Patient Sickness Prediction

SUBMITTED BY:

Priyam Pal

Roll – AIML/2021/005

For,

Advanced Machine Learning Laboratory

(PC – AIML691)

UNDER THE GUIDANCE OF

Dr. Shampa Sengupta

(Assistant Professor of IT Department)

DEPARTMENT OF INFORMATION TECHNOLOGY
MCKV INSTITUTE OF ENGINEERING
(NAAC ACCREDITED “A” GRADE AUTONOMOUS INSTITUTE)
243, G.T. ROAD(NORTH), LILUAH
HOWRAH-711204

1. Abstract

The project "Patient Sickness Prediction" focuses on leveraging machine learning algorithms to develop a predictive model for identifying and forecasting various diseases based on patient symptoms and medical data. With the increasing availability of healthcare data and advancements in data analytics, this project aims to revolutionize healthcare by enabling proactive intervention and personalized treatment plans. By analysing comprehensive datasets encompassing patient demographics, medical history, lifestyle factors, and clinical indicators, the predictive model employs sophisticated machine learning techniques such as decision trees, random forests, and k-nearest neighbours to accurately predict the likelihood of different illnesses.

Through the integration of machine learning algorithms, the project seeks to achieve several objectives, including early disease detection, risk stratification, and personalized healthcare interventions. By identifying patterns and correlations within the data, the predictive model can assist healthcare professionals in making informed decisions, prioritizing patient care, and allocating resources efficiently. Additionally, the project emphasizes the importance of ethical considerations, data privacy, and transparency in model development and implementation.

Overall, the "Patient Sickness Prediction" project holds significant promise in revolutionizing healthcare delivery by harnessing the power of machine learning to proactively predict and prevent diseases, ultimately leading to improved patient outcomes, reduced healthcare costs, and enhanced population health.

Keywords: Machine Learning (ML), Medical Science, Healthcare, Exploratory Data Analysis (EDA)

2. Introduction

2.1 Objective of this Project:

The primary objective of the project "Patient Sickness Prediction" is to develop an advanced machine learning-based system capable of predicting various diseases and health conditions based on patient data and symptoms. Through the utilization of machine learning algorithms and predictive analytics techniques, the project aims to achieve several key goals:

- **Early Disease Detection**: The project seeks to detect diseases at their earliest stages by analysing patient data, medical history, lifestyle factors, and symptoms. Early detection is crucial for timely intervention and improved patient outcomes.
- **Risk Stratification**: By assessing individual patient risk factors and analyzing patterns within the data, the project aims to stratify patients based on their likelihood of developing specific diseases. This risk stratification facilitates targeted preventive measures and personalized healthcare interventions.
- **Proactive Healthcare Interventions**: Through the predictive model developed in the project, healthcare providers can proactively identify high-risk patients and intervene with preventive measures, lifestyle modifications, and appropriate medical treatments to mitigate disease progression and improve health outcomes.
- **Resource Allocation**: By accurately predicting disease occurrence and identifying high-risk patient populations, the project assists healthcare organizations in optimizing resource allocation, including staffing, equipment, and healthcare services, to meet the needs of patients effectively.
- **Empowering Patients**: The project aims to empower patients by providing them with personalized insights into their health risks and enabling them to take proactive steps towards disease prevention and management.

Overall, the objective of the "Patient Sickness Prediction" project is to leverage cutting-edge machine learning technology to revolutionize healthcare delivery, improve patient care, and enhance population health outcomes.

2.2 Machine Learning Algorithm Used:

- Naïve Bayes Classifier
- Decision Tree Classifier
- Random Forest Algorithm
- Logistic Regression

