PROJECT REPORT

Name: Phebe Devadattam

Project Name: Breast Cancer Prediction

CONTENTS:

1. INTRODUCTION

OVERVIEW:

Breast cancer is a malignant cell growth in the breast. If left untreated, the cancer spreads to other areas of the body. Excluding skin cancer, breast cancer is the most common type of cancer in women in the United States, accounting for one of every three cancer diagnoses. Breast cancer ranks second among cancer deaths in women. This project aims at analyzing data of women residing in the state of Wisconsin, USA for to predict whether a case of breast cancer is malignant or benign.

PURPOSE:

The goal of this project is the application of several data mining and machine learning techniques to classify whether the tumor mass is benign or malignant in women residing in the state of Wisconsin, USA. This will help in understanding the important underlying importance of attributes thereby helping in predicting the stage of breast cancer depending on the values of these attributes. Through the understanding of the nature of attributes in cancer prediction, the healthcare community can perform additional research corresponding to these attributes to help prevent pervasion of breast cancer into the population of the USA.

2. LITERATURE SURVEY

EXISTING PROBLEM:

Breast cancer is one of the main causes of cancer death worldwide. Early diagnostics significantly increases the chances of correct treatment and survival, but this process is tedious and often leads to a disagreement between pathologists. Computer-aided diagnosis systems showed the potential for improving diagnostic accuracy. But early detection and prevention can

significantly reduce the chances of death. It is important to detect breast cancer as early as possible.

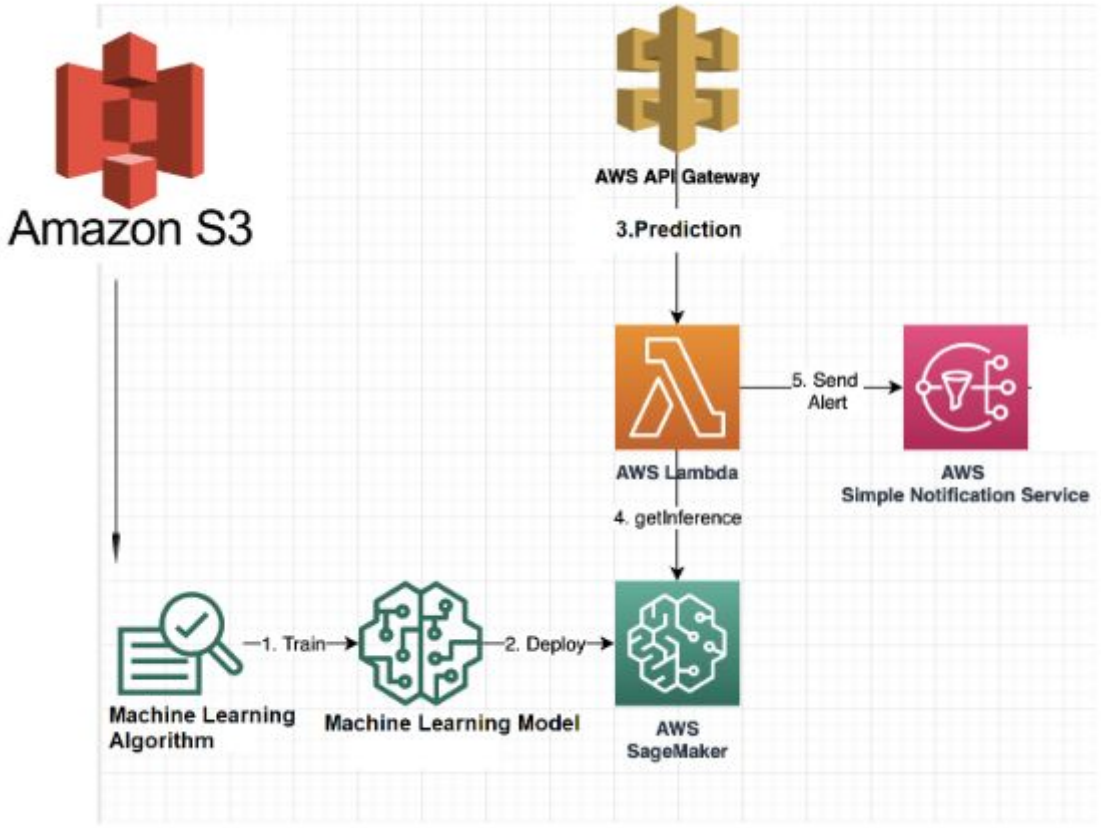
We will be building and deploying the model in AWS SageMaker and use SNS service to generate alerts about the risk.

