Project 7 Report

Jonah Harmon & Razin Issa
Github: https://github.com/JHamoni676/cs5830_project7
Presentation:https://docs.google.com/presentation/d/1K2PBNaYPXsPqAhwaOz7t9VJe9LSDmydxzxQ7fJzuTsY/edit?usp=sharing

Introduction:

In this project, we examined two key domains, music classification and medical diagnostics, using logistic regression and support vector machines (SVM) to classify data points in both contexts accurately. For the music genre classification dataset, which includes over 1000 audio tracks, we used audio features such as tempo, beats, spectral attributes, and mel-frequency cepstral coefficients (MFCCs) to capture tonal, rhythmic, and timbral characteristics of each track. The goal was to classify each audio file by genre, a task useful in music recommendation systems where accurately identifying genres can enhance user experience and music discovery. In the medical domain, the breast cancer diagnosis dataset provided cell nucleus measurements, including radius, texture, and concavity, known to differentiate malignant from benign samples. Correct classification here impacts patient diagnosis and treatment decisions, making accuracy important.

Music Genre Classification

Dataset:

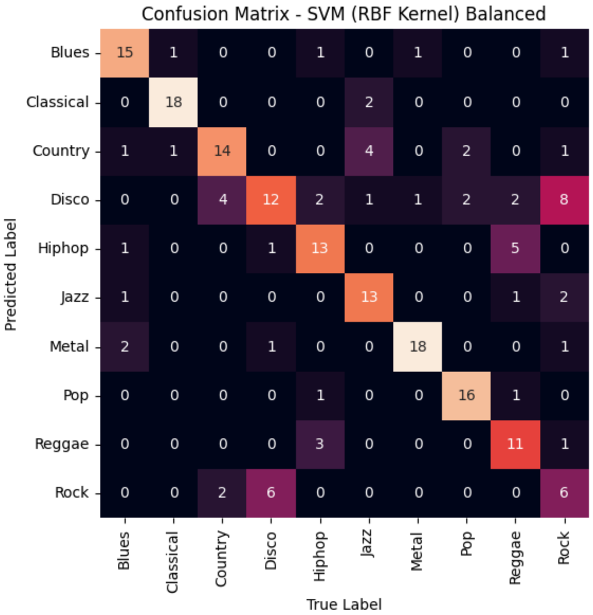
The dataset used for music classification consists of 1000 audio tracks each being 30 seconds long. Code was used to extract the following information from each 30 second audio file: tempo, beats, short time fourier transformation, root mean square error, spectral centroid, spectral bandwidth, roll-off, zero crossing rate, and the mel-frequency cepstral coefficients 1-20. The dataset also provided a label indicating the true genre of the song. Each of these explanatory variables is used to predict the genre of the file. Tempo and beats are more commonly known variables pertaining to music and it's structure, the rest of the variables provide insight into the acoustic and tonal qualities of each track. For instance, short-time Fourier transform captures changes in frequency over time, while root mean square error measures the energy or loudness of the audio. Spectral features, such as spectral centroid, bandwidth, and roll-off, reflect the distribution of frequencies, with higher values often indicating "brighter" sounds, common in genres like electronic music. Zero-crossing rate is useful for identifying more percussive sounds, helping to differentiate between smooth genres like jazz and sharper, beat-heavy genres like hip-hop. The MFCCs capture the timbral texture of each track, allowing the model to distinguish between genres with unique sound profiles. Together, these features offer a comprehensive view of each track's rhythm, tonal quality, and energy, providing a solid foundation for predicting its genre.

Analysis Techniques:

