## 622 Homework 1

Exploratory Data Analysis

[Brandon Cunningham](mailto:btc2628@rit.edu)

[Jean Jimenez](mailto:jeanj7722@gmail.com)

Chafiaa Nadour

Introduction:

The main goal of this assignment is to investigate two datasets to identify suitable machine learning algorithms learned so far that can be effectively applied to analyze the data. Furthermore, the assignment seeks to address various inquiries concerning the influence of correlations among variables, categorical labels on algorithm selection, the advantages and disadvantages of the chosen algorithms, the correlation between algorithm choice and dataset characteristics, the reliability of results for making business decisions, and the possibility of errors in the analysis.

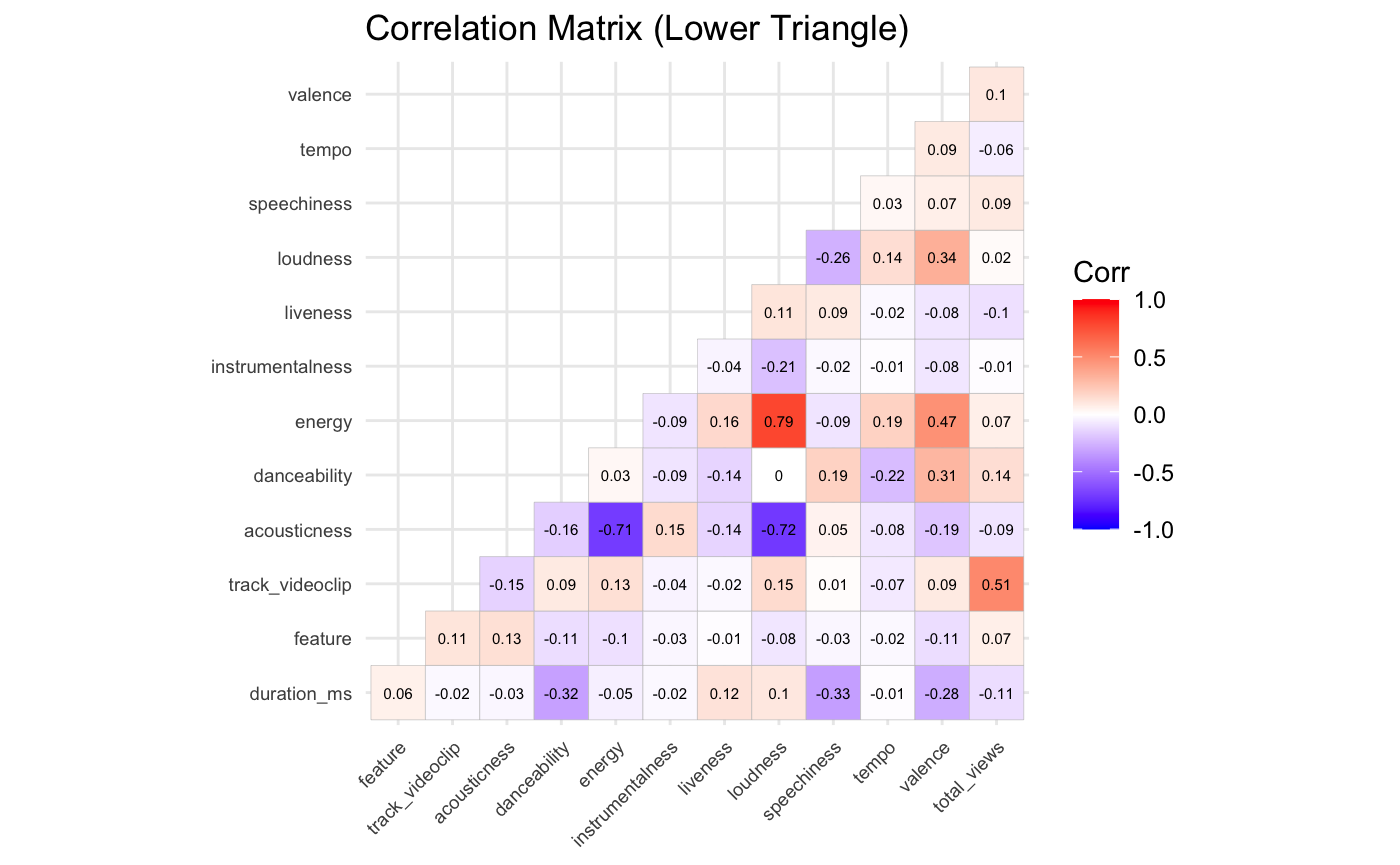
In this assignment, We have used two datasets collected from Kaggle.com, a small and large dataset which on their surface do not have much to do with one another, the first dataset is a smaller has 6567 row and 1 column: dataset on [Taylor Swift songs](https://www.kaggle.com/datasets/delfinaoliva/taylor-swift-discography), and the second is a larger 100K row, and 13 columns dataset on [diabetes](https://www.kaggle.com/datasets/priyamchoksi/100000-diabetes-clinical-dataset).

