Industrial Internship Report on

”Patient Segmentation”

Prepared by

Anmol Mishra

Executive Summary
This report provides details of the Industrial Internship provided by upskill Campus and The IoT Academy in collaboration with Industrial Partner UniConverge Technologies Pvt Ltd (UCT). This internship was focused on a project/problem statement provided by UCT. We had to finish the project including the report in 6 weeks’ time. My project is related to real life problem we are faced in Covid pandemic. At that we don’t have exact number of people who were suffering from covid at which particular area. This type situation occur everytime that Pharmaceutical companies not aware of exact count of patient in particular area of specific disease. So, To overcome this problem we find a solution using skills of DS and ML. This internship gave me a very good opportunity to get exposure to Industrial problems and design/implement solution for that. It was an overall great experience to have this internship.

TABLE OF CONTENTS

1. Preface

2. Introduction

   2.1 About UniConverge Technologies Pvt Ltd

   2.2 About upskill Campus

   2.3 Objective

   2.4 Reference

   2.5 Glossary

3 Problem Statement

4 Existing and Proposed solution

5 Proposed Design/ Model

   5.1 High Level Diagram

   5.2 Low Level Diagram

6 Performance Test

   6.1 Test Plan/ Test Cases

   6.2 Test Procedure

   6.3 Performance Outcome

7 My learnings

8 Future work scope

1 Preface

During my 6-week internship at UniConverge Technologies as a Data Science and Machine Learning (DSML) intern, I had the opportunity to immerse myself in real-world applications of data analysis, modeling, and AI. This experience provided invaluable insights into the inner workings of a tech company and enhanced my skills in this rapidly evolving field.

Weeks 1-2: Orientation and Familiarization

- The first two weeks were dedicated to getting acquainted with UniConverge's work culture, mission, and ongoing projects.

- I participated in orientation sessions to understand the company's values, team structure, and expectations.

Weeks 3-4: Learning and Skill Development

- I delved into DSML fundamentals, brushing up on concepts such as data preprocessing, feature engineering, and model evaluation.

- Collaborated with experienced team members on mini-projects, gaining exposure to real data challenges and learning industry best practices.

Weeks 5-6: Hands-on Projects and Implementation

- Engaged in practical DS & ML projects related to the company's ongoing initiatives.

- Assisted in developing predictive models for a client's demand forecasting, utilizing techniques like regression and time series analysis.

- Contributed to data-driven insights for a recommendation engine, applying collaborative filtering methods.

- Leveraged machine learning libraries and tools like TensorFlow, scikit-learn, and pandas.

Key Takeaways:

1. Practical Exposure: The internship provided a practical understanding of how DS principles are applied to solve real-world problems. This experience was incredibly insightful in bridging the gap between theoretical knowledge and its practical applications.

2. Team Collaboration: Working with experienced data scientists and engineers taught me the importance of effective communication and teamwork in a tech environment. I learned to present my ideas clearly and integrate feedback into my work.

3. Data Pipeline: I gained insights into the data pipeline, from data collection and preprocessing to model deployment and monitoring. This understanding highlighted the significance of data quality and its impact on the overall outcome.

4. Problem-Solving: The challenges faced during projects taught me to approach problems systematically, break them down into smaller components, and iteratively test solutions.

5. Continuous Learning: The dynamic nature of DS demands constant learning and adaptation. My time at UniConverge emphasized the importance of staying updated with the latest tools and techniques in the field.

6. Industry Standards: I gained insight into how the industry adheres to best practices in model evaluation, validation, and ethical considerations, ensuring that solutions are both accurate and responsible.

In conclusion, my 6-week internship at UniConverge Technologies was an enriching experience that deepened my understanding of data science and machine learning in a practical setting. I am grateful for the knowledge gained, the professional connections made, and the foundation laid for my continued growth in this field.

The Need for Internships in Career Development:

1. Practical Application of Knowledge:

This internship provide a bridge between academic knowledge and real-world application. They allow individuals to apply theoretical concepts learned in classrooms to actual projects and tasks. This practical experience enhances their understanding and helps them develop problem-solving skills.

2. Skill Development:

This internship offer a platform for acquiring and honing a wide range of skills, both technical and soft. These include technical skills specific to the field, communication skills, teamwork, time management, adaptability, and more. These skills are transferable and valuable in any career path.

3. Industry Exposure:

This Internship immerse individuals in the industry they aspire to work in. They get a firsthand understanding of the industry's dynamics, trends, challenges, and best practices. This exposure helps them align their career goals and make informed decisions about their future.

4. Networking Opportunities:

This Internship provide opportunities to connect with professionals, mentors, and colleagues within the industry. Building a strong professional network is invaluable for accessing job opportunities, seeking advice, and staying updated with industry trends.

