CSCI 4521
HOMEWORK 1

AKSHAT SHARMA
UNIVERSITY OF MINNESOTA

Problem 1

Load the CSV into a DataFrame. Add a new column named "Rating_Category" to the DataFrame to classify individuals based on their credit rating. Classify the credit rating as 'Good' if the credit rating is 670 or above, and 'Poor' if the credit rating is below 670.

Solution:

```
# load your data
df = pd.read_csv("https://github.com/sziccardi/CSCI4521_DataRepository/blob/main/Credit.csv?raw=true")
names = list(df.columns) #Save column names

# Create a new column for the rating category.
df['Rating_Category'] = df['Rating'].apply(lambda x: 'Good' if x >= 670 else 'Poor')
```

Date: February 13, 2025.

Problem 2

Choose **four** features in the dataset (other than *Rating*) that you think maybe most indicative of whether an individual has a Good or Poor credit rating. Use the Seaborn pairplot function to create a collection of pairwise scatter plots for these four features. Color the scatter plot points by the *Rating_Category* criteria.

Solution: Based on hit and trials with differnt pairplots of different features, I have chosen the following 4 variables.

- Income (in thousands of dollars)
- Limit
- Balance
- Cards

Next we create a pairplot

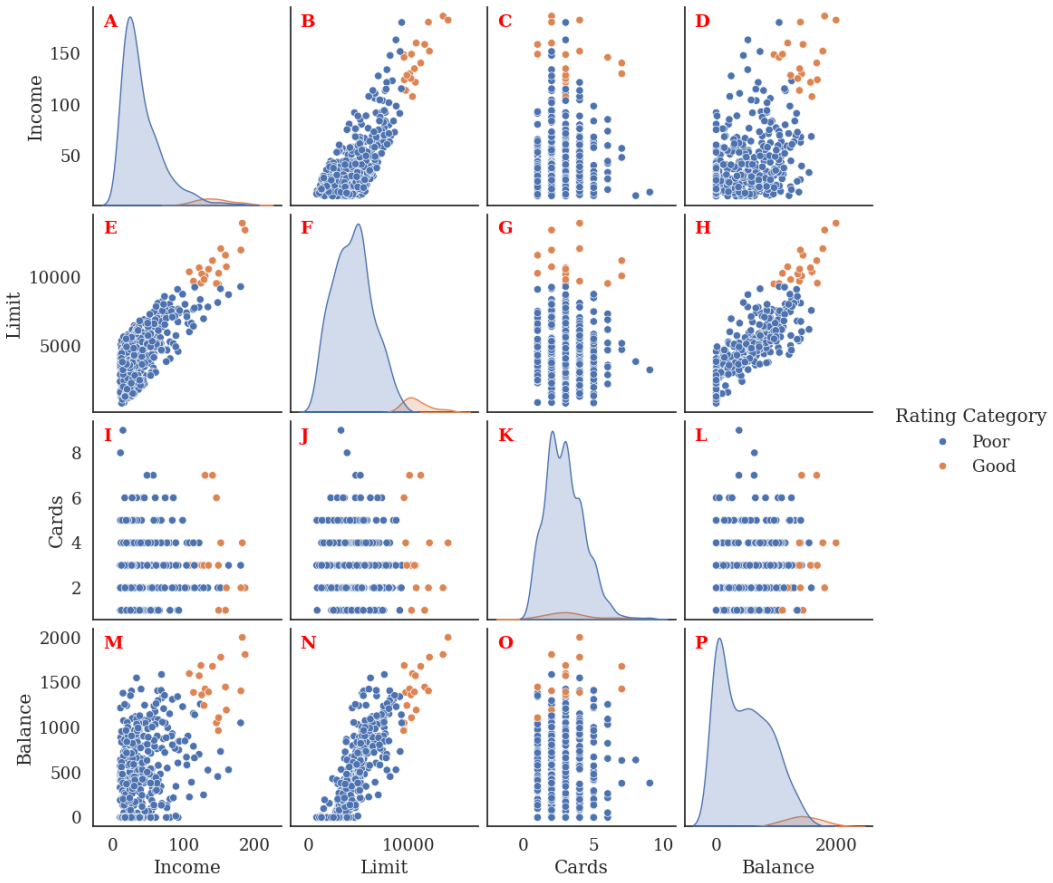

Figure 1. Represents the Pairplot of Financial Attributes (Income, Limit, Balance and Cards) categorized by Rating Category

Problem 3

Based on your plot from Q2, write a paragraph analyzing the resulting graphs.

