### CRISP-ML(Q) Project Report

#### 1. \*\*Business Understanding\*\*

- \*\*Business Problem\*\*: Cancer diagnosis is often challenging, with physicians sometimes encountering difficulties in distinguishing between benign (non-cancerous) and malignant (cancerous) tumors. The need for an AI-assisted diagnostic tool has become evident as a supplementary method to improve diagnostic accuracy and support radiologists, pathologists, and oncologists in their assessments.

- \*\*Business Objective\*\*: Develop a machine learning solution capable of confidently predicting the likelihood of cancer, thus enhancing the accuracy of diagnoses while also reducing treatment costs and improving patient convenience.

- \*\*Success Criteria\*\*:

- \*\*Business Success Criteria\*\*: Correctly diagnose cancer in at least 96% of cases.

- \*\*Machine Learning Success Criteria\*\*: Achieve a model accuracy of at least 98%.

- \*\*Economic Success Criteria\*\*: Reduce treatment costs, fostering patient trust and increasing hospital revenue by at least 12%.

---

#### 2. \*\*Data Understanding\*\*

- \*\*Data Collection\*\*: The dataset consists of medical records from 569 patients, containing 30 features describing characteristics of cell nuclei and one label indicating whether the tumor is benign or malignant.

- \*\*Features\*\*:

- Ten real-valued features represent different aspects of each cell nucleus, including metrics such as \*\*radius (mean distance from center to perimeter)\*\*, \*\*texture (standard deviation of gray-scale values)\*\*, \*\*perimeter\*\*, \*\*area\*\*, \*\*smoothness\*\*, \*\*compactness\*\*, \*\*concavity\*\*, \*\*concave points\*\*, \*\*symmetry\*\*, and \*\*fractal dimension\*\*.

- These features are divided into measurements on both mean values (e.g., `radius\_mean`) and worst-case values (e.g., `radius\_worst`).

---

#### 3. \*\*Data Preparation and EDA (Exploratory Data Analysis)\*\*

\*\*Descriptive Analysis Summary\*\*:

- \*\*Central Tendency\*\*:

- The average values for primary metrics such as `radius\_mean` (14.13), `texture\_mean` (19.29), and `perimeter\_mean` (91.97) offer a general sense of tumor sizes in the dataset.

- The worst-case metrics, such as `radius\_worst` (16.27), `texture\_worst` (25.68), and `perimeter\_worst` (107.26), reflect the maximum severity observed among tumors.

- \*\*Dispersion\*\*:

- Standard deviation shows substantial variation, with `area\_mean` having a high standard deviation (351.91), indicating considerable variance in tumor sizes.

- This variability is mirrored in features like `compactness\_mean` (0.0528) and `concavity\_mean` (0.0798), indicating differences in cell nucleus shape and structure between benign and malignant cases.

- \*\*Statistical Measures (Skewness and Kurtosis)\*\*:

- \*\*Skewness\*\*:

- Positively skewed metrics such as `radius\_mean` (0.94) and `area\_mean` (1.64) indicate a right-skewed distribution, suggesting some cases with significantly larger radius and area values.

- High skewness in features like `radius\_se` (3.09) and `area\_se` (5.45) reflects the presence of outliers, possibly indicative of aggressive cancer types.

- \*\*Kurtosis\*\*:

- Kurtosis in features like `area\_se` (49.21) and `dimension\_se` (26.28) highlights the presence of extreme values, indicating sharp peaks. This suggests that certain extreme cell features may be critical for distinguishing between benign and malignant tumors.

---

#### 4. \*\*Modeling\*\*

- \*\*Model Selection\*\*: Given the high-dimensional nature of the data and the need for reliable accuracy, algorithms like \*\*Random Forest Classifier\*\*, \*\*Support Vector Machine (SVM)\*\*, or \*\*Deep Neural Networks\*\* will be explored for this classification task.

- \*\*Target Metric\*\*: The primary metric is \*\*accuracy\*\*, aiming for at least 98% to satisfy machine learning success criteria. Secondary metrics, such as precision, recall, and F1-score, will be monitored to ensure the model’s reliability across different tumor types.

---

#### 5. \*\*Evaluation\*\*

- \*\*Model Performance\*\*:

- Once the model achieves a satisfactory accuracy level, validation on separate test data will confirm the results’ generalizability.

- The focus will remain on minimizing false negatives to avoid misdiagnosis, as these errors are particularly costly in cancer detection.

- \*\*Success Review\*\*:

- Business success is determined by achieving correct diagnoses in at least 96% of cases.

- The machine learning success criterion of achieving 98% accuracy will signal the model’s readiness for real-world use.

- Economic success will be indirectly assessed through patient trust and satisfaction metrics post-deployment, along with revenue growth analysis over time.

---

#### 6. \*\*Deployment and Monitoring\*\*

- \*\*Deployment\*\*:

- Once validated, the model will be integrated into the hospital’s diagnostic system as an assistive tool for physicians.

- Regular performance monitoring and retraining will be conducted to adapt to any new patterns or data drifts in patient populations.

- \*\*Feedback Loop\*\*:

- Continuous feedback from medical professionals will refine model performance and ensure alignment with clinical needs.

- Economic metrics such as cost reduction and patient satisfaction will be reviewed quarterly to assess the solution's long-term impact.

---

#### \*\*Conclusion\*\*

This AI-based cancer diagnosis tool is expected to provide physicians with a robust second opinion, reducing diagnostic errors and improving patient outcomes. The model’s ability to accurately classify tumor types will potentially lower treatment costs, enhance patient satisfaction, and increase the hospital’s revenue by at least 12%, marking a successful alignment of business, clinical, and economic objectives.