PROPOSED SOLUTION:

Develop a model that is capable of detecting the Breast Cancer in early stages. The model must send alerts using the SNS service. The Machine learning model is trained and deployed on Amazon Sage Maker. Create an API Endpoint for the model with the help of API Gateway and AWS Lambda Service.

3. THEORETICAL ANALYSIS

BLOCK DIAGRAM:



SOFTWARE DESIGNING:

- ➢ AWS S3 is used for storage purposes and we use this to store our dataset , training and testing dataset.
- ➢ Starting from the client side, a client script calls an Amazon Gateway API action and passes parameter values.
- ➢ API Gateway is a layer that provides API to the client. In addition, it seals the backend so that AWS Lambda stays and executes in a protected private network. API Gateway passes the parameter values to the Lambda function.
- ➢ The Lambda function parses the value and sends it to the SageMaker model endpoint. The model performs the prediction and returns the predicted value to AWS Lambda.
- ➢ The Lambda function parses the returned value and sends it back to API Gateway. API Gateway responds to the client with that value.
- ➢ The AWS lambda function through the AWS SNS generates alerts to the mail.

3. EXPERIMENTAL INVESTIGATION

Amazon S3:

Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance. This means customers of all sizes and industries can use it to store and protect any amount of data for a range of use cases, such as websites, mobile applications, backup and restore, archive, enterprise applications, IoT devices, and big data analytics. Amazon S3 provides easy-to-use management features so you can organize your data and configure finely-tuned access controls to meet your specific business, organizational, and compliance requirements

AWS API Gateway:

Amazon API Gateway is a fully managed service that makes it easy for developers to create, publish, maintain, monitor, and secure APIs at any scale. APIs act as the "front door" for applications to access data, business logic, or functionality from your backend services. Using API Gateway, you can create RESTful APIs and WebSocket APIs that enable real-time two-way communication applications. API Gateway supports containerized and serverless workloads, as well as web applications.

API Gateway handles all the tasks involved in accepting and processing up to hundreds of thousands of concurrent API calls, including traffic management, CORS support, authorization and access control, throttling, monitoring, and API version management. API Gateway has no minimum fees or startup costs. You pay for the API calls you receive and the amount of data transferred out and, with the API Gateway tiered pricing model, you can reduce your cost as your API usage scales.

AWS Lambda:

AWS Lambda lets you run code without provisioning or managing servers. With Lambda, you can run code for virtually any type of application or backend service – all with zero administration. Just upload your code and Lambda takes care of everything required to run and scale your code with high availability. You can set up your code to automatically trigger from other AWS services or call it directly from any web or mobile app.

AWS SNS:

Amazon Simple Notification Service (SNS) is a fully managed messaging service for both system-to-system and app-to-person (A2P) communication. It enables you to communicate between systems through publish/subscribe (pub/sub) patterns that enable messaging between decoupled microservice applications or to communicate directly to users via SMS, mobile push and email.

