# ASSIGNMENT 5 DIAML

## Report

Q1. Describe at least four steps to implementing a rule-based approach to decision-making and give an example. Is any domain knowledge required to establish a rule? Support your answer with an explanation

Introduction

A rule-based approach in artificial intelligence and expert systemthat is methodology used leverages a predetermined set of rules to guide decision-making and problem-solving processes. These rules are crafted based on the extensive knowledge of human experts within a specific domain. This approach is particularly adept at analysing and processing textual data by systematically applying predefined rules and patterns to identify and capture specific structures, ultimately leading to the extraction of valuable information.[1], where it is used in healthcare to provide medical diagnosis and treatment recommendations, in manufacturing to optimize the supply chain management, and can be applied also in finance for financial analysis.

Steps to implement Rule-based approach to decision making.

**Decision table**: A decision table is a structured and systematic way to represent complex decision-making processes, particularly in the context of business rules, logic, and decision support systems. It is a tabular representation that helps to simplify, analyze, and document decision rules and their outcomes

**Rule creation**: Develop domain-specific linguistic rules for the defined tasks. These rules encompass grammar rules, syntax patterns, semantic rules, and regular expressions. They guide the system in understanding and generating language correctly

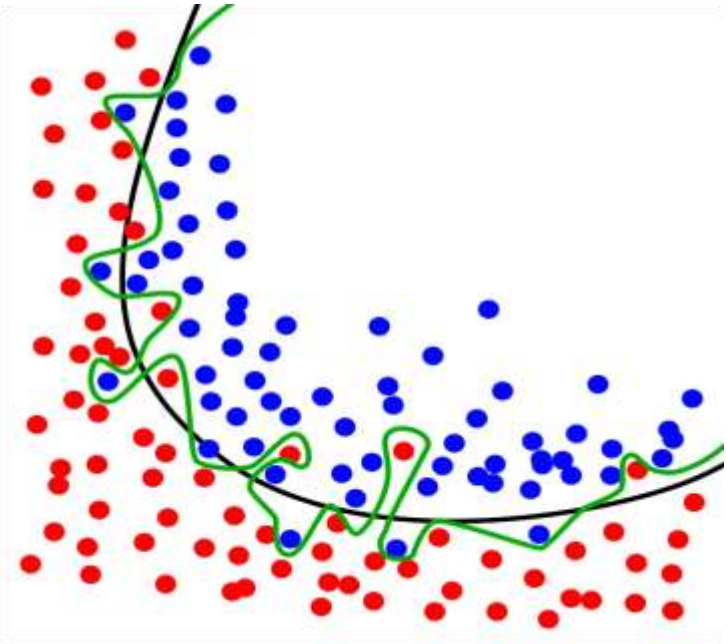
**Rule Processing**:is a fundamental component of rule-based systems, especially in the context of natural language processing and other rule-based applications. It involves the application of predefined rules to input data to extract, transform, or generate meaningful information.

**Rule refinement**: process that involves continuously improving and adjusting the rules based on input and real-world data. This iterative step is crucial for enhancing the system's effectiveness and making it adaptable to changing language patterns.[2]

1.2 Over-fitting and why is a problem in statistical learning.

By definition overfitting is undesirable machine learning behaviour that occur when the model gives accurate prediction for training data but not for a new data. Means to use this machine learning model for prediction first need to train model to data it will refer too whike making production. An overfit model can give in accurate prediction and cannot perform well for all new data[3]. By simplifying overfitting occur in data modelling as result of particular function aligning too closely to minimum set of data point. According to the lecture slide it is like memorizing answer to math problem instead of understanding the formula to solve the problem.

It is problem in statistical learning when you want to train to new dataset it can give poor performance and raise noise to the test set.



By Chabacano — Own work, CC BY-SA 4.0 Overfitting image source link

There are different sign that the model is overfitting which include low error rates for training set compared to test set, high standards of error, and having different parameter for subsets of the data.

Considering small dataset containing ten data point should you prefer a simple model with one parameter or a complex model with ten parameter. For this question I will prefer to simple model with one parameter because it is the one that can get me the best fit and the reason why I did not choose complex model its because complex model is one of the two reason that cause the model to go overfitting. To prevent getting error or integrating noise to the test set and generate the desired best fit I will prefer to have clean output.

The cause of overfitting is training on too many variables or too complex models are two of the reason model undergo overfitting.[4]

### 1.3 Approach used to avoid overfitting

Data augmentation is used to improve the performance of machine learning models by reducing overfitting through changing the sample data slightly every time the model processes it. It can be done by changing input data in small way, this will make the test data more unique to the model to prevent overfitting.

Feature selection is the process of identifying the most important ones within the training data and then eliminating irrelevant or redundant ones. By employing either forward or backward selection methods can helps in selecting the most relevant features for our model to prevent us from using features that are unrelated to our target variable.[5]

1.4 Metrics used to evaluate the performance of a model with example on each

R-Squared ($R^2$ or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit)

For example, an r-squared of 0.7 reveals that 0.7 of the variability observed in the target variable is explained by the regression model. Generally, a higher r-squared indicates more variability where the higher r-squared indicate the better fit of the model.