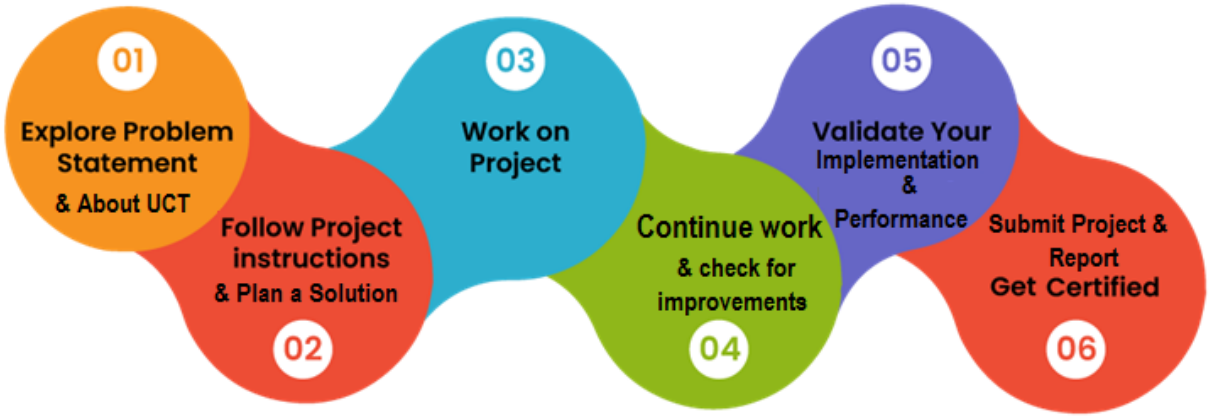
5. Resume Enhancement:

A relevant internship adds credibility to a resume. It demonstrates practical experience and a commitment to learning beyond the classroom. Employers often value candidates with internship experience, as it shows their readiness for the professional environment.

6. Confidence Building:

Successfully completing an internship boosts an individual's confidence in their abilities. It validates their skills and validates their career aspirations. This newfound confidence can positively impact their job search and overall career trajectory.

How Program was planned



2 Introduction

2.1 About UniConverge Technologies Pvt Ltd

A company established in 2013 and working in Digital Transformation domain and providing Industrial solutions with prime focus on sustainability and RoI.

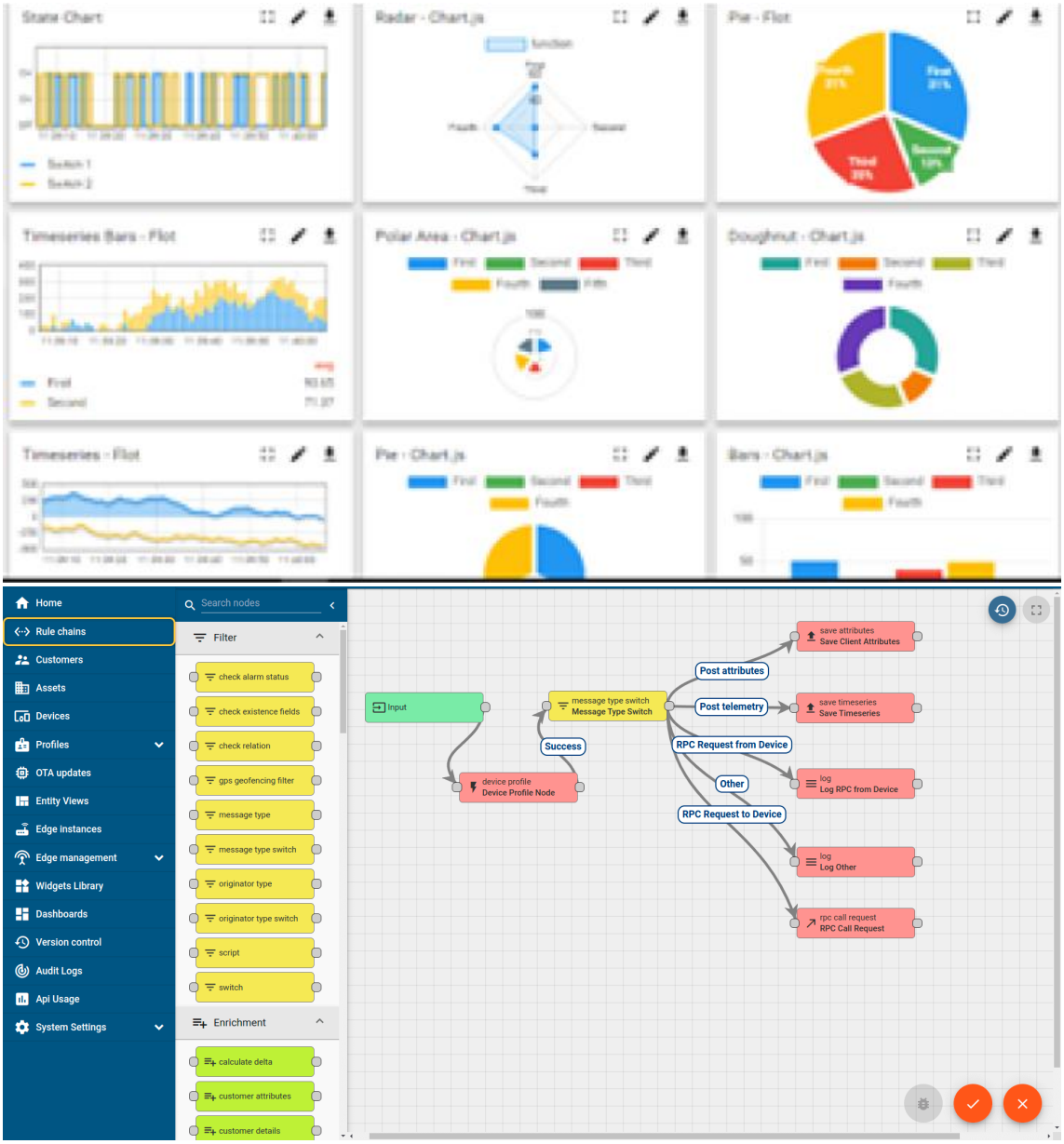
For developing its products and solutions it is leveraging various **Cutting Edge Technologies e.g. Internet of Things (IoT), Cyber Security, Cloud computing (AWS, Azure), Machine Learning, Communication Technologies (4G/5G/LoRaWAN), Java Full Stack, Python, Front end** etc.

