**Insurance**

This is a dataset about how age, BMI, and the number of children affect the insurance charge. Please use the age, BMI, and the number of children as independent variables and the insurance charge as the dependent variable and use a linear regression model to answer the following questions.

What impact will different factors have on insurance?

Does having children have a significant impact on insurance prices?

How will age affect insurance prices?

Which factor has the greatest impact on insurance prices?

**Life expectancy**

This dataset explores how factors such as the Year, Alcohol consumption, Percentage Expenditure, BMI, HIV/AIDS prevalence, Population, and GDP influence Life Expectancy.

Please use the Year, Alcohol consumption, Percentage Expenditure, BMI, HIV/AIDS prevalence, Population, and GDP as independent variables, with Life Expectancy as the dependent variable, and apply a linear regression model to answer the following questions:

Which factors significantly impact life expectancy?

Which factors are positively correlated with life expectancy?

What is the most crucial factor affecting life expectancy?

**Car selling price**

This dataset examines how the current price, the kilometers driven, and the car's production year impact its selling price.

Please employ the current price, kilometers driven, and car's production year as independent variables, with the selling price as the dependent variable, and utilize a linear regression model to address the following questions:

How strong is the relationship between the selling price and the aforementioned independent variables in this model?

What influence do different factors have on the selling price?

Which factor exerts the most significant impact on the selling price?

**cancer mortality rates**

This dataset focuses on how variables such as the average per capita (per 100,000) cancer diagnoses, median income per county, county population, percentage of the population living in poverty, and median age of county residents influence the average per capita (per 100,000) cancer mortality rates.

Please analyze the average per capita (per 100,000) cancer diagnoses, median income per county, county population, percentage of the population in poverty, and median age of county residents as independent variables, with the average per capita (per 100,000) cancer mortality rates as the dependent variable, using a linear regression model to explore the following questions:

How do various factors affect the cancer mortality rate?

Is there a significant impact of average income on the cancer mortality rate?

Which factor is the most influential in determining the cancer mortality rate?

**Flight**

This dataset explores how factors such as Flight Distance, Inflight Wifi Service, Passenger Age, Departure Delay in Minutes, and Arrival Delay in Minutes influence Inflight Service satisfaction.

Please analyze these factors—Flight Distance, Inflight Wifi Service, Age, Departure Delay, and Arrival Delay—as independent variables, with Inflight Service satisfaction as the dependent variable, employing a linear regression model to investigate the following questions:

What impact do different factors have on the satisfaction level of Inflight services?

How does the Departure Delay influence passenger satisfaction with Inflight services?

In what way does Arrival Delay impact satisfaction levels with Inflight services?

Which factor is most influential in determining satisfaction with Inflight services?

**Auto mpg**

This dataset examines the influence of variables such as cylinders, engine displacement, horsepower, vehicle weight, acceleration, and model year on the fuel efficiency (measured in miles per gallon, mpg).

Utilize these aspects—cylinders, displacement, horsepower, weight, acceleration, and model year—as independent variables, with mpg as the dependent variable. Apply a linear regression model to delve into the following queries:

How do different factors impact the vehicle's fuel efficiency (mpg)?

Is the model year significantly correlated with the vehicle's miles per gallon?

Among these factors, which has the most substantial influence on miles per gallon?

**Bicycles rent**

This dataset focuses on exploring how various factors such as holidays, weekdays, temperature, humidity, wind speed, and the season influence the total number of bicycles rented.

Please consider holidays, weekdays, temperature, humidity, wind speed, and season as independent variables, with the count of rented bicycles as the dependent variable. Utilize a linear regression model to address the following questions:

What effect do different factors have on the number of bicycles rented out?

How does the day of the week influence the bicycle rental numbers?

Which factor is the most significant in determining the number of bicycles rented?

**Concrete**

This dataset is designed to analyze how various components—Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, and the Age of the mixture (in days)—impact the Compressive Strength of Concrete.

Consider these elements: Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, and Age (days), as independent variables with Concrete Compressive Strength as the dependent variable. Employ a linear regression model to explore the following inquiries:

How do different components affect the compressive strength of concrete?

Does the age of the concrete significantly influence its compressive strength?

Among these factors, which one most significantly affects the concrete's compressive strength?

**Loan amount**

This dataset investigates the influence of various factors such as an individual's income, employment length, percentage of income, and length of credit history on the loan amount granted.

Consider the following as independent variables: the individual's income, the duration of their employment, their percentage of income, and the length of their credit history. Use these to predict the dependent variable, the loan amount, using a linear regression model. Address the following questions:

How do different factors affect the loan amount?

Does the length of employment play a significant role in determining the loan amount?

Among these variables, which one has the most substantial impact on the loan amount?

**Energy Heating Load**

This dataset is structured to analyze how eight variables—X1 (Relative Compactness), X2 (Surface Area), X3 (Wall Area), X4 (Roof Area), X5 (Overall Height), X6 (Orientation), X7 (Glazing Area), and X8 (Glazing Area Distribution)—affect Y1 (Energy Heating Load). This analysis is conducted using a linear regression model.

Please evaluate X1 to X8 as independent variables in relation to Y1, the dependent variable. Utilize a linear regression approach to investigate the following questions:

How strong is the correlation between the independent variables (X1 to X8) and the dependent variable (Y1)?

What influence do the various factors (X1 to X8) have on Y1 (Energy Heating Load)?

Among these factors, which one has the most significant impact on Y1?

**Energy Cooling Load**

This dataset investigates how variables X1 (Relative Compactness), X2 (Surface Area), X3 (Wall Area), X4 (Roof Area), X5 (Overall Height), X6 (Orientation), X7 (Glazing Area), and X8 (Glazing Area Distribution) influence Y2 (Energy Cooling Load) using a linear regression model.

Please analyze X1 through X8 as independent variables to assess their impact on Y2, the dependent variable. The linear regression model will help answer the following questions:

How robust is the relationship between the independent variables (X1 to X8) and the dependent variable (Y2)?

What effects do the different factors (X1 to X8) have on Y2 (Energy Cooling Load)?

Which among these factors exerts the most significant influence on Y2?

**Forest fire**

This dataset is focused on analyzing the effects of various factors on forest fire damage area. These factors include X (x-axis spatial coordinate within Montesinho park), Y (y-axis spatial coordinate), FFMC (Fine Fuel Moisture Code), DMC (Duff Moisture Code), DC (Drought Code), ISI (Initial Spread Index), temperature, relative humidity (RH), wind speed, and rainfall.

Please consider X, Y, FFMC, DMC, DC, ISI, temperature, RH, wind, and rain as independent variables and the forest fire damage area as the dependent variable. Apply a linear regression model to investigate these questions:

Is the correlation between the independent variables and the forest fire damage area strong?

How do each of these independent variables influence the extent of fire damage?

Which of these variables has the most significant impact on the area affected by forest fires?

**news**

This dataset investigates how various elements of online content influence its shareability. These elements include the number of words in the title (n\_tokens\_title), the number of words in the content (n\_tokens\_content), the proportion of unique words in the content (n\_unique\_tokens), the proportion of non-stop words in the content (n\_non\_stop\_unique\_tokens), the number of hyperlinks (num\_hrefs), the number of internal links (num\_self\_hrefs), the number of images (num\_imgs), the number of videos (num\_videos), the average word length in the content (average\_token\_length), and the number of keywords in the metadata (num\_keywords).

Please analyze these factors as independent variables in relation to the number of shares (dependent variable) using a linear regression model. The study focuses on the following questions:

How strong is the relationship between these content characteristics and the number of shares?

In what ways do these various factors influence the share count?

Among these factors, which one exerts the greatest influence on the number of shares?

**Car price**

This dataset is designed to explore how various car characteristics such as symboling, wheelbase, length, width, height, curb weight, engine size, bore ratio, stroke, compression ratio, horsepower, peak rpm, city mpg, and highway mpg influence the price of a car.

Utilize these variables - symboling, wheelbase, car length, width, height, curb weight, engine size, bore ratio, stroke, compression ratio, horsepower, peak rpm, city mpg, and highway mpg - as independent variables to analyze their effect on car price, the dependent variable. Employ a linear regression model to address these inquiries:

What impact do the different factors have on car prices?

Which of these factors significantly influence car pricing?

Amongst these variables, which one has the most substantial impact on the price of a car?

