CC2025-ML-PROJECT

PROJECT REPORT

-T.MANIKANTA SAI

#TASK1 “CampusPulse Initiative”:

#LEVEL_1:Variable Identification Protocol

=>For this **Level-1 and 2,3** Used Libraries Are **“pandas”,”numpy”,”matplotlib”,”seaborn”.**
=>Firstly Take another dataset using iloc() function for getting the dataset which contains only 3 unknown features
=>by using info() foand describe() functions to know about the unknown Features
=>heat map is drawn using seaborn(sns.heatmap()).
=>FEATURE_1:
=>For identifying the Feature_1 plot the histogram between Feature_1 and Frequency by using plt.hist() we got a structre like a bell shaped
=>after drawing a correlation heat map by using sns.heatmap()
=> we can get know that Feature_1 is positively correlated with failures,absences and negatively correlated with G1,G2,G3
=>The values are ranged between 15 and 22 from describe() option we can get this data By this we can guess it can be “**AGE**”
=>Age histogram will be in bell shape in general (for big data) and as age increases there is possibilty to decrease of grades and increase of failures and absences.
=>FEATURE_2:
=>The range of feature_2 was from 1 to 4 from the data we can conclude
=>so it can be a categorical data
=>From correlation heatmap Feature_2 is negatively correlated with Failures and absences and positively correlated with grades
=> boxplots between grades and Feature_2 shows grades decreases as feature 2 decreases and from 1 to 4 as it goes the grades are increasing (boxplots compared by the median line)
=>And failures also decreases as feature_2 increases so it is related to “**studytime** “
=> From the histogram plotted for 2 there is a peak
=> so 1-less than 2hrs ,2- less than 4hrs and >2hrs,3-less than 7hrs and >4, 4-greater than 7hrs (per day study time)
=>FEATURE_3:
=>The range of feature_3 was from 1 to 5 from the data we can conclude
=>so it can be also categorical data
=>histogram as a peak at 1(more no of students)
=>From correlation heatmap feature_3 is negative correlated with grades and strongly positive correlative with goout and Dalc
=>by plotting bloxplots we can get to know that with increasing from 1 to 5 dalc and goout increases and grades are decreasing
=> so it can be “**skipping the classes**” or “**disengaged in work**” .Because by skipping classes grades can decrease and goout and dalc will increase
=> by increasing from 1 to 5 grades decreasing so 1 can be skipping of low no of classes (or it can be given rating for disengaged in a work)

#FEATURE_1:AGE
#FEATURE_2:STUDY TIME
#FEATURE_3:SKIPPING CLASSES

#LEVEL_2:Data Integrity Audit

=>First we can find the no of NaN enteries in all data columns by using .isnull().sum()
=> In these some have **dtype=object** columns which have NaN enteries
=>For Object Type Data
=> So to fill them we use the "**Mode**" of the object data column because We cant compute the mean and median for object type data so we take "Mode" into consideration
=>mode is a most frequent value in the data column so we select Mode as the filling method
=>filling of the data column can be done using the function df[''].fillna.(df[''].mode()[0])
=>For Numerical Type Data
=>from isnull().sum() we get to know that numerical data type is int64
=>Numerical Data types**: Fedu,traveltime,freetime,absences,G2,Feature_1,Feature_2,Feature_3**
=>plot the histogram for Feature_3 it shows 1 was dominant .
=> so we can fill for **Feature 3** null enteries using **Mode** of the data column because When one value appears more times in a column, that suggests it is the most common or typical category in the data.so if we don't know the value it make sense to take more frequently occurring answer to fill null entries.
=>for **Feature_2** we use **Mode** to fill as feature_2 is a categorical data like feature 3 and when we calculate the mean and median of feature_2 by using .median() and .mean() function we get non int64 so it is best to take Mode into consideration
=> "**Fedu"** is also categorical data but it is symmetrical and no outliers present this can be find out by plotting the boxplot
=>in this case it is good to take "**Mean**" for symmetrical data but the mean is also not a int64 so we can fill using median or mode of data
=>but by plotting histogram b/w fedu and freq we get to know no category so much dominant so it is safe to take " **Median** "and also **Median** is int64 for Fedu
=> for remaining columns as they are non categorical we use mean or median but mean for all of them are not int64 so we choose the "**Median**" to fill then remaining NaN entries.
=>median generally used for skewed data(can be confirmed from boxplots where there will be longer whisker).

#LEVEL_3:Exploratory Insight Report

=>Question_1:
=>the question is whether students having freetime will goout or not
=>by using matplotlib plotted a boxplot between freetime and goout .It shows that with increase in free time students tends to go out but the tendency is more varying when students freetime was high as boxplot has wider width .
=>with increase in free time the median of goout also increasing .low freetime tendency to goout was less
=>Question_2:
=> the question is does study time affects the final grade or not.
=>by plotting a boxplot between studytime and final grade we come to know as study time increases the final grade also increases
=> As studytime is categorical it varies from 1 to 4 the final grade peak was increasing.
=> median of the data also increases with study time.
=>in graph there is slight decrease in median for 3 to 4 may be due to some students due to rigorous studying they might have not performed well this can be a reason.
=>Question_3:
=>the question is daily alcohol consumption affects grades or not.
=>by plotting the boxplot between Dalc and G3 we come to know that as dalc goes to extreme the final grade definetly decreases

=>the median of grades also decreases as Dalc increases from 1 to 5. And at higher level like 4 the boxplot width was also narrow that means there is low varying among the students all are getting low grades who are at level 5 of Dalc.
=>by plotting the scatter plot we come to know more clear about that almost all the getting low grades at the level 5.
=>Question_4:
=>the question is does the students location was affecting the studies ,grades or not.
=> by plotting the boxplot between address and G3 we come to know that the median decreases for final grade while comparing urban to rural
=> but there are few outliers in rural who are getting higher grades as well when compared to urban regions
=>Question_5:
=>the question is about the relationship status of the student effect the studies,grades or not.
=>by plotting the boxplot between romantic and G3 we come to know that median of both yes and no options are same from the data
=> for further clarity we plotted scatter plot there we can confirm that there no median change for both options yes and no.
=> from the plots we can confirm that there is slight distraction for some students that’s why some got less grades