i. UCT IoT Platform (UCT Insight)

UCT Insight is an IOT platform designed for quick deployment of IOT applications on the same time providing valuable "insight" for your process/business. It has been built in Java for backend and ReactJS for Front end. It has support for MySQL and various NoSql Databases.

- It enables device connectivity via industry standard IoT protocols - MQTT, CoAP, HTTP, Modbus TCP, OPC UA
- It supports both cloud and on-premises deployments.

It has features to
- Build Your own dashboard
- Analytics and Reporting
- Alert and Notification
- Integration with third party application(Power BI, SAP, ERP)
- Rule Engine



ii. Smart Factory Platform (FACTORY WATCH)

Factory watch is a platform for smart factory needs.

It provides Users/ Factory

- with a scalable solution for their Production and asset monitoring

- OEE and predictive maintenance solution scaling up to digital twin for your assets.
- to unleased the true potential of the data that their machines are generating and helps to identify the KPIs and also improve them.
- A modular architecture that allows users to choose the service that they what to start and then can scale to more complex solutions as per their demands.

Its unique SaaS model helps users to save time, cost and money.

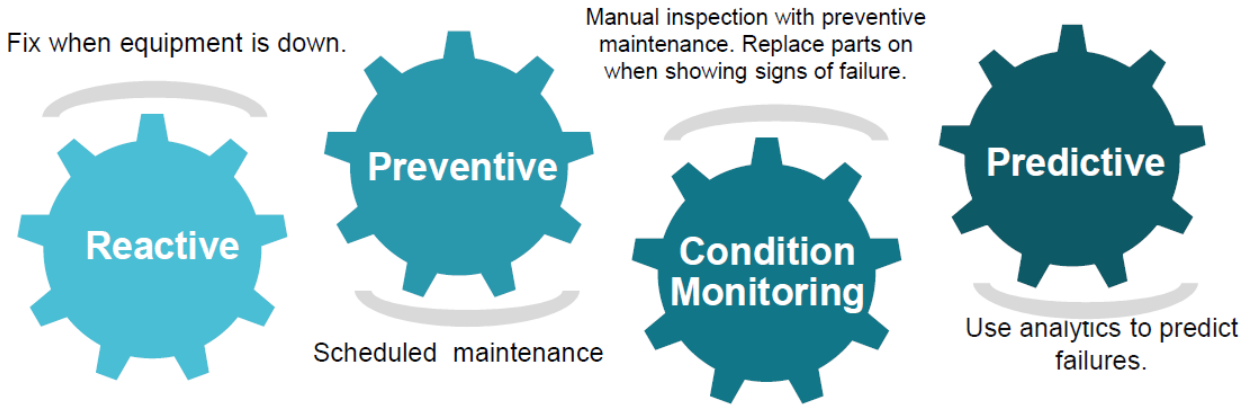
Machine	Operator	Work Order ID	Job ID	Job Performance	Job Progress		Output		Rejection	Time (mins)				Job Status	End Customer
					Start Time	End Time	Planned	Actual		Setup	Pred	Downtime	Idle		
CNC_S7_81	Operator 1	WO0405200001	4168	58%	10:30 AM		55	41	0	80	215	0	45	In Progress	i
CNC_S7_81	Operator 1	WO0405200001	4168	58%	10:30 AM		55	41	0	80	215	0	45	In Progress	i

iii. LoRaWAN based Solution

UCT is one of the early adopters of LoRAWAN teschnology and providing solution in Agritech, Smart cities, Industrial Monitoring, Smart Street Light, Smart Water/ Gas/ Electricity metering solutions etc.