**house values**

This dataset provides an opportunity to explore how various factors such as geographical location (longitude and latitude), the median age of houses, the total number of rooms and bedrooms, the population within a block, the number of households, and the median household income within a block influence housing values in California.

In this analysis, consider longitude, latitude, median age of houses, total rooms and bedrooms, population, total households, and median household income as independent variables to assess their impact on house values (the dependent variable). Utilize a linear regression model to investigate the following questions:

How do different factors affect housing prices in California?

Is there a significant correlation between the number of bedrooms and housing prices?

Among these factors, which one has the most pronounced impact on the housing prices?

**cereal rating**

This dataset is centered around understanding how various nutritional and packaging factors like calories, protein, fat, sodium, fiber, complex carbohydrates, sugars, potassium, vitamins, display shelf level, serving weight in ounces, and serving size in cups influence cereal ratings.

Employ these factors - calories, protein, fat, sodium, fiber, complex carbohydrates, sugars, potassium, vitamins, shelf level, serving weight, and serving size - as independent variables to assess their effect on the dependent variable, which is the cereal rating. A linear regression model will be used to explore the following questions:

What is the influence of different nutritional and packaging attributes on cereal ratings?

Which of these components significantly affect the cereal ratings?

Among all these factors, which one has the most pronounced impact on the cereal ratings?

**abalone**

This dataset aims to analyze how various physical and weight attributes of an abalone—namely Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, and Shell weight—affect its age, as indicated by the number of rings.

Utilize these variables: Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, and Shell weight as independent variables to study their influence on the dependent variable, the number of rings (which correlates with the age of the abalone). Apply a linear regression model to address the following queries:

How do different physical and weight-related factors affect the age (as denoted by the number of rings) of an abalone?

Which of these factors have a significant impact on the age determination of abalone?

Among these variables, which one exerts the most substantial influence on the age of the abalone?

**electrical energy**

This dataset focuses on analyzing how various environmental variables, such as Temperature (AT), Exhaust Vacuum (V), Ambient Pressure (AP), and Relative Humidity (RH), influence the net hourly electrical energy output (PE).

Consider Temperature (AT), Exhaust Vacuum (V), Ambient Pressure (AP), and Relative Humidity (RH) as independent variables, and the net hourly electrical energy output (PE) as the dependent variable. Employ a linear regression model to investigate these specific questions:

How do different environmental factors (AT, V, AP, RH) impact the net hourly electrical energy output (PE)?

Which of these factors play a significant role in influencing the hourly electrical energy output?

Among these environmental variables, which one has the most substantial effect on the hourly electrical energy output (PE)?

**Real estate**

This is a dataset about how the transaction date, the house age, the distance to the nearest MRT station, and the number of convenience stores affect the house price of one unit area. Please use the transaction date, the house age, the distance to the nearest MRT station, and the number of convenience stores as independent variables and the house price of one unit area as the dependent variable and use a linear regression model to answer the following questions. Is the relationship between the dependent variable and the independent variable strong in this model?

Is there a risk of overfitting in the model?

Which factor has the greatest impact on real estate prices?

**atomic**

This dataset delves into how various atomic and field characteristics like mean atomic mass, field values, geometric means, entropy, ranges, and critical temperature influence the composition of superconductors in terms of their element count.

In this analysis, variables such as mean atomic mass, mean field, geometric means of atomic masses and field, entropy measures, ranges, standard deviations, and critical temperature will be treated as independent variables, with the focus on how they determine the number of elements in superconductors. A decision tree classifier model will be employed to answer the following questions:

What effect do different eigenvalues of these variables have on determining the number of elements in superconductors?

Which specific feature or variable has, on average, the most significant influence on determining the composition of superconductors?

Considering the accuracy of the model, can the impacts of these variables on the number of elements in superconductors be considered reliable?

**iris**

This dataset focuses on analyzing the characteristics of iris flowers, specifically Sepal Length, Sepal Width, Petal Length, and Petal Width, to classify the species of the iris.

In this analysis, a decision tree classifier model will be used, with Sepal Length, Sepal Width, Petal Length, and Petal Width serving as independent variables, and the iris species as the dependent variable. The model will help address the following questions:

What is the accuracy level of the decision tree classifier in correctly identifying the species of iris based on its sepal and petal measurements?

Does the decision tree model exhibit any signs of overfitting, implying that it might not generalize well to new, unseen data?

Which attribute among Sepal Length, Sepal Width, Petal Length, and Petal Width tends to have the greatest influence on the decision tree's classification of iris species?

**glass**

This dataset provides an analysis of how various chemical components like Refractive Index (RI), Sodium (Na), Magnesium (Mg), Aluminum (Al), Silicon (Si), Potassium (K), Calcium (Ca), Barium (Ba), and Iron (Fe) determine the classification of glass types.

In the proposed study, these chemical constituents will be treated as independent variables, and the glass type as the dependent variable. A decision tree classifier model will be employed to explore the following questions:

What level of accuracy does the decision tree classifier achieve in correctly classifying the glass type based on its chemical composition?

Is there a potential risk that the decision tree model might be overfitting the training data, leading to decreased generalization to new data?

Which of these chemical elements (RI, Na, Mg, Al, Si, K, Ca, Ba, or Fe) has, on average, the most significant influence on the decision tree's classification of glass types?

**redwine**

This dataset provides an in-depth look at how various chemical attributes of wine, including citric acid, chlorides, free sulfur dioxide, total sulfur dioxide, sulphates, and alcohol content, contribute to its overall quality.

In the analysis, these wine characteristics will be examined as independent variables, with the quality of the wine as the dependent variable. A decision tree classifier model will be used to explore the following key questions:

What is the level of accuracy achieved by the decision tree classifier in predicting the quality of wine based on these chemical properties?

Does the model show signs of being overtrained on the dataset, which might affect its performance on new, unseen data?

Among the factors of citric acid, chlorides, free and total sulfur dioxide, sulphates, and alcohol, which one, on average, plays the most crucial role in determining the quality of wine?

**student mat**

This dataset aims to explore the impact of various factors such as gender, age, travel time to school, study habits, past academic failures, availability of educational support (both school and family), family relationship quality, leisure time activities, alcohol consumption habits, health status, and class absences on students' final grades (G3).

To analyze these relationships, the following variables will be used as independent factors: sex, age, home to school travel time, study time, number of past class failures, extra educational support (schoolsup), family educational support (famsup), family relationships (famrel), free time after school (freetime), social outings (goout), workday and weekend alcohol consumption (Dalc and Walc), health status, and number of absences. The dependent variable in this analysis is the students' Grade (G3). A decision tree classifier model will be employed to address the following questions:

How reliable is the accuracy of the decision tree classifier in predicting students' grades based on these variables?

Is there evidence suggesting overfitting in the model, which could compromise its performance on new, unseen data?

Among these various conditions, which one exerts the most significant influence on the students' final grades?

**payment delays**

This dataset provides an analysis of how various financial factors, such as age, revolving utilization of unsecured lines, debt ratio, monthly income, number of open credit lines and loans, and the number of real estate loans or lines, influence the frequency of short-term payment delays (30-59 days past due).

In this study, these financial factors will be used as independent variables to assess their impact on the dependent variable, which is the frequency of being 30-59 days past due but not worse. A decision tree classifier model will be utilized to explore the following questions:

How do different borrower characteristics and financial behaviors impact the frequency of being overdue by 30-59 days?

Which specific piece of borrower information is most influential in predicting short-term overdue occurrences?

How reliable is the accuracy of the classifier in predicting these short-term payment delays based on the given financial factors?

**hotel bookings**

This is a dataset about how the lead time, the arrival date week, the arrival date day of the month, the adults number, the children number, the babies number, if it is repeated guest, the agent, the average daily rate(adr), if required car parking spaces, and the total of special requests affect if the hotel booking is cancelled.

Please use the lead time, the arrival date week, the arrival date day of the month, the adults number, the children number, the babies number, if it is repeated guest, the agent, the average daily rate(adr), if required car parking spaces, and the total of special requests as independent variables and the cancelled booking as the dependent variable and use a Logistic Regression model to answer the following questions.

How accurate is the model?

What impact will different factors have on whether a reservation is cancelled?

Which factor has the greatest impact?

