SI 370 Final Project Report: Heart Disease Analysis

Statement of Purpose

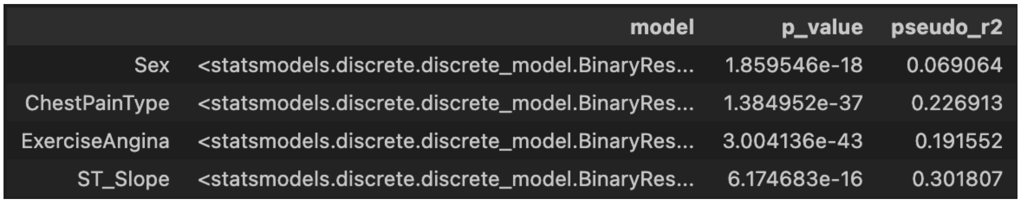
Heart disease is the leading cause of death for men and women worldwide across a variety of ethnic groups. According to the CDC, on average, 1 person dies every 33 seconds from heart disease and accounts for roughly 1 in 5 deaths (CDC). Over the past 3 years alone, treatments and resources related to heart disease have cost an estimated $250 billion. As such, managing and understanding underlying causes and risk factors of heart disease is important to ensure a healthy population and reduce these financial burdens among patients. This project utilizes data science methods to analyze key predictors of heart disease, such as cholesterol levels, chest pain, and blood pressure. By developing predictive models, this project can help to identify the most significant factors contributing to heart disease. These insights could inform targeted treatment strategies and help prioritize interventions to address the most critical risk factors.

Question 1: What variable is the strongest predictor of heart disease?

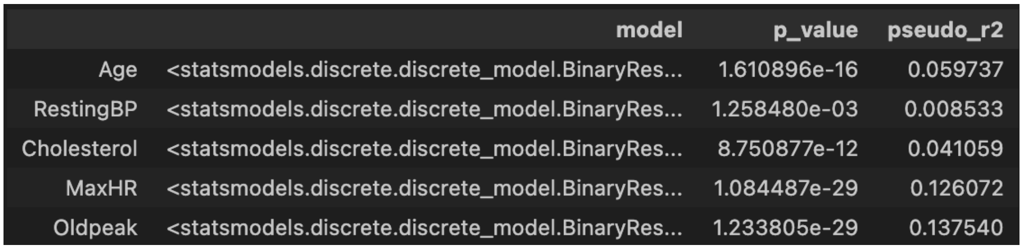
Analysis

This question aims to determine which variable can predict the incidence of heart disease the most. With each row representing a patient, the dataset includes numerous variables relating to heart disease. Numerical variables included: Age, RestingBP (resting blood pressure), Cholesterol, MaxHR (max heart rate), and Oldpeak (depression of the ST curve of an ECG during activity relative to rest). Categorical variables included Sex, ChestPainType, ExerciseAngina (angina refers to chest pain), ST_Slope (slope of the ST curve). Originally, a linear regression model was intended to be created between each variable and heart disease and the summary of the model characteristics to be compared; however, because the outcomes of heart disease are either 0 (no incidence of heart disease) or 1 (incidence of heart disease) for each row. Due to this binary outcome, a linear regression model would not necessarily be appropriate for this question or yield meaningful results. Therefore, logistic regression from statsmodels was used instead, which is a type of regression analysis used when the categorical dependent variable only takes discrete values. Although we did not formally learn about them in class, this source was used to understand logistic regression more.

The numerical and categorical variables were explored separately. For each variable type, a function was applied to each individual variable to create a logistic regression model and output the resulting p-value and pseudo R-squared value. The following p-values and R-squared values were determined:

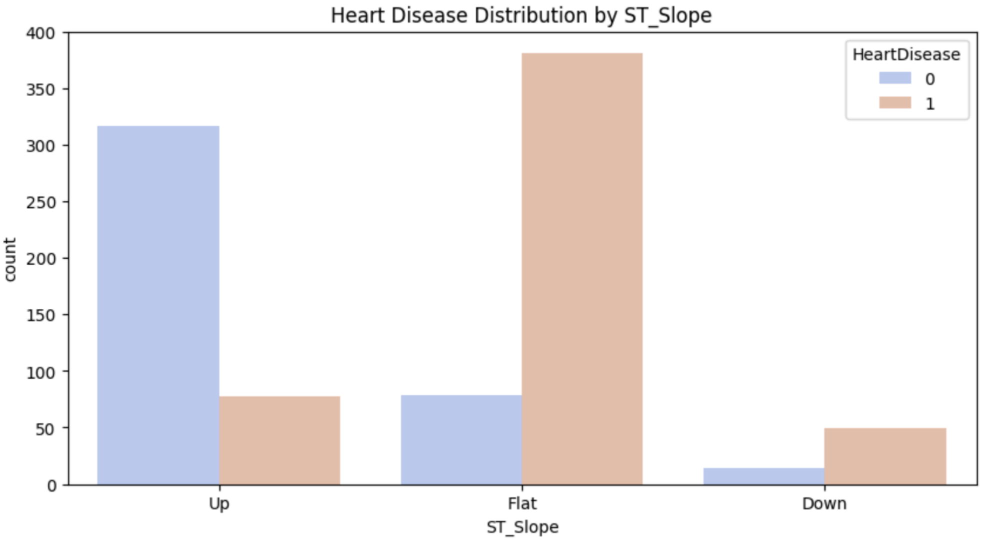


	model	p_value	pseudo_r2
Sex	<statsmodels.discrete.discrete_model.BinaryRes...	1.859546e-18	0.069064
ChestPainType	<statsmodels.discrete.discrete_model.BinaryRes...	1.384952e-37	0.226913
ExerciseAngina	<statsmodels.discrete.discrete_model.BinaryRes...	3.004136e-43	0.191552
ST_Slope	<statsmodels.discrete.discrete_model.BinaryRes...	6.174683e-16	0.301807



	model	p_value	pseudo_r2
Age	<statsmodels.discrete.discrete_model.BinaryRes...	1.610896e-16	0.059737
RestingBP	<statsmodels.discrete.discrete_model.BinaryRes...	1.258480e-03	0.008533
Cholesterol	<statsmodels.discrete.discrete_model.BinaryRes...	8.750877e-12	0.041059
MaxHR	<statsmodels.discrete.discrete_model.BinaryRes...	1.084487e-29	0.126072
Oldpeak	<statsmodels.discrete.discrete_model.BinaryRes...	1.233805e-29	0.137540

To visualize the importance of ST_slope graphically, the following bar chart was created:



Explanation of Findings

Among the categorical variables, all of them had p-values less than 0.05, indicating that all variables within the logistic regression models are statistically significant. Pseudo-R^2 values depict the proportion of variance in the dependent variable that is due to the independent variable, with larger pseudo-R^2 values closer to 1 indicating more variance that can be explained by the independent variable. Among the categorical variables, ST_slope, or the slope of the ST curve on an electrocardiogram, had the largest pseudo-R^2 value of 0.30. This suggests that the slope of the ST curve is the strongest categorical predictor of heart disease in the dataset. Among the numerical variables, Oldpeak, or the depression of the ST curve of an ECG during activity relative to rest, had the greatest pseudo-R^2 value of 0.14. This indicates that Oldpeak is the strongest numerical predictor of heart disease in the dataset. Furthermore, as illustrated in the bar chart, the vast majority of patients with a flat ST slope have heart disease, and the vast majority of patients with an upwards ST slope do not have heart disease. This further supports that the slope of the ST curve in an ECG can possibly predict if a patient has heart disease or not.

Limitations

One limitation of this analysis lies with the feature representation. Specifically, here the categorical variables were not encoded with any further feature engineering. More advanced techniques such as one-hot encoding could potentially improve the logistic regression model's performance. In a similar sense, one key limitation is that the logistic regression model used is relatively simple and might not depict complex relationships between variables. Similarly, there were some challenges in choosing the right model for this analysis, as linear regression was the original plan, but further exploration of the data demonstrated that this plan was not feasible. While logistic regression allowed for relationships between various variables and a binary outcome to be compared, exploring other models and comparing their performance would be beneficial. For this question specifically, there were also challenges in determining which features would be the most important to be explored; as such, this question was largely constructed so as to explore each variable separately rather than diving into a specific variable more deeply which limits the extent to which deeper relationships between individual variables are understood.