(a) Are there any clear trends or insights?
(b) What features or pairs of features are most correlated with having a Good or Poor credit rating?
(c) Based on these results, what recommendations would you give to individuals looking to improve their credit rating effectively (e.g., strategies to maintain a good credit rating)?

Solution: The pairplot provides insights into the relationships between income, credit limit, the number of credit cards, and balance, with credit rating categories labeled as 'Poor' and 'Good'.

(a) As shown in Fig: 1, the pairplot reveals several key trends. There is a strong positive correlation between Income and Credit Limit, meaning that individuals with higher incomes tend to have higher credit limits as seen in Fig: 1B. Additionally, Balance is positively correlated with both Income and Credit Limit from Fig: 1M and 1N, suggesting that individuals with higher credit limits tend to carry higher balances. The distribution of Cards appears more discrete, with individuals having varying numbers of credit cards regardless of their income or credit limit from Fig: 1I and 1J.

(b) Aoording to the pairplots in Fig:1, the strongest distinguishing features between 'Good' and 'Poor' credit rating categories are Income and Credit Limit. Individuals categorized as 'Good' tend to cluster towards higher values of both, while those labeled 'Poor' are more concentrated in the lower ranges. The Balance feature also shows a visible distinction, with 'Good' individuals maintaining somewhat higher balances but still in proportion to their credit limits as seen in Fig: 1N. The number of Cards does not show a strong separation between the two groups, indicating that merely holding more credit cards does not necessarily correlate with a better credit rating from Fig: 1[I-L].

(c) To improve credit ratings effectively, individuals should Increase their income if possible, as it is strongly correlated with a higher credit limit, which may improve creditworthiness as suggested by Fig: 1[A-D]. Even though there is a positive correlation between the credit limit and Good rating as seen in Fig: 1[E-H], the variable is not controlled by the consumer. Alternatively, according to data in Fig: 1[M-P], a higher balance also means a better chance of a person having a Good rating. However, I think that this is a False relation due to the Fallacy of Causation Vs. Correlation. The factors of Balance and Credit Rating are correlated but having good Balance doesn't cause a Good Credit Rating. The confounding variables between the two are limit and income. Having a higher income causes an increase in credit card limit, which causes more credit card spending, which then causes higher balance. So, I think having a higher balance has nothing to do with a good Credit Rating. In fact, if we think about it logically, having a lower balance would easily allow someone to clear their debts on time and hence, increase their credit score. Nonetheless, since the data doesn't support that, I will only suggest to **increase your income**.

Problem 4

Write and evaluate two classifier models that predict if an individual has a good credit rating given some of the inputs of your choice (from Q2). Your models should include one model from each of the two categories below:

Category 1 (Choose one)

- A KNN with a small k of your choice
- A KNN with a large K of your choice

Category 2 (Choose one)

- A classifier that says all individuals have good credit
- A classifier which says no individuals have good credit

When evaluating your models, make sure you normalize the input data and split the data into testing and training datasets. Evaluate your models in paragraph form. Be sure to include answers to ***all*** of the following:

(A) For both training and testing datasets, what is the precision, recall, accuracy, and F1-score of each of your classifiers?

(B) For each metric (precision, recall, accuracy, and F1-score), which of the two classifiers did the best and worst on the training data? Why?

(C) For each metric (precision, recall, accuracy, and F1-score), which of the two classifiers did the best and worst on the testing data? Why?

(D) Which of the two classifiers would you most recommend for real-world questions on this dataset?

Solution: I will choose the following models out of the given options:
⇒A KNN with a small K of your choice (K=1)