**Income**

This is a dataset about how the age, the educational-num (educational level), the fnlwgt(weight assigned by the Census Bureau), the capital-gain, the capital-loss, and the hours-per-week affect the income. Please use the age, the educational-num (educational level), the fnlwgt(weight assigned by the Census Bureau), the capital-gain, the capital-loss, and the hours-per-week as independent variables and the income as the dependent variable and use logistic regression classifier to answer the following questions.

How accurate is it?

What are the effects of different factors on whether the income is greater than 50000?

What is the most important factor for whether the income is greater than 50000?

**diabetes**

This is a dataset about how pregnancy, glucose level, blood pressure, the triceps skin fold thickness, insulin level, pedigree, BMI, and age affect the diagnosis of diabetes. Please use the pregnancy, glucose level, blood pressure, the triceps skin fold thickness, insulin level, pedigree, BMI, and age as independent variables and the diabetes as the dependent variable and use a Logistic Regression model to answer the following questions.

Is the accuracy of the classifier trustworthy?

What impact will different factors have on the classification of diabetes?

What is the most influential factor for diabetes?

**Titanic**

This dataset provides insights into the Titanic tragedy by analyzing factors such as Ticket class, Age, the number of spouses aboard, the number of children aboard, and the Fare, and how they influenced passengers' survival.

For this analysis, these factors - Ticket class, Age, number of spouses, number of children, and Fare - will be used as independent variables to examine their impact on the dependent variable, 'Survived', using a logistic regression classifier. The study will aim to answer the following questions:

How did different factors like ticket class, age, family aboard, and fare affect the likelihood of survival for passengers on the Titanic?

Among these factors, which one had the most significant influence on whether a passenger survived?

How reliable and accurate is the model in predicting survival outcomes based on these variables?

**client**

This dataset investigates the influence of various client characteristics - age, account balance, marital status, education level, and previous engagement - on their likelihood of subscribing to a term deposit (denoted as 'y').

To understand these relationships, the following variables will be analyzed using a logistic regression classifier: client age, balance, marital status, education level, and past interactions. The model will focus on predicting the dependent variable, 'y', which represents whether the client subscribed to a term deposit. The study aims to answer these key questions:

Can the accuracy of the model in predicting term deposit subscriptions be considered reliable?

How do individual factors like age, balance, marital status, education, and previous interactions affect the decision to purchase a term deposit?

Among these variables, which one plays the most critical role in a client's decision to subscribe to a term deposit?

**candy**

This dataset is centered on analyzing what makes a Halloween candy popular, particularly looking at whether it contains chocolate. Key variables include the candy's sugar percentile, unit price percentile, and overall win percentage in popularity.

The study will employ logistic regression classifier to determine the relationship between these independent variables (sugar percentile, unit price percentile, and win percentage) and the dependent variable (presence of chocolate in the candy). The analysis aims to answer the following questions:

Can the accuracy of the classifier in predicting the presence of chocolate in the candy be considered reliable?

How do factors like sugar content, price, and popularity impact the likelihood of a candy containing chocolate?

Among these factors, which one most strongly influences whether a candy contains chocolate?

**churn**

This dataset aims to explore factors that may influence customer churn, specifically focusing on whether the customer is a senior citizen, their tenure with the company, and their monthly charges.

In this analysis, these variables - senior citizen status, tenure, and monthly charges - will be used as independent variables, and the dependent variable will be whether the customer left the company in the last month. A logistic regression classifier will be employed to investigate the following questions:

How do factors like being a senior citizen, tenure, and monthly charges impact the likelihood of a customer leaving?

Among these factors, which one is the most influential in determining customer churn?

Can the accuracy of the model in predicting customer churn be considered reliable?

**credit card**

This dataset provides a comprehensive look into various transaction-related variables—V1 to V28—alongside the transaction amount to determine their influence on the likelihood of credit card fraud.

In this study, the variables V1 to V28 will be used as independent variables to assess their impact on the dependent variable, which is the occurrence of fraud. A logistic regression classifier model will be employed to explore the following key questions:

How accurate is the model in identifying and predicting instances of credit card fraud?

What is the influence of each factor (V1 to V28) on the likelihood of a transaction being fraudulent?

Among these variables, which one plays the most crucial role in the occurrence of fraud?

**fraud**

This dataset examines how the following borrower characteristics influence serious delinquency within two years: revolving utilization of unsecured lines, age, number of times 30–59 days past due, debt ratio, number of open credit lines and loans, number of times 90 days late, number of real estate loans or lines, and number of times 60–89 days past due. Please treat these as the independent variables and serious delinquency within two years as the dependent variable, and use a logistic regression classifier to answer.

How reliable is the model’s predictive performance for SeriousDlqin2yrs?

Which single variable has the largest impact on SeriousDlqin2yrs?

When RevolvingUtilizationOfUnsecuredLines increases, does the predicted probability of SeriousDlqin2yrs increase or decrease?

**Bank**

This dataset examines how the following customer characteristics influence whether the client subscribes to a term deposit: client age, account balance, duration of the last contact (in seconds), number of contacts during the current campaign, days since the previous contact from an earlier campaign, and number of contacts made in previous campaigns.

Please treat these as the independent variables and term-deposit subscription within the campaign as the dependent variable (column y), and use a logistic regression classifier to answer:

How accurate is the model in predicting?

Which single variable has the greatest impact on the prediction?

When duration of the last contact increases, does the predicted probability of term-deposit subscription increase or decrease?

**Insurance**

This is a dataset about how age, BMI, and the number of children affect the insurance charge. Please use the age, BMI, and the number of children as independent variables and the insurance charge as the dependent variable and use a random forest regressor to answer the following questions.

What impact will different factors have on insurance?

Does having children have a significant impact on insurance prices?

How will age affect insurance prices?

Which factor has the greatest impact on insurance prices?

**Life expectancy**

This dataset explores how factors such as the Year, Alcohol consumption, Percentage Expenditure, BMI, HIV/AIDS prevalence, Population, and GDP influence Life Expectancy.

Please use the Year, Alcohol consumption, Percentage Expenditure, BMI, HIV/AIDS prevalence, Population, and GDP as independent variables, with Life Expectancy as the dependent variable, and apply a random forest regressor to answer the following questions:

Which factors significantly impact life expectancy?

Which factors are positively correlated with life expectancy?

What is the most crucial factor affecting life expectancy?

**Car selling price**

This dataset examines how the current price, the kilometers driven, and the car's production year impact its selling price.

Please employ the current price, kilometers driven, and car's production year as independent variables, with the selling price as the dependent variable, and utilize a random forest regressor to address the following questions:

How strong is the relationship between the selling price and the aforementioned independent variables in this model?

What influence do different factors have on the selling price?

Which factor exerts the most significant impact on the selling price?

**cancer mortality rates**

This dataset focuses on how variables such as the average per capita (per 100,000) cancer diagnoses, median income per county, county population, percentage of the population living in poverty, and median age of county residents influence the average per capita (per 100,000) cancer mortality rates.

Please analyze the average per capita (per 100,000) cancer diagnoses, median income per county, county population, percentage of the population in poverty, and median age of county residents as independent variables, with the average per capita (per 100,000) cancer mortality rates as the dependent variable, using a random forest regressor to explore the following questions:

How do various factors affect the cancer mortality rate?

Is there a significant impact of average income on the cancer mortality rate?

Which factor is the most influential in determining the cancer mortality rate?

**Flight**

This dataset explores how factors such as Flight Distance, Inflight Wifi Service, Passenger Age, Departure Delay in Minutes, and Arrival Delay in Minutes influence Inflight Service satisfaction.

Please analyze these factors—Flight Distance, Inflight Wifi Service, Age, Departure Delay, and Arrival Delay—as independent variables, with Inflight Service satisfaction as the dependent variable, employing a random forest regressor to investigate the following questions:

What impact do different factors have on the satisfaction level of Inflight services?

How does the Departure Delay influence passenger satisfaction with Inflight services?

In what way does Arrival Delay impact satisfaction levels with Inflight services?

Which factor is most influential in determining satisfaction with Inflight services?

**Auto mpg**

This dataset examines the influence of variables such as cylinders, engine displacement, horsepower, vehicle weight, acceleration, and model year on the fuel efficiency (measured in miles per gallon, mpg).

Utilize these aspects—cylinders, displacement, horsepower, weight, acceleration, and model year—as independent variables, with mpg as the dependent variable. Apply a random forest regressor to delve into the following queries:

How do different factors impact the vehicle's fuel efficiency (mpg)?