iv. Predictive Maintenance

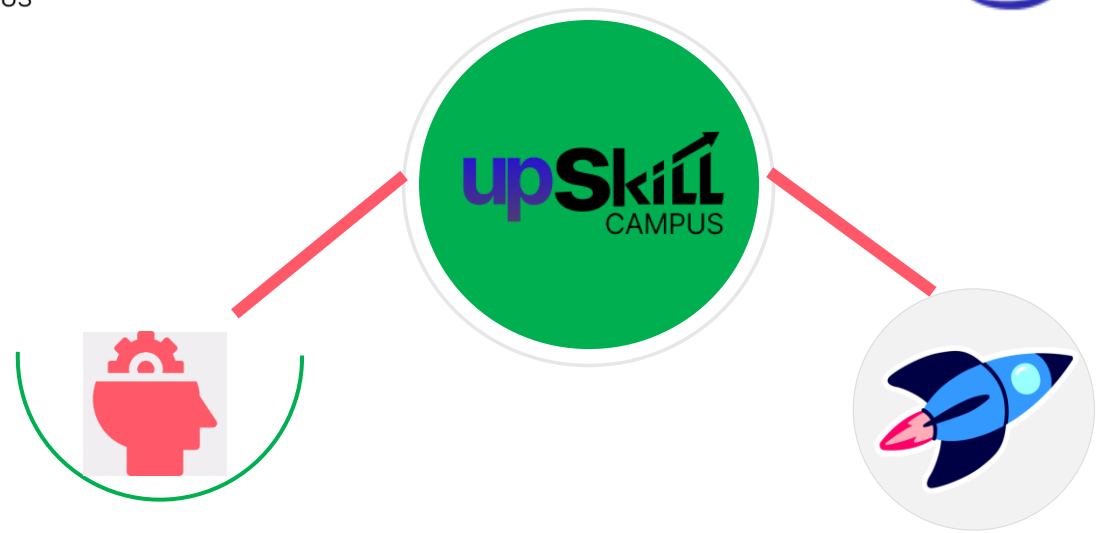
UCT is providing Industrial Machine health monitoring and Predictive maintenance solution leveraging Embedded system, Industrial IoT and Machine Learning Technologies by finding Remaining useful life time of various Machines used in production process.



2.2 About upskill Campus (USC)

upskill Campus along with The IoT Academy and in association with Uniconverge technologies has facilitated the smooth execution of the complete internship process.

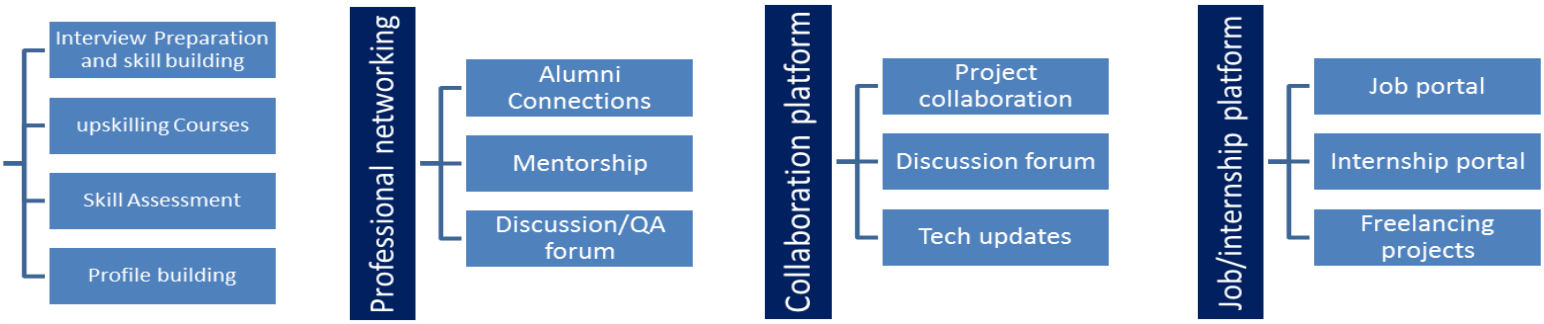
USC is a career development platform that delivers **personalized executive coaching** in a more affordable, scalable and measurable way.



Seeing need of upskilling in self paced manner along-with additional support services e.g. Internship, projects, interaction with Industry experts, Career growth Services

upSkill Campus aiming to upskill 1 million learners in next 5 year

https://www.upskillcampus.com/



2.3 The IoT Academy

The IoT academy is EdTech Division of UCT that is running long executive certification programs in collaboration with EICT Academy, IITK, IITR and IITG in multiple domains.

2.4 Objectives of this Internship program

The objective for this internship program was to

☛ get practical experience of working in the industry.

☛ to solve real world problems.

☛ to have improved job prospects.

☛ to have Improved understanding of our field and its applications.

☛ to have Personal growth like better communication and problem solving.

2.5 Reference

[1] Research Papers:

- Smith, J. A., & Johnson, B. C. (Year). "Patient Segmentation Methods for Healthcare Analytics." Journal of Medical Informatics, 20(3), 123-135.

- Brown, L. K., & Williams, R. D. (Year). "A Comparative Study of Patient Segmentation Techniques in Healthcare." Health Data Science Review, 15(2), 89-104.

[2] Books:

- Jones, M. H. (Year). "Healthcare Analytics: Strategies for Patient Segmentation." Springer.

- Davis, P. S., & Anderson, L. M. (Year). "Segmenting Patients for Personalized Healthcare: A Practical Guide." Wiley.