1.5 Benchmark in machine learning is a type of model used to compare performance of other models. There are different types of benchmarks. Sometimes, it is a so-called state-of-the-art model, i.e., the best one on a given dataset for a given problem. The goal of benchmarking is then to see if we can create a better model and beat published results.[6]

Useful of benchmark

- **Performance benchmarking** is the process of gathering data to assess your performance in achieving various outcomes, which could range from financial metrics like revenue growth to measures of customer satisfaction. This assessment can involve comparing your performance either within your organization or against external standards. It can also encompass evaluating functional areas within your organization, such as assessing the HR team's performance through metrics like employee net promoter score or staff engagement surveys or examining the marketing team's performance through measures like net promoter score or brand awareness.

- Strategic benchmarking involves the comparison of strategies, business methodologies, and business models to enhance your own strategic planning and identify your strategic priorities. The goal is to gain insights into the strategies that drive success in other companies, teams, or business units and then analyse how these strategies can be integrated into your own operations to enhance competitiveness.[7]

Q2. Machine Learning

2.1 What is Machine learning

Machine learning is a branch of artificial intelligence (AI) and computer science that emphasizes the utilization of data and algorithms to imitate the process of human learning, with the goal of progressively enhancing its accuracy over time. In simpler terms, machine learning enables computers to learn from data and make decisions or predictions without being explicitly programmed to do so.[8]

Discuss its evolution over time and why is it popular?

Machine learning technology has been around since 1952, but it has evolved significantly over the past decade, with several transitional periods in the mid-90s. In the 1990s, the data-driven approach to machine learning emerged, with a particular focus on natural language processing, search, and information retrieval from 1995 to 2005. During this period, machine learning tools were relatively simpler compared to today's advanced tools.

Neural networks, which are computer systems inspired by the human brain and nervous system, gained popularity in the 1980s and made a resurgence around 2005. Machine learning has become a prominent and trending technology in recent years, with Gartner's 2016 Hype Cycle for Emerging Technologies placing it in the "peak of inflated expectations," expecting mainstream adoption in the next 2 to 5 years. To support this growth, technological capabilities, including infrastructure and technical skills, must advance.

Machine learning has been a highly active and rewarding area of research, finding widespread applications across various fields. It has brought about a significant shift in technology and its practical applications. Some key advancements that have positively impacted real-world problem-solving are highlighted in the following sections.[9]

Reason why is popular

- The modern challenges are high-dimensional in nature: The modern challenges in various fields are often high-dimensional in nature, meaning that they involve many variables, features, or data points. This high dimensionality is one of the key reasons why machine learning is popular and particularly effective in addressing contemporary problems example handling complex data and enhancing accuracy.
- With rich data sources, it is important to build models that solve problems in high-dimensional space. The ability of machine learning to tackle problems in high-dimensional space, especially when rich data sources are involved, is another pivotal reason for its widespread popularity. It empowers organizations and researchers to harness the potential of complex, multidimensional data, fostering innovation and providing practical solutions in different field or sectors it can be employed through
- Through it, the models can be integrated into working software. It supports the kinds of products. The integration of machine learning into software and products leads to significant enhancements in automation, personalization, efficiency, safety, and the quality of services and experiences across various industries. It empowers innovative solutions and improves existing offerings, making machine learning so popular nowadays.[10]

2.2 Three examples of machine learning techniques that can be viewed as either supervised or unsupervised approaches.

- **Decision tree:** is a supervised learning algorithm employed for both classification and regression tasks. It serves as a predictive modelling technique, enabling the classification of data or the prediction of future outcomes. Decision trees are visualized as flowcharts, commencing at the root node with a particular data-related question. This question directs the flow of the tree, branching into potential answers. Subsequently, the branches lead to decision nodes, which further inquire, yielding additional outcomes. This process continues until the data arrives at a terminal or "leaf" node, concluding the decision-making process.[11]

- **Random Forest** is a supervised machine learning technique. It is a widely used ensemble learning algorithm that combines multiple decision trees to make predictions or reach a final result. Random Forest is classified as a supervised learning technique, as it requires labelled training data to train the model.

- K-Means clustering is an unsupervised machine learning algorithm, setting it apart from supervised learning, which relies on labelled data. K-Means excels at partitioning objects into clusters based on their inherent similarities, distinguishing them from objects in other clusters. This process of grouping data into clusters without prior labels makes K-Means a valuable tool for exploring patterns and structure within data and identifying natural groupings, even when the categories are not predefined.[12]

2.3 Difference between classification and regression

Regression and classification algorithms are both types of supervised learning techniques used in machine learning for prediction tasks. They both operate on labeled datasets, where the input data is associated with corresponding output labels. However, their main difference lies in how they approach and solve machine learning problems.

Differences

**Classification**

Classification is a type of algorithm designed to determine functions that can partition a dataset into distinct classes based on various parameters. When using a classification algorithm, a computer program undergoes training on a dataset to learn and subsequently categorize data into different classes based on the knowledge it acquires.

These algorithms are focused on identifying the mapping function that associates input "x" with discrete output "y." They estimate these discrete values, such as binary outcomes (e.g., 0

and 1, yes and no, true, or false), relying on specific independent variables. In simpler terms, classification algorithms gauge the likelihood of an event occurring by fitting the data to a logistic function.