Is the model year significantly correlated with the vehicle's miles per gallon?

Among these factors, which has the most substantial influence on miles per gallon?

**Bicycles rent**

This dataset focuses on exploring how various factors such as holidays, weekdays, temperature, humidity, wind speed, and the season influence the total number of bicycles rented.

Please consider holidays, weekdays, temperature, humidity, wind speed, and season as independent variables, with the count of rented bicycles as the dependent variable. Utilize a random forest regressor to address the following questions:

What effect do different factors have on the number of bicycles rented out?

How does the day of the week influence the bicycle rental numbers?

Which factor is the most significant in determining the number of bicycles rented?

**Concrete**

This dataset is designed to analyze how various components—Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, and the Age of the mixture (in days)—impact the Compressive Strength of Concrete.

Consider these elements: Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, and Age (days), as independent variables with Concrete Compressive Strength as the dependent variable. Employ a random forest regressor to explore the following inquiries:

How do different components affect the compressive strength of concrete?

Does the age of the concrete significantly influence its compressive strength?

Among these factors, which one most significantly affects the concrete's compressive strength?

**Loan amount**

This dataset investigates the influence of various factors such as an individual's income, employment length, percentage of income, and length of credit history on the loan amount granted.

Consider the following as independent variables: the individual's income, the duration of their employment, their percentage of income, and the length of their credit history. Use these to predict the dependent variable, the loan amount, using a random forest regressor. Address the following questions:

How do different factors affect the loan amount?

Does the length of employment play a significant role in determining the loan amount?

Among these variables, which one has the most substantial impact on the loan amount?

**Energy Heating Load**

This dataset is structured to analyze how eight variables—X1 (Relative Compactness), X2 (Surface Area), X3 (Wall Area), X4 (Roof Area), X5 (Overall Height), X6 (Orientation), X7 (Glazing Area), and X8 (Glazing Area Distribution)—affect Y1 (Energy Heating Load). This analysis is conducted using a random forest regressor.

Please evaluate X1 to X8 as independent variables in relation to Y1, the dependent variable. Utilize a linear regression approach to investigate the following questions:

How strong is the correlation between the independent variables (X1 to X8) and the dependent variable (Y1)?

What influence do the various factors (X1 to X8) have on Y1 (Energy Heating Load)?

Among these factors, which one has the most significant impact on Y1?

**Energy Cooling Load**

This dataset investigates how variables X1 (Relative Compactness), X2 (Surface Area), X3 (Wall Area), X4 (Roof Area), X5 (Overall Height), X6 (Orientation), X7 (Glazing Area), and X8 (Glazing Area Distribution) influence Y2 (Energy Cooling Load) using a random forest regressor.

Please analyze X1 through X8 as independent variables to assess their impact on Y2, the dependent variable. The random forest regressor will help answer the following questions:

How robust is the relationship between the independent variables (X1 to X8) and the dependent variable (Y2)?

What effects do the different factors (X1 to X8) have on Y2 (Energy Cooling Load)?

Which among these factors exerts the most significant influence on Y2?

**Forest fire**

This dataset is focused on analyzing the effects of various factors on forest fire damage area. These factors include X (x-axis spatial coordinate within Montesinho park), Y (y-axis spatial coordinate), FFMC (Fine Fuel Moisture Code), DMC (Duff Moisture Code), DC (Drought Code), ISI (Initial Spread Index), temperature, relative humidity (RH), wind speed, and rainfall.

Please consider X, Y, FFMC, DMC, DC, ISI, temperature, RH, wind, and rain as independent variables and the forest fire damage area as the dependent variable. Apply a random forest regressor to investigate these questions:

Is the correlation between the independent variables and the forest fire damage area strong?

How do each of these independent variables influence the extent of fire damage?

Which of these variables has the most significant impact on the area affected by forest fires?

**news**

This dataset investigates how various elements of online content influence its shareability. These elements include the number of words in the title (n\_tokens\_title), the number of words in the content (n\_tokens\_content), the proportion of unique words in the content (n\_unique\_tokens), the proportion of non-stop words in the content (n\_non\_stop\_unique\_tokens), the number of hyperlinks (num\_hrefs), the number of internal links (num\_self\_hrefs), the number of images (num\_imgs), the number of videos (num\_videos), the average word length in the content (average\_token\_length), and the number of keywords in the metadata (num\_keywords).

Please analyze these factors as independent variables in relation to the number of shares (dependent variable) using a random forest regressor. The study focuses on the following questions:

How strong is the relationship between these content characteristics and the number of shares?

In what ways do these various factors influence the share count?

Among these factors, which one exerts the greatest influence on the number of shares?

**Car price**

This dataset is designed to explore how various car characteristics such as symboling, wheelbase, length, width, height, curb weight, engine size, bore ratio, stroke, compression ratio, horsepower, peak rpm, city mpg, and highway mpg influence the price of a car.

Utilize these variables - symboling, wheelbase, car length, width, height, curb weight, engine size, bore ratio, stroke, compression ratio, horsepower, peak rpm, city mpg, and highway mpg - as independent variables to analyze their effect on car price, the dependent variable. Employ a random forest regressor to address these inquiries:

What impact do the different factors have on car prices?

Which of these factors significantly influence car pricing?

Amongst these variables, which one has the most substantial impact on the price of a car?

**house values**

This dataset provides an opportunity to explore how various factors such as geographical location (longitude and latitude), the median age of houses, the total number of rooms and bedrooms, the population within a block, the number of households, and the median household income within a block influence housing values in California.

In this analysis, consider longitude, latitude, median age of houses, total rooms and bedrooms, population, total households, and median household income as independent variables to assess their impact on house values (the dependent variable). Utilize a random forest regressor to investigate the following questions:

How do different factors affect housing prices in California?

Is there a significant correlation between the number of bedrooms and housing prices?

Among these factors, which one has the most pronounced impact on the housing prices?

**cereal rating**

This dataset is centered around understanding how various nutritional and packaging factors like calories, protein, fat, sodium, fiber, complex carbohydrates, sugars, potassium, vitamins, display shelf level, serving weight in ounces, and serving size in cups influence cereal ratings.

Employ these factors - calories, protein, fat, sodium, fiber, complex carbohydrates, sugars, potassium, vitamins, shelf level, serving weight, and serving size - as independent variables to assess their effect on the dependent variable, which is the cereal rating. A random forest regressor will be used to explore the following questions:

What is the influence of different nutritional and packaging attributes on cereal ratings?

Which of these components significantly affect the cereal ratings?

Among all these factors, which one has the most pronounced impact on the cereal ratings?

**abalone**

This dataset aims to analyze how various physical and weight attributes of an abalone—namely Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, and Shell weight—affect its age, as indicated by the number of rings.

Utilize these variables: Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, and Shell weight as independent variables to study their influence on the dependent variable, the number of rings (which correlates with the age of the abalone). Apply a random forest regressor to address the following queries:

How do different physical and weight-related factors affect the age (as denoted by the number of rings) of an abalone?

Which of these factors have a significant impact on the age determination of abalone?

Among these variables, which one exerts the most substantial influence on the age of the abalone?

**electrical energy**

This dataset focuses on analyzing how various environmental variables, such as Temperature (AT), Exhaust Vacuum (V), Ambient Pressure (AP), and Relative Humidity (RH), influence the net hourly electrical energy output (PE).

Consider Temperature (AT), Exhaust Vacuum (V), Ambient Pressure (AP), and Relative Humidity (RH) as independent variables, and the net hourly electrical energy output (PE) as the dependent variable. Employ a random forest regressor to investigate these specific questions:

How do different environmental factors (AT, V, AP, RH) impact the net hourly electrical energy output (PE)?

Which of these factors play a significant role in influencing the hourly electrical energy output?

Among these environmental variables, which one has the most substantial effect on the hourly electrical energy output (PE)?

**Real estate**

This is a dataset about how the transaction date, the house age, the distance to the nearest MRT station, and the number of convenience stores affect the house price of one unit area. Please use the transaction date, the house age, the distance to the nearest MRT station, and the number of convenience stores as independent variables and the house price of one unit area as the dependent variable and use a random forest regressor to answer the following questions. Is the relationship between the dependent variable and the independent variable strong in this model?

Is there a risk of overfitting in the model?