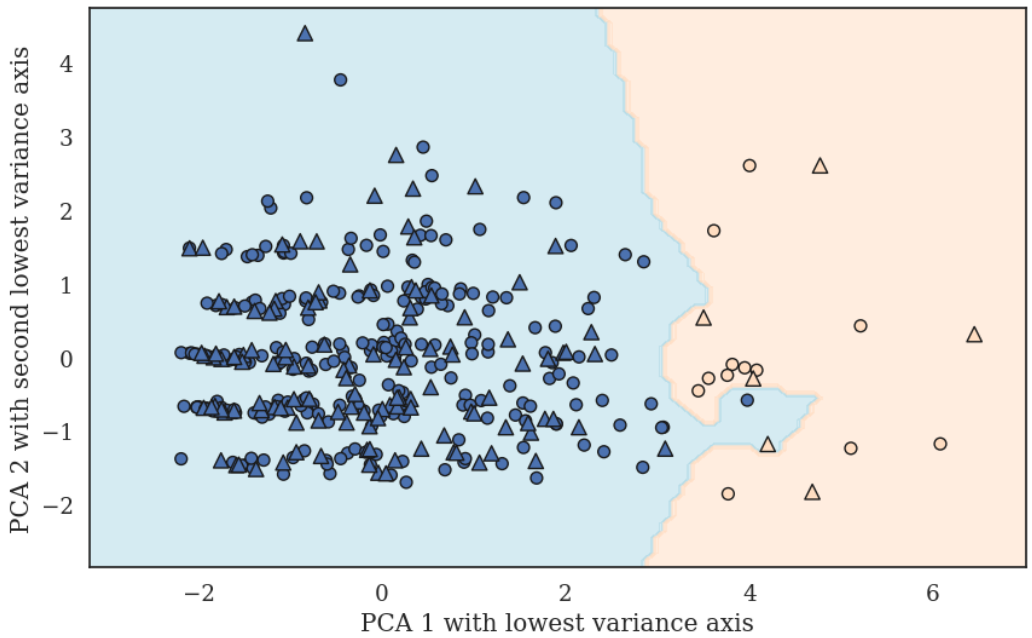

Figure 2. Plot of a KNN graph (K = 1), where the triangles represent testing data while the circle represents training data. The blue backgroud and muted orange background show the region of poor and Good category rating respectively.

⇒ **Classifier which says no individuals have good credit**

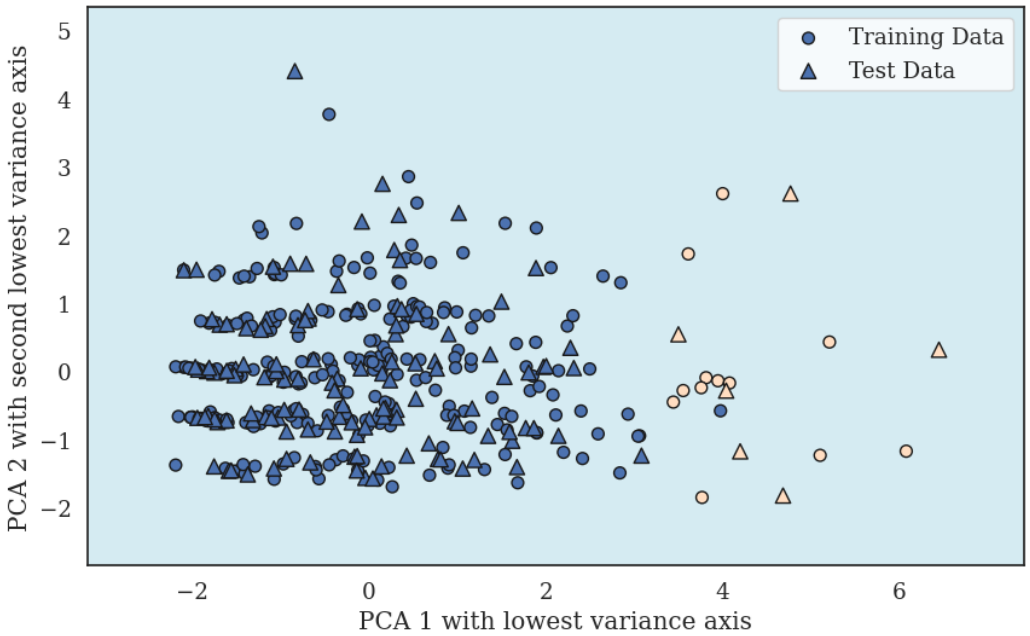

Figure 3. Plot of Decision Boundary with Bad Classifier (always 0). The triangles represent testing data while the circle represents training data

I have chosen a PCA components to reduce the number of dimensions from 4 to 2 in order to get a better visualization of the 2 classifiers. As we can see from Fig: 2 and Fig: 3, the KNN classifier overfits and the Bad Classifier puts all the data points in the Bad credit section. Now, let's answer the Questions

(A) For K=1 classifier, the metric is

Dataset	Accuracy	Precision	Recall	F1-Score
Test	0.99	0.92	1.00	0.95
Train	1.00	1.00	1.00	1.00

Table 1. Macro average metrics for KNN Classifier (Test and Train)

For Bad Classifier, the metric is

Dataset	Accuracy	Precision	Recall	F1-Score
Test	0.95	0.48	0.50	0.49
Train	0.96	0.48	0.50	0.49

Table 2. Macro average metrics for Bad classifier (test and train data)

(B) We can make the following observations about Precision, Recall, Accuracy and F1-score of the KNN ($K = 1$) and Bad Classifier from Table 1 and Table 2:

- **Precision:** K=1 Classifier has a higher average Precision. This is because precision is based on proximity to correct result. With the Bad classifier, we didn't train the model on any data at all and hence, we have high False positive in comparison to K = 2 Classifier.

