Prior to analyzing this evaluation, please note that all referred output numbers should be visible after running the provided Python code and many ratios (as well as some numbers) were calculated manually. Here is the list of explained methods of calculations that were done outside the output provided by the Python code:

- Occurrences were calculated by opening the csv files and via search window. When a given value was searched, the number of instances was also shown. For example, the 5th attribute of "breast-cancer.csv" was copied, pasted in an new sheet, "no" value was searched, the number of occurrences was noted in the same search bar, and that number was used as the total number occurrences of "no" in that attribute. Same rule was applied when counting occurrences of values in the 6th attribute of "car.csv".
- The averages mentioned in the evolution under Question 1 and Question 2 was calculated by summing the row from B8 to G8, and then divided by the total number of elements. The excel formula was as following: "=(SUM(B8:G8)/(6))".' Please check the bottom right-most value in the Excel sheets — those values outside the matrices are the averages.
- Excel sheet with Information Gain values have been printed in the output when one runs the Python code. However, the values were manually copy and pasted into the prepared excel sheet for clarity.

1. The Naive Bayes classifiers can be seen to vary, in terms of their effectiveness on the given datasets (e.g. in terms of Accuracy). Consider the Information Gain of each attribute, relative to the class distribution — does this help to explain the classifiers' behaviour? Identify any results that are particularly surprising, and explain why they occur.

Naive Bayes' accuracy certainly vary. For "car.csv" dataset, the accuracy was 87.384%, for "breast-cancer.csv" was 75.524%, and for "cmc.csv" it was just 50.577%. The difference between the highest accuracy and the lowest accuracy is startling 36.807 percent points, which is a significant value. Such huge gap could be a pivotal point for a learner — some projects may approve the 87.384% accuracy for it is relatively close to 90% accuracy, and 50.577% may be way too unsatisfactory for some researchers.

As for correlation between Naive Bayes' behavior and Information Gain, there seems to be a positive relationship — as Naive Bayes' becomes more accurate, the Information Gain values for each pair (of an attribute and the class) seems to increase. The averages of each pairs (attribute and the class) were calculated in the Excel sheet. For "car.csv" dataset the average was 0.129901, for "breast-cancer.csv" the average was 0.03473121, and for "cmc.csv" the average was 0.03799498.

This is rather surprising, because as accuracy increases, one of the reasons for that could be more predictable values. Moreover, one should also notice that "car.csv" dataset contains the most instances (1728 to be precise) and that "cmc.csv" had a bit less (1473 instances), yet the accuracy between those two datasets is huge — by nearly 40 percent points, so the number of instances does not seem to contribute to the accuracy. It is "breast-cancer.csv" with neither the highest nor the lowest accuracy, which has the most predictable dataset out of all three of them. A reason for that could be the fact that "breast-cancer.csv" had missing values which could significantly impact Information Gain and as the result, predicability overall. After all, missing values will always trigger some unpredictability due to obvious lack of values.

Yet another point worth to be raised is that when one looks at the Information Gain values for "car.csv", the Information Gains between pairs of attribute values is usually 0 (only two out of 15 pairs have a value other than 0; the values hav been rounded to three digits after the dot). Meanwhile, in both datasets, there is only one — and only one — pair that has a value of 0, which means that all other pairs may be considered to have a correlation of some kind, suggesting some dependence and violating Naive Bayes' assumptions.

The witnessed values point toward this conclusion: the more predictable dataset (predictability based on the pairs consisting of an attribute value and the class) the more independent attribute values are (the pairs consisting of two attribute values). This also suggests that the more predictable the dataset, the less Naive Bayes' assumptions are violated, and the accuracy of the learner in that dataset is higher.

2. The Information Gain can be seen as a kind of correlation coefficient between a pair of attributes: when the gain is low, the attribute values are uncorrelated; when the gain is high, the attribute values are correlated. In supervised ML, we typically calculate the Information Gain between a single attribute and the class, but it can be calculated for any pair of attributes. Using the pair-wise IG as a proxy for attribute interdependence, in which cases are our NB assumptions violated? Describe any evidence (or indeed, lack of evidence) that this is has some effect on the effectiveness of the NB classifier.

Naive Bayes' base assumption is that all attributes are independent — this can be seen in very the formula of Naive Bayes. Since Information Gain can be perceived as a correlation coefficient, that would mean that if Information Gain is equal or close to 0, then the given two attributes would be independent because their values have no correlation.

The attached excel sheet shows the Information Gain values for each pair of attributes, and pairs of each attribute in relation to the class. As one can see in "car.csv", only two pairs of attributes out of 15 other pairs has any kind of correlation . That means that majority of attribute pairs have no correlation, which suggests independence and in turn it proves the assumption to be true.