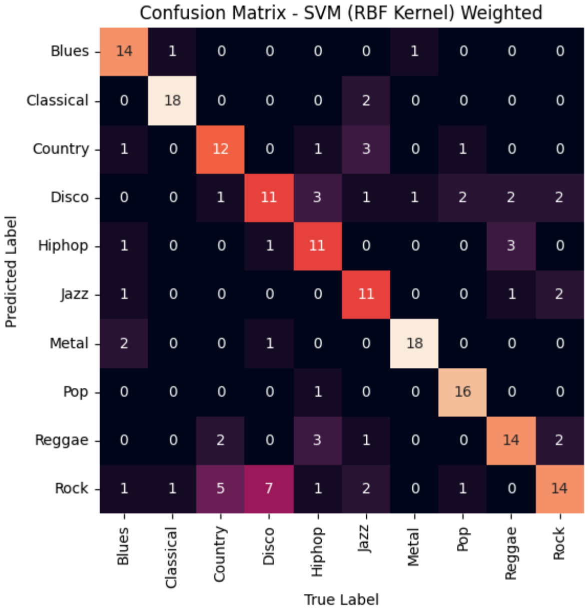
For this project, SVMs with both linear and nonlinear (RBF) kernels were used to classify music genres because they work well in high-dimensional spaces and can capture complex patterns in the data. SVMs are a good fit for this dataset because they can find the best boundaries between genres, even when some genres overlap in characteristics like rhythm or tone. The RBF kernel specifically helps capture non-linear relationships, which is useful for distinguishing genres with subtle acoustic differences. Logistic regression was also tested as a simpler, baseline model to see how well a linear approach could handle genre classification. Both SVM and logistic regression allow for efficient and accurate classification, making them suitable for this type of audio data.

Results:

The results of this analysis showed that the SVM model with an RBF kernel and custom class weights performed best for classifying music genres based on audio features. Logistic regression served as a decent baseline model, especially for genres with distinct characteristics like *Classical* and *Metal*, where it achieved high precision and recall. However, it struggled with more complex and overlapping genres like *Rock* and *Reggae*, where linear boundaries were insufficient. This limitation highlighted the need for a more flexible model that could handle subtle differences in audio features across genres. The SVM with an RBF kernel outperformed the other models, especially for genres with overlapping characteristics. By using a non-linear kernel, the RBF SVM was able to capture complex patterns within the audio features, leading to higher accuracy in challenging cases. Custom weights further enhanced the model's performance by boosting recall for genres that were frequently misclassified, like *Rock* and *Reggae*. The confusion matrix visualization confirmed that the RBF SVM was more effective at distinguishing genres with similar tonal and rhythmic characteristics, showing fewer misclassifications for these difficult cases compared to the linear models.



	Blues	Classical	Country	Disco	Hiphop	Jazz	Metal	Pop	Reggae	Rock
Blues	15	1	0	0	1	0	1	0	0	1
Classical	0	18	0	0	0	2	0	0	0	0
Country	1	1	14	0	0	4	0	2	0	1
Disco	0	0	4	12	2	1	1	2	2	8
Hiphop	1	0	0	1	13	0	0	0	5	0
Jazz	1	0	0	0	0	13	0	0	1	2
Metal	2	0	0	1	0	0	18	0	0	1
Pop	0	0	0	0	1	0	0	16	1	0
Reggae	0	0	0	0	3	0	0	0	11	1
Rock	0	0	2	6	0	0	0	0	0	6



	Blues	Classical	Country	Disco	Hiphop	Jazz	Metal	Pop	Reggae	Rock
Blues	14	1	0	0	0	0	1	0	0	0
Classical	0	18	0	0	0	2	0	0	0	0
Country	1	0	12	0	1	3	0	1	0	0
Disco	0	0	1	11	3	1	1	2	2	2
Hiphop	1	0	0	1	11	0	0	0	3	0
Jazz	1	0	0	0	0	11	0	0	1	2
Metal	2	0	0	1	0	0	18	0	0	0
Pop	0	0	0	0	1	0	0	16	0	0
Reggae	0	0	2	0	3	1	0	0	14	2
Rock	1	1	5	7	1	2	0	1	0	14

Genre	**Precision**	**Recall**	**F-score**
blues	0.8000	0.6000	0.6857
classical	0.9000	0.9000	0.9000
country	0.5882	0.5000	0.5405
disco	0.5455	0.6000	0.5714
hiphop	0.6087	0.7000	0.6512
jazz	0.6667	0.7000	0.6829
metal	0.7917	0.9500	0.8636
pop	0.7895	0.7500	0.7692
reggae	0.5882	0.5000	0.5405
rock	0.5455	0.6000	0.5714