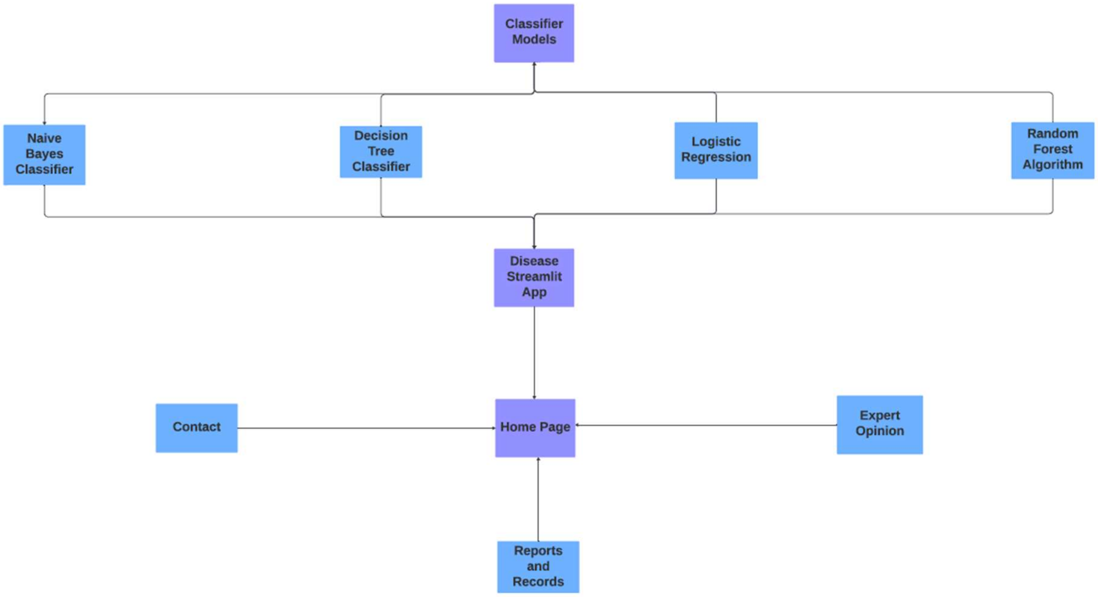
2.3 Data Analytics Software Used:

- **Numpy** - for Complex Mathematical operations.
- **Pandas** - for data manipulation and analysis.
- **Matplotlib** and **Seaborn** - for Data Visualization.
- **Scikit-learn** - for Implementing Machine Learning Algorithms.
- **Streamlit** - for Website Design and Deployment

2.4 Concept of the Project:

In the realm of healthcare, predicting diseases before they manifest can be a game-changer. It can lead to early interventions, better management of health conditions, and improved patient outcomes. To this end, we propose the development of a Disease Prediction Model using Machine Learning (ML) techniques. This model will analyse various health parameters of an individual and predict the likelihood of them developing a specific disease. The parameters could include age, gender, lifestyle habits, genetic factors, and existing health conditions, among others.

2.5 Block Diagram of the Project



2.6 Purpose of the Project:

The purpose of the project "Patient Sickness Prediction" is to develop a robust system that utilizes machine learning algorithms to predict diseases based on patient symptoms or other relevant factors. By leveraging data science and predictive analytics, the project aims to enhance healthcare outcomes by enabling early detection and intervention for various illnesses. The ultimate goal is to create a tool that healthcare professionals can use to assist in diagnosing diseases promptly, improving patient care, and potentially saving lives. Additionally, such a system could empower individuals to monitor their health proactively, leading to better management of chronic conditions and overall well-being.

3. Experimental Results

3.1 Description of the Datasets Used:

1. Breast Cancer Dataset:

The Breast Cancer dataset comprises instances used for the classification of breast cancer. It includes features computed from digitized images of fine needle aspirates (FNA) of breast masses. The dataset contains a total of 699 instances, with 458 being benign cases. Each instance is described by attributes, and the dataset provides information about the attributes and their values. Additionally, the dataset is commonly used for machine learning tasks, particularly in binary classification problems where the objective is to distinguish between benign and malignant tumors. This dataset is valuable for developing predictive models to aid in the diagnosis and prognosis of breast cancer, contributing to advancements in medical research and healthcare.

2. Diabetes Dataset

The Diabetes dataset typically consists of healthcare statistics, patient information, and medical measurements related to diabetes. These datasets often contain various features such as age, gender, body mass index (BMI), blood pressure, insulin level, and blood glucose level. They are commonly used in machine learning and data analysis projects aimed at predicting diabetes risk, diagnosis, and management. Datasets may vary in size, with some containing thousands of instances collected over several years from healthcare facilities, surveys, or research studies. Researchers and data scientists analyze these datasets to develop predictive models, identify risk factors, and understand trends in diabetes prevalence and management. Some datasets are publicly available and freely accessible online, facilitating research and collaboration in the field of diabetes healthcare. They may be used for research purposes, academic studies, or to develop applications and tools for diabetes prevention, diagnosis, and treatment. Diabetes datasets play a crucial role in advancing medical research, healthcare analytics, and public health interventions related to diabetes prevention and management.