Which factor has the greatest impact on real estate prices?

**hotel bookings**

This is a dataset about how the lead time, the arrival date week, the arrival date day of the month, the adults number, the children number, the babies number, if it is repeated guest, the agent, the average daily rate(adr), if required car parking spaces, and the total of special requests affect if the hotel booking is cancelled.

Please use the lead time, the arrival date week, the arrival date day of the month, the adults number, the children number, the babies number, if it is repeated guest, the agent, the average daily rate(adr), if required car parking spaces, and the total of special requests as independent variables and the cancelled booking as the dependent variable and use a Linear Discriminant Analysis to answer the following questions.

How accurate is the model?

What impact will different factors have on whether a reservation is cancelled?

Which factor has the greatest impact?

**Income**

This is a dataset about how the age, the educational-num (educational level), the fnlwgt(weight assigned by the Census Bureau), the capital-gain, the capital-loss, and the hours-per-week affect the income. Please use the age, the educational-num (educational level), the fnlwgt(weight assigned by the Census Bureau), the capital-gain, the capital-loss, and the hours-per-week as independent variables and the income as the dependent variable and use Linear Discriminant Analysis (LDA) model to answer the following questions.

How accurate is it?

What are the effects of different factors on whether the income is greater than 50000?

What is the most important factor for whether the income is greater than 50000?

**diabetes**

This is a dataset about how pregnancy, glucose level, blood pressure, the triceps skin fold thickness, insulin level, pedigree, BMI, and age affect the diagnosis of diabetes. Please use the pregnancy, glucose level, blood pressure, the triceps skin fold thickness, insulin level, pedigree, BMI, and age as independent variables and the diabetes as the dependent variable and use a Linear Discriminant Analysis to answer the following questions.

Is the accuracy of the classifier trustworthy?

What impact will different factors have on the classification of diabetes?

What is the most influential factor for diabetes?

**Titanic**

This dataset provides insights into the Titanic tragedy by analyzing factors such as Ticket class, Age, the number of spouses aboard, the number of children aboard, and the Fare, and how they influenced passengers' survival.

For this analysis, these factors - Ticket class, Age, number of spouses, number of children, and Fare - will be used as independent variables to examine their impact on the dependent variable, 'Survived', using a Linear Discriminant Analysis (LDA) model. The study will aim to answer the following questions:

How did different factors like ticket class, age, family aboard, and fare affect the likelihood of survival for passengers on the Titanic?

Among these factors, which one had the most significant influence on whether a passenger survived?

How reliable and accurate is the model in predicting survival outcomes based on these variables?

**client**

This dataset investigates the influence of various client characteristics - age, account balance, marital status, education level, and previous engagement - on their likelihood of subscribing to a term deposit (denoted as 'y').

To understand these relationships, the following variables will be analyzed using a Linear Discriminant Analysis (LDA) model: client age, balance, marital status, education level, and past interactions. The model will focus on predicting the dependent variable, 'y', which represents whether the client subscribed to a term deposit. The study aims to answer these key questions:

Can the accuracy of the model in predicting term deposit subscriptions be considered reliable?

How do individual factors like age, balance, marital status, education, and previous interactions affect the decision to purchase a term deposit?

Among these variables, which one plays the most critical role in a client's decision to subscribe to a term deposit?

**candy**

This dataset is centered on analyzing what makes a Halloween candy popular, particularly looking at whether it contains chocolate. Key variables include the candy's sugar percentile, unit price percentile, and overall win percentage in popularity.

The study will employ Linear Discriminant Analysis (LDA) model to determine the relationship between these independent variables (sugar percentile, unit price percentile, and win percentage) and the dependent variable (presence of chocolate in the candy). The analysis aims to answer the following questions:

Can the accuracy of the classifier in predicting the presence of chocolate in the candy be considered reliable?

How do factors like sugar content, price, and popularity impact the likelihood of a candy containing chocolate?

Among these factors, which one most strongly influences whether a candy contains chocolate?

**churn**

This dataset aims to explore factors that may influence customer churn, specifically focusing on whether the customer is a senior citizen, their tenure with the company, and their monthly charges.

In this analysis, these variables - senior citizen status, tenure, and monthly charges - will be used as independent variables, and the dependent variable will be whether the customer left the company in the last month. A Linear Discriminant Analysis (LDA) model will be employed to investigate the following questions:

How do factors like being a senior citizen, tenure, and monthly charges impact the likelihood of a customer leaving?

Among these factors, which one is the most influential in determining customer churn?

Can the accuracy of the model in predicting customer churn be considered reliable?

**credit card**

This dataset provides a comprehensive look into various transaction-related variables—V1 to V28—alongside the transaction amount to determine their influence on the likelihood of credit card fraud.

In this study, the variables V1 to V28 will be used as independent variables to assess their impact on the dependent variable, which is the occurrence of fraud. A Linear Discriminant Analysis (LDA) model model will be employed to explore the following key questions:

How accurate is the model in identifying and predicting instances of credit card fraud?

What is the influence of each factor (V1 to V28) on the likelihood of a transaction being fraudulent?

Among these variables, which one plays the most crucial role in the occurrence of fraud?

**fraud**

This dataset examines how the following borrower characteristics influence serious delinquency within two years: revolving utilization of unsecured lines, age, number of times 30–59 days past due, debt ratio, number of open credit lines and loans, number of times 90 days late, number of real estate loans or lines, and number of times 60–89 days past due. Please treat these as the independent variables and serious delinquency within two years as the dependent variable, and use a Linear Discriminant Analysis (LDA) model to answer.

How reliable is the model’s predictive performance for SeriousDlqin2yrs?

Which single variable has the largest impact on SeriousDlqin2yrs?

When RevolvingUtilizationOfUnsecuredLines increases, does the predicted probability of SeriousDlqin2yrs increase or decrease?

**Bank**

This dataset examines how the following customer characteristics influence whether the client subscribes to a term deposit: client age, account balance, duration of the last contact (in seconds), number of contacts during the current campaign, days since the previous contact from an earlier campaign, and number of contacts made in previous campaigns.

Please treat these as the independent variables and term-deposit subscription within the campaign as the dependent variable (column y), and use a Linear Discriminant Analysis (LDA) model to answer:

How accurate is the model in predicting?

Which single variable has the greatest impact on the prediction?

When duration of the last contact increases, does the predicted probability of term-deposit subscription increase or decrease?

**atomic**

This dataset delves into how various atomic and field characteristics like mean atomic mass, field values, geometric means, entropy, ranges, and critical temperature influence the composition of superconductors in terms of their element count.

In this analysis, variables such as mean atomic mass, mean field, geometric means of atomic masses and field, entropy measures, ranges, standard deviations, and critical temperature will be treated as independent variables, with the focus on how they determine the number of elements in superconductors. A Linear Discriminant Analysis (LDA) model will be employed to answer the following questions:

What effect do different eigenvalues of these variables have on determining the number of elements in superconductors?

Which specific feature or variable has, on average, the most significant influence on determining the composition of superconductors?

Considering the accuracy of the model, can the impacts of these variables on the number of elements in superconductors be considered reliable?

**iris**

This dataset focuses on analyzing the characteristics of iris flowers, specifically Sepal Length, Sepal Width, Petal Length, and Petal Width, to classify the species of the iris.

In this analysis, a Linear Discriminant Analysis (LDA) model will be used, with Sepal Length, Sepal Width, Petal Length, and Petal Width serving as independent variables, and the iris species as the dependent variable. The model will help address the following questions:

What is the accuracy level of the decision tree classifier in correctly identifying the species of iris based on its sepal and petal measurements?

Does the decision tree model exhibit any signs of overfitting, implying that it might not generalize well to new, unseen data?

Which attribute among Sepal Length, Sepal Width, Petal Length, and Petal Width tends to have the greatest influence on the decision tree's classification of iris species?

**glass**

This dataset provides an analysis of how various chemical components like Refractive Index (RI), Sodium (Na), Magnesium (Mg), Aluminum (Al), Silicon (Si), Potassium (K), Calcium (Ca), Barium (Ba), and Iron (Fe) determine the classification of glass types.

In the proposed study, these chemical constituents will be treated as independent variables, and the glass type as the dependent variable. A Linear Discriminant Analysis (LDA) model will be employed to explore the following questions:

What level of accuracy does the decision tree classifier achieve in correctly classifying the glass type based on its chemical composition?

