INTRODUCTION

In the realm of cancer research, people are constantly seeking ways to enhance treatment outcomes and provide hope for patients and their loved ones. This project presents an opportunity to evaluate the effectiveness of a promising new cancer treatment known as Miraculon-B, specifically for solid tumors.

Our primary objective is to gain a comprehensive understanding of which patient subgroups might experience more favorable responses to this innovative treatment, thereby offering them the potential for improved outcomes. To facilitate our analysis, GSK has generously provided us with two datasets: clinical-study.csv and protein-levels.csv. The clinical-study.csv dataset encompasses detailed information on 772 patients, including essential variables such as age, weight, and, notably, the treatment response itself. Through careful examination, we aim to determine whether patients who received Miraculon-B exhibited more favorable treatment outcomes in comparison to those in the control group. Moreover, the protein-levels.csv dataset provides valuable insights into the concentration of a specific protein that could serve as a predictive biomarker for solid tumors. By thoroughly exploring this dataset, we hope to uncover whether factors like age, weight, or protein concentration can serve as indicators of a patient's likelihood to respond positively to Miraculon-B.

DATA CLEANING

The biggest challenge with data is that data are always filled with errors because data in its raw form is always poorly managed. The process of data cleaning is necessary to ensure a consistent dataset that provides accurate information. You can find the detailed process of cleaning the dataset in the attached Jupiter notebook. Each feature in the dataset was reviewed for errors, missing data, and inconsistent information.

- Duplicate value removal: In the dataset, a duplicate value was identified for SUBJ_001. To ensure data integrity, the duplicate value was dropped from the dataset.
- Exclusion of pediatric data: Since the analysis focuses on adult patients, rows where the age was less than 18 were excluded from the dataset. This decision was made to remove pediatric data and maintain consistency within the analysis.

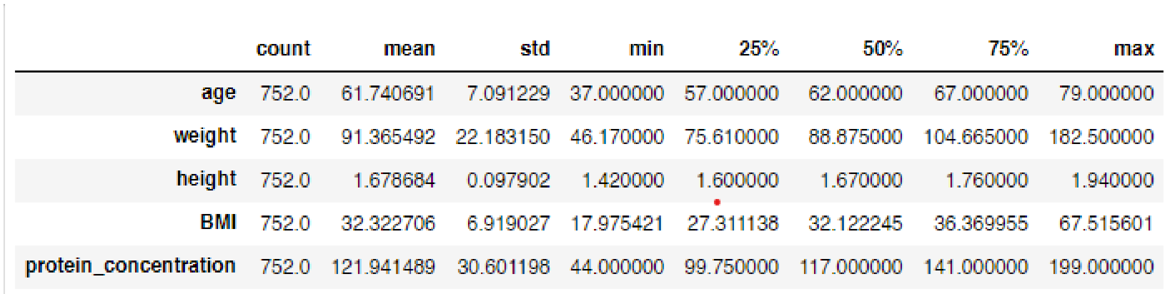
- Missing value removal: In the weight variable, certain entries had missing values. To maintain the accuracy of the analysis, all rows with missing values in the weight variable were dropped from the dataset.

FEATURE ENGINEERING

After completing the data cleaning stage, the dataset has been improved to provide more accurate and useful information. To enable further analysis and modeling, additional features have been added to the clinical data.