Classification algorithms find application in tasks like differentiating between email and spam, predicting the likelihood of bank customers repaying loans, and detecting cancer tumor cells. Its types is K-Nearest Neighbors, and logistic regression

While

**Regression**

Regression finds correlations between dependent and independent variables. Therefore, regression algorithms help predict continuous variables such as house prices, market trends, weather patterns, oil and gas prices (a critical task these days!), etc.

The Regression algorithm's task is finding the mapping function so we can map the input variable of "x" to the continuous output variable of "y."[13] Its types is Random forest, and Decision Tree


Source : Regression vs Classification image address

The figure shows how differ both types of supervised techniques where Regression give straight line and other curved.

2.4 Difference between supervised learning and unsupervised learning

**Supervised Learning**

Supervised learning is a machine learning technique, characterized by its reliance on labelled datasets. These datasets are specifically curated to instruct or "supervise" algorithms in effectively categorizing data or making accurate predictions. By employing datasets containing labelled inputs and corresponding outputs, the model can evaluate its performance and iteratively enhance its ability to make precise classifications and predictions as it learns from the provided supervision.

It is grouped into two types: Classification and Regression

In supervised learning, the algorithm “learns” from the training dataset by iteratively making predictions on the data and adjusting for the correct answer. Which tend to increase the accuracy than unsupervised learning

**Unsupervised Learning**

Unsupervised learning is machine learning algorithms used examine and group unlabeled datasets. These algorithms uncover concealed patterns within the data autonomously, without the necessity of human guidance or supervision. As a result, they are referred to as "unsupervised" learning techniques.

It is grouped into Three types: Clustering, Association, and Dimensionality reduction

Unsupervised learning models, in contrast, work on their own to discover the inherent structure of unlabelled data, which cause requirement of human intervention to increase the accuracy and performance.[14]

2.5 Examples of successful applications of machine learning and explain what technique appropriate and what type of learning is involved

**Speech recognition**

When using Google, you have the option to Search by voice, which is a prominent application of machine learning. This feature falls under the category of speech recognition, and it is indeed a popular and widely used application of machine learning technology.[15]

Speech recognition, also known as Speech to text or Computer speech recognition, is the process of converting spoken words and voice instructions into text. In contemporary applications, machine learning algorithms play a crucial role in achieving accurate and efficient speech recognition. Virtual assistants like Google Assistant, Siri, Cortana, and Alexa heavily rely on speech recognition technology to understand and respond to voice instructions from users. These virtual assistants use machine learning to continuously improve their ability to comprehend and respond to natural language commands and queries.

The technique used is both Supervised learning to train data with pair of example and use unsupervised learning to discover pattern or labelled voice data.

Q3 Analysing Diabetes data

For approaching programming question I used Jupyter notebook as IDE for coding and on the first cell was for library's if I want to use one I will import them on the first cell.

3.1 Asked to load the diabetes dataset into notebook and produce a correlation matrix of the explanatory variables and then make heatmap of the matrix and describe the relationship between variables.

Starting I imported pandas as the python library used to extract the dataset and then I analyse the dataframe to extract the independent variables by removing dependent variable which is mentioned as Y.

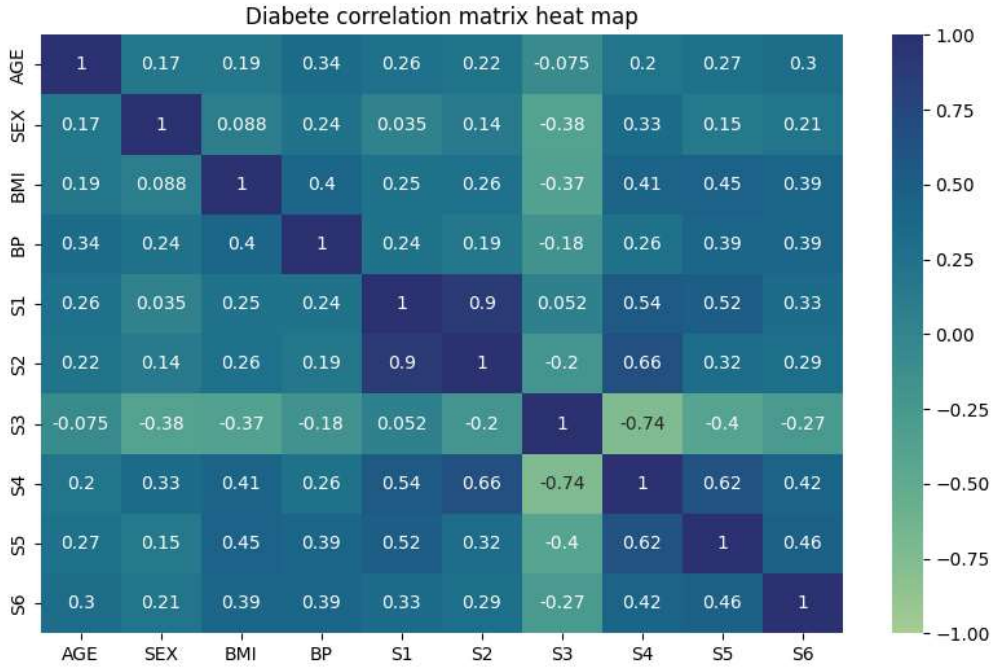
By using corr () python function to create the correlation matrix for the diabate dataset with 10 rows and 10 column.