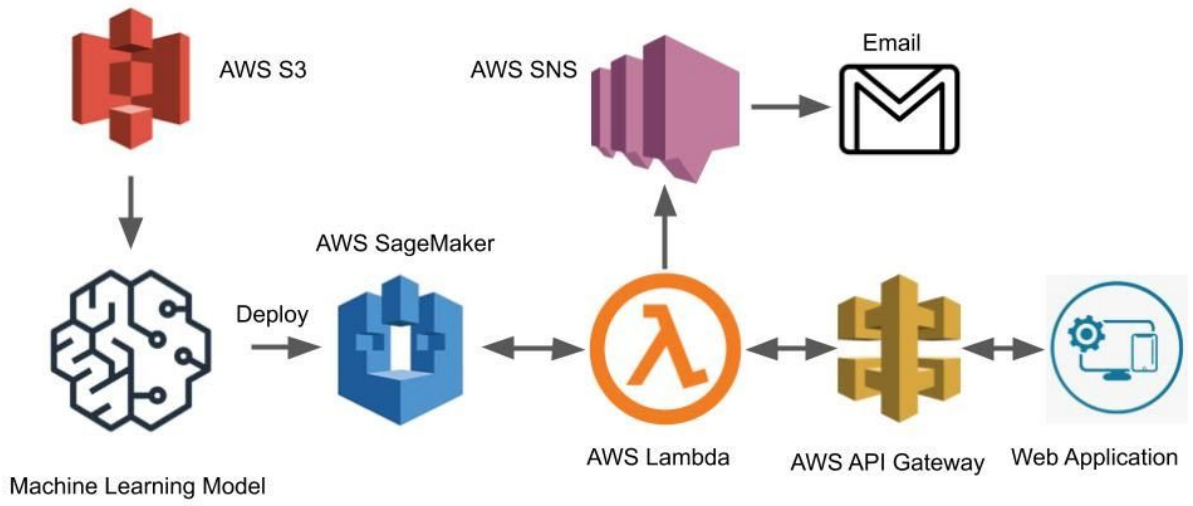
The system-to-system pub/sub functionality provides topics for high-throughput, push-based, many-to-many messaging. Using Amazon SNS topics, your publisher systems can fanout messages to a large number of subscriber systems or customer endpoints including Amazon SQS queues, AWS Lambda functions and HTTP/S, for parallel processing. The A2P messaging functionality enables you to send messages to users at scale using either a pub/sub pattern or direct-publish messages using a single API.

Amazon SageMaker:

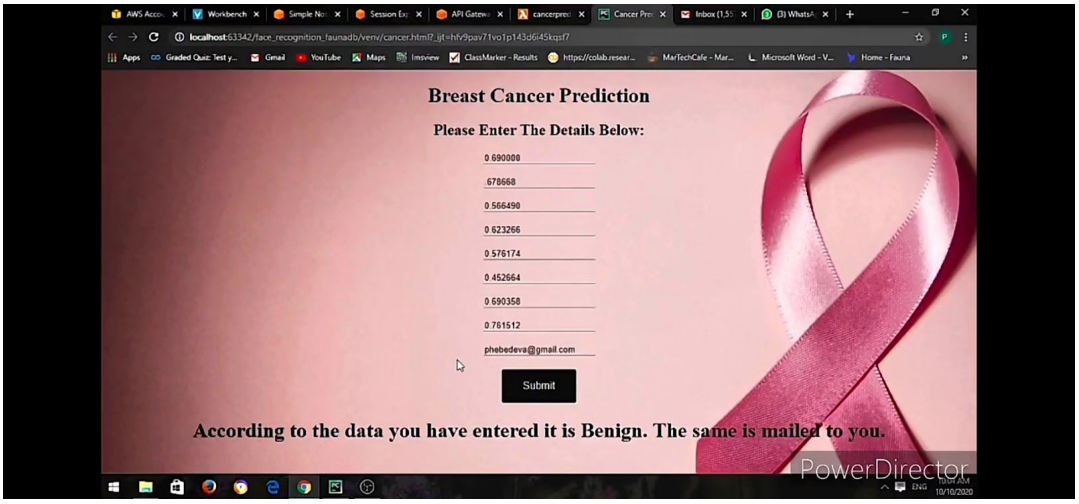
Amazon SageMaker is a fully managed service that provides every developer and data scientist with the ability to build, train, and deploy machine learning (ML) models quickly. SageMaker removes the heavy lifting from each step of the machine learning process to make it easier to develop high quality models.