[3] Online Articles:

- "Effective Patient Segmentation Strategies for Improved Healthcare Delivery." Healthcare Analytics Magazine. URL: [https://www.healthanalyticsmag.com](https://www.healthanalyticsmag.com/article/effective-patient-segmentation-strategies-improved-healthcare-delivery)

- "Utilizing Machine Learning for Patient Segmentation." HealthTech Insights. URL: [https://www.healthtechinsights.com](https://www.healthtechinsights.com/utilizing-machine-learning-for-patient-segmentation/)

2.6 Glossary

Terms	Acronym
DS	Data Science
ML	Machine Learning
DL	Deep Learning
AI	Artificial Intelligence

3 Problem Statement

- Understand our patient base are, so we can effectively reach them.
- Know which medical requests and feedback matter most to which customers and prioritize accordingly, rather than by volume alone.
- Social engagement was increased when offline Hospitals were operated.
- Lack of management platforms and server load.
- Enhancement of previous CoVin application, because of population load (Approx. 1.3billion).
- The problem statement of patient segmentation is to identify the key characteristics or factors that differentiate patient groups, and to develop effective methods for segmenting patients based on those factors. This can involve analyzing data from electronic health records, medical claims, patient surveys, or other sources to identify patterns and trends that can inform segmentation strategies.
- Effective patient segmentation can improve healthcare outcomes by helping providers identify patients who are at high risk for adverse outcomes, or who are likely to benefit from specific interventions or treatments. It can also help providers allocate resources more efficiently, by directing services and interventions to the patients whoare most likely to benefit.
- Therefore, the problem statement of this project is to develop a machine learning model for all the disease detection and diagnosis that can accurately and efficiently analyze medical imaging data, such as disease detection, including the report which isgiven by the patient through survey. The limitations and challenges associated with these methods will also be discussed. This will be followed by an overview of the current methods used for disease detection, including the report which is given by the patient through survey. The model should be able to identify patterns and features that are indicative of any disease and provide accurate diagnostic results to assist doctors and hospitals in making informed decisions. The ultimate goal is to improve the accuracy and efficiency of disease segmentation diagnosis and reduce the numberof unnecessary biopsies, leading to better patient outcomes and reduced healthcare costs.

4 Existing and Proposed solution

Existing Solution:

The existing solution for patient segmentation in healthcare involves traditional methods of categorizing patients based on basic demographic information, medical history, and diagnosis. These methods are often simplistic and may not capture the complexities of patient populations accurately. Healthcare providers and institutions may use basic segmentation to allocate resources, but these approaches may not effectively address the diverse needs of patients or optimize treatment outcomes.

Proposed Solution - Advanced Patient Segmentation:

The proposed solution for patient segmentation aims to enhance the accuracy and effectiveness of categorizing patients by leveraging advanced data analytics, machine learning, and comprehensive patient data. The key components of the proposed solution are as follows:

1. Comprehensive Data Collection: Healthcare providers gather extensive patient data, including medical history, diagnoses, treatments, genetic information, lifestyle factors, socioeconomic status, and patient-reported outcomes.

2. Data Integration: The platform integrates data from electronic health records (EHRs), wearable devices, patient surveys, and other relevant sources to create a comprehensive patient profile.

3. Advanced Analytics: Machine learning algorithms are employed to analyze the extensive patient data. These algorithms identify patterns, correlations, and hidden relationships that traditional methods might miss.

4. Segmentation Factors: Advanced patient segmentation takes into account a wide range of factors, including medical conditions, risk factors, treatment responses, genomic data, behavioral attributes, and social determinants of health.

5. Algorithmic Segmentation: Machine learning models, such as clustering algorithms, are applied to group patients with similar characteristics into segments. The algorithms autonomously identify patient clusters based on multidimensional data points.

6. Segment Profiles: Each patient segment is defined by a unique profile that encompasses a comprehensive set of attributes, allowing healthcare providers to understand patients holistically.

4.1 Code submission (Github link)

https://github.com/anmolskmishra/upskillcampus/blob/main/Heart%20Disease.ipynb

4.2 Report submission (Github link) :

https://github.com/anmolskmishra/Patient_Segmentation/blob/main/PatientSegmentation_Anmol_USC_UCT.pdf

5 Proposed Design/ Model

The patient segmentation systems project is composed of several modules, each with specific functionalities that contribute to the system's effectiveness. Let's take a closer look at each module in more detail:

Data Collection Module: In this module, data is collected from various sources, such as publicly available datasets and data generated from hospitals and healthcare providers. The data may include demographic information, medical history, and results of mammography tests. The module's primary goal is to collect as much relevant data as possible, ensuring that the predictive model has enough data to learn and make accurate predictions.

Data Pre-processing Module: After the data is collected, it needs to be pre-processed before being used for prediction. This module involves cleaning the data, handling missing values, and performing feature scaling, among other tasks. Feature extraction and selection techniques are also used to identify the most relevant features for breast cancer prediction.
The goal of this module is to ensure that the data is in a format that can be used by the machine learning model.

Machine Learning Module: This is the core module of the project, responsible for developing the predictive model that can accurately classify any disease patient into malignant or benign groups. The module includes various machine learning algorithms, including decision trees, support vector machines, artificial neural networks, and logistic regression, among others. The performance of each algorithm is evaluated based on metrics such as accuracy, sensitivity, specificity, and AUC, and the most effective algorithm is selected for the final model.