Is there a potential risk that the decision tree model might be overfitting the training data, leading to decreased generalization to new data?

Which of these chemical elements (RI, Na, Mg, Al, Si, K, Ca, Ba, or Fe) has, on average, the most significant influence on the decision tree's classification of glass types?

**redwine**

This dataset provides an in-depth look at how various chemical attributes of wine, including citric acid, chlorides, free sulfur dioxide, total sulfur dioxide, sulphates, and alcohol content, contribute to its overall quality.

In the analysis, these wine characteristics will be examined as independent variables, with the quality of the wine as the dependent variable. A Linear Discriminant Analysis (LDA) model will be used to explore the following key questions:

What is the level of accuracy achieved by the decision tree classifier in predicting the quality of wine based on these chemical properties?

Does the model show signs of being overtrained on the dataset, which might affect its performance on new, unseen data?

Among the factors of citric acid, chlorides, free and total sulfur dioxide, sulphates, and alcohol, which one, on average, plays the most crucial role in determining the quality of wine?

**student mat**

This dataset aims to explore the impact of various factors such as gender, age, travel time to school, study habits, past academic failures, availability of educational support (both school and family), family relationship quality, leisure time activities, alcohol consumption habits, health status, and class absences on students' final grades (G3).

To analyze these relationships, the following variables will be used as independent factors: sex, age, home to school travel time, study time, number of past class failures, extra educational support (schoolsup), family educational support (famsup), family relationships (famrel), free time after school (freetime), social outings (goout), workday and weekend alcohol consumption (Dalc and Walc), health status, and number of absences. The dependent variable in this analysis is the students' Grade (G3). A Linear Discriminant Analysis (LDA) model will be employed to address the following questions:

How reliable is the accuracy of the decision tree classifier in predicting students' grades based on these variables?

Is there evidence suggesting overfitting in the model, which could compromise its performance on new, unseen data?

Among these various conditions, which one exerts the most significant influence on the students' final grades?

**payment delays**

This dataset provides an analysis of how various financial factors, such as age, revolving utilization of unsecured lines, debt ratio, monthly income, number of open credit lines and loans, and the number of real estate loans or lines, influence the frequency of short-term payment delays (30-59 days past due).

In this study, these financial factors will be used as independent variables to assess their impact on the dependent variable, which is the frequency of being 30-59 days past due but not worse. A Linear Discriminant Analysis (LDA) model will be utilized to explore the following questions:

How do different borrower characteristics and financial behaviors impact the frequency of being overdue by 30-59 days?

Which specific piece of borrower information is most influential in predicting short-term overdue occurrences?

How reliable is the accuracy of the classifier in predicting these short-term payment delays based on the given financial factors?

**Insurance**

This is a dataset about how age, BMI, and the number of children affect the insurance charge. Please use the age, BMI, and the number of children as independent variables and the insurance charge as the dependent variable and use a gradient boosting regressor model to answer the following questions.

What impact will different factors have on insurance?

Does having children have a significant impact on insurance prices?

How will age affect insurance prices?

Which factor has the greatest impact on insurance prices?

**Life expectancy**

This dataset explores how factors such as the Year, Alcohol consumption, Percentage Expenditure, BMI, HIV/AIDS prevalence, Population, and GDP influence Life Expectancy.

Please use the Year, Alcohol consumption, Percentage Expenditure, BMI, HIV/AIDS prevalence, Population, and GDP as independent variables, with Life Expectancy as the dependent variable, and apply a gradient boosting regressor model to answer the following questions:

Which factors significantly impact life expectancy?

Which factors are positively correlated with life expectancy?

What is the most crucial factor affecting life expectancy?

**Car selling price**

This dataset examines how the current price, the kilometers driven, and the car's production year impact its selling price.

Please employ the current price, kilometers driven, and car's production year as independent variables, with the selling price as the dependent variable, and utilize a gradient boosting regressor model to address the following questions:

How strong is the relationship between the selling price and the aforementioned independent variables in this model?

What influence do different factors have on the selling price?

Which factor exerts the most significant impact on the selling price?

**cancer mortality rates**

This dataset focuses on how variables such as the average per capita (per 100,000) cancer diagnoses, median income per county, county population, percentage of the population living in poverty, and median age of county residents influence the average per capita (per 100,000) cancer mortality rates.

Please analyze the average per capita (per 100,000) cancer diagnoses, median income per county, county population, percentage of the population in poverty, and median age of county residents as independent variables, with the average per capita (per 100,000) cancer mortality rates as the dependent variable, using a gradient boosting regressor model to explore the following questions:

How do various factors affect the cancer mortality rate?

Is there a significant impact of average income on the cancer mortality rate?

Which factor is the most influential in determining the cancer mortality rate?

**Flight**

This dataset explores how factors such as Flight Distance, Inflight Wifi Service, Passenger Age, Departure Delay in Minutes, and Arrival Delay in Minutes influence Inflight Service satisfaction.

Please analyze these factors—Flight Distance, Inflight Wifi Service, Age, Departure Delay, and Arrival Delay—as independent variables, with Inflight Service satisfaction as the dependent variable, employing a gradient boosting regressor model to investigate the following questions:

What impact do different factors have on the satisfaction level of Inflight services?

How does the Departure Delay influence passenger satisfaction with Inflight services?

In what way does Arrival Delay impact satisfaction levels with Inflight services?

Which factor is most influential in determining satisfaction with Inflight services?

**Auto mpg**

This dataset examines the influence of variables such as cylinders, engine displacement, horsepower, vehicle weight, acceleration, and model year on the fuel efficiency (measured in miles per gallon, mpg).

Utilize these aspects—cylinders, displacement, horsepower, weight, acceleration, and model year—as independent variables, with mpg as the dependent variable. Apply a gradient boosting regressor model to delve into the following queries:

How do different factors impact the vehicle's fuel efficiency (mpg)?

Is the model year significantly correlated with the vehicle's miles per gallon?

Among these factors, which has the most substantial influence on miles per gallon?

**Bicycles rent**

This dataset focuses on exploring how various factors such as holidays, weekdays, temperature, humidity, wind speed, and the season influence the total number of bicycles rented.

Please consider holidays, weekdays, temperature, humidity, wind speed, and season as independent variables, with the count of rented bicycles as the dependent variable. Utilize a gradient boosting regressor model to address the following questions:

What effect do different factors have on the number of bicycles rented out?

How does the day of the week influence the bicycle rental numbers?

Which factor is the most significant in determining the number of bicycles rented?

**Concrete**

This dataset is designed to analyze how various components—Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, and the Age of the mixture (in days)—impact the Compressive Strength of Concrete.

Consider these elements: Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, and Age (days), as independent variables with Concrete Compressive Strength as the dependent variable. Employ a gradient boosting regressor model to explore the following inquiries:

How do different components affect the compressive strength of concrete?

Does the age of the concrete significantly influence its compressive strength?

Among these factors, which one most significantly affects the concrete's compressive strength?

**Loan amount**

This dataset investigates the influence of various factors such as an individual's income, employment length, percentage of income, and length of credit history on the loan amount granted.

Consider the following as independent variables: the individual's income, the duration of their employment, their percentage of income, and the length of their credit history. Use these to predict the dependent variable, the loan amount, using a gradient boosting regressor model. Address the following questions:

How do different factors affect the loan amount?

Does the length of employment play a significant role in determining the loan amount?

Among these variables, which one has the most substantial impact on the loan amount?

**Energy Heating Load**

This dataset is structured to analyze how eight variables—X1 (Relative Compactness), X2 (Surface Area), X3 (Wall Area), X4 (Roof Area), X5 (Overall Height), X6 (Orientation), X7 (Glazing Area), and X8 (Glazing Area Distribution)—affect Y1 (Energy Heating Load). This analysis is conducted using a gradient boosting regressor model.

Please evaluate X1 to X8 as independent variables in relation to Y1, the dependent variable. Utilize a linear regression approach to investigate the following questions:

How strong is the correlation between the independent variables (X1 to X8) and the dependent variable (Y1)?

What influence do the various factors (X1 to X8) have on Y1 (Energy Heating Load)?

Among these factors, which one has the most significant impact on Y1?