Then I imported matplotlib and seaborn library to plot the heatmap that display and for easy visualization of the correlation matrix and labelled it with title has 1 in their cell.

Explaining the heatmap

The diagonal of the heatmap is which is dark blue shows that the independent or explanatory are correlated which is true every variable is correlated to itself.

The correlation matrix is in range of -1 to 1 whenever the cell has the value close to 1 it will get the strong blue color or dark blue and if it is close to -1 I will get light blue for example we seen that diagonal is all 1 we have strong dark blue, and for S3&S4 and for S4&S3 is where I have correlation of -0.74 which is very light indicating that variable tend to be in opposite directions and there is no relationship between the variables.




Insight

From the heatmap it become easy to visualize and analyse the matrix with the help of color bar that show the range the cell is in based on the color easly. By remembering that the positive correlation denote that the two variable has relationship and moves in the same direction where increase in one variable tend to increase other. While for negative variable the correlation is in opposite direction apart from variable being correlated to each other S1&S2 and S2&S1 are strong correlated on the level of 0.9 which is close to one means the increase of one influence the increase of another. And become inverse to S3&S4 and for S4&S3 which are negatively correlated means tend to be in opposite direction.

In conclusion performing heatmap to analyse the correlation matrix enhanced more my understanding and better interpretation of the correlation matrix.

3.2. collinearity is special case when two or more variables are closely related to one another correlated[16]. For regression analysis if the independent variable is correlated to each other produces standard error of coefficient estimated which reduce readability of the model's and for that reason we should not trust p value to determine whether the independent variable is significant or not. Tod determine the collinearity in regression model is to first check the correlation matrix of independent variables, if two variables have correlation above 0.9 there are highly correlated with each other.[17]

Studying the effect of collinearity among the predictor variables

I started by extracting the predictor or independent variables and dependent variable from diabetes dataset. By importing the stats model library I have added intercept term to independent variable to intercept linear regression then use linear regression model function to predict the dependent variable and fit the model. After I calculated variance inflation factor

(VIF) to measure the collinearity in regression analysis for each independent variables I measured collinearity.

| | Variable(Predictor) | VIF |
|---|---|---|
| 0 | const | 685.773603 |
| 1 | AGE | 1.217307 |
| 2 | SEX | 1.278071 |
| 3 | BMI | 1.509437 |
| 4 | BP | 1.459428 |
| 5 | S1 | 59.202510 |
| 6 | S2 | 39.193370 |
| 7 | S3 | 15.402156 |
| 8 | S4 | 8.890986 |
| 9 | S5 | 10.075967 |
| 10 | S6 | 1.484623 |

**Interpretation of the output**

The R-squared value in the linear regression model summary tells us that approximately 51.8% of the variance in the dependent variable is explained by the independent variables included in the model. In other words, it indicates the proportion of the variation in the dependent variable (Y) that can be accounted for by the independent variables (AGE, SEX, BMI, BP, S1, S2, S3, S4, S5, S6) collectively. A higher R-squared value suggests a better fit of the model to the data.

Additionally, the correlation between the independent variables and the dependent variable can be assessed through the coefficients, standard errors, t-statistics, and p-values provided in the summary. These statistics help determine the strength and significance of the relationships between each independent variable and the dependent variable which are show in the figure.

For VIF dataframe S1 & S2 These variables have high VIF values, indicating significant multicollinearity with other independent variables. Their VIF values are particularly high, with S1 having a VIF of approximately 59.20 and S2 having a VIF of about 39.19. While AGE, SEX, BMI, BP, S6: These independent variables have VIF values close to 1, indicating very little multicollinearity with other variables. They are relatively independent of each other in the model.

The low VIF indicate that variable is strongly independent there is no multicollinearity where I find AGE, SEX BMI, BP, and S6 are low multicollinearity whike for S1 and S2 there show significant multicollinearity with each other.

In conclusion from this task I learned more about collinearity and how to check if the variable are multicollinearity using variance inflation factor.

3.3 Create a multivariate linear model using all ten variables and a constant and then calculate e Mean Squared Error and the adjusted R2 for model1

In this code, a multivariate linear regression model is created to determine the Mean Squared Error (MSE). The independent variable is represented as 'X', and the dependent variable is 'y' (previously defined in a question before this code). The `sm.OLS` function from a statistics library (presumably StatsModels) is used to create a linear regression model between 'y' and 'X'. The model is then fitted using the `fit()` method, and predictions are made for 'y' using this model. The Mean Squared Error is calculated by taking the mean of the squared differences between the actual 'y' values and the predicted values. The adjusted R-squared value, a measure of the model's goodness of fit, is also computed. Finally, the Mean Squared Error, the adjusted R-squared value, and a summary of the regression model's statistics are printed.

```
The Mean Square Error: 2859.6963475867506

Adjusted R-squared for model1: 0.5065592904853231
```