Future Work

More complex algorithms and intensive data collection efforts could improve this model. As the logistic regression model is quite simple, in future explorations, more advanced models such as decision trees, random forests, or other classifiers involving machine learning could potentially capture more complex relationships better. Because this analysis virtually did not include any approaches to feature selection, more systematic approaches to determine which features to explore, like feature importance, or reading contemporary research about the most pervasive risk factors could benefit this analysis further in future work.

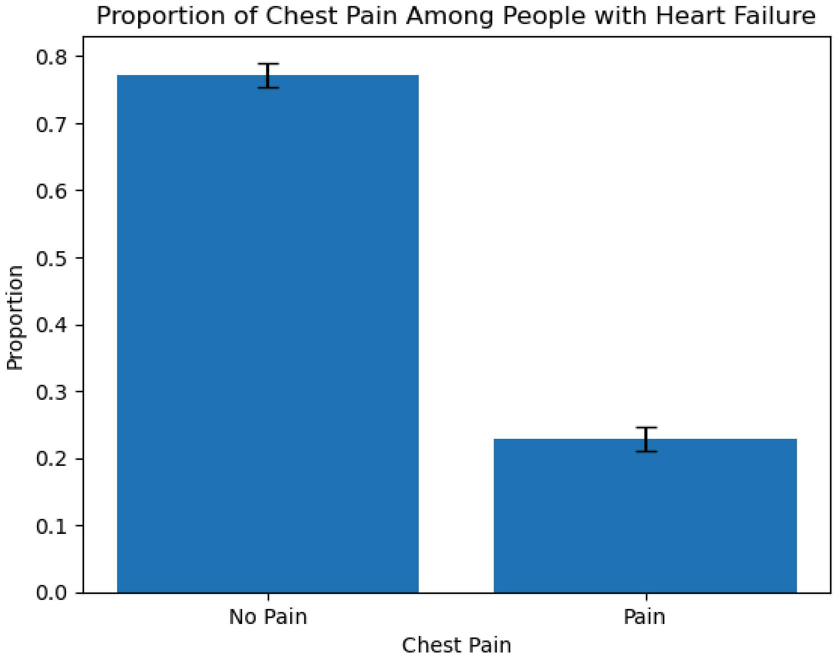
Question 2: What proportion of people with heart failure have chest pain?

Analysis

To analyze the proportion of people with heart failure that have chest pain, we needed to perform a variety of data analysis techniques. The dataset that we chose contains a “HeartDisease” column, which is our variable of interest. A value of 1 suggests that the individual has heart disease and a value of 0 suggests that the individual is normal. Additionally, to answer this question, we also needed to analyze the “ChestPainType” column. The values are either TA, ATA, NAP, or ASY, which represent typical angina, atypical angina, non-anginal pain, and asymptomatic. Typical angina refers to chest pain due to heart-related issues, which is typically staggered by exertion or stress. People with atypical angina and non-anginal pain still experience discomfort. Due to these various classifications, we decided to create a column “pain,” which would contain “pain” if the chest pain type was typical angina, atypical angina, or non-anginal pain. If the ChestPainType was asymptomatic, then we categorized that as “no pain” because the individual is not experiencing any symptoms. I needed to apply a basic function to translate these values into “pain” and “no pain” statements. To find the proportion of people with health failure that have chest pain, I counted the number of people with pain and heart failure and divided that by the number of people with heart failure.

Findings

From the analysis, I found that 22.8% of people with heart failure have chest pain. This aligns with expectations because other studies report a prevalence between 23-85% depending on the study population. To see if there were significant differences between the proportion of people with heart failure that experienced pain versus those who didn't, I conducted a chi-square test. I found a p-value less than 0.01, which suggests statistically significant differences in the proportion of individuals who experience chest pain among people with heart failure. To analyze this graphically, I created a bar graph with error bars.



Limitations

This analysis presents many limitations. The sample size and population diversity may limit the generalizability of the findings. There could be characteristics of individuals that are not fully represented in the dataset, which can introduce bias. Additionally, the sample size of this dataset that we are using is relatively small at 918 entries.

The simplification of "ChestPainType" into binary categories of "pain" and "no pain" may reduce the accuracy of the results. People have different tolerances of pain and simplifying it may overlook any insights from pain type. This can obscure meaningful distinctions between different pain types and their correlation with heart failure severity. Additionally, this analysis does not account for potential confounding variables, including other conditions like diabetes or hypertension. Medication usage could also influence heart failure prevalence.

