Title Page

- Title of project: **Breast Cancer Prediction**
- Name of students: Mst. Sumiya Siddika, Md. Shakil Ahmed, Amit Azim Amit
- Roll: 1975, 2013, 2090
- **Group-19**

Abstract

- Brief overview of the project:
  The project aims to develop a breast cancer risk prediction model using advanced data analytics techniques. The primary focus is on accurately predicting the risk of breast cancer in individuals based on their demographic information, medical history, genetic markers, and lifestyle factors. This predictive model could significantly assist healthcare professionals in identifying individuals at higher risk, enabling early intervention and personalized care strategies.

- Problem statement:
  Breast cancer is the most prevalent cancer in women, accounting for about one-third of all cancer diagnoses in this population, and it is also the second greatest cause of cancer-related mortality in women. Breast tissue tumors, or aberrant cell development in the breast tissue, are what cause breast cancer. Tumors can be benign (not cancerous), pre-malignant (pre-cancerous), or malignant (cancerous), therefore a tumor does not necessarily indicate cancer. Breast cancer is frequently diagnosed through procedures including MRIs, mammograms, ultrasounds, and biopsies.

- Goals:
  The purpose of this study is to create a model for predicting breast cancer risk that may be used to pinpoint women who are most at risk for the condition. By attaining this objective, the project hopes to equip medical professionals with a tool that improves their capacity to recognize people who are more likely to acquire breast cancer, resulting in more specialized and efficient preventative measures and early detection procedures.

- Proposed methods:
  Apply the fundamental concepts of machine learning from an available dataset. Evaluate and interpret results and justify interpretation based on the observed data set. Create notebooks that serve as computational records and document

my thought process. The analysis is divided into four sections, saved in Jupiter notebooks in this repository. Firstly, Identifying the problem and Data Sources. Secondly, Exploratory Data Analysis, Thirdly, Pre-Processing of the Data and finally model to predict whether breast cell tissue is malignant or Benign.

Introduction

- Background information on the problem domain
  A machine learning method called support vector machine (SVM) can be applied to classification and regression tasks. SVM can be used to pinpoint patients who are at a high risk of contracting the illness or who are most likely to experience a recurrence of the illness in the setting of breast cancer prediction.
  One type of cancer that begins in the breast is breast cancer. The second most significant cause of cancer death for women is the most prevalent cancer among females. The creation of an accurate and dependable approach for utilizing SVMs to predict breast cancer falls under the problem domain of breast cancer prediction using SVM. This is a difficult task because breast cancer is a complex disease with numerous risk factors.

- Discussion of why this problem is important to solve with machine learning:
  Machine learning can be used to develop more accurate breast cancer risk prediction models. This is because machine learning algorithms can learn from large amounts of data to identify patterns that may not be obvious to humans. This can lead to the development of models that can accurately predict a woman's risk of developing breast cancer, even if she has no family history of the disease.

- Description of the goals and objectives of the project
  The goals of this project:
  - ➢ Develop a machine learning model that can accurately predict a woman's risk of developing breast cancer.
  - ➢ Identify the factors that are most strongly associated with breast cancer risk.
  - ➢ Validate the model using a separate dataset of women diagnosed with breast cancer.
  - ➢ Make the model available to healthcare providers and women so that they can make informed decisions about screening and prevention.

The objectives of this project :

- ➢ Collect data on a large number of women, including their personal and family history of breast cancer, their reproductive history, and their lifestyle factors.
- ➢ Use machine learning algorithms to identify the factors that are most strongly associated with breast cancer risk.
- ➢ Validate the model using a separate dataset of women who have been diagnosed with breast cancer.

Literature Review

- Summary of relevant prior work and existing methods/techniques in this problem domain

Title	Methodology	Related work	Result
Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine	SVM	1. Khan et al. (2018) demonstrated a system for the identification of breast cancer using GA feature selection and Rotation Forest (RF).	high sensitivity (0.89), specificity (0.73), positive predictive values (0.75), and negative predictive values (0.89).
Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors	SVM, KNN	1. Azar et al. (2017) proposed a novel technique for the detection of breast cancer.	SVM 99.68%, KNN 98.25% accuracy
A Gene Signature for Breast Cancer Prognosis Using Support Vector Machine	support vector machine-based recursive feature elimination (SVM-RFE)	1. Van't Veer, L. J., H. Dai, et al. "Gene expression profiling predicts clinical outcome of breast cancer."	Accuracy 34%, Sensitivity 48% and Specificity 3%

- Discussion of limitations of current approaches that the project aims to address
  Limitations:
  1. The study was conducted in a single institution, so the results may not be generalizable to other populations.
  2. The study was retrospective, so there is a risk of bias.
  3. The follow-up period was relatively short, so the long-term predictive performance of the model is not known.

