Thyroid Cancer Risk Predictor

Ryaan Zahidani, rhz9, Section 08   Tanav Sureddy, ts1218, Section 08

1. Project Definition

The purpose of this project is to implement an end-to-end data management and machine learning pipeline for predicting thyroid cancer risk using patient health, demographic, and clinical data. Our main research question that we seek to answer is: can machine learning enhance the accuracy and consistency of thyroid-related risk assessment by leveraging demographic, clinical, and health-history factors? The project integrates database engineering, data cleaning, feature preprocessing, supervised learning, and model output tracking into a single reproducible system. Our goal is to build a structured and consistent approach to risk assessment that can supplement traditional physician-based screening methods, which are often subjective, inconsistent across providers, and dependent on expensive imaging procedures. Using a dataset of more than 200,000 patient records, we aim to classify individuals into low, medium, or high thyroid cancer risk categories by leveraging both supervised and unsupervised learning techniques. The project also incorporates a full data-management pipeline through a custom SQLite database, allowing us to store, clean, process, and evaluate patient information in a systematic way. Ultimately, this project seeks to demonstrate how data-driven tools can support early detection efforts and help standardize clinical decision-making.

2. Introduction

Thyroid cancer rates have risen noticeably in recent years, with more than 40,000 new cases diagnosed each year in the United States. Catching the disease early is essential for improving treatment outcomes, but risk evaluations can vary widely from one clinician to another due to incomplete electronic health records, scattered data systems, and the unavoidable subjectivity involved in medical decision making. As more patient-level metabolic, demographic, and clinical data become available, machine learning offers a way to spot risk patterns that may be hard to recognize through manual review. When risk assessments are inconsistent, important opportunities for prevention can be missed, and high-risk patients may not receive timely diagnoses. This project aims to address these issues by exploring how machine learning models such as K-Means clustering and Logistic Regression can turn subjective thyroid cancer risk assessments into a more consistent and data-driven process. Our work centers on the Kaggle Thyroid Cancer Risk dataset, which includes demographic details, medical history, and information on radiation exposure for a large group of patients. We built a clean, query-optimized SQLite database, processed and encoded all variables, and evaluated several learning algorithms. Our goal is to measure how much machine learning can improve both predictive accuracy and the clarity of clinical risk scoring, and the novelty of this project comes from combining database engineering with explainable machine learning to create a full pipeline that reflects how medical data is handled in real-world settings.

3. Methodology

Our methodology brings together data management, preprocessing, machine learning, and model evaluation into one streamlined pipeline. We started by downloading the Thyroid Cancer Risk dataset from Kaggle and loading the raw CSV file into a custom SQLite database. The database uses three main tables: PATIENT_PROFILES, which holds all patient attributes along with the target risk label; DATA_QUALITY_LOGS, which records missing values, removed duplicates, and an overall data quality score; and MODEL_PREDICTIONS, which stores predicted risk levels, confidence scores, and timestamps for each model run. We also added indexing to key fields such as patient_id, risk_category, and country so the system could query the data efficiently and work smoothly with the machine learning components.