Future Work

To address these limitations, future work can include data improvements. Expanding the dataset to include a larger and more diverse population can help increase generalizability and power of the analysis. Additionally, introduction of other confounding variables that are already present might be helpful, such as diabetes and hypertension. Additionally, with more entries, we could explore associations between specific types of pain and heart failure. With this additional data, we would perform more advanced modeling, including logistic regression to evaluate the relationship between heart failure and chest pain while adjusting for confounding conditions.

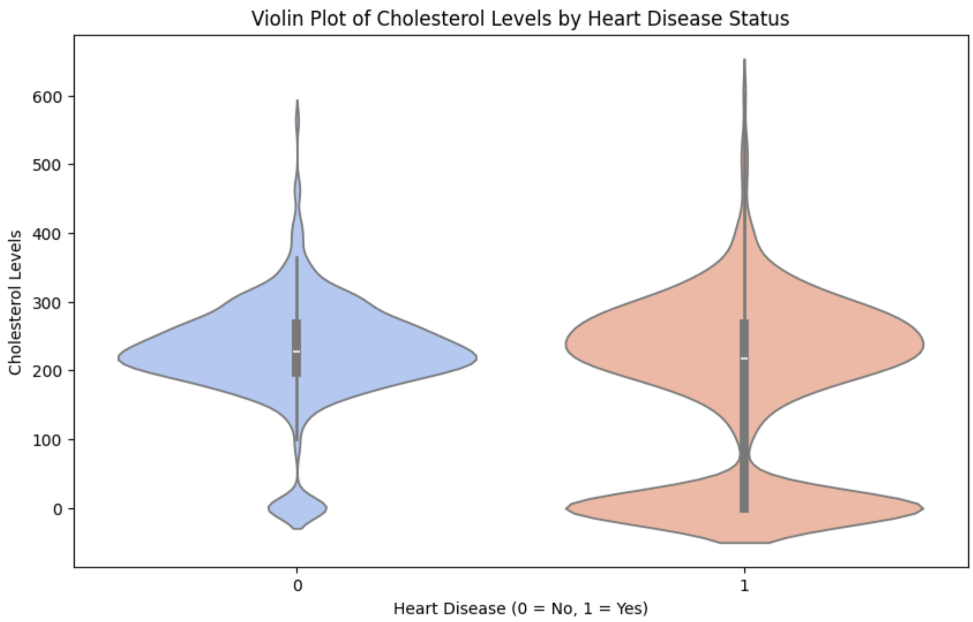
Pain could be evaluated next time using a numeric scale. While people have different pain tolerances, using such a scale would allow us to see if there are any associations between the level of pain and heart failure.

Question 3: How can cholesterol predict heart failure?

Analysis

To explore this question, two different machine learning models were used. In the first model, cholesterol levels were used as the only feature to predict heart disease incidence. A Voting Ensemble Classifier was implemented, combining three different machine learning models: Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. Before fitting the models, the cholesterol data was standardized using StandardScaler to ensure that the feature's scale did not negatively affect the model performance. After training the model, its accuracy was computed to assess how well it could predict heart disease based on cholesterol levels alone.

In the second model, all available variables, not just cholesterol, from the dataset were used to predict heart disease, and these variables were either scaled or one hot encoded if they were numerical or categorical, respectively. Like the first model, a Voting Ensemble Classifier was used to combine the three classifiers: Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. The full dataset, including variables such as age, blood pressure, etc, was inputted into the model. After fitting the model, its accuracy was also calculated and compared to the first model to evaluate the impact of additional features on prediction accuracy. The following visualization was created to further understand cholesterol and heart disease:



Findings

The model that used only cholesterol levels to predict heart disease had an accuracy of 57.6%. This suggests that while cholesterol may play a role in heart disease, it is not a highly reliable predictor when used alone. In contrast, the model that accounted for all the available features from the dataset had an accuracy of 88%. This indicates that factors beyond just cholesterol are needed in order to accurately predict the incidence of heart disease. Ultimately, the comparison of these two models demonstrates that cholesterol alone is not enough to predict heart disease with high accuracy and that more variables are needed for more reliable and accurate predictions. Based on the violin plot, across those with and without heart disease, there is a similar distribution of cholesterol levels and the median cholesterol levels are close to each other. This further supports the idea that generally, the distribution of cholesterol levels are similar between those with and without heart disease, reinforcing the fact that cholesterol levels alone might not be a good predictor of heart disease.