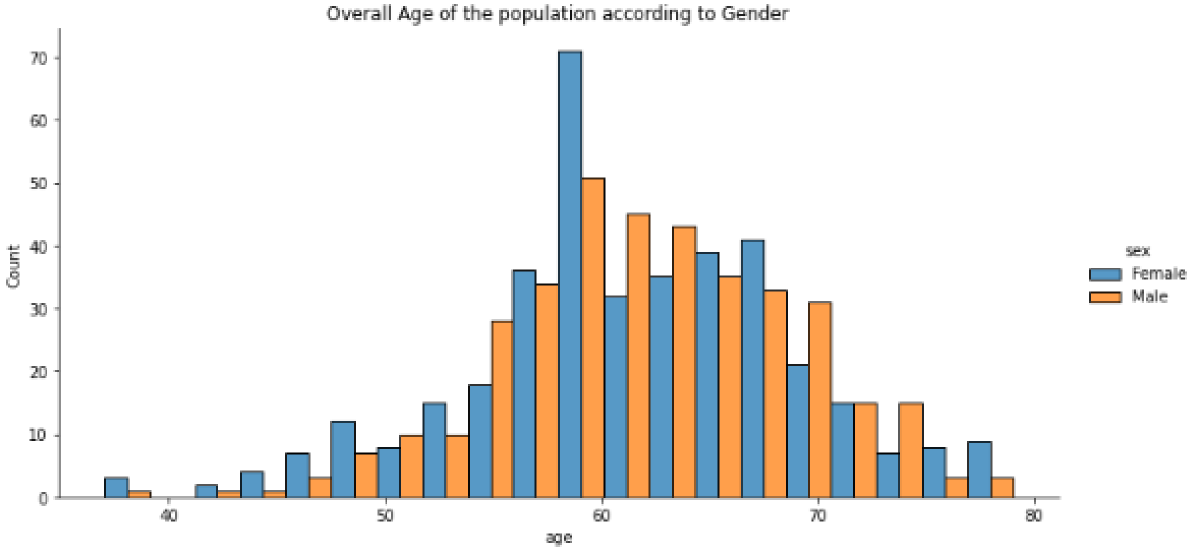
- BMI calculation: To enhance the analysis, a new feature called BMI (Body Mass Index) was derived by dividing the weight of each patient by the square of their height. This calculation provides a standardized measure of body composition and can contribute valuable insights to the analysis.
- Data merging: In order to enrich the dataset, the clinical data and the protein level data were merged together using the subject ID as the common identifier. This merging process consolidates the information from both datasets, allowing for a more comprehensive analysis that incorporates clinical variables and protein concentration levels.

DESCRIPTIVE ANALYSIS AND VISUALIZATION

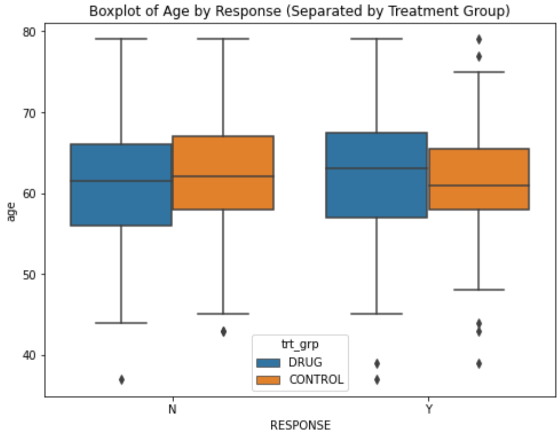


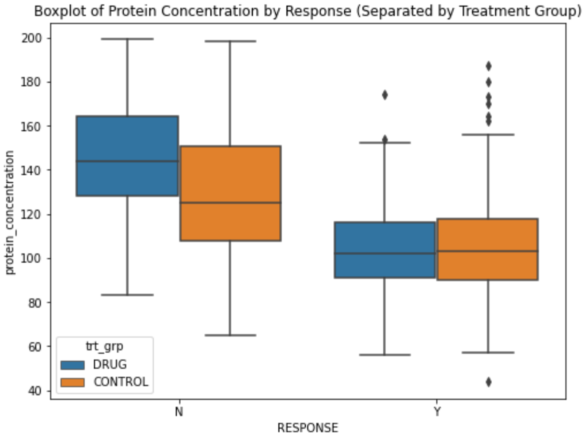
	count	mean	std	min	25%	50%	75%	max
age	752.0	61.740691	7.091229	37.000000	57.000000	62.000000	67.000000	79.000000
weight	752.0	91.365492	22.183150	46.170000	75.610000	88.875000	104.665000	182.500000
height	752.0	1.678684	0.097902	1.420000	1.600000	1.670000	1.760000	1.940000
BMI	752.0	32.322706	6.919027	17.975421	27.311138	32.122245	36.369955	67.515601
protein_concentration	752.0	121.941489	30.601198	44.000000	99.750000	117.000000	141.000000	199.000000