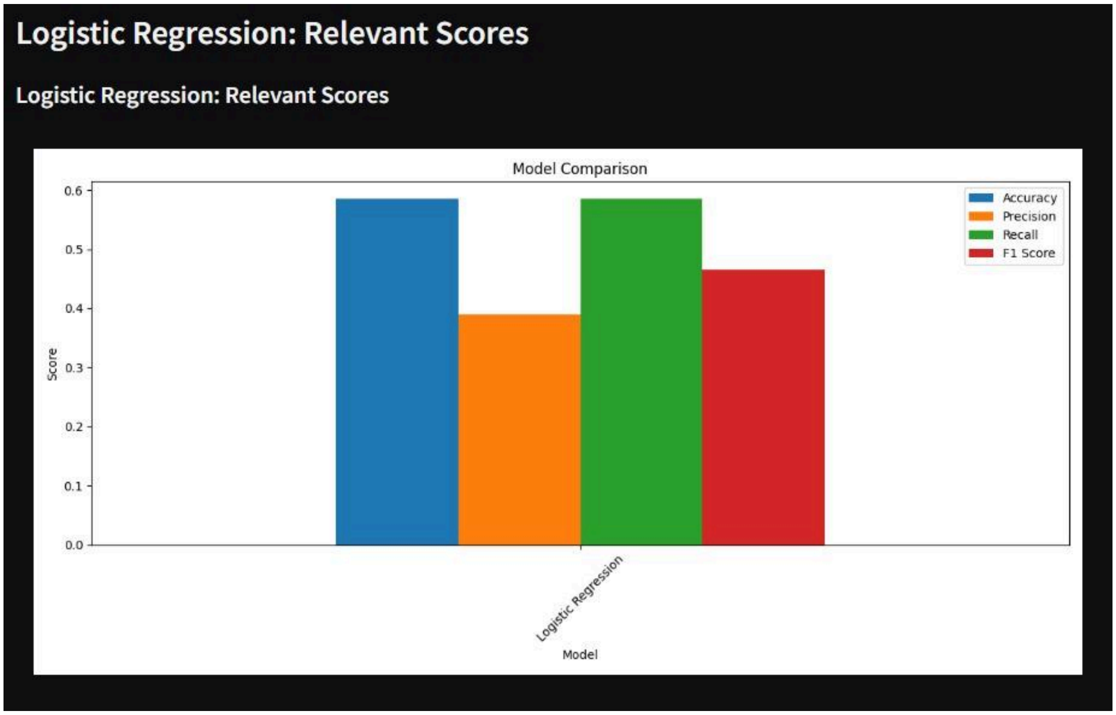
During preprocessing, we handled missing data using median imputation for numeric features and mode imputation for categorical ones, followed by removing duplicate entries to ensure each patient appeared only once. Although the dataset reported that no values were missing, we verified this ourselves to maintain good data-cleaning practice. After confirming all data types were correct, we used OneHotEncoding for categorical variables and LabelEncoding for the target label. We then scaled all numerical features with StandardScaler to give each feature an equal influence during model training for both K-Means and Logistic Regression.

We implemented two machine learning models for our analysis. K-Means Clustering was used as an unsupervised approach to uncover natural groupings based on similarities in patient features, independent of the labeled risk categories. This helps us see whether the dataset contains underlying patterns that align with clinical expectations. Logistic Regression served as our main supervised model, classifying patients into Low, Medium, or High risk groups. We chose Logistic Regression because it is highly interpretable and allows us to see which demographic or clinical factors contribute most to the predicted risk.

To evaluate the full pipeline, we calculated Silhouette Scores and Davies-Bouldin Index values for the K-Means model, and we measured accuracy, precision, recall, F1-scores, confusion matrices, and feature importance for Logistic Regression. We tested the entire system, from data ingestion through prediction, against criteria that required strong model performance, reliable database operations, and prediction patterns that made sense in a clinical context.

4. Results

Logistic Regression:

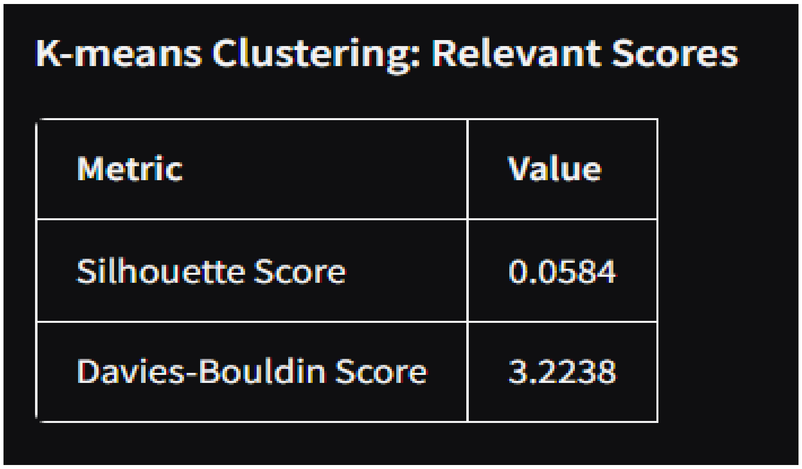


Logistic Regression: Relevant Scores

Logistic Regression: Relevant Scores

Our Logistic Regression model reached an overall accuracy of about 59 percent, which is roughly seven points higher than the majority-class baseline. This indicates that the model is learning real patterns rather than simply predicting the most common class. The macro-averaged precision across all three risk categories was around 0.39, which suggests a relatively high number of false positives. In practice, this means that some patients were labeled as higher risk than they actually were. For medical screening, this trade-off is generally acceptable because flagging too many patients is safer than missing those who truly need attention, though reducing unnecessary high-risk classifications is still an area for future improvement. The model's macro-averaged recall was approximately 0.58, which is strong compared with its precision. High recall is especially important in healthcare because failing to identify patients who are genuinely high or medium risk carries significant consequences, while moderate false positive rates can be managed through follow-up clinical assessment. The F1 score, which balances precision and recall, was about 0.47. This moderate performance reflects the difficulty of predicting three imbalanced classes. It also suggests that while the model is capturing useful risk patterns, more advanced methods such as class-weighted logistic regression, resampling strategies, or ensemble models could improve results. We did not apply class weighting or resampling in this version of the model, so incorporating these techniques is a clear opportunity for enhancement. Overall, the model shows that machine learning can moderately predict thyroid cancer risk using demographic and clinical features. It outperforms the baseline, identifies meaningful patterns, and supports our hypothesis that ML can aid in risk stratification. While not a replacement for medical judgment, the model works well as an early-warning tool to help guide clinical decision making.

K-Means Clustering:

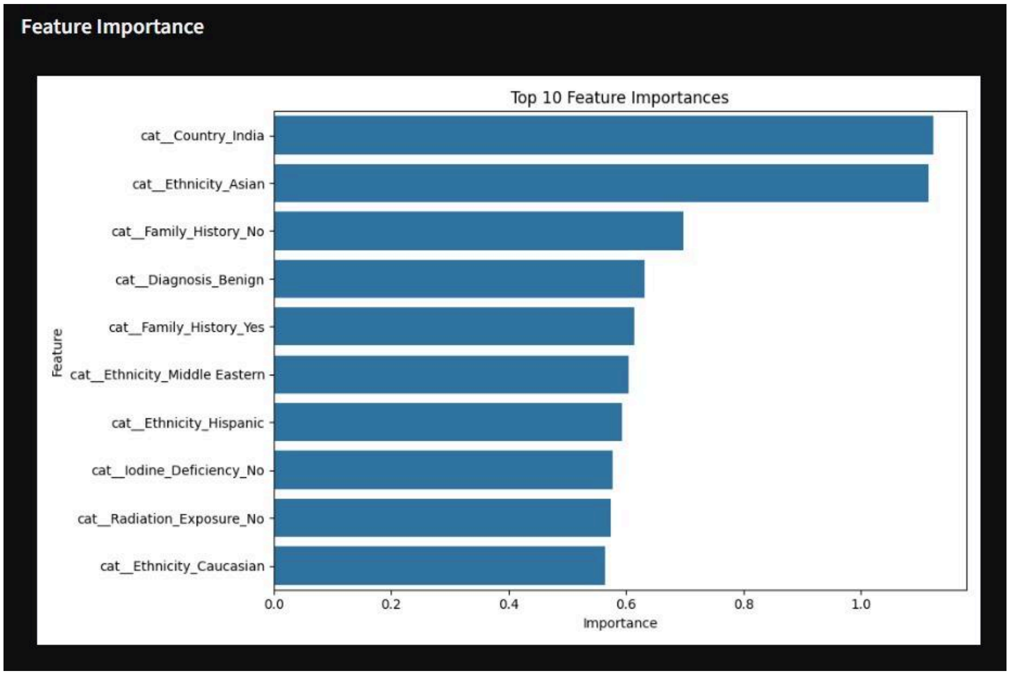


K-means Clustering: Relevant Scores

Metric	Value
Silhouette Score	0.0584
Davies-Bouldin Score	3.2238

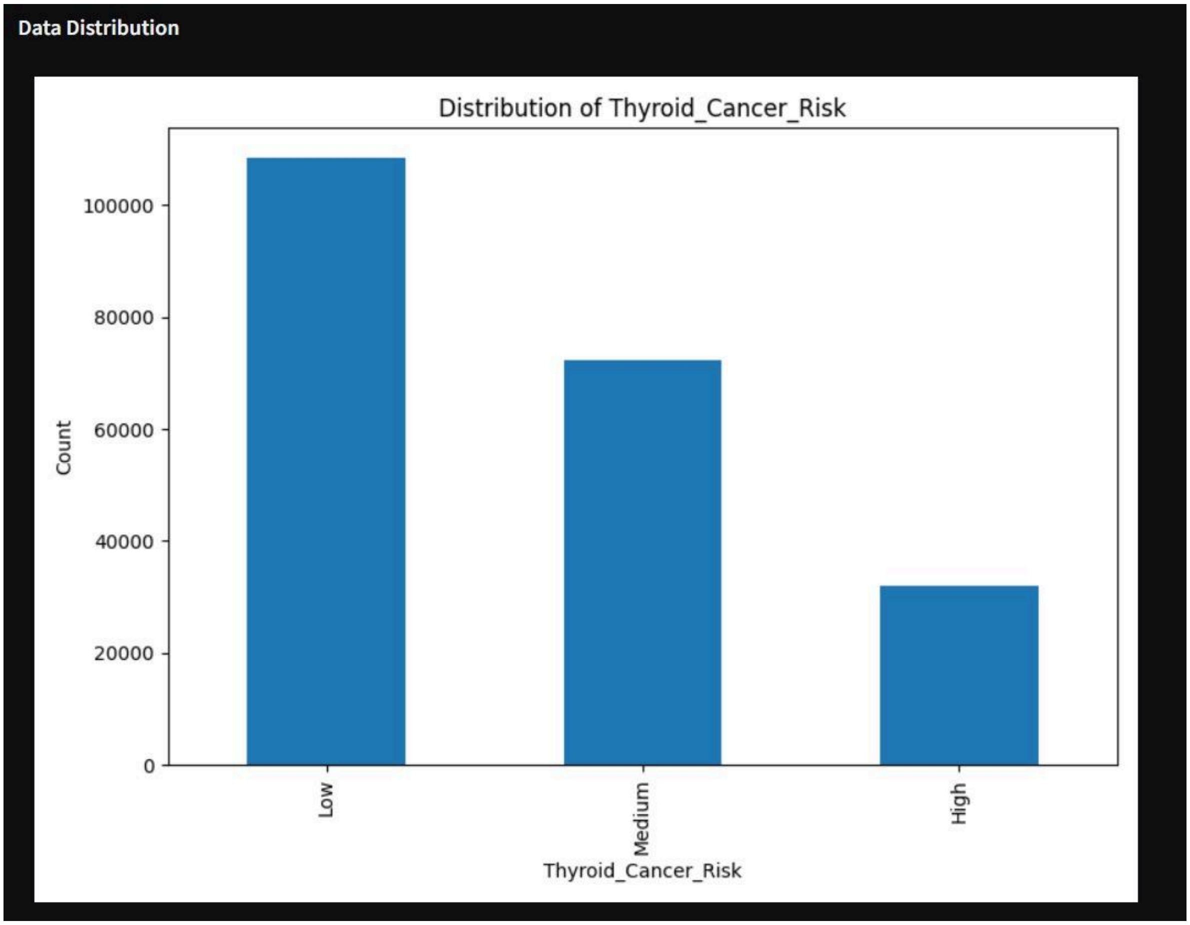
We used K-Means clustering as an unsupervised method to identify underlying patient groups without relying on the labeled risk categories. The Silhouette Score, which ranges from −1 to +1 and reflects how well clusters are separated, was approximately 0.0584, which is very close to 0 and indicates that clusters are only weakly separated; patients assigned to a given cluster are not much more similar to each other than to patients in neighboring clusters. The Davies–Bouldin Score was approximately 3.2238, and because lower values indicate better separated and more compact clusters, this relatively high value further confirms substantial overlap between clusters. Together, these metrics suggest that although there is some underlying structure in the dataset, the partitions produced by K-Means are noisy and not suitable for direct clinical risk assignment. Instead, K-Means is more useful for broad exploratory analysis, helping to reveal general population patterns rather than making patient-level decisions. Overall, K-Means detected weak but observable natural groupings that support the idea that risk-related structure exists in the dataset, yet the clustering strength is not sufficient for clinical decision-making without additional supervised modeling.