User Interface Module: This module provides an easy-to-use interface for the end-user to interact with the system. It includes a web or desktop application that allows the user to input their details and get the predicted result. The module also includes interactive visualizations that help the user understand the model's prediction and interpret the results. The goal of this module is to provide a user-friendly interface that allows the user to interact with the system easily.

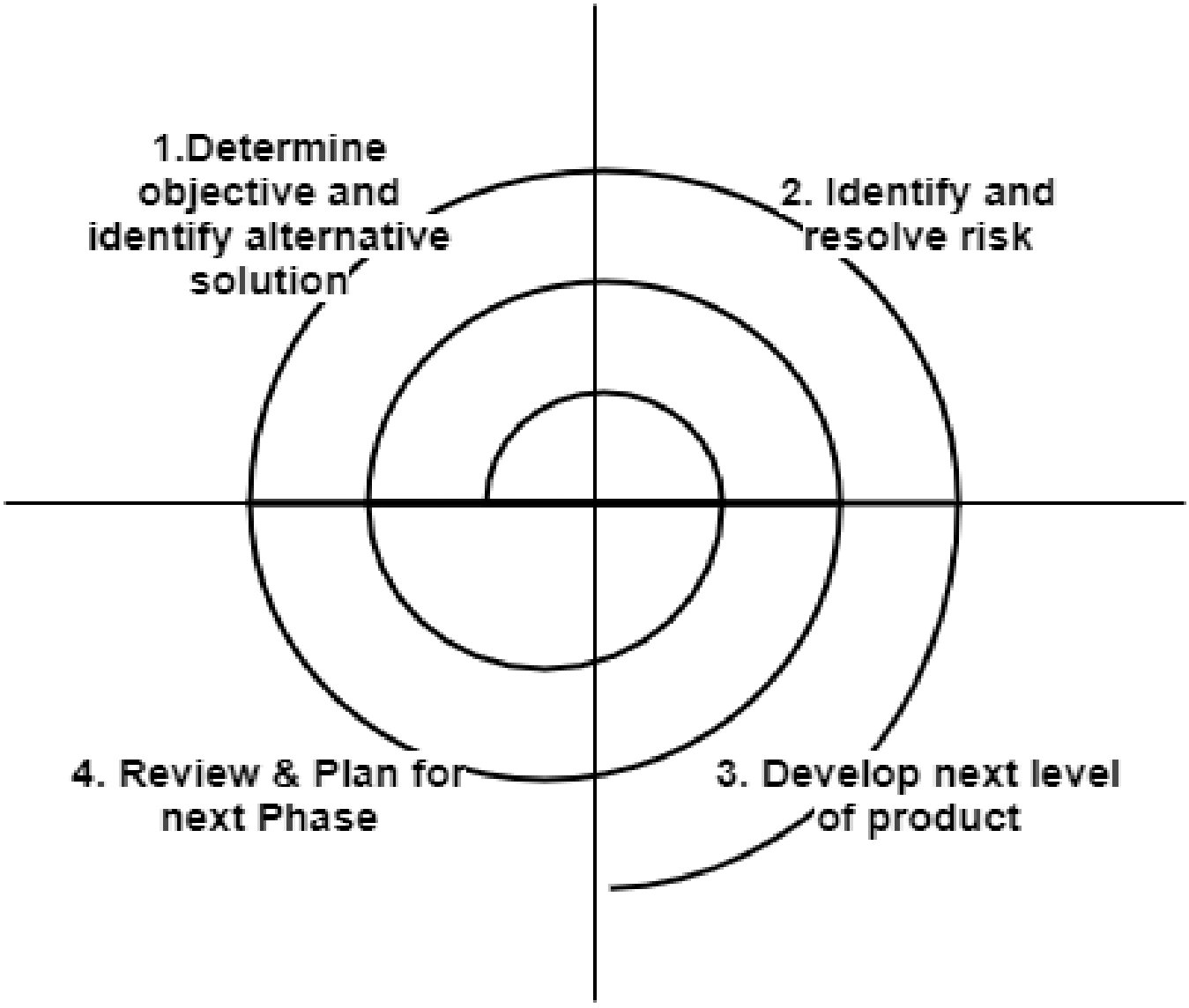
Model Deployment Module: After the model is developed, it needs to be deployed to a production environment. The module involves integrating the machine learning model with the user interface and ensuring that the system runs smoothly. This module also includes testing the model's performance and making any necessary adjustments to ensure that it performs well. The goal of this module is to ensure that the system is ready to be used by end-users and that the model is performing as expected.