- **Recall:** K=1 Classifier has a higher average Recall. This is because Recall is based on False Negatives and with Bad classifiers, we had a high number of false negatives in comparison.
- **Accuracy:** K=1 Classifier has a higher average Accuracy. Even though the distinction, between them is not that much, it is only because, most of the points of the csv data file had poor credit. Also, Accuracy deals with the entire confusion matrix, so the extreme issues with Recall and Precision and hard to notice with just accuracy.
- **F1-score:** Since, it is a combination of both Precision and Recall, It stands to reason that k=1 Classifier had a higher F1 score.

(C) Overall, the comparative results of the metric of both testing and training data is similar. Hence, the reason of comparison that is used for training data in part (B) can also be used for testing data. However, notable differences that can be observed from Table 1 and 2 are that for K=1 Classifier, Testing metrics are lower in score than trianing Metric. K = 1, classifier has performed well in both training and testing. But it has performed much better in training than testing.

On the other hand, Bad classifier has perfomed equally terribly in both testing and training. For bad classifier, training is not better than testing.

And that makes sense, since we didn't train Bad classifier on anything whatsoever. So, it just gave equally poor results for both testing and training. But the K=1 Classifier was trained on an over-fitting classification model, so it was much better with training than testing.

(D) For real world application, I would recommend using K = 1 classifier. It might overfit, but looking at all the metrics from Table 1 and 2, I can clearly deduce that K = 1 is far superior to Bad Classifier.

Problem 5

Consider a hypothetical consumer applying for a credit card. Their pre-approval application indicates the following information:

- Personal income: $60,000 a year
- Age: 23 years old, recent graduate (16 years of education completed)
- Marital status: Single (not married)
- Housing: Renting an apartment near their office
- Current credit cards: Two credit cards with a combined outstanding balance of $1,500 out of their combined $7,500 credit limit

(a) What region would you expect this applicant resides in (east, west, or south)? Why?

(b) Using a KNN analysis, would you predict this person to have a good credit score (e.g., above 670)? Why or why not?
   - What k are you using and why?
   - What features are you using?
   - How have you processed the data?

(c) **Extra Credit:** What is the smallest change to the consumer's stats that would flip your prediction? (Use whatever definition of small change makes sense to you.)

Solution:

(a) To find the region of the person, I tried using some machine learning techinique. I used ['Own', 'Balance', 'Age', 'Married', 'Income', 'Limit'] as features and tried to predict the region of th hypothetical consumer. According to my analysis, I can say with almost 40 - 45 % accuracy, that the region of the hypothetical person is **South**. Please refer to the region Section of the Jupyter Notebook for more details.