3. Heart Disease Dataset

The Heart Disease dataset is widely used in machine learning and healthcare research. It typically includes patient data related to heart health, such as demographic information, medical history, and various physiological measurements. The dataset often contains attributes like age, gender, cholesterol levels, blood pressure, presence of heart conditions, and other relevant factors. Different versions of the dataset exist, with varying numbers of attributes and instances. For instance, the Cleveland Heart Disease dataset from the UCI Machine Learning Repository contains 76 attributes, but most experiments focus on a subset of 14 attributes. Another version may include comprehensive patient data from multiple sources. Researchers and data scientists use this dataset to develop predictive models for heart disease risk assessment, diagnosis, and treatment planning. By analyzing the dataset, they aim to identify patterns, risk factors, and correlations that can aid in early detection and prevention of heart-related conditions. The dataset's availability and standardization facilitate collaboration among researchers and enable the development of tools and technologies to improve cardiovascular healthcare outcomes.

4. HyperTension Dataset:

The Hypertension dataset typically comprises data related to blood pressure measurements, demographic information, lifestyle factors, medical history, and other relevant clinical variables. It aims to capture information about individuals with varying degrees of blood pressure levels, including normal (healthy), prehypertension, hypertension, and related conditions. Datasets may vary in terms of the number of attributes and instances, with some focusing specifically on blood pressure data and others incorporating a broader range of factors contributing to hypertension. Researchers and healthcare professionals use these datasets to study the prevalence, risk factors, treatment outcomes, and predictive modeling of hypertension. Machine learning algorithms are often applied to analyze the data, aiming to develop accurate classification models for hypertension detection, risk assessment, and personalized treatment planning. These datasets play a crucial role in advancing our understanding of hypertension, facilitating the development of preventive strategies, and improving clinical management practices to mitigate the adverse health effects associated with high blood pressure

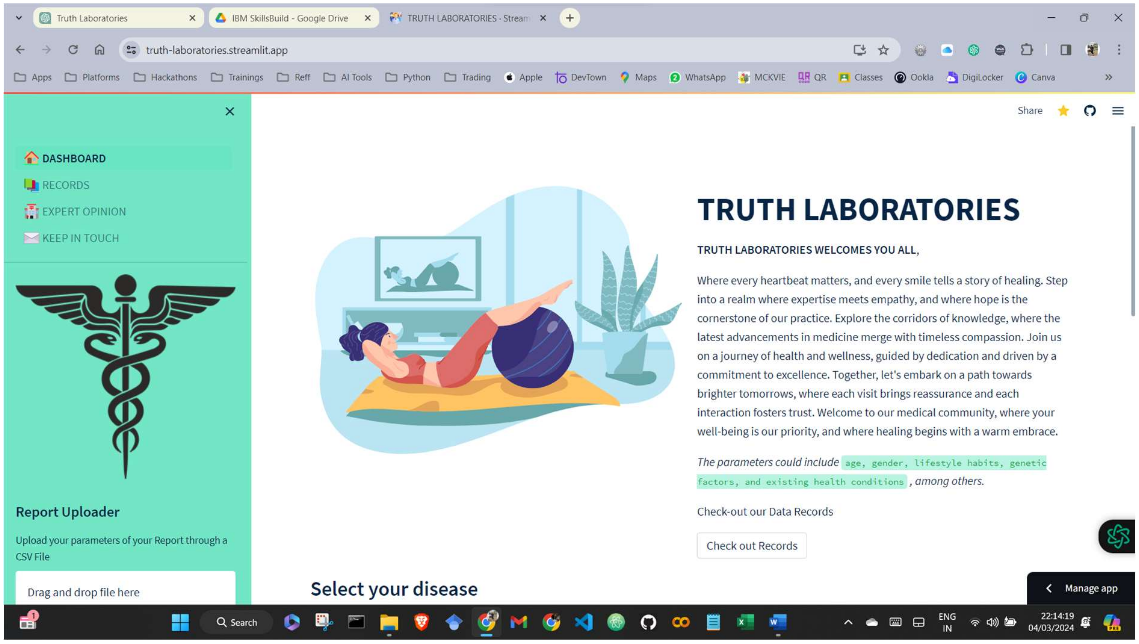
5. Kidney Disease Dataset:

The Chronic Kidney Disease (CKD) dataset contains clinical and laboratory data relevant to predicting chronic kidney disease in patients. It typically includes features such as age, blood pressure, specific gravity, albumin, sugar levels, red blood cell count, pus cell count, and other indicators obtained from urinalysis and blood tests. These datasets aim to predict the presence or progression of chronic kidney disease based on various patient characteristics and medical parameters. They are collected from hospitals or clinical settings and are used for research purposes to develop machine learning models for early detection, prognosis, and management of chronic kidney disease. Researchers analyze the CKD datasets to identify patterns, risk factors, and predictive markers associated with kidney disease. Machine learning algorithms, including classification, regression, and clustering techniques, are commonly applied to these datasets to build predictive models and assist healthcare professionals in diagnosing and treating chronic kidney disease effectively. The CKD dataset is valuable for advancing our understanding of kidney disease, improving patient outcomes, and informing clinical decision-making.

6. Stroke Dataset:

The Stroke dataset typically comprises clinical and demographic features used to predict the occurrence of stroke in patients. These features commonly include attributes such as age, gender, hypertension status, heart disease history, smoking status, and BMI. Additionally, some datasets may include medical indicators such as average glucose levels, cholesterol levels, and BMI. Researchers and data scientists utilize Stroke datasets to develop predictive models aimed at identifying individuals at risk of stroke. Machine learning algorithms, including logistic regression, decision trees, and neural networks, are commonly employed to analyze these datasets and build predictive models. The goal of utilizing Stroke datasets is to facilitate early detection, prevention, and management of strokes, a critical health concern worldwide. By identifying individuals at high risk of stroke, healthcare professionals can implement targeted interventions and lifestyle modifications to reduce the incidence and severity of strokes. Stroke datasets contribute to advancing medical research and improving patient outcomes by providing valuable insights into stroke risk factors, prognosis, and potential interventions

4. Interface



DASHBOARD

RECORDS

EXPERT OPINION

KEEP IN TOUCH

Report Uploader

Upload your parameters of your Report through a CSV File

Drag and drop file here

TRUTH LABORATORIES

TRUTH LABORATORIES WELCOMES YOU ALL,

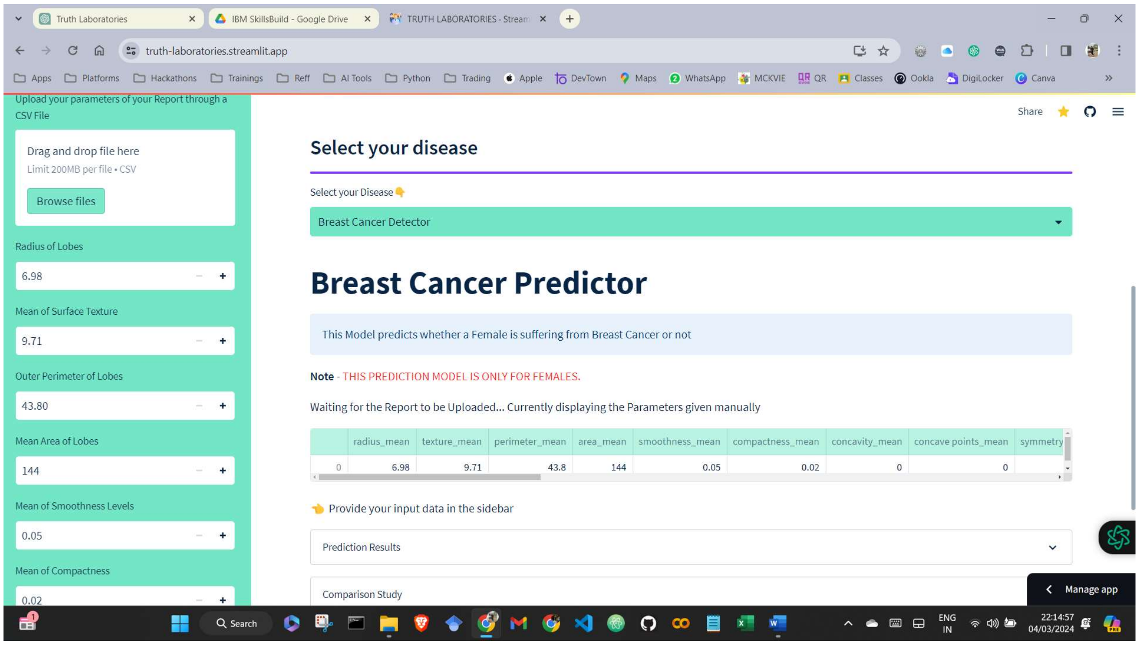
Where every heartbeat matters, and every smile tells a story of healing. Step into a realm where expertise meets empathy, and where hope is the cornerstone of our practice. Explore the corridors of knowledge, where the latest advancements in medicine merge with timeless compassion. Join us on a journey of health and wellness, guided by dedication and driven by a commitment to excellence. Together, let's embark on a path towards brighter tomorrows, where each visit brings reassurance and each interaction fosters trust. Welcome to our medical community, where your well-being is our priority, and where healing begins with a warm embrace.