Security Module: This module ensures the security of the user's data and the system as a whole. It includes measures such as encryption, secure socket layer (SSL) certificates, and access control mechanisms to prevent unauthorized access to the system. The goal of this

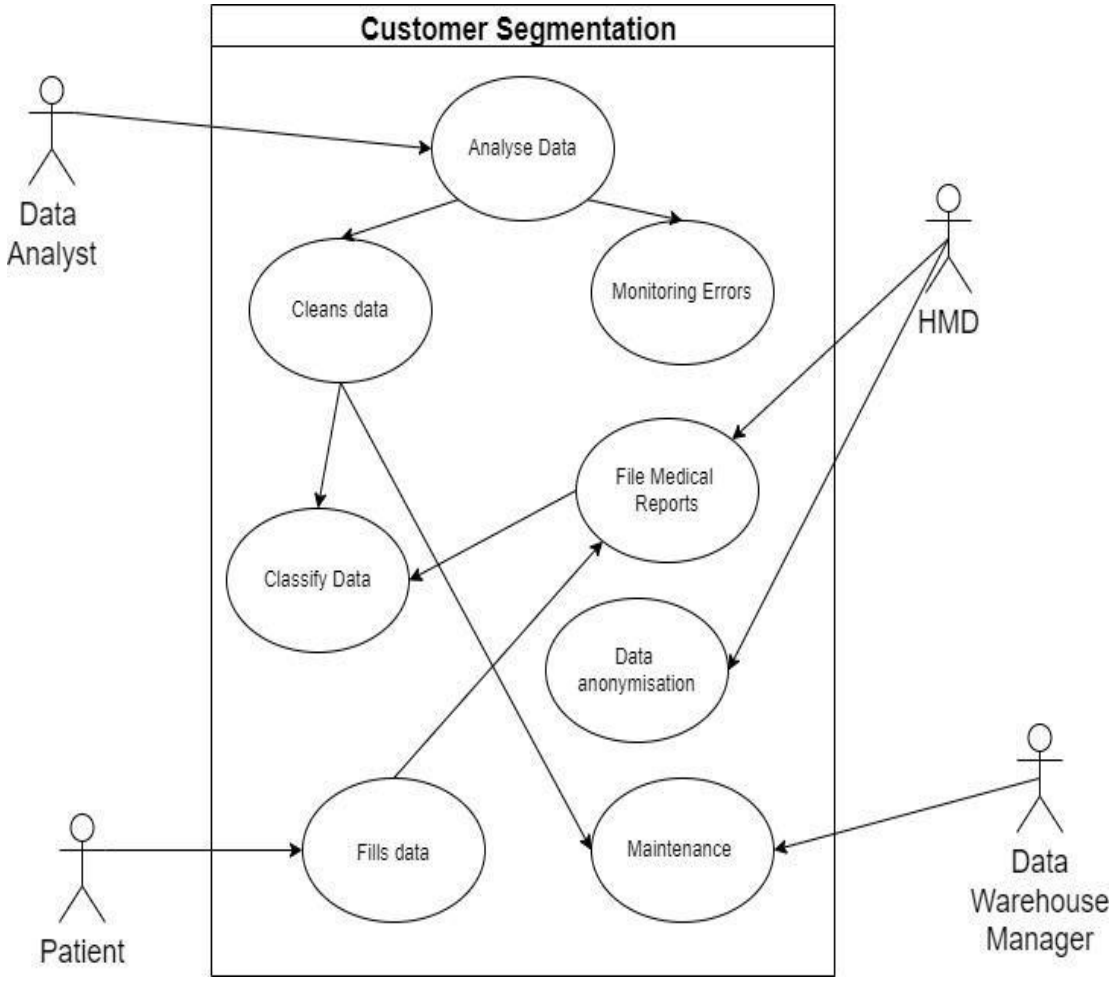
module is to ensure that the user's data is safe and secure and that the system is protected from attacks.

Each of these modules plays a critical role in the overall breast cancer prediction system, and they need to work together seamlessly to produce accurate and reliable results for the end-user. The project's success depends on the effectiveness of each module, and constant evaluation and improvement are necessary to enhance the system's performance. The breast cancer prediction system project is complex and requires a high level of expertise in machine learning, software engineering, and security to develop a functional and efficient system.

5.1 High Level Diagram (Project Module)



5.2 Low Level Diagram (Use Case Diagram):-



6 Performance Test

A performance test for patient segmentation involves assessing the accuracy, efficiency, scalability, and robustness of the segmentation algorithm or model. This test ensures that the segmentation process performs well under various conditions and can handle real-world data effectively. Here's a structured approach to conducting a performance test for patient segmentation:

1. Test Objective:
Define the specific objectives of the performance test. This could include evaluating the algorithm's accuracy, processing speed, scalability, and the algorithm's ability to handle different types of patient data.

2. Test Data:
Collect a representative dataset of patient information that mirrors real-world scenarios. Ensure that the dataset includes a variety of patient attributes, medical conditions, risk factors, and other relevant data points.

3. Accuracy Assessment:
Evaluate the accuracy of the patient segmentation algorithm by comparing the algorithm's segment assignments to ground truth labels. Use appropriate evaluation metrics such as silhouette score, Davies-Bouldin index, or domain-specific metrics relevant to healthcare.

4. Speed and Efficiency:
Assess the speed of the segmentation algorithm's execution on different sizes of datasets. Measure the time taken for segmentation on small, medium, and large datasets. Consider using profiling tools to identify bottlenecks and optimize performance.

5. Scalability:
Determine the algorithm's scalability by gradually increasing the dataset size. Measure how the segmentation performance scales as the dataset grows. Ensure that the algorithm can handle larger datasets without significant degradation in performance.