Proposed Methods

- Detailed description of the proposed machine learning methods to be used
  Create predictive algorithms to foretell the detection of breast cancer. Create a prediction model for the diagnosis of a breast tumor using the SVM machine learning method. Breast tumors can be classified as benign or malignant based on a binary variable. Additionally, examine the model using the receiver operating curves' (ROC) confusion matrix, which is crucial for analyzing and understanding the fitted model.

- Reasoning behind the choice of methods and how they will achieve the goals of the project
  Here are some of the reasons why SVM is a good choice for breast cancer prediction:
  - ➔ SVM is a very powerful machine-learning algorithm that can be used for a variety of tasks.
  - ➔ SVM is very accurate, even for small datasets.
  - ➔ SVM is robust to noise and outliers.
  - ➔ SVM can be used for both linear and nonlinear classification problems.
  - ➔ SVM can be used to handle high dimensional data.

  Here are some of the ways that SVM can be used to achieve the goals of the breast cancer prediction project:

  - ➔ SVM can be used to identify the factors that are most strongly associated with breast cancer risk. This information can be used to develop personalized screening and prevention recommendations.
  - ➔ SVM can be used to develop a mathematical model that can be used to predict a woman's risk of developing breast cancer based on her individual risk factors. This model can identify women at high risk of developing breast cancer and who may benefit from early screening or other interventions.

➔ SVM can be used to validate the accuracy of breast cancer risk prediction models. This is important to ensure that the models are accurate and can be used to make informed decisions about screening and prevention.

- Any dataset requirements and how the data will be obtained/generated
  The University of California, Irvine's machine learning repository makes the Breast Cancer datasets available. 569 samples of both malignant and benign tumor cells are included in the collection.
  The first two columns of the dataset contain, respectively, the samples' distinctive ID numbers and the accompanying diagnosis (M = malignant, B = benign). Columns 3-32 contain 30 real-value traits that may be used to create a model to determine whether a tumor is benign or malignant. These features were computed using digital photographs of the cell nuclei.

Project Plan

Breakdown of tasks to be completed	**Timeline estimating**	**Milestones**	**Deliverables**
Identify the types of information contained in the data set.	1 day	Gather data sets and review documentation.	List of data types and descriptions.
Explore the variables to assess how they relate to the response variable.	2 days	Perform descriptive statistics, correlation analysis, and other exploratory data analysis techniques.	Report on variable relationships.
Find the most predictive features of the data and filter it so it will enhance the predictive power of the analytics model.	3 days	Apply feature selection techniques, such as univariate selection, recursive feature elimination, and random forest.	List of selected features.

Construct predictive models to predict the diagnosis of a breast tumor.	5 days	Build and evaluate different machine learning models, such as logistic regression, decision trees, and support vector machines.	Model performance metrics, such as accuracy, precision, and recall.
Optimizing the Support Vector Classifier.	2 days	Tune the hyperparameters of the support vector classifier to improve its performance.	Model with optimized hyperparameters.

Expected Results

Given the findings of a breast fine needle aspiration (FNA) test, which involves using a thin needle-like to one used for blood samples to take some fluid or cells from a breast lesion or cyst (a lump, sore, or swelling). Because of this, I built a model that uses two training classes to categorize breast cancer tumors:

- 1= Malignant (Cancerous) - Present
- 0= Benign (Not Cancerous) -Absent

References

- List of cited literature and sources

[1] Van't Veer, L. J., H. Dai, et al. "Gene expression profiling predicts clinical outcome of breast cancer." Nature 415(6871): 530-536, 2002..

[2] Ahmad Taher Azar, Shaimaa Ahmed El-Said, “Probabilistic neural network for breast cancer classification,” Neural Computing and Applications, Springer, vol. 23, pp.1737-1751, 2013.

[3] Weigelt, B., Z. Hu, et al. "Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer." Cancer research 65(20): 9155, 2005.

[4] Na KY, Kim KS, LeeJE, Kim HJ, Yang JH, Ahn SH,et al. The 70-gene prognosticsignaturefor Korean breastcancer patients. J Breast Cancer 2011;14:33-8