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Y | R-squared: | 0.518 |
| Model: | OLS | Adj. R-squared: | 0.507 |
| Method: | Least Squares | F-statistic: | 46.27 |
| Date: | Mon, 06 Nov 2023 | Prob (F-statistic): | 3.83e-62 |
| Time: | 23:05:13 | Log-Likelihood: | -2386.0 |
| No. Observations: | 442 | AIC: | 4794. |
| Df Residuals: | 431 | BIC: | 4839. |
| Df Model: | 10 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -334.5671 | 67.455 | -4.960 | 0.000 | -467.148 | -201.986 |
| AGE | -0.0364 | 0.217 | -0.168 | 0.867 | -0.463 | 0.390 |
| SEX | -22.8596 | 5.836 | -3.917 | 0.000 | -34.330 | -11.389 |
| BMI | 5.6030 | 0.717 | 7.813 | 0.000 | 4.194 | 7.012 |
| BP | 1.1168 | 0.225 | 4.958 | 0.000 | 0.674 | 1.560 |
| S1 | -1.0900 | 0.573 | -1.901 | 0.058 | -2.217 | 0.037 |
| S2 | 0.7465 | 0.531 | 1.406 | 0.160 | -0.297 | 1.790 |
| S3 | 0.3720 | 0.782 | 0.475 | 0.635 | -1.166 | 1.910 |
| S4 | 6.5338 | 5.959 | 1.097 | 0.273 | -5.178 | 18.245 |
| S5 | 68.4831 | 15.670 | 4.370 | 0.000 | 37.685 | 99.282 |
| S6 | 0.2801 | 0.273 | 1.025 | 0.306 | -0.257 | 0.817 |

| | | | |
|---|---|---|---|
| Omnibus: | 1.506 | Durbin-Watson: | 2.029 |
| Prob(Omnibus): | 0.471 | Jarque-Bera (JB): | 1.404 |
| Skew: | 0.017 | Prob(JB): | 0.496 |
| Kurtosis: | 2.726 | Cond. No. | 7.24e+03 |

Notes:

The MSE value suggests that, on average, the squared difference between your model's predictions and the actual data is approximately 2859.70. which is very high error and shows that there is no significant relationship between the variable.

The adjusted R-squared value of approximately 0.5066 indicates that your model explains around 50.66% of the variation in the dependent variable. This means that your model accounts for more than half of the variability in the data.

3.4 Difference between stepwise forward selection and backward selection.

Stepwise selection is the iterative construction of a regression model that involves the step-by-step selection of independent variables to be used in the regression model. It entails adding or removing explanatory variables in succession and testing their statistical significance after each iteration[18].In other words, stepwise regression is the process of determining the importance of independent variables, distinguishing the significant ones from the non-significant ones. Two common approaches for performing stepwise regression are forward selection and backward selection.[19]

Stepwise forward selection is a variable selection method that starts with no variables in the model. It tests each variable as it is added to the model, retaining those that are deemed statistically significant and yield the best results. The process is repeated iteratively until an optimal result is achieved. While this method is simple and easy to implement, it can be computationally expensive.

Backward selection is a variable selection method that begins with a set of independent variables and then proceeds to remove one variable at a time. It tests whether the removed variable is significant or if its removal provides the least improvement in the model. The variable with the highest p-value is removed from the model, and a new model is fitted. This process is repeated until all variables in the model have p-values below a predefined threshold, typically alpha = 0.05. Unlike forward selection, backward selection is generally more computationally efficient.[20]

In conclusion, both forward and backward selection are methods used for feature selection. They achieve this by iterating step by step, with one method starting with no variables in the model and gradually adding the best-fit features, while the other begins with all variables in the model and successively removes the least significant ones.

3.4 Explaining how stepwise it work.

Stepwise is method of selecting the best fit features with the least significancy level alpha = 0.05. There is method that can be employed to perform this task Forward selection and backward selection.

Forward selection

Forward selection starting from the null model which has no covariates, at each iteration step of forward selection, a new variable is added to the current model based on some criterion such as Residual sums of squares (RSS). This provides sequence of p models and the model minimizing a criterion, such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC)

Backward selection: works in the same way as forward selection but the difference is run in backward by starting with full model containing independent variables and remove one variable at time based on the increase in  Residual Sum of Squares (RSS)[21]

```
The selected variable: BMI ---> 3.4660064451673995e-42

The selected variable: S5 ---> 3.0396348492618457e-20

The selected variable: BP ---> 3.742619620837897e-05

The selected variable: S1 ---> 0.0014544305422726858

The selected variable: SEX ---> 0.009230559696370681

The selected variable: S2 ---> 0.00027230239927345684


The Selected Features: ['const', 'BMI', 'S5', 'BP', 'S1', 'SEX', 'S2']


Mean Squared Error (MSE) for the new model: 2876.683251787016

R-squared (R^2) for the new model: 0.5148837959256445
```

The code performs a stepwise feature selection technique called "forward selection" to identify the most relevant features for a linear regression model. It systematically evaluates the addition of each feature to the model, considering their p-values. If a feature is statistically significant (with a p-value less than the significance level of 0.05), it is included in the list of selected predictors. This process continues until no more significant features can be added to the model. The selected features include 'BMI,' 'S5,' 'BP,' 'S1,' 'SEX,' and 'S2,' each with its corresponding p-value. Once the feature selection is complete, the code builds a new linear regression model using these selected features and calculates key performance metrics.

