Breast Cancer Outcome Prediction Using Multi-Dataset Comparative Analysis

Group D8: Triin-Elis Kuum, Greete Siemann, Hanna Samelselg

Repository: https://github.com/greetesiemann/DataScience.git

Task 2 – Business understanding

Breast cancer is one of the most widely studied medical conditions, and early detection is critical for effective treatment. The Wisconsin Breast Cancer Diagnostics dataset is a benchmark dataset containing measurements of tumor cell nuclei obtained from digitized images of breast tissue. Each sample is labeled as *benign* or *malignant*, allowing development of classification models that support diagnostic decision-making.

This project uses this dataset to train and evaluate machine-learning models that can predict tumor malignancy based on measurable cell features. The goal is not to produce a clinical tool, but to understand how data mining techniques perform on biomedical classification problems and to practice the CRISP-DM framework.

The primary business goal is to build a machine-learning model that can accurately classify breast tumors as benign or malignant. Such a model could hypothetically support medical professionals in identifying high-risk cases more rapidly. From an academic perspective, the project aims to demonstrate understanding of data preprocessing, feature exploration, model development and evaluation.

Business Success Criteria

For our project, business success criteria are for starters, a functioning and validated machine-learning model for tumor diagnosis. In addition, achieving strong predictive performance, whereas our internal target is accuracy > 90% and AUC > 0.95. We also want to produce clear visualisations, metrics and explanations useful for decision-making.

Inventory of Resources

Our project requires many resources such as a **dataset** (Wisconsin Diagnostic Breast Cancer dataset with 569 samples, 30 features + target), **tools** (Python, Pandas, NumPy, Scikit-Learn, Matplotlib, Seaborn, Jupyter Notebook), **hardware** (personal laptops capable of running ML algorithms) and **human resources** (3 team members contributing equally).

Requirements, Assumptions and Constraints

Firstly, the data must be cleaned and encoded before modeling and no patient-identifying information should be included, so the dataset would be safe to use. The project must stay within the scope of the course and CRISP-DM requirements and interpretation should be limited to data science perspective which means no clinical claims are to be made.

Risks and Contingencies

Risk: Overfitting due to small dataset → *Mitigation:* using cross-validation and simpler baseline models

Risk: Class imbalance could affect sensitivity/recall → *Mitigation:* using stratified splitting and appropriate metrics

Risk: Feature correlation could distort model interpretation → *Mitigation:* reviewing correlation matrix, optionally trying PCA

Terminology

Benign: non-cancerous tumor

Malignant: cancerous tumor

Feature: numerical measurement describing cell nuclei properties

AUC: area under ROC curve; measures model's ability to distinguish classes

Costs and Benefits

Costs: project time for data cleaning, analysis, modeling, tuning and reporting
Benefits: deeper understanding of supervised classification, performance evaluation and biomedical ML workflows

Data-Mining Goals

First step in data-mining is cleaning and preparing the dataset for modeling before exploring distributions, correlations and feature relationships. Next step is to train at least three models: Logistic Regression, RandomForest and SVM and perform hyperparameter tuning. From there we can evaluate performance using accuracy, precision, recall, F1 and ROC-AUC. Last goal is to identify most influential features.

Data-Mining Success Criteria

Data-mining is successful if models run without issues and produce interpretable results. In addition, if ROC-AUC > 0.95, which is desirable for a high-quality diagnostic model. Another success criteria is explaining results clearly and supporting the explanations with visualizations.

Task 3 – Data understanding

Data Requirements

For this project, we require a labeled dataset suitable for binary classification. Essential requirements include numerical input features that allow the use of a wide range of machine-learning algorithms, a clearly defined binary target variable, sufficient number of samples to train and validate models reliably, minimal or no missing values to preserve data quality and no features that directly leak information about the target (e.g., post-diagnosis variables). The chosen dataset – *the Wisconsin Diagnostic Breast Cancer (WDBC) dataset* – satisfies all of these conditions. It is widely used for benchmarking classification algorithms and is known for its clean structure.