Logistic Regression Results

Genre	Precision	Recall	F-score
blues	0.8750	0.7000	0.7778
classical	0.9000	0.9000	0.9000
country	0.6667	0.6000	0.6316
disco	0.4783	0.5500	0.5116
hiphop	0.6875	0.5500	0.6111
jazz	0.7333	0.5500	0.6286
metal	0.8571	0.9000	0.8780
pop	0.9412	0.8000	0.8649
reggae	0.6364	0.7000	0.6667
rock	0.4375	0.7000	0.5385

SVM with Weighted Classes Results

Overall, the SVM with RBF kernel and custom weights proved to be the most suitable model for this dataset, offering robust performance across a variety of genres. This model's ability to capture non-linear relationships between features makes it particularly useful for real-world applications like music recommendation systems, where accurately distinguishing between genres improves user satisfaction and music discovery.

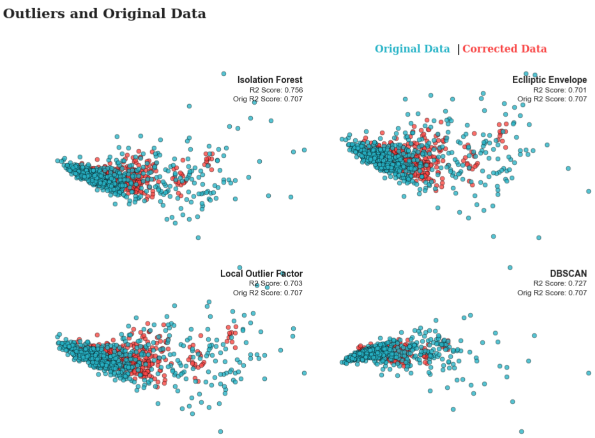
Breast Cancer Identification

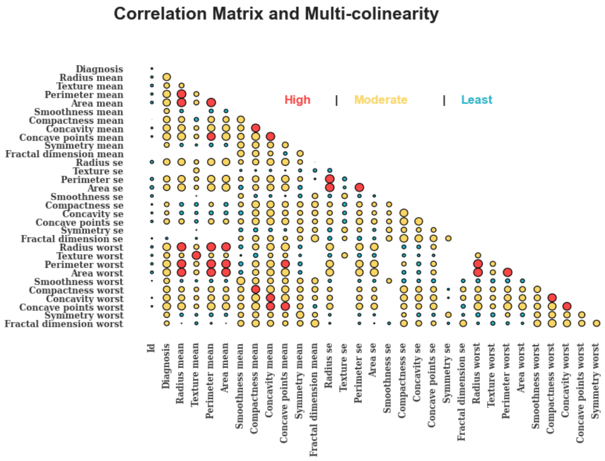
Dataset:

The dataset in this project is highly suitable for breast cancer analysis, containing essential attributes predictive of cancerous conditions. The target variable, Diagnosis, indicates whether each sample is malignant (M) or benign (B), alongside ten real-valued features that describe cell nuclei characteristics, crucial for distinguishing between the two categories. Key features include radius (mean distance from center to perimeter), texture (variation in gray-scale intensity), perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. These attributes capture critical structural and textural patterns in cell nuclei, supporting effective classification of cancer samples. To prepare the dataset for analysis, preprocessing steps included removing non-predictive columns (e.g., ID) and addressing missing values to maintain data integrity. The dataset was scaled and standardized to accommodate algorithms sensitive to feature magnitudes, such as logistic regression and support vector machines. Additionally, outlier detection methods (Isolation Forest, Local Outlier Factor) were used to correct anomalies, and skewness and kurtosis analyses were conducted to align feature distributions with model requirements. Overall, several features—such as radius, texture, perimeter, area, concavity, and concave points—showed strong predictive power for Diagnosis, correlating well with cancerous and non-cancerous outcomes. The dataset is now clean, standardized, and optimized for building reliable classification models, making it a robust foundation for predictive analysis.

Analysis Techniques:

A structured approach of EDA, outlier detection, feature selection, and classification modeling was applied to predict malignant or benign diagnoses in the breast cancer dataset. Key features such as radius_mean, perimeter_mean, concavity_mean, and texture_mean were identified as predictive through EDA, showing clear separability between classes in pair plots. Outlier detection techniques, including Isolation Forest and DBSCAN, were used to remove anomalies, improving model accuracy and stability. A correlation matrix helped address multicollinearity, ensuring only unique, relevant features were used in model training.