The code reports two crucial metrics for the selected model. First, it calculates the Mean Squared Error (MSE), which quantifies the average squared difference between the actual target values and the predicted values from the model. In this case, the MSE is approximately 2876.68, implying the model's accuracy in predicting the target variable. Second, the code computes the R-squared (R^2), a statistical measure that reveals the proportion of variance in the target variable explained by the chosen features. The R-squared value is around 0.515, indicating that approximately 51.5% of the variance in the target variable is accounted for by the selected features. A higher R-squared value suggests a better fit of the model to the data, while a lower MSE signifies better predictive accuracy. In this case the R-square show we have accuracy above 50%

Q4.

4.1 what is the difference between logistic regression and linear regression

Introduction

Linear regression and logistic regression are powerful machine learning techniques that leverage historical data to make predictions and inform decision-making in various domains. Linear regression, for instance, is well-suited for modeling continuous numeric outcomes, such as sales forecasts or temperature predictions, by identifying and quantifying

relationships between variables. On the other hand, logistic regression is a valuable tool for classification tasks, like predicting whether an email is spam or not. Continuous refinement and advancements in these regression methods, coupled with the increasing availability of data, enable us to continually improve their predictive accuracy and applicability across a wide range of real-world scenarios, ultimately contributing to more informed and data-driven decision-making processes.

Linear regression

Linear Regression is a supervised machine learning algorithm for data science learners that predicts continuous values. Linear Regression assumes that there is a linear relationship present between dependent and independent variables. In simple words, it finds the best fitting line/plane that describes two or more variables.[22]

Regression finds the relationship between the input and output data by plotting a line that fits the input data and maps it onto the output. This line represents the mathematical relationship between the independent input variables and is called The Line of Best Fit. Ideally, it covers as many input variables as possible while leaving out the outliers or the noise. For your given data, the best fit is a straight line.

Linear regression is used to predict the continuous dependent variable which is integer using a given set of independent variables, where it predicts the value of continuous variables, and the best fit line is straight line. The accuracy in linear regression is measured using least square estimation. Linear regression is mostly used in business domain, forecasting stock.[23]

Logistic regression

Logistic Regression is a classification algorithm, used to classify elements of a set into two groups (binary classification) by calculating the probability of each element of the set Logistic Regression is the appropriate regression analysis to conduct when the dependent variable has a binary solution, we predict the values of categorical variables.

In Logistic Regression, the input data belongs to categories, which means multiple input values map onto the same output values. Using Logistic Regression, you can find the category that a new input value belongs to. Unlike Linear regression, Logistic Regression does not assume that the values are linearly correlated to one other. Consider the data below, which shows the input data mapped onto two output categories, 0 and 1.[24]

Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables, is mostly used for solving classification problem. The accuracy is measured using maximum likelihood estimation. Unlike the linear regression for logistic the best fit is given by a curve and output is always binary value between 0 and 1 value. It is applied in classification like image processing.[23]

4.2 Calculate the probability of survival for a passenger on the titanic

The code calculates the probability of passenger survival on the Titanic by dividing the total number of survivors by the overall count of passengers. The result, approximately 0.382 or

38.2%, represents the probability that a passenger would survive the Titanic disaster. This figure indicates that, based on the dataset used, about 38.2% of the passengers aboard the Titanic survived. Interpreting this probability, we can conclude that during the tragic event, there was a substantial loss of life, with a majority of passengers not surviving. This underscores the severity of the disaster the people in boat faced.

```
The Probability of passenger survival on Titanic:0.3819709702062643
```

4.3 Table giving survival probabilities broken down by passenger class, gender and age

The code generates a table that breaks down the probability of survival on the Titanic based on different passenger categories, including passenger class (Pclass) and gender (male and female). It calculates the probability of survival for each category by taking the mean of the 'survived' column within the corresponding subsets of the Titanic dataset. The resulting table shows that passengers in Pclass 1 had the highest probability of survival (approximately 61.9%), followed by Pclass 2 (42.96%) and Pclass 3 (25.5%). Additionally, the table reveals a significant difference in survival probabilities between genders, with females having a much higher likelihood of survival (72.75%) compared to males (19.1%). This analysis provides valuable insights into the impact of passenger class and gender on survival rates during the Titanic disaster, highlighting the priority given to passengers in higher classes and the "women and children first" protocol during evacuation.

| | Category | Probability of Survival |
|---|---|---|
| 0 | Pclass 1 | 0.619195 |
| 1 | Pclass 2 | 0.429603 |
| 2 | Pclass 3 | 0.255289 |
| 3 | male | 0.190985 |
| 4 | female | 0.727468 |