These statistical measures provide an overview of the central tendency, variability, and distribution of the variables in the dataset, enabling a better understanding of their characteristics and potential insights for further analysis. There are 752 observations for all features after cleaning. The average age is 61.74. The age data has a standard deviation of 7.09, indicating a moderate amount of variability or spread around the mean age.The minimum and maximum age are 37 and 79 respectively. The median age is 62. The mean and the median age are closely the same which means the age of the patients are almost symmetrically distributed. The average weight and height are 88.87kg and 1.67 meters respectively.

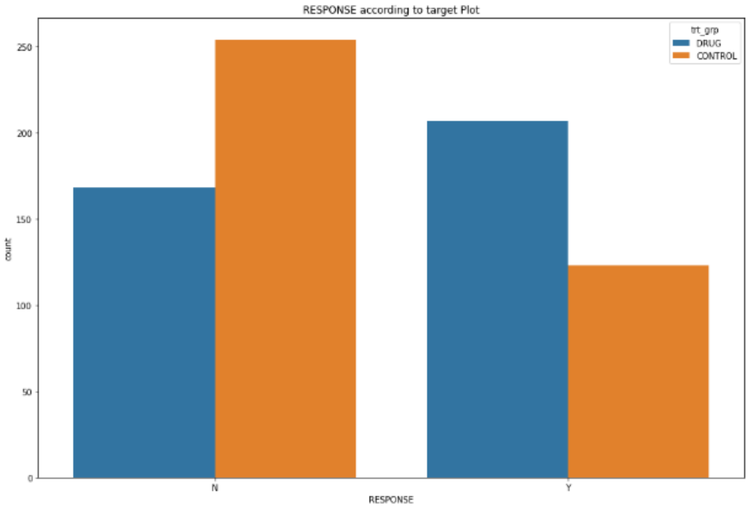


Based on the plot above, it is evident that a significant portion of the respondents belong to the age group ranging from 55 to 70 years. This age range represents the majority of the individuals who participated in the study or survey. Additionally, the plot indicates that there is a higher number of female respondents compared to male respondents. This suggests that the female population has a greater representation or participation in the data being analyzed.