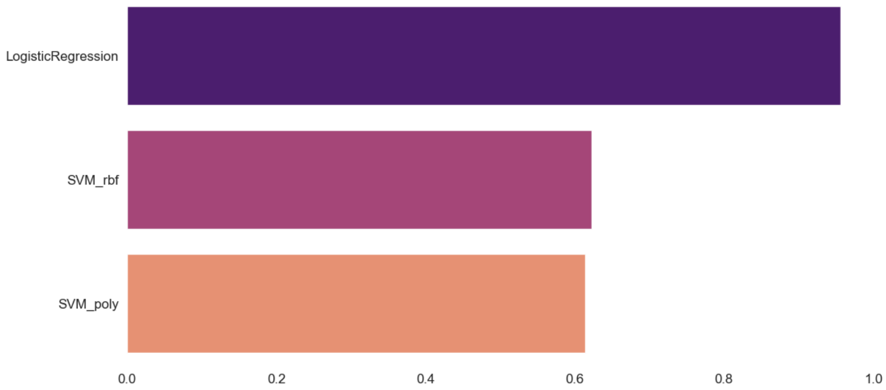
Outlier Detection Techniques

Correlation Matrix

Various classification models were evaluated, with logistic regression achieving the highest accuracy and linear SVM offering effective, interpretable boundaries. Decision boundary plots demonstrated that both logistic regression and linear SVM provided simpler, linear boundaries, while the RBF SVM's flexible nonlinear boundary better captured complex relationships. However, this flexibility introduced higher variance, suggesting overfitting risks in RBF SVM without proper regularization. Balancing the bias-variance tradeoff, logistic regression showed low variance and strong generalizability, while RBF SVM, though adaptable, risked overfitting. Adjusting class_weight in SVM models to balance improved sensitivity to malignant cases, which is essential for accurate detection. This dataset preparation ensures reliable, accurate cancer classification with models tailored to the data's characteristics.

Results:

The analysis revealed that several features, including radius_mean, texture_mean, perimeter_mean, and concavity_mean, are highly predictive of the target variable, Diagnosis. Logistic regression, due to its simplicity and high accuracy, was found to be the most effective model for this dataset, outperforming both SVM models in terms of accuracy. The decision boundary visualizations indicated that the dataset is well-suited for classification models, showing clear separability in selected feature pairs. Additionally, the impact of outlier removal and feature scaling was reflected in improved skewness and kurtosis values, supporting the dataset's readiness for predictive analysis. The RBF SVM, while providing more flexible boundaries, exhibited a slight accuracy decrease, which could indicate overfitting or heightened sensitivity to class imbalance. The custom correlation matrix also highlighted multicollinearity, which influenced feature selection and provided insights into redundant features, ensuring that the final models remained both interpretable and effective.



Accuracy Comparison between different models

Technical:

The dataset used for music genre classification was already in a pretty decent shape, there were no null values that needed to be dropped. The data manipulation consisted of encoding the label column to values 0-9, as well as standardizing the explanatory variables. SVM using RBF with weighted classes ended up being the best model to classify the music genre, it allowed for more complicated boundaries to be detected. Custom class weights allowed for a better f-score for more difficult genres to predict with little sacrifice to the more distinguishable genres. The analysis process began with analyzing and understanding the explanatory variables. Most of the research was done understanding what each of the variables actually meant. We started with logistic regression, then moved to SVM linear, SVM RBF balanced, and found SVM RBF with custom weights gave the best results overall.

Integrating EDA, outlier detection, feature selection, and modeling to distinguish malignant from benign samples. Key predictive features (radius_mean, perimeter_mean, concavity_mean, texture_mean) were identified through EDA with clear separability in pairwise plots. Outlier correction via Isolation Forest and DBSCAN enhanced data integrity, and a custom correlation matrix reduced multicollinearity, optimizing input features for model training. Logistic regression and SVM models, including linear and RBF kernels, were assessed for accuracy, boundary complexity, and bias-variance tradeoff. Logistic regression provided the highest accuracy and stability, while the RBF SVM's flexible boundary captured complex relationships, albeit with higher variance.