**Energy Cooling Load**

This dataset investigates how variables X1 (Relative Compactness), X2 (Surface Area), X3 (Wall Area), X4 (Roof Area), X5 (Overall Height), X6 (Orientation), X7 (Glazing Area), and X8 (Glazing Area Distribution) influence Y2 (Energy Cooling Load) using a gradient boosting regressor model.

Please analyze X1 through X8 as independent variables to assess their impact on Y2, the dependent variable. The gradient boosting regressor model will help answer the following questions:

How robust is the relationship between the independent variables (X1 to X8) and the dependent variable (Y2)?

What effects do the different factors (X1 to X8) have on Y2 (Energy Cooling Load)?

Which among these factors exerts the most significant influence on Y2?

**Forest fire**

This dataset is focused on analyzing the effects of various factors on forest fire damage area. These factors include X (x-axis spatial coordinate within Montesinho park), Y (y-axis spatial coordinate), FFMC (Fine Fuel Moisture Code), DMC (Duff Moisture Code), DC (Drought Code), ISI (Initial Spread Index), temperature, relative humidity (RH), wind speed, and rainfall.

Please consider X, Y, FFMC, DMC, DC, ISI, temperature, RH, wind, and rain as independent variables and the forest fire damage area as the dependent variable. Apply a gradient boosting regressor model to investigate these questions:

Is the correlation between the independent variables and the forest fire damage area strong?

How do each of these independent variables influence the extent of fire damage?

Which of these variables has the most significant impact on the area affected by forest fires?

**news**

This dataset investigates how various elements of online content influence its shareability. These elements include the number of words in the title (n\_tokens\_title), the number of words in the content (n\_tokens\_content), the proportion of unique words in the content (n\_unique\_tokens), the proportion of non-stop words in the content (n\_non\_stop\_unique\_tokens), the number of hyperlinks (num\_hrefs), the number of internal links (num\_self\_hrefs), the number of images (num\_imgs), the number of videos (num\_videos), the average word length in the content (average\_token\_length), and the number of keywords in the metadata (num\_keywords).

Please analyze these factors as independent variables in relation to the number of shares (dependent variable) using a gradient boosting regressor model. The study focuses on the following questions:

How strong is the relationship between these content characteristics and the number of shares?

In what ways do these various factors influence the share count?

Among these factors, which one exerts the greatest influence on the number of shares?

**Car price**

This dataset is designed to explore how various car characteristics such as symboling, wheelbase, length, width, height, curb weight, engine size, bore ratio, stroke, compression ratio, horsepower, peak rpm, city mpg, and highway mpg influence the price of a car.

Utilize these variables - symboling, wheelbase, car length, width, height, curb weight, engine size, bore ratio, stroke, compression ratio, horsepower, peak rpm, city mpg, and highway mpg - as independent variables to analyze their effect on car price, the dependent variable. Employ a gradient boosting regressor model to address these inquiries:

What impact do the different factors have on car prices?

Which of these factors significantly influence car pricing?

Amongst these variables, which one has the most substantial impact on the price of a car?

**house values**

This dataset provides an opportunity to explore how various factors such as geographical location (longitude and latitude), the median age of houses, the total number of rooms and bedrooms, the population within a block, the number of households, and the median household income within a block influence housing values in California.

In this analysis, consider longitude, latitude, median age of houses, total rooms and bedrooms, population, total households, and median household income as independent variables to assess their impact on house values (the dependent variable). Utilize a gradient boosting regressor model to investigate the following questions:

How do different factors affect housing prices in California?

Is there a significant correlation between the number of bedrooms and housing prices?

Among these factors, which one has the most pronounced impact on the housing prices?

**cereal rating**

This dataset is centered around understanding how various nutritional and packaging factors like calories, protein, fat, sodium, fiber, complex carbohydrates, sugars, potassium, vitamins, display shelf level, serving weight in ounces, and serving size in cups influence cereal ratings.

Employ these factors - calories, protein, fat, sodium, fiber, complex carbohydrates, sugars, potassium, vitamins, shelf level, serving weight, and serving size - as independent variables to assess their effect on the dependent variable, which is the cereal rating. A gradient boosting regressor model will be used to explore the following questions:

What is the influence of different nutritional and packaging attributes on cereal ratings?

Which of these components significantly affect the cereal ratings?

Among all these factors, which one has the most pronounced impact on the cereal ratings?

**abalone**

This dataset aims to analyze how various physical and weight attributes of an abalone—namely Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, and Shell weight—affect its age, as indicated by the number of rings.

Utilize these variables: Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, and Shell weight as independent variables to study their influence on the dependent variable, the number of rings (which correlates with the age of the abalone). Apply a gradient boosting regressor model to address the following queries:

How do different physical and weight-related factors affect the age (as denoted by the number of rings) of an abalone?

Which of these factors have a significant impact on the age determination of abalone?

Among these variables, which one exerts the most substantial influence on the age of the abalone?

**electrical energy**

This dataset focuses on analyzing how various environmental variables, such as Temperature (AT), Exhaust Vacuum (V), Ambient Pressure (AP), and Relative Humidity (RH), influence the net hourly electrical energy output (PE).

Consider Temperature (AT), Exhaust Vacuum (V), Ambient Pressure (AP), and Relative Humidity (RH) as independent variables, and the net hourly electrical energy output (PE) as the dependent variable. Employ a gradient boosting regressor model to investigate these specific questions:

How do different environmental factors (AT, V, AP, RH) impact the net hourly electrical energy output (PE)?

Which of these factors play a significant role in influencing the hourly electrical energy output?

Among these environmental variables, which one has the most substantial effect on the hourly electrical energy output (PE)?

**Real estate**

This is a dataset about how the transaction date, the house age, the distance to the nearest MRT station, and the number of convenience stores affect the house price of one unit area. Please use the transaction date, the house age, the distance to the nearest MRT station, and the number of convenience stores as independent variables and the house price of one unit area as the dependent variable and use a gradient boosting regressor model to answer the following questions. Is the relationship between the dependent variable and the independent variable strong in this model?

Is there a risk of overfitting in the model?

Which factor has the greatest impact on real estate prices?

**atomic**

This dataset delves into how various atomic and field characteristics like mean atomic mass, field values, geometric means, entropy, ranges, and critical temperature influence the composition of superconductors in terms of their element count.

In this analysis, variables such as mean atomic mass, mean field, geometric means of atomic masses and field, entropy measures, ranges, standard deviations, and critical temperature will be treated as independent variables, with the focus on how they determine the number of elements in superconductors. A Random Forest regressor classifier will be employed to answer the following questions:

What effect do different eigenvalues of these variables have on determining the number of elements in superconductors?

Which specific feature or variable has, on average, the most significant influence on determining the composition of superconductors?

Considering the accuracy of the model, can the impacts of these variables on the number of elements in superconductors be considered reliable?

**iris**

This dataset focuses on analyzing the characteristics of iris flowers, specifically Sepal Length, Sepal Width, Petal Length, and Petal Width, to classify the species of the iris.

In this analysis, a Random Forest regressor classifier will be used, with Sepal Length, Sepal Width, Petal Length, and Petal Width serving as independent variables, and the iris species as the dependent variable. The model will help address the following questions:

What is the accuracy level of the decision tree classifier in correctly identifying the species of iris based on its sepal and petal measurements?

Does the decision tree model exhibit any signs of overfitting, implying that it might not generalize well to new, unseen data?

Which attribute among Sepal Length, Sepal Width, Petal Length, and Petal Width tends to have the greatest influence on the decision tree's classification of iris species?

**glass**

This dataset provides an analysis of how various chemical components like Refractive Index (RI), Sodium (Na), Magnesium (Mg), Aluminum (Al), Silicon (Si), Potassium (K), Calcium (Ca), Barium (Ba), and Iron (Fe) determine the classification of glass types.

In the proposed study, these chemical constituents will be treated as independent variables, and the glass type as the dependent variable. A Random Forest regressor classifier will be employed to explore the following questions:

What level of accuracy does the decision tree classifier achieve in correctly classifying the glass type based on its chemical composition?

Is there a potential risk that the decision tree model might be overfitting the training data, leading to decreased generalization to new data?

Which of these chemical elements (RI, Na, Mg, Al, Si, K, Ca, Ba, or Fe) has, on average, the most significant influence on the decision tree's classification of glass types?

**redwine**

This dataset provides an in-depth look at how various chemical attributes of wine, including citric acid, chlorides, free sulfur dioxide, total sulfur dioxide, sulphates, and alcohol content, contribute to its overall quality.