The values seem to be credible. The diagonal that starts from top left corner and ends on the bottom right, has eye-catching values of over 1, which makes sense because an attribute compared to itself should have high correlation for its values are the same. It is also worth to note that the two pairs that have a correlation bigger than 0, have very small Information Gain values — just 0.116 and 0.011. Both of those values are lower than the average Information Gain values for pairs that consist of an attribute and the class (the average was 0.129901)

Yet, in each “cmc.csv” and “breast-cancer.csv” datasets, there is only one pair of attribute values pairs that have a value of 0, suggesting that every other pair has violated Naive Bayes’ base assumptions. Hence, given those three datasets, datasets usually will lead to Naive Bayes’ learner which violates its own assumptions, which in turn makes dataset more credible because in the real world attributes usually are correlated because they are often there for a reason rather than out of randomness.

3. Since we have gone to all of the effort of calculating Infomation Gain, we might as well use that as a criterion for building a “Decision Stump” (1-R classifier). How does the effectiveness of this classifier compare to Naive Bayes? Identify one or more cases where the effectiveness is notably different, and explain why.

1-R classifier was less accurate in making predictions in comparison to Naive Bayes. For instance, Naive Bayes’ accuracy for “breast-cancer.csv” was 75.524% while 1-R’s was 72.027%. The difference between their accuracies was 3.497 percent points. However, “car.csv” dataset had a more drastic difference of 17.361 percent points — Naive Bayes’ accuracy was 87.384%. Meanwhile, 1-R’s accuracy was only 70.023%.

One of the reasons why the “car.csv” dataset had such significantly bigger differences could be that “breast-cancer.csv” data is more uniform and predictable. This implies that the variety of choices for each attribute in the datasets is smaller. “Breast-cancer.csv” has two attributes with only two possible values (5th and 9th attribute seem to be accepting values of either “yes” or “no”). Meanwhile, “car.csv” dataset has attributes of at least 3 different values.

That means that the most popular value of an instance accounts for a bigger proportion — for example, the 5th attribute in “breast-cancer.csv” has 222 occurrences of “no” and only 56 occurrences of “yes”, which makes the total error of that attribute equal to 22.377% (this includes the 8 missing values). In contrast, in the 6th attribute in “car.csv” dataset all values account for equal share — “high”, “med”, and “low” each account for exactly a third of all values. That makes predicting unpredictable because each value has equal chances (of 33%) of occurring. This is a stark difference, especially when one recalls that the mentioned 5th attribute of “breast-cancer.csv” has 77.623% chances of being “no” versus mere 33% for either “high”, “med”, and “low” for the 6th attribute of “car.csv”.

6. Naive Bayes is said to elegantly handle missing attribute values. For the datasets with missing values, is there any evidence that the performance is different on the instances with missing values, compared to the instances where all of the values are present? Does it matter which, or how many values are missing? Would an imputation strategy have any effect on this?

For this question, ”breast-cancer.csv” dataset was considered because out all three datasets, it was the only with missing values. First, accuracy was calculated for Naive Bayes and the dataset with missing values was fed to the classifier. The accuracy was equal to 75.524%.

Then, an imputation strategy was implemented. The only missing values were laying in the 5th attribute (it had 8 missing values) and in the 8th attribute from the left (it had 1 missing value). The mode imputation strategy was implemented — the value with the most occurrences was input in the place of the missing values. The missing values in the 5th attribute were replaced with ”no” (”no” accounted for 222 instances, ”yes” accounted for 56) and in the 8th attribute the missing value was replaced with ”left_low” (”left_low” accounted for 110 instances, left_up accounted fro 97, right_up accounted for 33, and right_low accounted for 24 instances).

The resulting accuracy have been changed, because Naive Bayes' accuracy has increased to 75.874%. One possible explanation for this rather minimal change of just 0.35 percent points could be the fact that missing values accounted for a very small portion of all values. A potential reason why the accuracy has increased is because the values of the 5th and 8th attribute could have been positively correlated to the correct classes, which resulted in a higher accuracy of the classifier.

Mode imputation strategy was chosen because the values in the attributes were not numerical — they were nominal; hence, no average or median could have been calculated. The ordering could not be objectively found and the distances between the values could not be estimated or calculated in any way. Perhaps with different missing values, the change in accuracy could have been better.

Average imputation strategy is often used for the resulting occurrence of values looks more like a classic normal distribution. Meanwhile, mode imputation makes that normal distribution thinner, which results in bias towards the most popular group. One of the drawback of mode imputation is the bias, which may favor wrong/incorrect/undesirable answers (or rather values) despite popularity of that value in the attribute.