Feature Importance:



The Logistic Regression coefficients, shown as importance magnitudes in the feature-importance chart, indicate that Family History, diagnosis-related features (especially “Diagnosis_Benign”), and Ethnicity are among the strongest predictors in the model. This pattern aligns well with clinical expectations, since family history and previous diagnostic findings are well-established factors in thyroid disease risk. In contrast, several geographic features, such as Country_India, appear with unexpectedly high importance scores. These signals are likely artifacts of dataset bias or confounding variables rather than true biological risk factors. Such geographic effects may reflect differences in data collection, healthcare access, or screening practices across regions instead of meaningful clinical distinctions. This points to a key limitation of the Kaggle Thyroid Cancer Risk dataset, which is a curated and possibly synthetic dataset that may not fully represent real-world populations. Any clinical use of models trained on this data would therefore require careful validation across diverse demographic groups and a thorough assessment of potential biases. Overall, the feature-importance results highlight the need for domain expertise when interpreting model outputs and reinforce the importance of bias auditing in healthcare machine learning applications.

Data Distribution:



Our dataset contains 210,000 patient records spread across three risk categories: roughly 110,000 labeled Low Risk, 70,000 labeled Medium Risk, and 30,000 labeled High Risk. This distribution shows a clear class imbalance, with the Low Risk group holding almost four times as many samples as the High Risk group. For reference, a simple baseline model that always predicts "Low" would achieve about 52 percent accuracy, since 110,000 out of 210,000 patients fall into that category. Imbalances like this are common in clinical datasets, where low-risk outcomes naturally occur more frequently and reflect real patterns in thyroid cancer prevalence. Because accuracy alone overstates performance in such settings, we placed greater emphasis on metrics like precision, recall, and F1 score to better evaluate how well the model handles the underrepresented classes.

Overall Summary of Model Performance:

In summary, the logistic regression model shows moderate predictive performance, achieving an accuracy about seven points higher than the majority-class baseline and a recall of 0.58, which indicates that it identifies most true medium- and high-risk patients. The K-Means clustering results suggest that although some underlying patient structure is present, it is not strongly separated, making the unsupervised analysis useful for exploration but inadequate for assigning clinical risk categories on its own. Taken together, these results support the feasibility of a machine-learning–enhanced risk assessment pipeline while also highlighting several limitations, including class imbalance, the lack of imaging or advanced laboratory features, modest F1 performance, and possible geographic biases in the dataset. Future work should explore class-weighting techniques, incorporate multimodal data such as imaging and lab results, and validate the approach on diverse and representative patient populations.

5. Contributions

Ryaan - Final Report, K-means clustering, data cleaning

Tanav - Logistic Regression model, SQLlite database creation/loading, model evaluation

6. References

https://www.geeksforgeeks.org/sql/introduction-to-sqlite/

https://www.kaggle.com/datasets/bhargavchirumoamilla/thyroid-cancer-riskdataset/

https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/