Exploratory Data Analysis:

Taylor Swift (Small) Dataset:

First talking about the smaller of the two datasets we have the Taylor Swift dataset, in this dataset the first thing we did was figure out how the columns were formatted initially and then transform them to be the proper type, ie making number columns numeric and date columns dates. The next thing we did before getting into the main exploration was convert the columns feature and track\_videoclip from Yes/No text columns into 1/0 numeric columns.

For the Taylor Swift dataset the label in the dataset is a combination of two columns, those being spotify plays and video views which combined give us total views across both platforms. For finding correlations between columns even after transformations we are still left with many columns that either contain redundant/irrelevant data such as the song uri, id, or lyrics, or data which if we wanted to predict how a song would perform before that song is released is data which we would be unable to know such as spotify global peak, track number, and album physical sales.



Taking a look at the correlation graph we can see that there are a few highly correlated spots in that both acousticness and loudness are highly correlated with energy, so one thing I would do knowing this is eliminate both acousticness and loudness as their information is already captured in energy. The next highest correlation is between trackvideo\_clip and total\_views, this would suggest that if a music video is made for the song then it is likely to get more total views, an extension to this project could be looking into the chicken and egg problem for this in trying to figure out if a track is popular and therefore it gets a music video or if a music video is made which then makes the track popular.

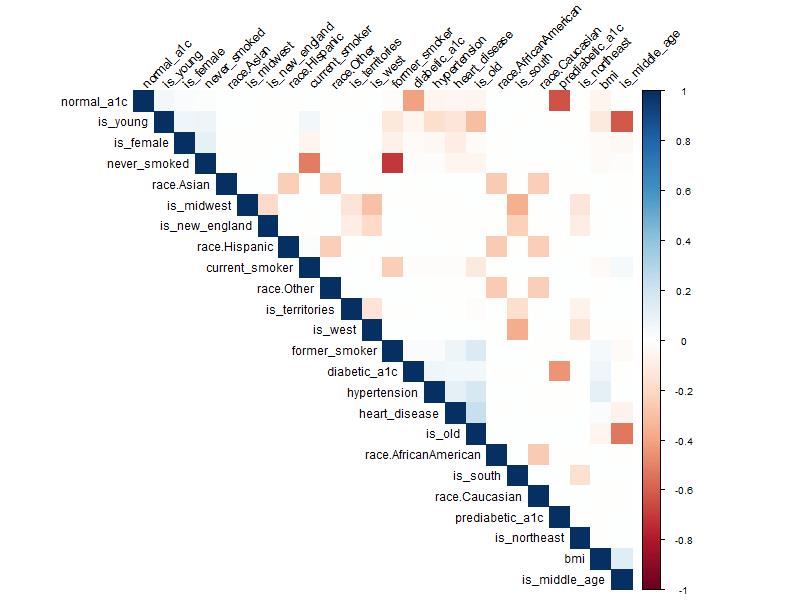
### Diabetes (Large) Dataset:

For our large dataset, we selected a diabetes dataset with 100,000 clinical records of diabetes patients. After reviewing the dataset, I identified ‘diabetes’ as our predictor variable, a binary outcome indicating whether or not a patient has diabetes. Since our predictor variable is binary and this is a healthcare dataset, I chose logistic regression. Logistic regression is frequently used in medical research for binary outcomes due to its interpretability, allowing us to calculate odds ratios to understand the likelihood of developing diabetes based on different risk factors.

To prepare the dataset for logistic regression, I analyzed each column, examining its contents, distributions, and statistics. I made necessary modifications to ensure the data was ready for modeling. For instance, the "year" column had uneven distributions across different years, so I decided to drop it to treat each data point as an individual observation, removing any time-series components.