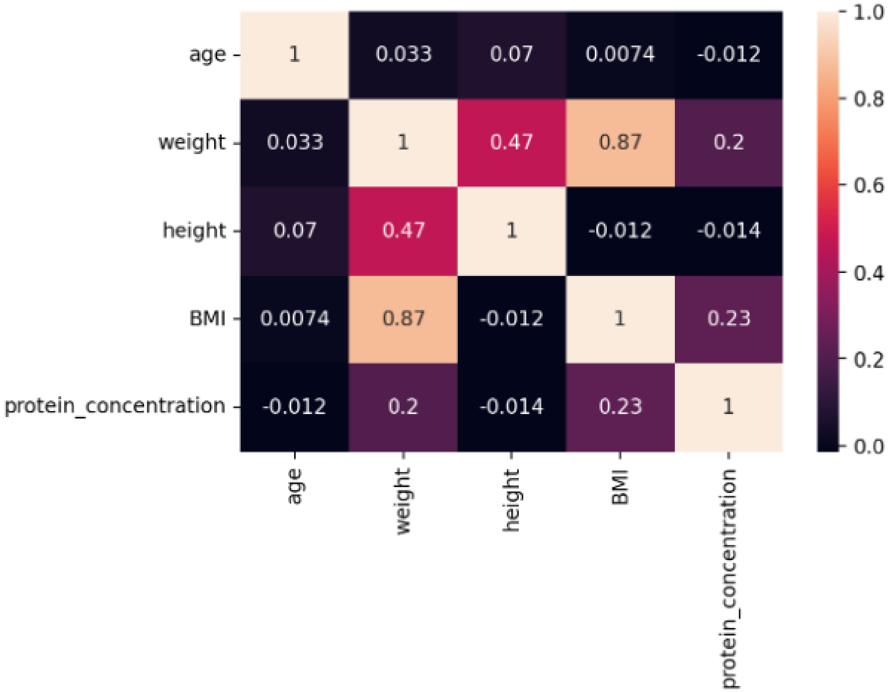


The average age of individuals in the control target group is 61.81, slightly higher than the drug target group's average age of 61.67. This indicates a small age difference between the two groups, suggesting some overlap in their age distributions. Regarding responder type, the mean age of "Yes" responders is 61.85, slightly higher than the mean age of "No" responders at 61.66. The difference is again small, indicating some age distribution overlap between the two responder groups. Moving on to weight, "Yes" responders have a higher average weight of 93.19 compared to "No" responders, who have an average weight of 89.94. The difference in mean weight is notable, with "Yes" responders having a mean weight 3.25 units higher than "No" responders. Lastly, in terms of protein concentration, "No" responders have a significantly higher mean protein concentration of 135.71, compared to "Yes" responders with a mean of 104.34. This indicates a substantial difference of 31.37 units between the two groups' mean protein concentrations.



In the Drug target group, there are 207 individuals who responded with "YES," while 168 individuals responded with "NO." On the other hand, in the Control target group, there are 123 individuals who responded with "YES," and 254 individuals responded with "NO."

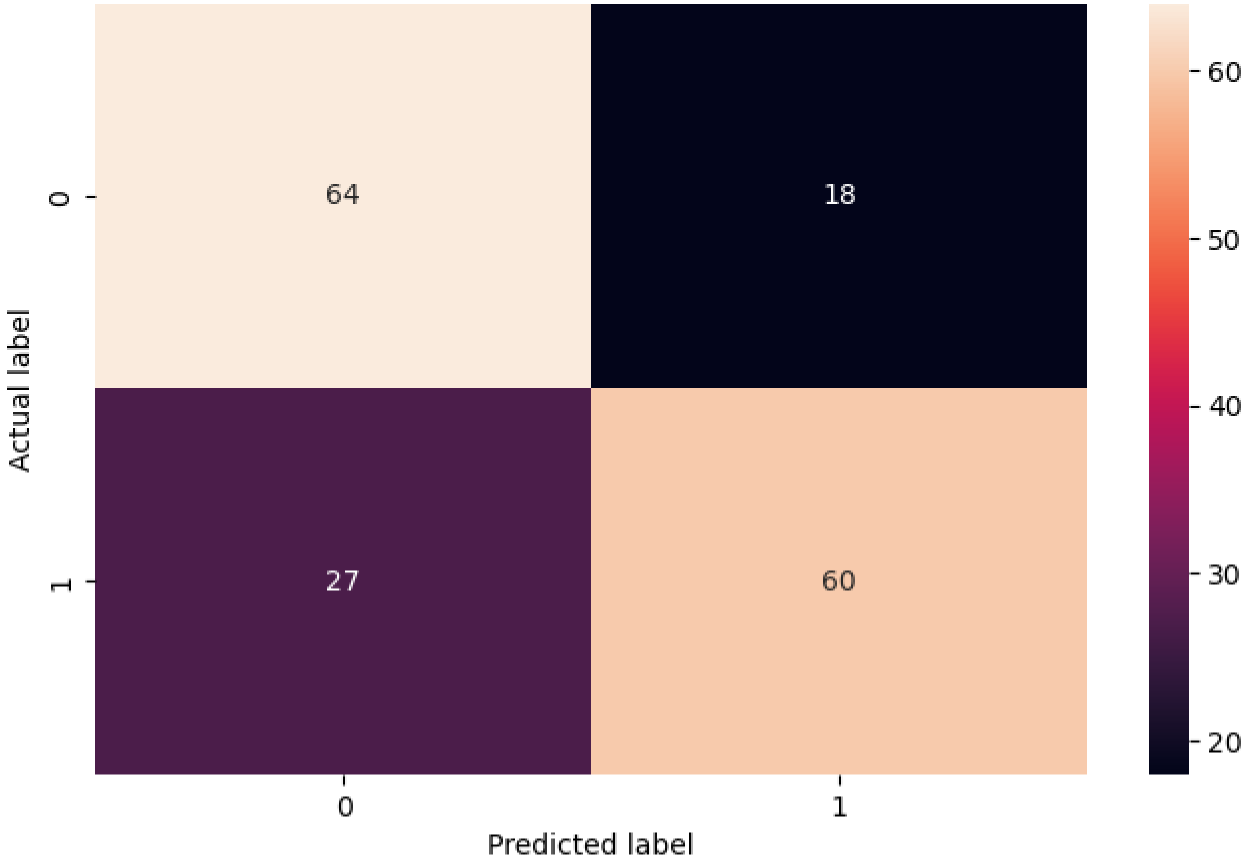
These numbers provide insights into the distribution of responses within each target group. In the Drug target group, a higher number of individuals responded with "YES" compared to "NO," indicating a greater inclination towards the positive response. Conversely, in the Control target group, the majority of individuals responded with "NO," outnumbering those who responded with "YES."