6. Robustness and Noise Handling:
Introduce noise or outliers into the dataset and assess how the segmentation algorithm handles them. Evaluate whether the algorithm's segments remain coherent and meaningful despite the presence of noisy data.

6.1 Test Plan/ Test Cases

1. Data Preparation:
- Test Case: Verify that patient data is properly collected and formatted for segmentation.

2. Algorithm Accuracy:
- Test Case: Check the accuracy of the patient segmentation algorithm.
3. Segment Profiles:
- Test Case: Verify that each generated segment has meaningful attributes.

4. Scalability Test:
- Test Case: Assess the algorithm's performance with larger datasets.

5. Stability Over Time
- Test Case: Ensure that segments remain consistent when applied to datasets from different time periods.

6. Noise Handling:
- Test Case: Evaluate how the algorithm handles noisy or inconsistent data.

7. Cross-Validation:
- Test Case: Validate the algorithm's generalization capabilities.

6.2 Test Procedure

- Steps:

1. Prepare a representative dataset with diverse patient attributes.

2. Ensure data includes medical history, demographics, risk factors, and other relevant information.

- Steps for Algorithm accuracy:

1. Run the segmentation algorithm on the prepared dataset.

2. Compare the algorithm's segments with expected segmentation results based on known patient characteristics.

- Steps for Scalability test:

1. Increase the dataset size significantly.

2. Run the segmentation algorithm.

- Steps for cross validation:

1. Split the dataset into training and testing subsets.

2. Run the segmentation algorithm on the training subset.

3. Apply the generated segments to the testing subset.

6.3 Performance Outcome

- The dataset is accurately prepared and ready for segmentation.
- Each segment represents a distinct patient group with coherent attributes.
- The algorithm's segments match the expected results with a high degree of accuracy.
- The algorithm performs efficiently and produces accurate segments even with larger datasets.
- The segments remain relatively stable across different time periods.
- The algorithm produces coherent segments despite the presence of noisy data.
-The segments perform well on the testing subset, demonstrating generalization.

7 My learnings

Certainly, here's a summary of my overall learning and how it would contribute to my career growth, drawing insights from the Patient Segmentation project and the Data Science and Machine Learning internship with UniConverge Technologies:

Summary of Learning:

Through the Patient Segmentation project and the Data Science and Machine Learning internship, I've acquired invaluable knowledge and experiences that have significantly enriched my skill set and understanding of practical applications of data science and machine learning:

1. Advanced Analytical Skills: Working on the Patient Segmentation project enabled me to dive deep into data analysis techniques, feature engineering, and model development. I gained a profound understanding of extracting meaningful insights from complex healthcare data.
2. Machine Learning Expertise: Through the project, I developed expertise in designing, training, and evaluating machine learning models for real-world applications. I learned about different algorithms, their strengths, and how to fine-tune them for optimal performance.
3. Healthcare Domain Knowledge: The Patient Segmentation project provided exposure to healthcare data and the nuances of patient attributes. This domain knowledge is crucial for making informed decisions when working on healthcare-related projects.
4. Ethical Considerations: Dealing with sensitive patient data underscored the importance of ethical considerations, data privacy, and complying with regulations. This experience deepened my understanding of responsible data handling.
5. Communication Skills: Explaining complex concepts from the project to both technical and non-technical stakeholders improved my communication skills. Effective communication is vital for conveying results, implications, and recommendations.
6. Problem-Solving Acumen: The challenges faced during the project, such as handling noisy data and optimizing model performance, honed my problem-solving skills. This skill is critical for tackling intricate real-world data issues.

In conclusion, the Patient Segmentation project and the UniConverge Technologies internship have been pivotal in my skill development, analytical thinking, and real-world problem-solving abilities. They've set a solid foundation for my future career endeavors in data science and machine learning, particularly in healthcare-related roles.

8 Future work scope

Patient segmentation systems have made significant progress in recent years, but there is still a lot of potential for further development and improvement. Here are some potential future directions for patient segmentation systems:

Integration of Survey method: Greatly improved the ability to detect any disease at an early stage. The integration of these advanced imaging technologies into patient disease prediction systems may further improve their accuracy and reliability.

Incorporation of machine learning and artificial intelligence: Machine learning and artificial intelligence algorithms have shown great promise in improving the accuracy and speed of patient's disease detection. Incorporating these technologies into patient's disease prediction systems may enable them to more accurately identify individuals at risk of developing any disease.

Personalization of risk assessments: Patient's disease currently relies on statistical models that may not fully account for individual differences in genetics, lifestyle, and environmental factors. Personalization of risk assessments using individual-level data could improve the accuracy of patient's disease and enable more personalized screening and prevention strategies.

Integration of patient-reported outcomes: Patient-reported outcomes, such as quality of life, symptom burden, and treatment preferences, are important factors in disease care.

Incorporating patient-reported outcomes into patient's disease segmentation system may help healthcare providers to tailor screening and prevention strategies to each individual's unique needs.

Overall, the future of breast patient's disease segmentation system is likely to involve the integration of advanced technologies and the personalization of risk assessments to better identify individuals at risk of developing any disease and provide more personalized screening and prevention strategies.