The "gender" variable, initially coded as “Male” and “Female,” was recorded into a binary "is\_female" variable to indicate whether a patient was female. The "age" variable followed a normal distribution, so I binned the ages into three categories and created dummy variables for each. Similarly, I grouped states into regions of the country to simplify the location variable and created dummy variables. Other variables such as race, ethnicity, hypertension, and heart disease were already in binary dummy format, so no additional modifications were necessary.

The smoking history variable contained errors, including misspelled data, which I corrected. After cleaning the data, I binned and categorized the smoking history variable and created dummy variables. I applied the same binning process to the blood biomarker variables. These biomarkers, especially A1C levels, are critical indicators of diabetes, as diabetes is defined as having an A1C level above 6.4%. Since A1C is strongly correlated with diabetes, we may consider excluding it from the final model to avoid multicollinearity. A1C, being a continuous value, may be predicted as well depending on the algorithm.



After preprocessing the dataset, I generated a correlation matrix to assess multicollinearity between our features. While no extreme multicollinearity was found, some correlations existed between variables. To further analyze the relationships among the features, I conducted a principal component analysis. The first principal component accounted for only 8.16% of the total variance, indicating that no single factor dominated the prediction of diabetes. This makes sense given the complex nature of diabetes, where multiple factors are involved.

In fact, we needed 11 principal components to explain 66% of the variance, and it took 19 components to explain nearly 100%. The gradual increase in variance observed in our cumulative variance plot highlights that many factors contribute to diabetes risk. This suggests that we should avoid oversimplifying our model, as most variables provide unique information.

For the logistic regression model, we should keep most of the features since they all contribute meaningful information about diabetes risk. However, to manage multicollinearity, we might consider applying regularization techniques such as Lasso or Ridge regression, which help control for correlated predictors while maintaining model accuracy.

## Model Selection and Discussion:

Taylor Swift (Small) Dataset:

Given the complexity of the data knowing that even after eliminating many columns we still have over a dozen left for a relatively small number of observations we would want to use an ensemble algorithm, and in this case we believe that random forest regression would be the best choice. Random forest is a good choice for this dataset because the stakes are not as high as with our other dataset which is a medical dataset on Diabetes. With this dataset we do not need high interpretability, and instead the best predictive power is what is most important. With that said, with the right fine tuning we believe that random forest regression would give the best results given the complexity of the problem and the fact that we are trying to predict a discrete number rather than a specific label.

### Diabetes (Large) Dataset:

When it comes to predicting diabetes, we believe the best algorithm to use is logistic regression. Our PCA revealed that most of our components/variables are important influencers of the model. Logistic regression is regularly used in medical research for this exact reason to develop algorithms. This is because logistic regression is easy to explain/understand and very interpretable. The results of a logistic regression can be in odds ratio.

Our results support this choice. The logistic regression model, trained using 5-fold cross-validation, achieved an average RMSE of 0.2713552 and an R-squared value of 0.283191 on the training data. When applied to the test data, the model achieved an RMSE of 0.2700682, which is consistent with the training results, indicating good generalization. In comparison, the decision tree model (using the rpart function) resulted in a higher RMSE of 0.2835493 on the test data. This suggests that the logistic regression model outperforms the decision tree in terms of prediction accuracy for this dataset.

The logistic regression model's performance, while not perfect, is reasonable given the complexity of predicting diabetes, which involves multiple risk factors and interactions. The model's ability to consider all variables simultaneously aligns with our understanding that many factors contribute to diabetes risk. The interpretability of logistic regression is crucial in a medical context. Unlike the decision tree, which only used a subset of variables (bmi, diabetic\_a1c, hypertension, is\_middle\_age, is\_young, prediabetic\_a1c), the logistic regression model considers all available predictors.

Conclusion:

For the first dataset, the analysis reveals that the random forest model is deemed more reliable for business use due to its superior predictive performance. Conversely, in the case of the second dataset, we prefer a logistic regression model for business purposes because it considers all the important factors necessary for predicting diabetes. We have also tried to use linear regression and decision tree but it performed very poorly which is a proof that the models we’ve chosen are the best. Overall, this analysis has given us significant insights of building, selecting, and analyzing predicting models to make informed business decisions.