Data Mining Paper

Data Mining:

Data mining is the process of extracting useful information, knowledge, hidden pattern and relationships from large amount of dataset using techniques like machine learning, statistics and Artifical intelligence. It helps organizations to make beller decision and Prediction.

Traditional data analysis focuses on summarize data and answering "what happen" using basic statistics. Data mining goes beyond this by discovering hidden pattern, Predicting the futures trends and explaing why "it happend" and "what happend next" in future. That's why data mining is more automated, intelligent and Predictive than the traditional data analysis.

Traditional data analysis focuses on examing historical data to compute summeries such as averages total and percentage. It is mainly descriptive in nature and help to understand what happend already. In contrast data minning

one hot encoding and label encoding are the two methods that used to convert the categorical data to numerical data.

jordan burden relax.

Label Encoding :- label encoding assign each category by a unique integer value. - for example if the categories are (Red, Blue, Green) they may be encoded as 0, 1, 2. This method is useful when the categorical data have a natural order (ordinal data) like small medium, large. However, when use on nominal data which has no order it may creat a false sense of ranking which can mislead the model.

zipline dragon swimming south north west

One hot encoding :- create a separate binary column for each category for example for the same colors three new columns created. and the presence of category is marked by 1 while others are 0. This method is appropriate for nomial data. where categories do not have any order. It prevent model from assuming any ranking among categories however, it increase the no of colums especially when category is many. cylinder slender cone L.E for ordered data while one for nomial data

(b)

When one-hot encoding is applied to dataset that contain categorical variable with a large number of categories. Several challenges arise OHE create a separate binary colum of each category which can result in very large no of column this leads to problem known as high dimensionality

Firstly a large no of colum increase the memory and storage requirement of dataset.

Secondly it produce a sparse matrix means most value are zero. which makes computation slow and less efficient

Thirdly:- high dimensionality can negatively effect the performance of ML model cause overfitting. where the model learn noise instead of patterns

Therefore OHE become difficult to use when it increase dimensional computation cost & risk of overfit.

(C) (4)

filter method and wrapper method are two main approaches for feature selection in data Pre processing

Filter method :- evaluate the relevance of feature by using statistical test such as correlation, chi-square test etc. These method do not involve any ML model while selecting feature. Therefore they are fast, computational efficient and suitable for large dataset, however they may not always produce the best feature for specific model they do not consider model performance

Wrapper method use the performance of ML model to evaluate the subset of feature. The model is trained and tested repeatedly with diff combination of feature eg forward selection backward elimination These method generally produce a more accurate feature subset but they are computationally expensive slow - not suitable for big dataset

→ filter method are faster but less customized while wrapper are slower but produce more accurate feature.

(d)

Feature selection improves the performance of ML model by removing irrelevant or less informative features from the dataset. When unnecessary feature are included the model become more complex or may learn noise instead of meaningful Pattern this lead to overfitting. By selecting only the most useful feature the model become simple & more generalizable.

Also feature selection also reduce the dimensionality of data which decrease the computational cost of training the model. This result is faster training & Prediction time.

Therefore feature selection lead to improve model accuracy, reduce overfitting, low computation time & more efficient & understerdable

(c)

The k-Nearest Neighbors kNN algorithm is a simple and commonly used classification method based on similarity b/w the datapoint. kNN does not produce a mathematical model in advanced. Instead it stores a entire training dataset and make decision only when new data point needs to classify.

To classify a new datapoint kNN measure the distance b/w the new point and all existing point in dataset using the Euclidean distance

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

After calculating distance algorithm select the k closest datapoint called neighbour. The class of new data point is determined based on majority class among k neighbour.

Thus kNN determin neighbour based on distance & classification is made by majority voting among the nearest point.

Normalize study hours

max = 14 min = 6

= study_hour - 6
14 - 6

→ Normalize Exam_Score. min = 65 max = 96

Examscor - 65
96 - 65

Normalize table.

Normalization is important because it scale all values to same range. usually b/w 0 and 1 when featur like study_hour and exam score have different unit and scale the larger number may dominate the smaller one in analysis. Normalization ensure that each feature contribute equally, Prevent biased result, improve the performance of statistical and ML model and speed up calculation. Therefore normalization make data easier to compare and more suitable for accurate model training.