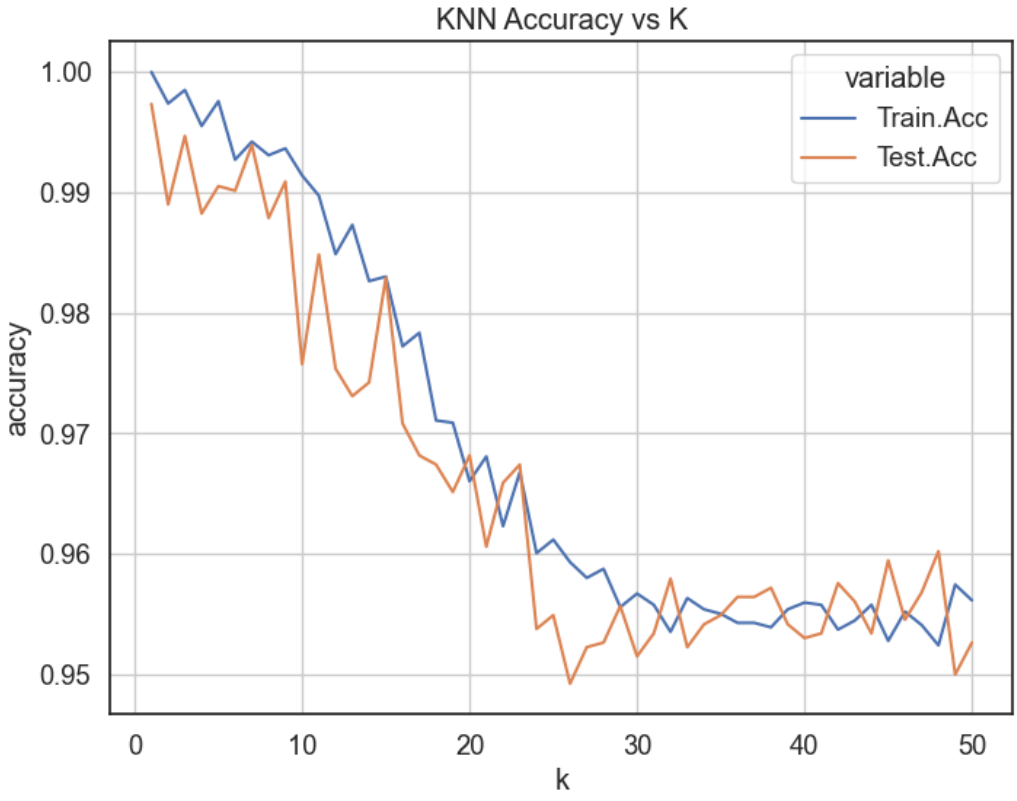

Figure 4. Plot of KNN (for different values of K) vs accuracy (both Training - Blue and Testing - Red)

(b) Using KNN analysis ($K = 1$), I make the prediction that the hypothetical person has a **Poor** credit Rating
- To find my k-value, I run all the k-values from 1 to 50 in a loop to find the value with maximum accuracy. Whatever I end up getting, I use that as my k-value. Usually, my k-value is in the range of 1-4 depending on the randomness of the split of the data. According to Fig: 4, the Best K-value for this run is 1.
- I am using all 4 features from Q2. Because I can see that these features are all directly linked to money and hence one's credit rating. Additionally, it can be seen that almost all features (except number of cards) show a poitive correlation between then and credit rating, as seen in the pair plot.
- After finding the Best K value, I used my entire data to train the model and skipped the testing phase because I already have the best K-value and its accuracy. After training the model, I used it to predict the Rating of the Hypothetical consumer.

(c) This question is vague, so I have made some assumptions and have also provided multiple answers based on the what context you use with the question:

Assumption: I assume that my smallest change, you mean smallest in a normalized sense, otherwise I will have to compare change in number or cards to change in Income, which has completely different units. So, I will normalize everything and then provide the change.

I assume that I will only be changing 4 variables that I discussed in question 2. There is no reason for this assumption, I am doing it for my own sanity. Otherwise, I will fry my computer with all these features in a nested for loop.

Now, let's look at multiple answers, If I consider smallest change in it's theoretical sense, then, I would have to consider number of cards as a floating point number. In this case, the smallest change is shown in Table: 3

Status	Income	Limit	Cards	Balance
Initial	60.000000	7500.000000	2	1500.000000
Final	91.680171	8883.187076	2.410868	1637.755146
Difference	31.680171	383.187076	0.410868	137.755146

Table 3. Comparison of Credit Card Data with non-integer value of cards

Second possibility is that I look at it practically, where I can't own 2.41 cards. In this case, I can either have 2 or 3 cards. Based on this: If I choose to get 2 cards only and don't change that number, then the final result is shown in Table: 4

Status	Income	Limit	Cards	Balance
Initial	60.000000	7500.000000	2	1500.000000
Final	88.160152	9574.780614	2.0	1637.755146
Difference	28.160152	1074.780614	0.0	137.755146

Table 4. Comparison of Credit Card Data with cards = 2

If I choose to take three cards, instead of 2, then the result is shown in Table: 5

Status	Income	Limit	Cards	Balance
Initial	60.000000	7500.000000	2	1500.000000
Final	84.640133	9113.718255	3.0	1729.591910
Difference	24.640133	613.718255	1.0	229.591910

Table 5. Comparison of Credit Card Data with cards = 3

I derived these results using nested for loop for each variable and then changing each of them by small amounts. Then, I saved all the values of changes and found which change was smallest. Based on that, I asnwered the question.