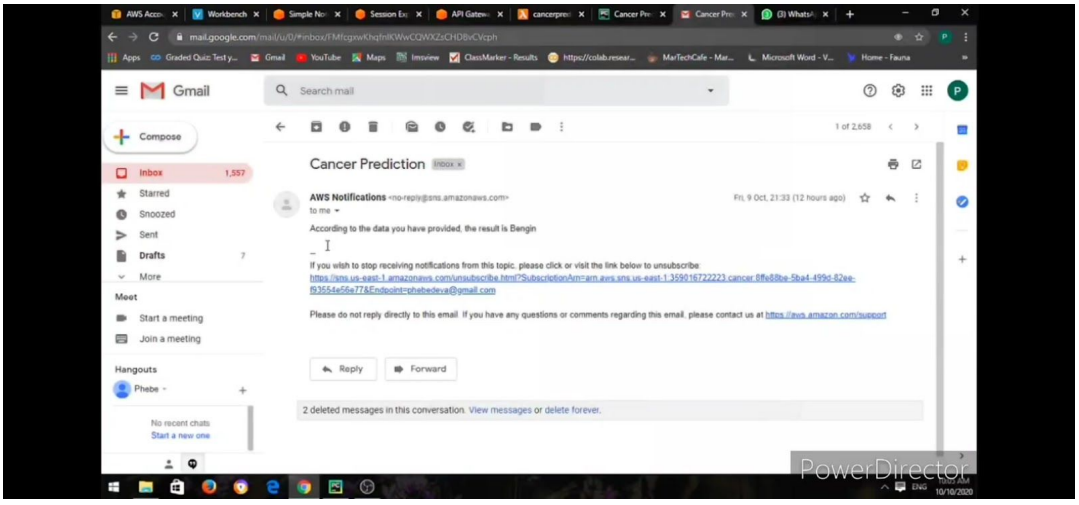
Traditional ML development is a complex, expensive, iterative process made even harder because there are no integrated tools for the entire machine learning workflow. You need to stitch together tools and workflows, which is time-consuming and error-prone. SageMaker solves this challenge by providing all of the components used for machine learning in a single toolset so models get to production faster with much less effort and at lower cost.

5.FLOWCHART



6. RESULT:





7. ADVANTAGES AND DISADVANTAGES:

The inclusion-exclusion selection criteria of the US and the Swiss datasets may have influenced the association between observed variance and outcomes. In the US population-based sample, YBCS had fewer affected relatives than their cancer-free relatives. Thus, the number of affected relatives was detected as an important variable but without external validity in interpretation. Interpretability of the function modeled by ML algorithms is only partially limited by the "black-box" nature of ML algorithms in our study because we included a limited number of well-established breast cancer risk factors. However, the inherent complexity of how risk factors interact with each other, their independent effect on the outcome, and how effect sizes are determined within each ML algorithm is not known.

Significant strengths of the study include the novelty of the approach, i.e., applying ML algorithms in individual breast cancer risk prediction and comparing predictive accuracy with existing models. The improvement achieved with ML algorithms in accurate classification of women with and without breast cancer compared to the state-of-the-art model-based approaches was striking. We demonstrated a range of ML algorithms with cross-validations, which is lacking in other applications of ML for cancer prognosis. Different ML algorithms for feature selection and classification showed great adaptability and discriminatory

accuracy in our study by handling multidimensional and heterogeneous data. Ranking variable importance may inform algorithm selection with diverse predictive risk factors for future development of new risk prediction models.

8. APPLICATIONS

Health care centers

9. CONCLUTION:

Predictive models are essential in personalized medicine because they contribute to early identification of high-risk individuals based on known epidemiological and clinical risk factors. Accurate breast cancer risk estimates can inform clinical care and risk management across the breast cancer continuum, e.g., behavioral changes, chemoprevention, personalized screening, and risk-stratified follow-up care.ML approaches offer the exciting prospect of achieving improved and more precise risk estimates. This is the first step in developing new risk prediction approaches and further explores diverse risk factors. ML algorithms are not limited to a specific number of risk factors but have the flexibility to change or incorporate additional ones. The improvement in predictive accuracy achieved in this study should be further explored and duplicated with prospective databases and additional risk factors, e.g., mammographic density, risk factors in IBIS Breast Cancer Risk Evaluation Tool, and polygenic genetic scores. Improvements in computational capacity and data management in healthcare systems can be followed by opportunities to exploit ML to enhance risk prediction of disease and survival prognosis in clinical practice.

10. FUTURE SCOPE

The analysis of the results signifies that the integration of multidimensional data along with different classification,features sel;ection and dimensionality reduction technique can provide auspicious tools for inference in this domain. Further research in this field should be carried out for the better performance of the classification techniques so that it can predict more variables. We should intend how to parameterize our classification techniques hence to achieve high accuracy. We should look into many datasets and how furter machine learning algorithms can be used to characterize breast cancer . We want to reduce the error rates with maximum accuracy.

11.BIBLOGRAPHY

K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methodsand Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).

APPENDIX:

Dataset:https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Source Code:https://github.com/SmartPracticeschool/SPS-3296-Breast-Cancer-Risk-Prediction-using-AWS-SageMaker/tree/master