, features were standardized, and the machine learning model was trained and tested accurately for predicting cancer diagnoses based on patient data.

## Machine Learning

The method used in the model is K-Nearest Neighbors (K-NN). K-NN is a supervised learning algorithm used for classification and regression tasks. In classification, such as this cancer prediction project, K-NN predicts the class of a data point based on the majority class of its nearest neighbors. For instance, it predicts whether a patient is likely to have cancer or not based on the characteristics of similar patients in the dataset.

K-NN was developed by analyzing the distances between data points in the dataset. It calculates the distance between the new data point and all other points in the dataset, then selects the K-nearest neighbors based on the shortest distances. The class of the new data point is determined by the majority class among its K-nearest neighbors. This approach makes K-NN simple to implement and understand, as it relies on the similarity of data points to make predictions.

The selection of K-NN for this project was justified by several factors. Firstly, K-NN is well-suited for classification tasks, which aligns with the objective of predicting cancer diagnoses (binary classification). Secondly, K-NN is non-parametric and does not make strong assumptions about the underlying data distribution, making it robust and flexible. Additionally, K-NN is easy to interpret and explain, which is beneficial for healthcare professionals who need to understand and trust the model's predictions. Overall, the simplicity, flexibility, and effectiveness of K-NN made it a suitable choice for developing the predictive model in this project.

## Validation

* The K-NN models accuracy was determined using the following confusion matrix to get a comprehensive evaluation of the models performance by including true negatives, true positives, false negatives and false positives.
* A screenshot of a computer program

  Description automatically generated

## Visualizations

1. A red and green dot chart

   Description automatically generated
2. A red and green chart

   Description automatically generated
3. A green and red graph

   Description automatically generated

## User Guide

1. Download my ‘CAPSTONE.ipynb’ file.
2. Download/ Run Jupyter Notebook
3. Download and upload the necessary ‘The\_Cancer\_data\_1500\_V2’ dataset onto Jupyter Notebook.
4. Click ‘Run’ in the header and click ‘Run All Cells’.
5. Scroll to the bottom and input patient information to determine predictive diagnosis.