**Breast‑Cancer Attributes Prediction**Authors: | Yehonatan Ezra | Shay Morad | Lior Zats | Avi Kfir |

We chose to do the second challenge, predicting clinical attributes of breast cancer, a meaningful and complex real-world challenge. In this short report, we’d like to walk you through our process and to explain how we cleaned the data, explored hidden patterns, and gradually built up a modeling pipeline we were proud of.

**Data Cleaning and Pre‑Processing**

Our first step was to prepare the raw data. We focused on cleaning and simplifying the features through manual checks, correlation analysis, and compact encodings.

**Initial Manual Screening:**

At the outset we attempted a column‑by‑column inspection of the 34 raw features provided in train.feats.csv. The goal was to catalogue data types, admissible values, and obvious anomalies. While this manual pass helped us familiarize ourselves with the dataset, it quickly became tedious and error‑prone - many variables contain dozens of heterogeneous string representations and many more rows made spreadsheet - style checks impractical.

**Automated Correlation‑Based Screening:**

To streamline the feature set we transformed every column into a provisional numeric form timestamps for dates, ordinal codes for categorical strings, and 0/1 for Booleans - then visualized the resulting Pearson - correlation matrix. We used this correlation analysis to remove columns that were not relevant for modelling, replacing exhaustive manual comparisons with a fast, objective filter.

In the graph below you can see our Pearson correlation Matrix:

תמונה שמכילה טקסט, צילום מסך, תרשים, קו

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.

**Embedding of High‑Cardinality Categorical:**

After finishing the correlation step, we addressed the few columns with dozens of unique values. Instead of sprawling one‑hot vectors, we learned compact 8–16‑dimension embeddings that summarize each category’s clinical signal while keeping the feature matrix dense. This was used for fields that made sense to apply this to, such as the types of surgeries a patient went and so on of which we do not want to have an order ratio between the values in such columns.

**Missing - Value Imputation:**

Numeric gaps were filled with each column’s median, and missing categorical received an explicit default/missing token. We also quantized values based on similar values such as "pos", "+", "100%pos" and such for a common value and so on.

**Predicting Metastases**

This task required predicting a **multi-label, multi-class** output: each patient could have zero to three metastasis sites. We framed the problem as a set of binary classification tasks (one-vs-rest), training a separate model per metastasis site.

We evaluated three main model families with various hyperparameters: Random Forest, Decision Tree, and K‑Nearest Neighbors. Each model type was tested with multiple configurations to explore its sensitivity and generalization capacity.

After exhaustively grid‑searching every hyper, we found that the deeper, tree‑rich Random Forest consistently delivered the highest scores on both F1‑micro (class‑frequency‑weighted) and F1‑macro (class‑balanced) metrics. As illustrated in the accompanying bar chart, its margin over the next‑best alternatives was clear, so we selected this Random Forest configuration as our final model for predicting metastasis sites.

In the graph below we can see the results:

תמונה שמכילה טקסט, צילום מסך, גופן, קו

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.

.