The parameters could include `age, gender, lifestyle habits, genetic factors, and existing health conditions` *, among others.*

Check-out our Data Records

Check out Records

Select your disease



Upload your parameters of your Report through a CSV File

Drag and drop file here
Limit 200MB per file • CSV

Browse files

Radius of Lobes: 6.98

Mean of Surface Texture: 9.71

Outer Perimeter of Lobes: 43.80

Mean Area of Lobes: 144

Mean of Smoothness Levels: 0.05

Mean of Compactness: 0.02

Select your disease

Select your Disease 👇

Breast Cancer Detector

Breast Cancer Predictor

This Model predicts whether a Female is suffering from Breast Cancer or not

Note - THIS PREDICTION MODEL IS ONLY FOR FEMALES.

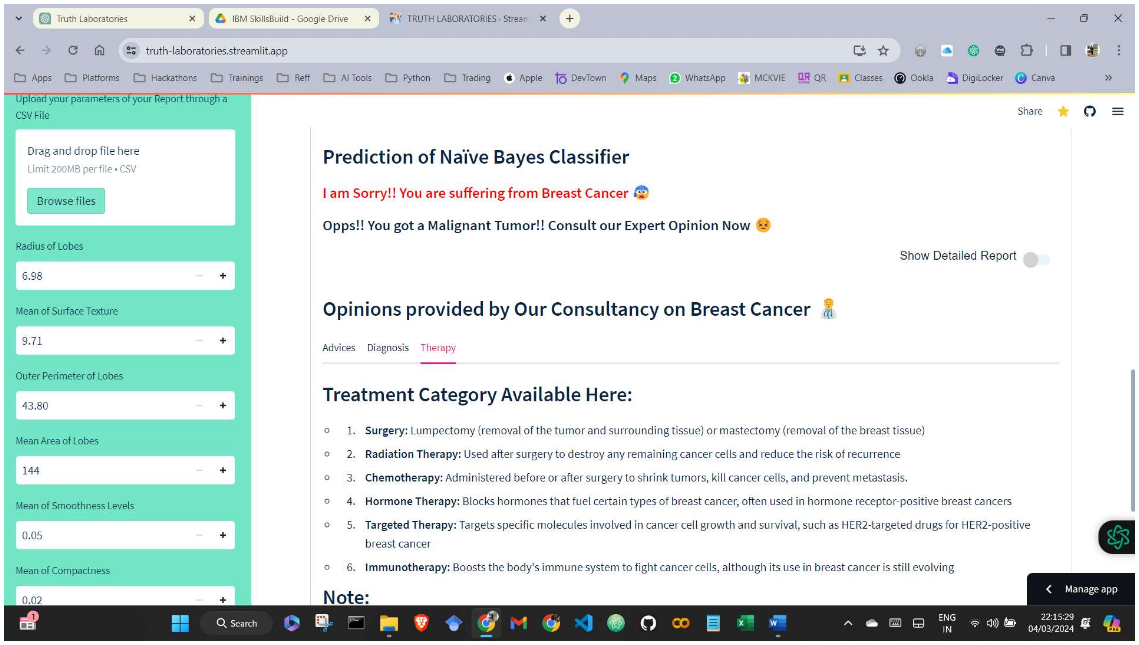
Waiting for the Report to be Uploaded... Currently displaying the Parameters given manually

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry
0	6.98	9.71	43.8	144	0.05	0.02	0	0	

👈 Provide your input data in the sidebar

Prediction Results

Comparison Study


Truth Laboratories
IBM SkillsBuild - Google Drive
TRUTH LABORATORIES - Stream
truth-laboratories.streamlit.app
Share
Manage app
22:15:29
04/03/2024