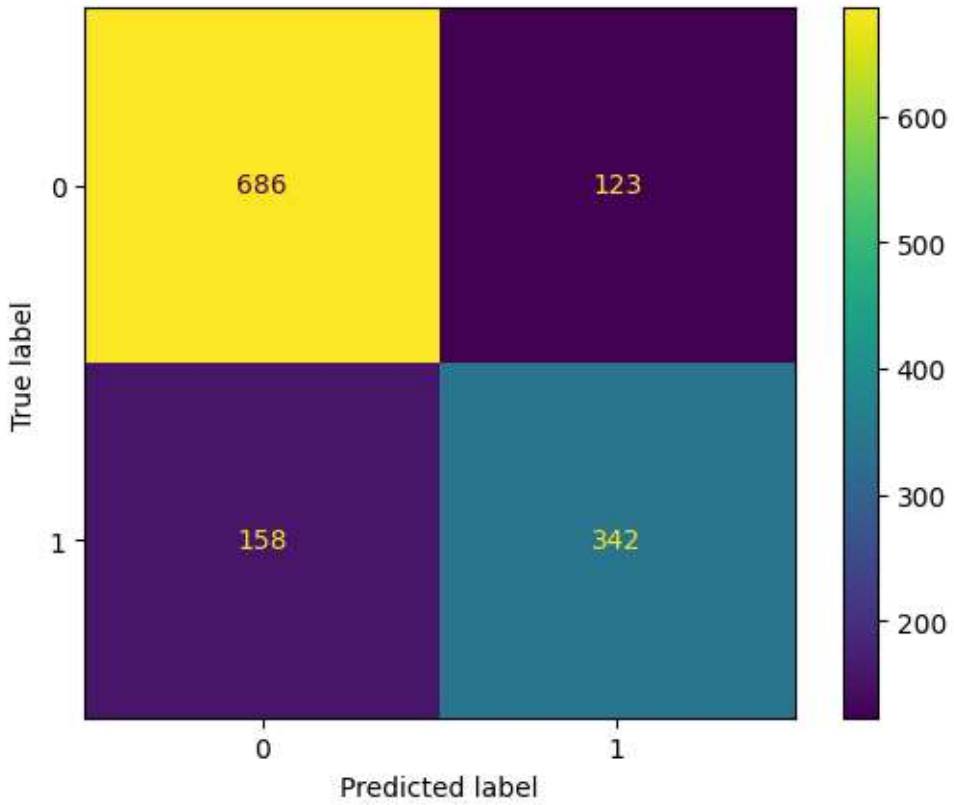
This study underscores the influence of socioeconomic status and gender on the survival outcomes of Titanic passengers. It sheds light on the historic events of the disaster and the ethical principles that guided the evacuation process, providing a valuable perspective on the tragic event.

4.4

I began by performing a logistic regression analysis using the Titanic dataset. The first part involves data preprocessing, which includes handling missing values in the 'age' column by replacing them with the mean age and one-hot encoding the 'sex' variable to convert it into binary columns ('sex_male' and 'sex_female'). The features (X) and the target variable (y) are defined, with 'pclass,' 'sex_male,' 'sex_female,' and 'age' as features and 'survived' as the target variable. A logistic regression model is created and fitted to the data, and predictions are made on the same dataset.

The estimated coefficients (parameters) of the logistic regression model are displayed, including 'pclass,' 'sex_male,' 'sex_female,' 'age,' and the intercept. The code also calculates various evaluation metrics such as accuracy, precision, recall, and the F1-score. Finally, the confusion matrix is displayed both as a raw matrix and using a visual representation created with ConfusionMatrixDisplay.

```
Parameter Estimates:
pclass: -1.0515340179630286
sex_male: -1.233599009013706
sex_female: 1.2335966864018313
age: -0.03163907334508645
Intercept: 3.0714412495546
Accuracy: 0.7853323147440795
Precision: 0.7354838709677419
Recall: 0.684
F1-score: 0.7088082901554404
Confusion Matrix:
[[686 123]
 [158 342]]
```




The results indicate that the logistic regression model achieved an accuracy of approximately 78.53%, implying that it correctly predicted passenger survival status for 78.53% of the cases. The precision of around 73.55% suggests that when the model predicted survival, it was correct about 73.55% of the time. The recall of approximately 68.4% indicates that the model captured 68.4% of the actual survivors. The F1-score, a harmonic mean of precision

and recall, is about 70.88%. The confusion matrix reveals that 686 true negatives (correctly predicted non-survivors), 123 false positives, 158 false negatives, and 342 true positives. In summary, the model provides valuable insights into passenger survival prediction, with room for potential improvements in precision and recall, depending on the specific use case and desired trade-offs between these metrics.

4.5 Performance of the model, measured by classification accuracy (number of correct classifications divided by total number of classifications) based on confusion matrix

Calculates the classification accuracy of a machine learning model applied to a dataset. Classification accuracy is a fundamental evaluation metric used to assess how well a model classifies instances correctly. In this code, the confusion matrix is computed, which is a tabular representation of the model's predictions against the actual outcomes. The matrix contains true positives (correct positive predictions), true negatives (correct negative predictions), false positives (incorrect positive predictions), and false negatives (incorrect negative predictions). The classification accuracy is then determined by taking the sum of true positives and true negatives and dividing it by the total number of instances in the dataset.

```
Classification Accuracy: 0.7853323147440795
```

The resulting accuracy score, approximately 78.53%, reveals the model's capability to correctly classify passengers in the Titanic dataset as survivors or non-survivors.

