DSBA Quiz:

Section 1:

a) Mean is calculated by adding all the values and dividing the sum by the total number of values. Median is the number found at the exact middle of the set of values. In the given set, the mean of the income is different from median because mean is not very robust as it is largely influenced by outliers. Also, the mean and median are different because the distribution is not even and there are many high-income customers or the outliers.

b) The mean and standard deviation are affected by extreme values. The purpose of Winsorization is to "robustify" classical statistics by reducing the impact of extreme observations.

c) Part 1) The Median is 4. By putting the value in the given formula: median ($|Xi - \tilde{X}|$) we will get (3,3,2,0,1,2,6). Sorting the answer will give us – (0,1,2,2,3,3,6). MAD of the series is 2.

   Part 2) MAD is a robust statistic which is more resilient to outliers in a data set than the standard deviation. In the standard deviation, the distances from the mean are squared, so large deviations are weighted more, hence outliers can heavily influence it.

d) Yes, Standardization is a scaling technique where the values are centered around the mean with a unit standard deviation. It is always better practice to fit the scaler on the training data and then use it to transform the testing data. This will avoid any data leakage during the model testing. Also, the scaling of target values is generally not required.
e) Feature 1 can be of more use without the transformation to predict the model as the bin plot somehow determines a linear model line

Section 2:

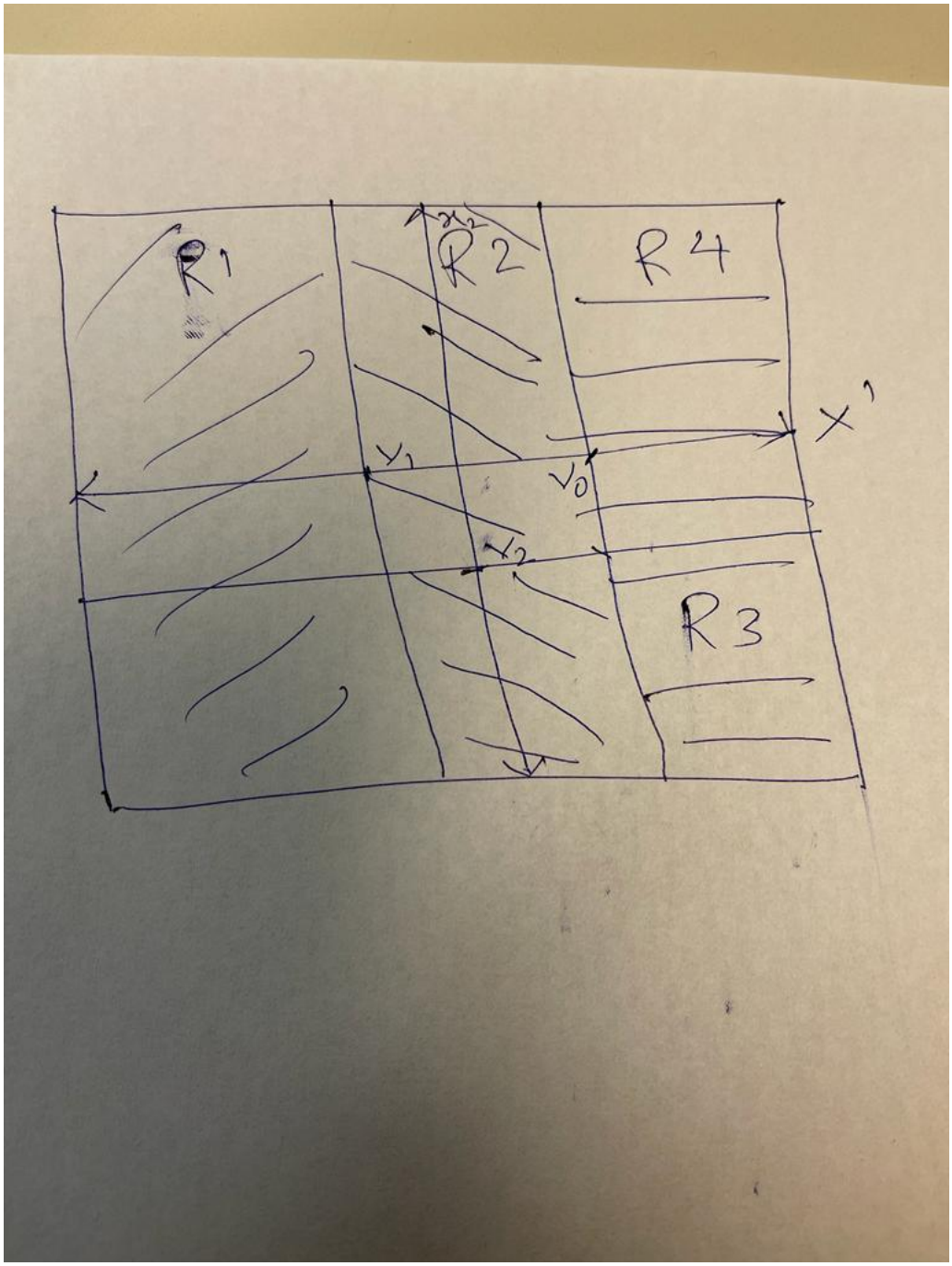
Part A

a) (ii) the sum of the absolute value of the coefficients
b) True
c) False
d) True
e) False
f) False
g) True

Part B – In this case if the log odds = 0 then the standard probability is 0.5 which signifies: Neither outcome is favored over the other.

Section 3:

a) Decision Tree Classifier
b) sklearn.model selection.StratifiedKFold, because it takes group information into account to avoid building folds with imbalanced class distributions.
c) i) False
   ii) True
   iii) True
   iv) False
d)



e) Random forest depends on multiple decision trees. It relies on feature importance of a single decision tree. Random forest works on a random basis selecting features/variables class with highest votes is considered the predicted class. It results in better accuracy by taking the subsets of the features. It reduces the variance part of the error and hence is suitable for a new dataset.

f) Part a) True
Part b) True
Part c) Large Number

g) Part a) Recall is 0.857
Part b) Precision is 0.4
Partc)
Between Model 1 and Model 2, We can prefer because as each model satisfies a different criterion. Based on the requirements, we can take decision on what information is relevant for our business use.
When, Recall = 1.0

TP= TP + FN -> FN=0
This means a patient who was told they doesn't have a tumor but turns out they had the one. They are the one with greater loss since it is a critical disease, and they missed an early detection.