Upload your parameters of your Report through a CSV File

Drag and drop file here
Limit 200MB per file • CSV

Browse files

Radius of Lobes: 6.98

Mean of Surface Texture: 9.71

Outer Perimeter of Lobes: 43.80

Mean Area of Lobes: 144

Mean of Smoothness Levels: 0.05

Mean of Compactness: 0.02

Prediction of Naïve Bayes Classifier

I am Sorry!! You are suffering from Breast Cancer 😱

Opps!! You got a Malignant Tumor!! Consult our Expert Opinion Now 😣

Show Detailed Report

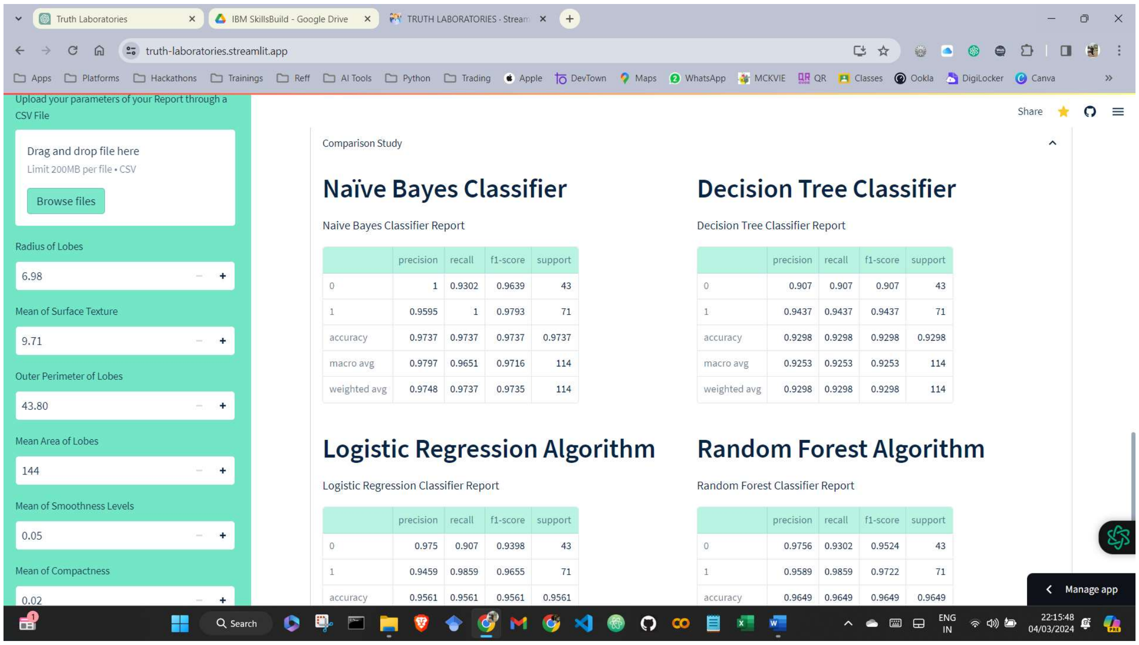
Opinions provided by Our Consultancy on Breast Cancer 🧑‍⚕️

Advices Diagnosis Therapy

Treatment Category Available Here:

1. **Surgery:** Lumpectomy (removal of the tumor and surrounding tissue) or mastectomy (removal of the breast tissue)
2. **Radiation Therapy:** Used after surgery to destroy any remaining cancer cells and reduce the risk of recurrence
3. **Chemotherapy:** Administered before or after surgery to shrink tumors, kill cancer cells, and prevent metastasis.
4. **Hormone Therapy:** Blocks hormones that fuel certain types of breast cancer, often used in hormone receptor-positive breast cancers
5. **Targeted Therapy:** Targets specific molecules involved in cancer cell growth and survival, such as HER2-targeted drugs for HER2-positive breast cancer
6. **Immunotherapy:** Boosts the body's immune system to fight cancer cells, although its use in breast cancer is still evolving

Note:


Truth Laboratories
IBM SkillsBuild - Google Drive
TRUTH LABORATORIES - Stream
truth-laboratories.streamlit.app
Share
Manage app
22:15:48
04/03/2024

Upload your parameters of your Report through a CSV File

Drag and drop file here
Limit 200MB per file • CSV

Browse files

Radius of Lobes: 6.98

Mean of Surface Texture: 9.71

Outer Perimeter of Lobes: 43.80

Mean Area of Lobes: 144

Mean of Smoothness Levels: 0.05

Mean of Compactness: 0.02

Comparison Study

Naïve Bayes Classifier

Naive Bayes Classifier Report

	precision	recall	f1-score	support
0	1	0.9302	0.9639	43
1	0.9595	1	0.9793	71
accuracy	0.9737	0.9737	0.9737	0.9737
macro avg	0.9797	0.9651	0.9716	114
weighted avg	0.9748	0.9737	0.9735	114