## Reference


[1] "Rule-based System In Artificial Intelligence Explained - Dataconomy." Accessed: Nov. 04, 2023. [Online]. Available: https://dataconomy.com/2023/04/25/rule-based-system-in-artificial-intelligence/

[2] "Rule Based Approach in NLP," GeeksforGeeks. Accessed: Nov. 06, 2023. [Online]. Available: https://www.geeksforgeeks.org/rule-based-approach-in-nlp/

[3] "What is Overfitting? - Overfitting in Machine Learning Explained - AWS," Amazon Web Services, Inc. Accessed: Nov. 06, 2023. [Online]. Available: https://aws.amazon.com/what-is/overfitting/

[4] J. Bogerd, "Signs of overfitting and how to avoid it," Medium. Accessed: Nov. 06, 2023. [Online]. Available: https://medium.com/@jonathanbogerd/signs-of-overfitting-and-how-to-avoid-it-7aa24c01f46f

[5] "What is Overfitting? | IBM." Accessed: Nov. 06, 2023. [Online]. Available: https://www.ibm.com/topics/overfitting

[6] zuzanna, "What is a benchmark and why do you need it?," MIM Solutions - We make artificial intelligence work for you. Accessed: Nov. 06, 2023. [Online]. Available: https://www.mim.ai/what-is-a-benchmark-and-why-do-you-need-it/

[7] B. Marr, "The Different Types Of Benchmarking – Examples And Easy Explanations," Bernard Marr. Accessed: Nov. 06, 2023. [Online]. Available: https://bernardmarr.com/the-different-types-of-benchmarking-examples-and-easy-explanations/

[8] "What is Machine Learning? | IBM." Accessed: Nov. 06, 2023. [Online]. Available: https://www.ibm.com/topics/machine-learning

[9] “The Evolution of Machine Learning - Synectics.” Accessed: Nov. 06, 2023. [Online]. Available: https://smdi.com/the-evolution-of-machine-learning/
[10] D. Team, “Why is Machine Learning so popular? - From a techno geek’s diary,” DataFlair. Accessed: Nov. 06, 2023. [Online]. Available: https://data-flair.training/blogs/why-machine-learning-is-popular/
[11] “Decision Trees in Machine Learning: Two Types (+ Examples),” Coursera. Accessed: Nov. 06, 2023. [Online]. Available: https://www.coursera.org/articles/decision-tree-machine-learning
[12] “K-means Clustering Algorithm: Applications, Types, and Demos [Updated] | Simplilearn,” Simplilearn.com. Accessed: Nov. 06, 2023. [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm
[13] “Regression vs. Classification in Machine Learning for Beginners | Simplilearn.” Accessed: Nov. 06, 2023. [Online]. Available: https://www.simplilearn.com/regression-vs-classification-in-machine-learning-article
[14] J. Delua, “Supervised vs. Unsupervised Learning: What’s the Difference?,” IBM Blog. Accessed: Nov. 06, 2023. [Online]. Available: https://www.ibm.com/blog/supervised-vs-unsupervised-learning/
[15] “Applications of Machine Learning - Javatpoint.” Accessed: Nov. 06, 2023. [Online]. Available: https://www.javatpoint.com/applications-of-machine-learning
[16] “Collinearity - an overview | ScienceDirect Topics.” Accessed: Nov. 04, 2023. [Online]. Available: https://www.sciencedirect.com/topics/mathematics/collinearity
[17] “A Beginner’s Guide to Collinearity: What it is and How it affects our regression model.” Accessed: Nov. 06, 2023. [Online]. Available: https://www.stratascratch.com/blog/a-beginner-s-guide-to-collinearity-what-it-is-and-how-it-affects-our-regression-model/
[18] “Stepwise Regression: Definition, Uses, Example, and Limitations,” Investopedia. Accessed: Nov. 06, 2023. [Online]. Available: https://www.investopedia.com/terms/s/stepwise-regression.asp
[19] “Stepwise Regression: Definition, Uses, Example, and Limitations,” Investopedia. Accessed: Nov. 06, 2023. [Online]. Available: https://www.investopedia.com/terms/s/stepwise-regression.asp
[20] “What Is Backward Elimination Technique In Machine Learning? | Simplilearn,” Simplilearn.com. Accessed: Nov. 06, 2023. [Online]. Available: https://www.simplilearn.com/what-is-backward-elimination-technique-in-machine-learning-article
[21] “Forward Selection - an overview | ScienceDirect Topics.” Accessed: Nov. 06, 2023. [Online]. Available: https://www.sciencedirect.com/topics/computer-science/forward-selection
[22] A. Kumar, “Linear regression hypothesis testing: Concepts, Examples,” Analytics Yogi. Accessed: Oct. 16, 2023. [Online]. Available: https://vitalflux.com/linear-regression-hypothesis-testing-examples/
[23] “Linear Regression vs Logistic Regression - Javatpoint,” www.javatpoint.com. Accessed: Nov. 06, 2023. [Online]. Available: https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning
[24] “Understanding The Difference Between Linear vs Logistic Regression,” Simplilearn.com. Accessed: Nov. 06, 2023. [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/linear-regression-vs-logistic-regression