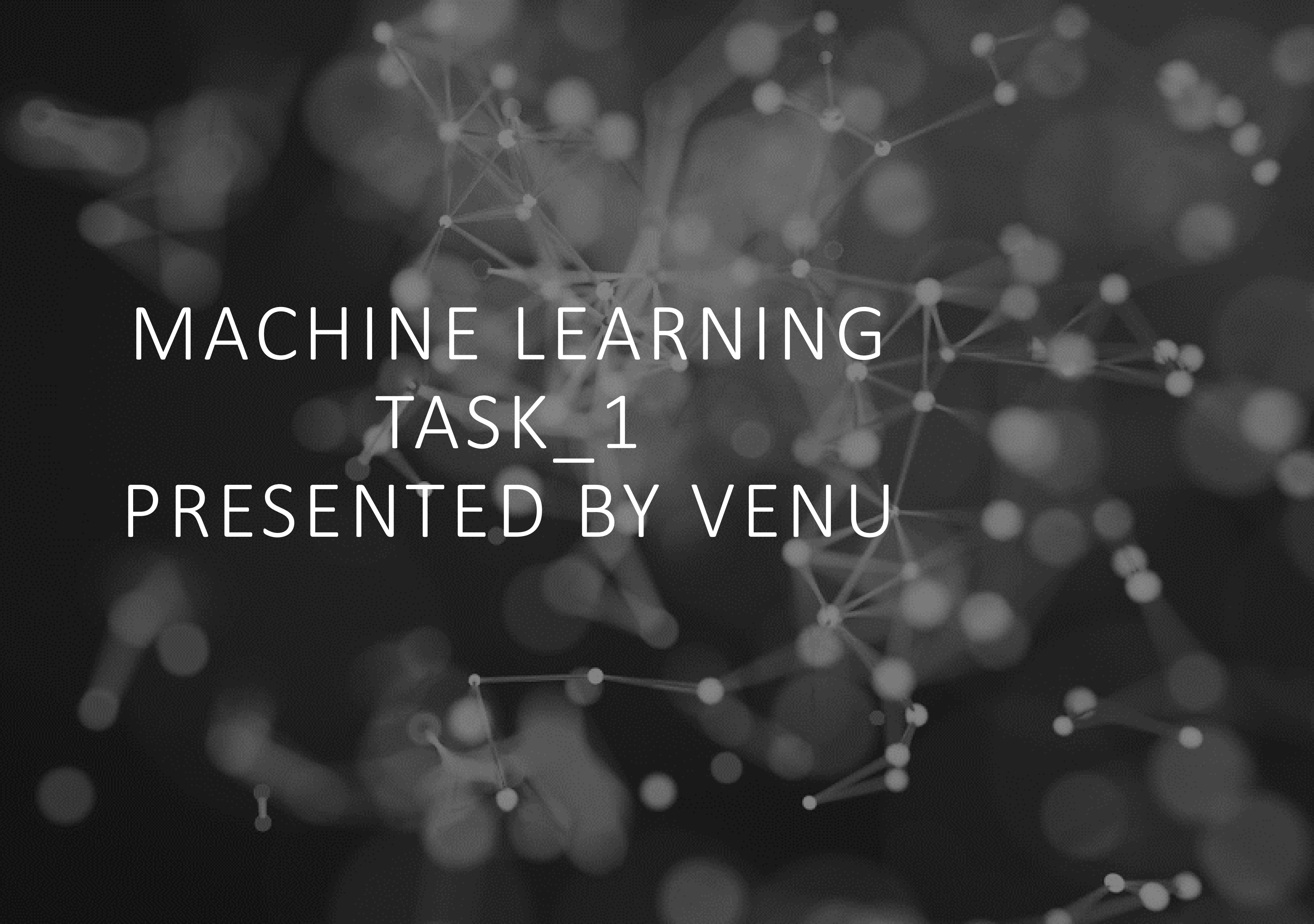
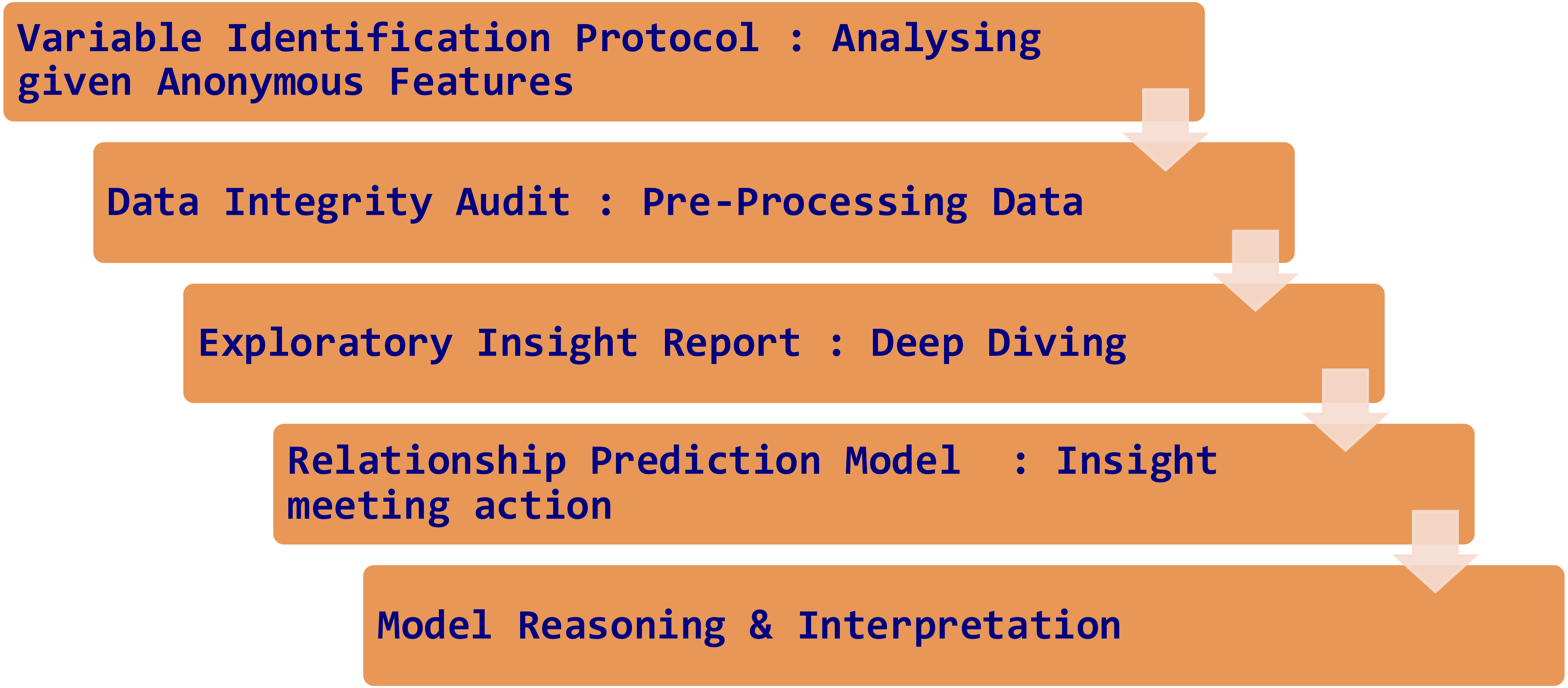


