STATS 302 Progress Report 1

Proposed topic 1: Classification of types of tumor based on RNA-seq data

Motivation:
In selecting the topic "Classification of Types of Tumor Based on RNA-seq Data" for my project, I was motivated by the critical intersection of machine learning and genomics in advancing medical diagnostics. The progress in RNA sequencing (RNA-seq) technology offers an expansive view of the transcriptome, providing detailed insights into the genetic expressions unique to different tumor types. Accurate classification of tumors using RNA-seq data can lead to more personalized treatment strategies, improving patient outcomes. Furthermore, this topic allows me to explore the cutting-edge field of machine learning in bioinformatics. The challenge of this topic is the number of features of the dataset. While there are only 801 cases, there are 20531 features, which I expect will cause trouble in terms of singularity. Since we will conduct a train-test split, the imbalance of the number of instances and the number of features is further worsened. Through this project, the team will understand better how to handle high-dimensional data in a multi-classification task, which will contribute to the expanding realm of personalized medical treatment based on machine learning algorithms.

Dataset introduction:
The dataset contains 801 cases with 20531 features. Each row contains the level of RNA-seq gene expression measured by Illumina HiSeq Platform. The labels have four categories: BRCA, KIRC, COAD, LUAD, and PRAD.

Proposed topic 2: Prediction of the promotion and discount of “Double Eleventh Day”

Motivation:
In the current digital era, the e-commerce industry is facing significant challenges, with one of the main issues being how to effectively predict the impact of promotional activities. This involves not only maximizing sales but also ensuring the efficient utilization of advertising and promotional resources. However, due to the complexity of market dynamics and user behavior, existing prediction methods have certain limitations. This project aims to address this issue by applying advanced statistical models and data science techniques to enhance the accurate prediction capability of e-commerce platforms regarding the effectiveness of promotional activities. By tackling this challenge, we can assist businesses in better formulating promotion strategies,

maximizing sales efficiency, and driving the sustainable development of the entire industry.
During the “Double Eleventh Day” (Double Eleven) shopping festival, which is the world's largest online shopping festival, major e-commerce platforms launch numerous discounts and promotions, attracting hundreds of millions of consumers. Consumers can take advantage of highly attractive prices to purchase a wide range of goods, spanning from clothing and electronics to home goods, among various other categories. This event typically sets record-breaking sales figures within a 24-hour period, surpassing even the sales of events like the United States' "Black Friday" and "Cyber Monday."
By analyzing the sales figures and promotional activities of various sectors during previous “Double Eleven” events, predicting the market trends of products in each sector for the next year can significantly enhance the success rate of Double Eleven marketing by setting more reasonable discount activities.

Dataset Required (For previous “Double Eleven”):
Sales: sales revenue, sales quantity
Promotion: promotion duration, promotion type, promotion intensity
Others: Advertising expenditure data, users behavior data

Model Required:
Linear regression, decision trees, random forests, gradient boosting, etc.

Proposed Topic 3: Classification of Soccer Team Playing Styles based on database of number of passes and the speed that the team progresses the ball up field.

Motivation: In football leagues, some teams are known for playing in a fast and direct style while others are playing in a slow and intricate way. These differences in playing styles might be noticed and classified by the spectator, but the classification can also be done by doing some data analysis.

In the database that is intended to be analyzed, there are 8 features for each team that are included in the dataset. All the 8 features above are either related to number of passes in games or the speed that the team plays the ball forward. So far, the basic idea is to classify teams into fast & direct playstyle (class 1) and slow & intricate playstyle (class 2) (Maybe more classes can be added such as balanced, an extra class of playstyle that is somewhere between fast & direct and slow & intricate) based on some or all of the 8 features above. Potential methods that may be involved include Naïve Bayes and KNN.

One problem that might affect the quality of the classification is the number of samples. Even if we include all major & minor soccer leagues worldwide that data related to passing and speed of playing forward are available, the scale of the data is still not large. Maybe the study should not be limited to one year and data from previous years should be also included in the samples, but further research on an appropriate database may be required.