Limitations

One major weakness is that the dataset primarily focuses on a relatively narrow spectrum of health factors, which does not necessarily take into account other potentially important variables such as diet, lifestyle factors, or genetic predispositions that could play a significant role in heart disease prediction. Another major limitation is that the dataset does not differentiate between low-density cholesterol and high-density cholesterol; the former has negative impacts on health whereas the latter has positive impacts on health. Without information as to what type of cholesterol was collected or if the variable in the dataset was an aggregate of the two, an accurate understanding of the impact of cholesterol on heart disease cannot fully be understood. Without additional information, the accuracy of the models could be limited. Further limitations are that it is possible for the classifiers to overfit the models if there is not proper tuning of hyperparameters.

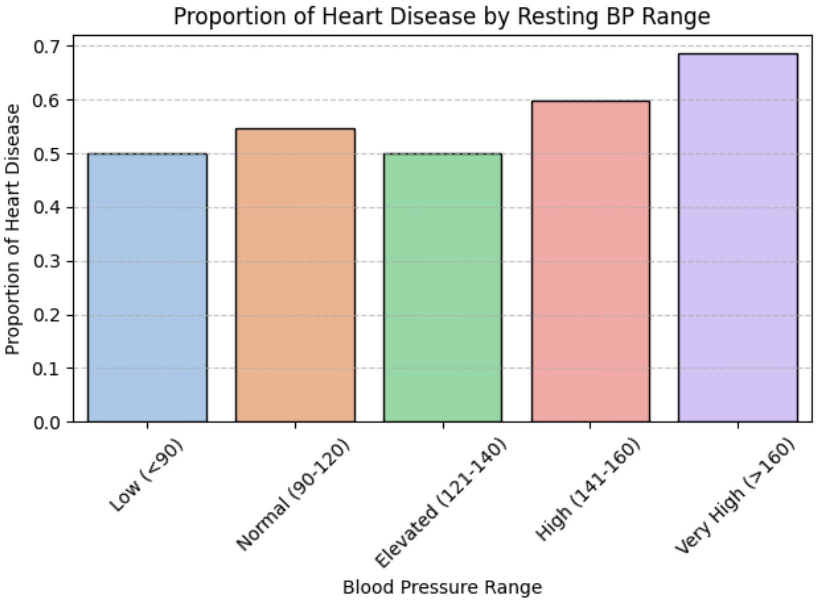
Future Work

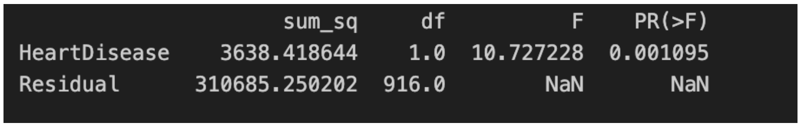
Firstly, adding more features through more thorough data collection could also perhaps improve this exploration. Additional variables that might have a more direct impact on heart disease could be measured, like family history, habits, or genetics. Differentiating between low density and high density cholesterol would also make the relationship between cholesterol and heart disease more insightful, as different types of cholesterol can have vastly different impacts on human health. These would provide a more comprehensive understanding of the factors contributing to heart disease. Additionally, searching for more optimal hyperparameters using Grid Search or Random Search like the depth of decision trees or the number of estimators in Random Forests could possibly help to improve the performance of the model. Additionally, while the current ensemble uses hard voting, exploring other ensemble strategies could possibly improve the strength of the model as well.

Question 4: What is the optimal blood pressure based on the heart failure outcomes? Does this match with the recommendations available online?