Decision Tree Classifier

Decision Tree Classifier Report

	precision	recall	f1-score	support
0	0.907	0.907	0.907	43
1	0.9437	0.9437	0.9437	71
accuracy	0.9298	0.9298	0.9298	0.9298
macro avg	0.9253	0.9253	0.9253	114
weighted avg	0.9298	0.9298	0.9298	114

Logistic Regression Algorithm

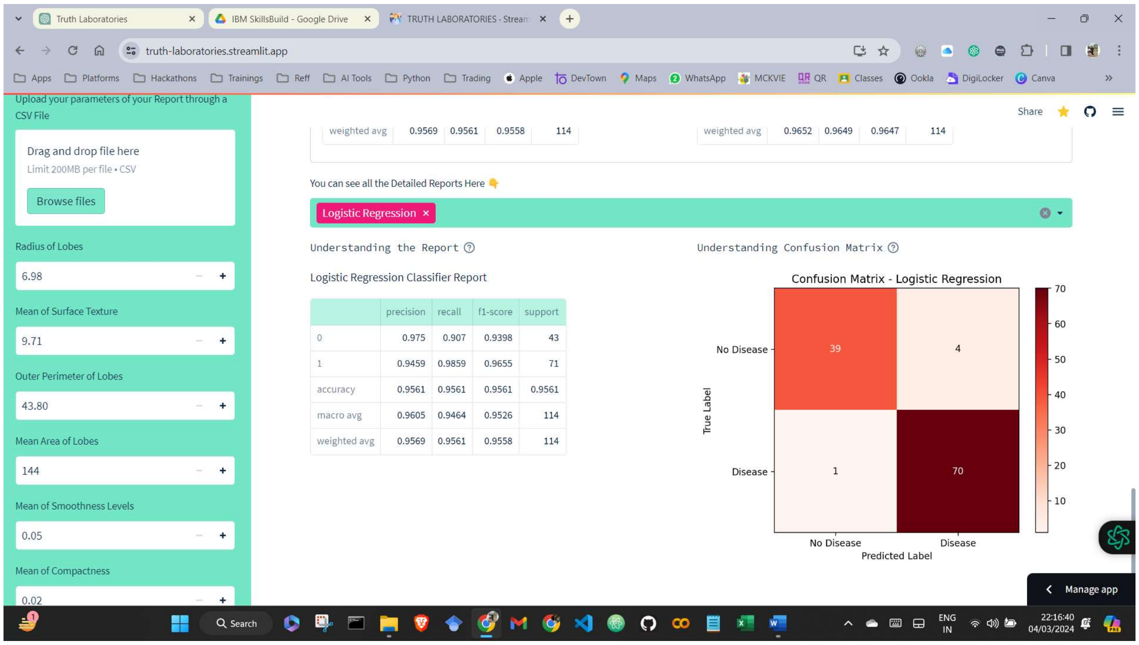
Logistic Regression Classifier Report

	precision	recall	f1-score	support
0	0.975	0.907	0.9398	43
1	0.9459	0.9859	0.9655	71
accuracy	0.9561	0.9561	0.9561	0.9561

Random Forest Algorithm

Random Forest Classifier Report

	precision	recall	f1-score	support
0	0.9756	0.9302	0.9524	43
1	0.9589	0.9859	0.9722	71
accuracy	0.9649	0.9649	0.9649	0.9649