MACHINE LEARNING TASK_1 PRESENTED BY VENU

LEVELS



Variable Identification Protocol : Analysing given Anonymous Features

Data Integrity Audit : Pre-Processing Data

Exploratory Insight Report : Deep Diving

Relationship Prediction Model : Insight meeting action

Model Reasoning & Interpretation

LEVEL 1: STEPS DONE BY ME

1.Observing data of three features individually

(ranges,mean,std etc.)

(Guessed Feature_1 to be age)

2.Plotting heatmap to find correlation of these features with others

3.Anaylsing the columns with more correlation with Features 3,2,1 individually

(For better understanding)

LEVEL 1 : RESULTS

1.We can assure that Feature_3 has strong relation with Dalc(0.62) and with goout(0.4)

So,Feature_3 can be **Social activity in a scale of 1-5**

2.*Feature_2 can be* **Study time in a scale(1-5)** *since it correlates positively with grades*

G1(0.26),G2(0.25),G3(0.25)

3.Feature_1 can be **Academic difficulty in a scale(18-21)** *due to its strong,positively correlation with failures(0.31)*

LEVEL 1 : EXPERIMENTS BY ME

Checking for supporting relations to check original guess of Feature_1 : *Age*

For that

1.Checking if one school offering PG courses,so high age students are present

So ,plotted box plot

2.Checking if high age more chance to be in relationship,plotted box plot

Conclusion : Good chance(not low) that Feature_1 to be *Age*(Values in same range)

LEVEL 2: STEPS & RESULTS

Checking for strange/inconsistent answers

Checked for all columns , but not found any strange/inconsistent answers :) :(

Filling Nan Values

Recall from correlation map Medu and Fedu have 0.65 correlation.So,

(1)Let's fill missing Fedu with corresponding Medu values

Similarly,due to strong correlation

(2)I prefer to fill *G2* with (*G1*+*G3*)/2 and rounding off for safety since all are integers in given data

Also,(3)Feature_3 based on Dalc and goout for the same reason(Dalc+goout/2)

LEVEL 2: STEPS & RESULTS

(4)Remaining categorical columns('higher') and

numerical columns('traveltime','famsize','Feature_2','freetime') with MODE

due to few distinct elements and small scale(1-5)

(5)Leftover 'Feature_1','absences' with median to avoid outliners

LEVEL 3 : Q & A

(1)Which school produces better student grades?

Interpretation:(By Box Plot)

If parents or students prioritize academic performance when choosing schools,GP has a stronger academic track record.As you can see GP has broader spread over higher grades.Median of GP too surpasses that of MS

(2)Does parental education impact grades?