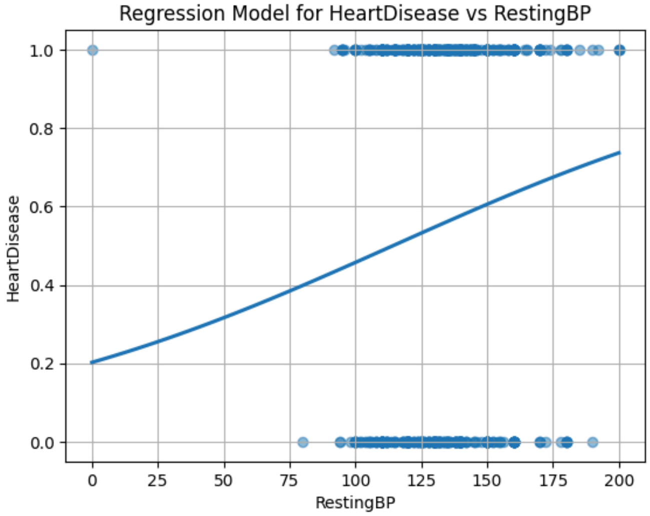
Analysis

To address this question, we need to first determine the relationship between the heart failure outcomes and the blood pressure variables. Then, we will determine if there is a threshold for the blood pressure level (“RestingBP”) that can divide the heart failure outcome from “0” to “1” using the “HeartDisease” column. We want to figure out which level of blood pressure for a person is too high which will result in a high chance of having a heart disease, and compare this level to the blood pressure recommendation online. We first divided the “RestingBP” variable into different ranges (bins) using the panda cut method and then found the respective Proportion_HeartDisease which shows the proportion of people having heart disease in each different blood pressure range. This will give us an understanding of what blood pressure range would have a relatively high heart disease occurrence. Besides, we fitted a linear regression model and performed an ANOVA test, trying to determine both the statistical significance between the “RestingBP” and the “HeartDisease” variable and how well the “RestingBP” could explain “HeartDisease”.





```
                  sum_sq     df          F    PR(>F)
HeartDisease    3638.418644    1.0  10.727228  0.001095
Residual      310685.250202  916.0        NaN       NaN
```



<u>Findings</u>

From the analysis, we can see a positive relationship between the heart disease incidence and the resting blood pressure using the regplot visualization, and there is a statistically significant difference in RestingBP between the two groups - with and without heart disease - using the ANOVA table. We also find that below the 140 blood pressure line, the proportion of people who had heart disease in the dataset is around 50%. However, for blood pressure range "141-160", the proportion of people who have heart diseases increases to 60%, and for people who have blood pressure greater than 160, their chance of having a heart disease is almost 70%. This indicates the blood pressure threshold from the data should be 160, and implies that having extremely high blood pressure could be associated with a higher chance of getting a heart disease. Comparing these values to the blood pressure recommendation on the heart.org website, the “141-160” range is counted as “High blood pressure (hypertension) stage 2”, and for blood pressure over 180, it would fit in the “Hypersensitive crises” range. These are all above the normal 120 blood pressure and associated with a high chance of heart disease. One thing noticeable is that the blood pressure range “121-140” examines a slight decrease in the chance of having heart disease compared to the bins that are below 120. This “121-140” blood pressure range is slightly above the normal blood pressure threshold online but didn’t show a higher chance of having heart disease. This shows a slight difference between the online normal blood pressure threshold and what we see as the threshold (140) in the data that separates a relatively high and low level of heart disease incidence.

Limitations

There is a limitation of the regression model. As one of the variable - “HeartDisease” - in the dataset is a categorical variable, all the data points are at the level of either 0 or 1 indicating whether that person has a heart disease. This leads to the result that the data did not fit the model perfectly, leaving a limitation to the model for prediction. This is also shown in the r-squared value (coefficient of determination). As the r-squared value is only 0.012, meaning that only 1.2% of the variability in HeartDisease can be explained by RestingBP. This indicates a very weak relationship between RestingBP and HeartDisease, suggesting that RestingBP alone is not a good predictor of HeartDisease in this dataset. Most of the variability in HeartDisease is likely explained by other factors not included in this model.

Future Work

From the data collection standpoint, it might lead to stronger results by having the same number of people in each range of blood pressure under the current proposed method, keeping the sample size in each blood pressure group the same. This is because when the number of people in each group of blood pressure levels is different, the level of randomness would occur more in the group that has a relatively small sample size. Thus, another round of data gathering with the same sample size in different blood pressure level groups could potentially improve the result under the current grouping method. Furthermore, as the HeartDisease is a categorical variable and the RestingBP is a numerical variable, it might be worthwhile to try logistic regression instead of linear regression. Logistic Regression, outputs probabilities between 0 and 1 using the sigmoid function, making it ideal for classification and is meaningful for interpreting relationships in classification tasks.