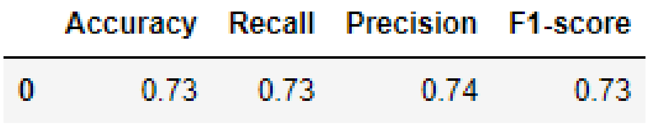


The heatmap provides insights into the relationships between these features. Positive correlation values indicate that as one feature increases, the other feature tends to increase as well, while negative correlation values indicate that as one feature increases, the other feature tends to decrease. The strength of the correlation is determined by the magnitude of the correlation coefficient, with values closer to 1 indicating a stronger relationship. The heatmap displays the correlation between different features. The correlation coefficient between BMI and height is -0.012, indicating a very weak negative correlation. The correlation between height and weight is 0.47, suggesting a moderate positive correlation. BMI and weight have a correlation coefficient of 0.87, indicating a strong positive correlation. The correlation between protein concentration and BMI is 0.23, representing a weak positive correlation. Height and protein concentration have a correlation coefficient of -0.014, indicating a very weak negative correlation. Lastly, weight and protein concentration have a correlation coefficient of 0.2, suggesting a weak positive correlation.

MODEL

To analyze the dataset, the Logistic Regression Classifier was. The RESPONSE variable served as the target variable for both models. By employing these classification models, effective prediction and classification of the dataset's response variable were achieved. Each model possesses its own strengths and weaknesses, and the selection of the appropriate model depends on several factors. Evaluating the models' performance using metrics like accuracy, precision, recall, and F1 score aids in determining the model that performs better in predicting the response variable for this specific dataset.





	Accuracy	Recall	Precision	F1-score
0	0.73	0.73	0.74	0.73

The classification report provides an assessment of the model's performance in terms of accuracy, precision, recall, and F1-score on the test set. The accuracy of the model is 0.75, meaning that it correctly classifies 75% of the instances in the dataset., a precision of 0.74, meaning that 74% of the instances predicted are correct. The recall is 0.73, indicating that the model identifies 73% of the actual instances. The F1-score for class N is 0.73, reflecting a harmonized measure of precision and recall.

The confusion matrix provides further insight into the model's predictions. It shows the counts of true negatives (correctly predicted as class N) which is 64, false positives (incorrectly predicted as class Y when they are actually class N) which is 18, false negatives (incorrectly predicted as class N when they are actually class Y) which is 27, and true positives (correctly predicted as class Y) which is 60.

The classification report and confusion matrix together provide a comprehensive assessment of the model's performance. The precision, recall, and F1-score indicate the model's ability to accurately classify instances for each class. The confusion matrix further reveals the model's predictions for each class and helps identify any potential misclassifications.

CONCLUSION

Based on our analysis, we can conclude that Miraculon-B shows promise as an effective treatment for solid tumors. The Logistic Regression Classifier model provided satisfactory performance in predicting treatment outcomes. However, further studies and validation are necessary to confirm the superiority of Miraculon-B over the standard of care.