(P_Edu=Medu+Pedu/2)

Interpretation:(By Regression plot)

There is a clear positive trend : as P_Edu increases, students tend to score higher final grades reflecting better guidance and supportive academic environment

Note:

This doesn't mean if there is less P_Edu there is less grades.This is just an estimation.As you can see there are other cases as well

LEVEL 3 : Q & A

(3.1)Does internet access affect grades?

Interpretation:(By Violin Plot)

From the graph we can say that,in contradiction internet access is making students to get higher grades as you can see broader distribution over higher grades

(3.2)Does internet access affect social activity?

Interpretation:(By Box plot)

There is no impact of internet access on social activity as you can median and mean are almost same in both cases

So,it is totally a misconception that internet access leads to low grades and limits social activity

LEVEL 3 : Q & A

(4)Is there any effect of Dalc on Absences?

Interpretation:(By Box plot)

As you can see as Dalc increases,the distribution is increasing over higher absences

So,We can assure that high Weekday alcoholic consumption leads to more absences

(5)What's relation between extraclasses and grades?

Interpretation:(By violin plot)

The students who attend paid classes are getting grades on par with students who aren't attending paid classes.Since,students who get low grades are more likely to join paid classes.Mean is almost same.So,It's better to join paid classes

LEVEL 4 : EVERYTHING

Let's breakdown the data into 3 parts i.e, academic,behavioral,social also not including the coulmns which have no relation to target

My plan is to first find the features in all the above three which most likely signal to be in relationship by plotting 3 heatmaps(after encoding) with each one and target column and also taking strongest predictors for building models

It seems there is no much corelation (meaning there exists non-linear,multi-factor relations)

So,there was no much use of heatmaps.Then let's directly dive into Random Forests to find feature importance which also covers building of prediction Model

let's find top 15 features important for predicting target from random tree classifier

LEVEL 4 : EVERYTHING

From graph and correlation graphs we got that

Academic pattern : [Less grades,more P_Edu] : more chance to be in relationship

Behavioral pattern : [more absences,more goout,more freetime,more Dalc] more chance to be in relationship

Social pattern : [less famrel] : more chance to be in relationship

Three features(Feature_1,_2,_3) play important role in prediction too

LEVEL 4 : EVERYTHING

Built different models

Random Forest - 67%

Logistic Regression 57%

Decision Tree - 64%

K-nearest-neighbour –59%

Revealed by Model

(1)Random Forest having maximum accuracy implying non-linear relations to predict target

(2)Students with more social activity tend to be in relationship

(3)Large number of 'YES' were predicted wrong as 'NO' of target implying it's difficult to find the people in relationship based on features :| also implying it's easy to find people who are not in relationship :)

LEVEL 4 : EVERYTHING

Not Revealed by Model

(1)Logistic regression can't reveal multiple columns collectively affecting target

(2)Hidden features (like emotional,not in dataset) which can be got by understanding remaining columns deeply can also effectively influence target but can't be revealed by model

(3)We can just get feature importance but not how it influences.We have to use our intuition and surroundings to conclude

Revealed by experimenting

(1)Taking only top important features to predict not always improve accuracy :(implying some factors effectively contributes when combined with other factors

(2)Correlation heatmaps aren't always to get top important features since there may be non-linear relation btw factors

LEVEL 4 : CONCLUSION:

1.Random Tree is best predictor for given data having an accuracy of 67%.Columns have non-linear relation with target and some factors collectively effect target

2.Removing non-important columns(got by model) not always improve accuracy

3.To conclude,less grades,more social activity,less parental education imply to be in relationship

LEVEL 5 : EVERYTHING

Let's start with making decision boundary plots(2D) for each classifier by selecting two columns

Let's take for

(1)Random Forest : plot with features 'goout' and 'P_Edu'

(2)Logistic Regression : plot with features 'G2' and 'Feature_3'

(3)Decision Tree : plot with features 'absences' and 'Feature_2'

(4)K - Nearest Neighbour : plot with features 'Feature_1' and 'freetime'

Chose the two features in each(from top important features got from Feature importance of RT) and also in two features one supporting and other non-supporting to be in relationship

LEVEL 5 : EVERYTHING

Now let's move to interesting part

(1)Computing shap values and plotting global feature importance using SHAP**

(2)Generating local explanations for two students one with 'yes' and one with 'no' using waterfall plots

Let's generate three waterfall plots since can't get all in single plot

We will use already segregated columns that are academic,behavioural,social

(Recall from level 4)

Then we will plot local explanation plots of 'yes' and 'no' using logistic regression classifier

LEVEL 5 : RESULTS

What really drives relationship prediction?

To summarise below features postively drive to 'yes' of target column

Dalc,Feature_3,internet,Pstatus,famrel,Fjob_services

To summarise below features postively drive to 'no' of target column

G1,G3,Fjob_teacher,Fjob_other,Fjob_services,internet