Data Availability

The dataset contains 569 samples and 30 numerical features, derived from digital images of breast tissue cell nuclei. These features describe characteristics such as **radius** (mean distance from the center), **texture** (standard deviation of gray-scale values), **perimeter**, **area**, and **smoothness**, **compactness**, **concavity**, **symmetry** and **fractal dimension**, representing structural irregularity. For each measurement, three values are included: **mean**, **standard error** (SE), and **worst** (largest value). This results in a rich and detailed representation of tumor morphology. The target variable **diagnosis** is initially encoded as M (malignant) and B (benign), which we later convert to 1 and 0 respectively.

Selection Criteria

We include all 30 numerical features in the analysis. Each feature represents a measurable and clinically relevant property of the tumor, and none of them contain direct diagnostic information. We remove two columns **id** (pure identifier) and **Unnamed: 32** (irrelevant column). These fields do not contribute predictive value and may interfere with model training.

Describing Data

After cleaning and encoding the target variable, 357 samples (62.7%) are benign, 212 samples (37.3%) are malignant and there are no missing values in any feature.

Statistical summary: The dataset includes features with different scales. For example, radius_mean ≈ 14, texture_mean ≈ 19, area_mean ≈ 650. Malignant tumors consistently show higher values of radius, perimeter, area and concavity-related features, which confirms clinical expectations and suggests that machine-learning models should be able to separate the two classes effectively.

Correlation: A correlation heatmap reveals strong internal structure in the dataset: **radius_mean**, **perimeter_mean**, and **area_mean** have correlations above 0.90, **compactness**, **concavity**, and **concave_points** form another highly correlated group. This multicollinearity indicates that dimensionality reduction methods (such as PCA) could be beneficial for some models, particularly logistic regression and SVM. Tree-based models, such as Random Forests, handle correlated features better due to their internal structure.

Exploring Data

We explore the dataset visually to better understand feature distributions and class separability:

- *histograms* – most numerical features show skewed distributions, malignant samples often occupy the upper tail of distributions for "worst" features, highlighting potentially informative separation
- *boxplots by diagnosis* – boxplots show clear differences between benign and malignant tumors in features such as **area_worst**, **radius_worst**, and **concavity_worst**, reinforcing the idea that these features have strong diagnostic value
- *correlation heatmap* – the heatmap reveals strong clustering of related features and confirms that multiple measurements capture similar aspects of cell morphology
- *pairplots* – pairplots of selected features show visually distinguishable clusters between benign and malignant tumors, this confirms the dataset is highly separable and suitable for classification tasks

Overall, exploratory analysis strongly suggests that the dataset contains meaningful structure that machine-learning algorithms can capture.

Verifying Data Quality

Data quality checks confirm that the dataset is well-prepared for modeling:

- *No missing values*: all 30 features contain complete measurements for every sample.
- *Consistent data types*: all input features are float64, which simplifies preprocessing.
- *No outliers removed*: although outliers exist, they likely represent real clinical variation and removing them could degrade model validity.
- *Target correctly encoded*: 'M' → 1 and 'B' → 0 ensures compatibility with ML algorithms.

Therefore the dataset is clean, complete, numerically consistent and fully ready for model training and evaluation.

Task 4 – Project plan

Our project is organized into structured work packages following the CRISP-DM methodology.

Task 1 – Data Preparation (all members)

Load the dataset, remove unnecessary columns, encode the target, clean the features and verify data quality. Estimated time: 6 hours per member.

Task 2 – Exploratory Data Analysis (all members)

Compute statistical summaries, visualize distributions, correlations and feature relationships. Produce the figures for the report. Estimated time: 5 hours per member.

Task 3 – Model Development (Triin)

Train Logistic Regression, RandomForestClassifier and SVM models. Perform hyperparameter tuning and cross-validation. Compute feature importances and ROC curves. Estimated time: 10 hours.

Task 4 – Evaluation and Interpretation (Greete)

Compare the models using accuracy, precision, recall, F1 and ROC-AUC. Create confusion matrices and analysis of strengths and weaknesses. Estimated time: 8 hours.

Task 5 – Final Report Compilation (Hanna)

Write conclusions, assemble visualizations, format the report, ensure clarity and correctness. Estimated time: 6 hours.

This plan ensures balanced contribution and clear workflow from data understanding to final evaluation.