Name :- Mandar Kishor More

Course :- DLBDSME01 Model Engineering Module

🙏 Acknowledgments

UCI ML Repository - Breast Cancer Wisconsin Dataset

IU Internationale Hochschule – DLBDSME01 Model Engineering Module

---

🧠 Breast Cancer Prediction – Interpretable Machine Learning

This project builds an **interpretable machine learning model** to predict whether a breast tumor is **benign or malignant**, using the Wisconsin Breast Cancer Dataset and IU csv file.

It follows the **CRISP-DM methodology** and includes a **Streamlit web app** for medical professionals to interact with the model and understand its predictions.

The dataset consists of 569 records representing samples from breast tissue biopsies, each identified by a unique ID. The primary goal is to classify whether a tumor is malignant (cancerous) or benign (non-cancerous) based on a set of diagnostic features derived from medical imaging.

Each record includes a diagnosis column, which contains either an M (malignant) or B (benign) label. There are 357 benign and 212 malignant cases, indicating a moderately imbalanced dataset favoring benign outcomes.

The dataset contains 30 numeric features grouped into three sets:

Mean Values (e.g., radius_mean, texture_mean) – representing the average measurement for each characteristic.

Standard Error values (with _se suffix) – indicating the variation or uncertainty of those measurements.

Worst-case values (with _worst suffix) – the most extreme observed values across the tumor sample.

These features are based on ten core physical properties of cell nuclei, including:

Radius (distance from center to edge)

Texture (variation in grey-level intensity)

Perimeter

Area

Smoothness (local variations in radius)

Compactness, Concavity, Concave Points, Symmetry, and Fractal Dimension.

Statistical analysis shows that benign tumors generally have lower values for features like radius, area, and concavity compared to malignant ones. For instance, the radius_mean ranges from approximately 7 to 28, and the average is around 14, with higher values more often associated with malignant cases.

Lastly, the dataset includes a column labeled Unnamed: 32, which contains no values and can be safely removed during preprocessing.

In summary, the data is clean (no missing values), well-structured, and rich in diagnostic information, making it highly suitable for building predictive models. The balance of interpretability and numerical depth makes it a strong candidate for applying explainable machine learning techniques in a medical context.

📌 Objective

- Develop a classification model with **F1 score > 0.95**
- Focus on **interpretability** using SHAP
- Provide a prototype **GUI** for prediction and explanation

Do not run template.py because it will create new files and folder again and all code will be lost use template.py if you want to create files in another project

---

libraries (you can see in requirement.txt)

pandas

numpy

scikit-learn

joblib

streamlit

shap

matplotlib

seaborn

pyyaml

Important Notes

src - **It is consist of all main files to run**

run.py - **It runs evaluate_model.py , explain_model.py , train_model.py it will run all three functions**

app - **It consists of two apps which are based on Logistic regression and RandomForestClassifier to run both apps you need to type streamlit run appname(logistic_reg.py in command prompt or terminal**

notebooks - **This notebook consist of eda.ipynb and model.ipynb**

data - **csv file for training model**

Logger - **I have created logger file to keep track of code which is in logs folder**

How to run

You need to install necessary **libraries** which i have written in requirement.txt to run the project or you can create a new **enviroment** To see result in a form of browser you need to go in **data/reports/breast_cancer_profiling_report.html** and from there you need to run live server

github project link :- https://github.com/mandylegend/mandylegend-IU-Breast-Cancer-Model_Project/tree/main - you can fetch project from here and updated files

---

📦 Project Structure

```
breast-cancer-prediction/
│
├── README.md                    # This file
├── LICENSE                      # Project license (optional)
├── .gitignore                   # Ignored files (e.g., data/, .pkl)
│
├── data/
│   ├── raw/                     # Original dataset (CSV)
│   └── reports/               # HTML file
│
├── notebooks/
│   ├── 01_eda.ipynb             # Exploratory Data Analysis
│
│
├── src/
│   ├── preprocessing/
│   │   └── clean_data.py        # Cleaning functions
│   ├── models/
│   │   ├── train_model.py       # Train and save model
│   │   ├── evaluate_model.py    # Evaluate metrics and F1
│   │   └── explain_model.py     # SHAP-based explanations
│   └── utils/
│       └── helpers.py
│   └── Logger
│
│
├── app/
│   └── Logistic_app.py         # GUI for predictions
    └── Ran_for_clf_app.py      # GUI for predictions
```

```
├── config/
│   └── config.yaml              # Model parameters and paths
│
├── requirements.txt             # Python dependencies
└── run.py                       # Main runner (train + evaluate)
```