Upload your parameters of your Report through a CSV File

Radius of Lobes: 6.98

Mean of Surface Texture: 9.71

Outer Perimeter of Lobes: 43.80

Mean Area of Lobes: 144

Mean of Smoothness Levels: 0.05

Mean of Compactness: 0.02

weighted avg	0.9569	0.9561	0.9558	114

weighted avg	0.9652	0.9649	0.9647	114

You can see all the Detailed Reports Here 👇

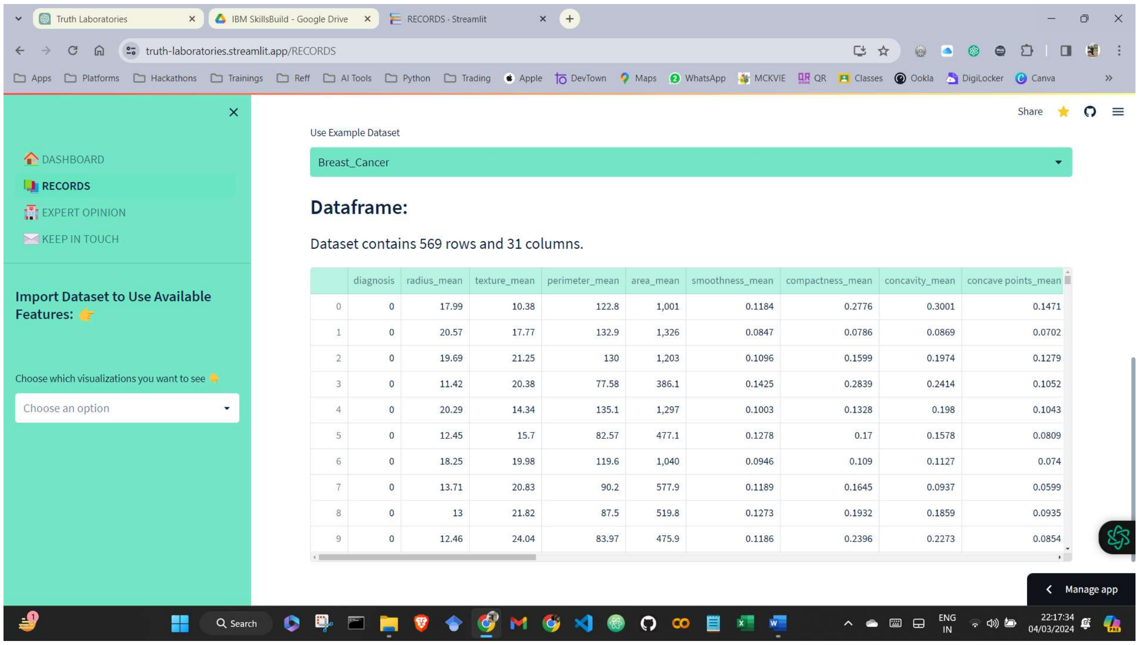
Logistic Regression ×

Understanding the Report ⓘ

Logistic Regression Classifier Report

	precision	recall	f1-score	support
0	0.975	0.907	0.9398	43
1	0.9459	0.9859	0.9655	71
accuracy	0.9561	0.9561	0.9561	0.9561
macro avg	0.9605	0.9464	0.9526	114
weighted avg	0.9569	0.9561	0.9558	114

Understanding Confusion Matrix ⓘ



Import Dataset to Use Available Features: 👉

Choose which visualizations you want to see 👇

Use Example Dataset: Breast_Cancer

Dataframe:

Dataset contains 569 rows and 31 columns.

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean
0	0	17.99	10.38	122.8	1,001	0.1184	0.2776	0.3001	0.1471
1	0	20.57	17.77	132.9	1,326	0.0847	0.0786	0.0869	0.0702
2	0	19.69	21.25	130	1,203	0.1096	0.1599	0.1974	0.1279
3	0	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052
4	0	20.29	14.34	135.1	1,297	0.1003	0.1328	0.198	0.1043
5	0	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.0809
6	0	18.25	19.98	119.6	1,040	0.0946	0.109	0.1127	0.074
7	0	13.71	20.83	90.2	577.9	0.1189	0.1645	0.0937	0.0599
8	0	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.0935
9	0	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.0854

5. Outcome and Discussions

The output of the Project Patient Sickness Disease appears to be a system designed to predict various diseases based on input symptoms. It likely utilizes machine learning algorithms to analyze patient data and generate predictions. The system may have a user interface where users can input symptoms, and the system then processes this data to provide predictions about the potential diseases a patient may have.

The dataset used in this project likely contains information about symptoms and corresponding diseases, enabling the machine learning model to learn patterns and associations between symptoms and diseases. Through data preprocessing, feature extraction, and model training, the system can generate accurate predictions about the likelihood of a patient having certain diseases based on their symptoms.

The output of the system could be presented as a list of predicted diseases along with their respective probabilities or confidence scores. This information can assist healthcare professionals in diagnosing patients and providing appropriate treatment plans. Additionally, the system may include features for continuous learning and improvement, allowing it to adapt to new data and enhance its predictive capabilities over time.

Overall, the Project Patient Sickness Disease output aims to provide a valuable tool for healthcare professionals to aid in disease diagnosis and management, ultimately improving patient outcomes and quality of care