In the analysis, these wine characteristics will be examined as independent variables, with the quality of the wine as the dependent variable. A Random Forest regressor classifier will be used to explore the following key questions:

What is the level of accuracy achieved by the decision tree classifier in predicting the quality of wine based on these chemical properties?

Does the model show signs of being overtrained on the dataset, which might affect its performance on new, unseen data?

Among the factors of citric acid, chlorides, free and total sulfur dioxide, sulphates, and alcohol, which one, on average, plays the most crucial role in determining the quality of wine?

**student mat**

This dataset aims to explore the impact of various factors such as gender, age, travel time to school, study habits, past academic failures, availability of educational support (both school and family), family relationship quality, leisure time activities, alcohol consumption habits, health status, and class absences on students' final grades (G3).

To analyze these relationships, the following variables will be used as independent factors: sex, age, home to school travel time, study time, number of past class failures, extra educational support (schoolsup), family educational support (famsup), family relationships (famrel), free time after school (freetime), social outings (goout), workday and weekend alcohol consumption (Dalc and Walc), health status, and number of absences. The dependent variable in this analysis is the students' Grade (G3). A Random Forest regressor classifier will be employed to address the following questions:

How reliable is the accuracy of the decision tree classifier in predicting students' grades based on these variables?

Is there evidence suggesting overfitting in the model, which could compromise its performance on new, unseen data?

Among these various conditions, which one exerts the most significant influence on the students' final grades?

**payment delays**

This dataset provides an analysis of how various financial factors, such as age, revolving utilization of unsecured lines, debt ratio, monthly income, number of open credit lines and loans, and the number of real estate loans or lines, influence the frequency of short-term payment delays (30-59 days past due).

In this study, these financial factors will be used as independent variables to assess their impact on the dependent variable, which is the frequency of being 30-59 days past due but not worse. A Random Forest regressor classifier will be utilized to explore the following questions:

How do different borrower characteristics and financial behaviors impact the frequency of being overdue by 30-59 days?

Which specific piece of borrower information is most influential in predicting short-term overdue occurrences?

How reliable is the accuracy of the classifier in predicting these short-term payment delays based on the given financial factors?

**hotel bookings**

This is a dataset about how the lead time, the arrival date week, the arrival date day of the month, the adults number, the children number, the babies number, if it is repeated guest, the agent, the average daily rate(adr), if required car parking spaces, and the total of special requests affect if the hotel booking is cancelled.

Please use the lead time, the arrival date week, the arrival date day of the month, the adults number, the children number, the babies number, if it is repeated guest, the agent, the average daily rate(adr), if required car parking spaces, and the total of special requests as independent variables and the cancelled booking as the dependent variable and use a Ridge Classifier model to answer the following questions.

How accurate is the model?

What impact will different factors have on whether a reservation is cancelled?

Which factor has the greatest impact?

**Income**

This is a dataset about how the age, the educational-num (educational level), the fnlwgt(weight assigned by the Census Bureau), the capital-gain, the capital-loss, and the hours-per-week affect the income. Please use the age, the educational-num (educational level), the fnlwgt(weight assigned by the Census Bureau), the capital-gain, the capital-loss, and the hours-per-week as independent variables and the income as the dependent variable and use Ridge Classifier model to answer the following questions.

How accurate is it?

What are the effects of different factors on whether the income is greater than 50000?

What is the most important factor for whether the income is greater than 50000?

**diabetes**

This is a dataset about how pregnancy, glucose level, blood pressure, the triceps skin fold thickness, insulin level, pedigree, BMI, and age affect the diagnosis of diabetes. Please use the pregnancy, glucose level, blood pressure, the triceps skin fold thickness, insulin level, pedigree, BMI, and age as independent variables and the diabetes as the dependent variable and use a Ridge Classifier model to answer the following questions.

Is the accuracy of the classifier trustworthy?

What impact will different factors have on the classification of diabetes?

What is the most influential factor for diabetes?

**Titanic**

This dataset provides insights into the Titanic tragedy by analyzing factors such as Ticket class, Age, the number of spouses aboard, the number of children aboard, and the Fare, and how they influenced passengers' survival.

For this analysis, these factors - Ticket class, Age, number of spouses, number of children, and Fare - will be used as independent variables to examine their impact on the dependent variable, 'Survived', using a Ridge Classifier model. The study will aim to answer the following questions:

How did different factors like ticket class, age, family aboard, and fare affect the likelihood of survival for passengers on the Titanic?

Among these factors, which one had the most significant influence on whether a passenger survived?

How reliable and accurate is the model in predicting survival outcomes based on these variables?

**client**

This dataset investigates the influence of various client characteristics - age, account balance, marital status, education level, and previous engagement - on their likelihood of subscribing to a term deposit (denoted as 'y').

To understand these relationships, the following variables will be analyzed using a Ridge Classifier model: client age, balance, marital status, education level, and past interactions. The model will focus on predicting the dependent variable, 'y', which represents whether the client subscribed to a term deposit. The study aims to answer these key questions:

Can the accuracy of the model in predicting term deposit subscriptions be considered reliable?

How do individual factors like age, balance, marital status, education, and previous interactions affect the decision to purchase a term deposit?

Among these variables, which one plays the most critical role in a client's decision to subscribe to a term deposit?

**candy**

This dataset is centered on analyzing what makes a Halloween candy popular, particularly looking at whether it contains chocolate. Key variables include the candy's sugar percentile, unit price percentile, and overall win percentage in popularity.

The study will employ Ridge Classifier model to determine the relationship between these independent variables (sugar percentile, unit price percentile, and win percentage) and the dependent variable (presence of chocolate in the candy). The analysis aims to answer the following questions:

Can the accuracy of the classifier in predicting the presence of chocolate in the candy be considered reliable?

How do factors like sugar content, price, and popularity impact the likelihood of a candy containing chocolate?

Among these factors, which one most strongly influences whether a candy contains chocolate?

**churn**

This dataset aims to explore factors that may influence customer churn, specifically focusing on whether the customer is a senior citizen, their tenure with the company, and their monthly charges.

In this analysis, these variables - senior citizen status, tenure, and monthly charges - will be used as independent variables, and the dependent variable will be whether the customer left the company in the last month. A Ridge Classifier model will be employed to investigate the following questions:

How do factors like being a senior citizen, tenure, and monthly charges impact the likelihood of a customer leaving?

Among these factors, which one is the most influential in determining customer churn?

Can the accuracy of the model in predicting customer churn be considered reliable?

**credit card**

This dataset provides a comprehensive look into various transaction-related variables—V1 to V28—alongside the transaction amount to determine their influence on the likelihood of credit card fraud.

In this study, the variables V1 to V28 will be used as independent variables to assess their impact on the dependent variable, which is the occurrence of fraud. A Ridge Classifier model model will be employed to explore the following key questions:

How accurate is the model in identifying and predicting instances of credit card fraud?

What is the influence of each factor (V1 to V28) on the likelihood of a transaction being fraudulent?

Among these variables, which one plays the most crucial role in the occurrence of fraud?

**fraud**

This dataset examines how the following borrower characteristics influence serious delinquency within two years: revolving utilization of unsecured lines, age, number of times 30–59 days past due, debt ratio, number of open credit lines and loans, number of times 90 days late, number of real estate loans or lines, and number of times 60–89 days past due. Please treat these as the independent variables and serious delinquency within two years as the dependent variable, and use a Ridge Classifier model to answer.

How reliable is the model’s predictive performance for SeriousDlqin2yrs?

Which single variable has the largest impact on SeriousDlqin2yrs?

When RevolvingUtilizationOfUnsecuredLines increases, does the predicted probability of SeriousDlqin2yrs increase or decrease?

**Bank**

This dataset examines how the following customer characteristics influence whether the client subscribes to a term deposit: client age, account balance, duration of the last contact (in seconds), number of contacts during the current campaign, days since the previous contact from an earlier campaign, and number of contacts made in previous campaigns.

Please treat these as the independent variables and term-deposit subscription within the campaign as the dependent variable (column y), and use a Ridge Classifier model to answer:

How accurate is the model in predicting?

Which single variable has the greatest impact on the prediction?

When duration of the last contact increases, does the predicted probability of term-deposit subscription increase or decrease?