CSCI4150U: Data Mining
Lab 04

Syed Naqvi
100590852

October 16, 2024

Part I:

1. Preprocessing (German Credit Data)

This dataset contains a mixture of ordinal, nominal and numeric features. The ordinal and nominal features must be processed such that Minkowski distance provides a meaningful metric during k nearest neighbors classifications while retaining as much information as possible. We begin with the following features which have either a completely arbitrary ordering or contain only 2 unique values:

- attribute 4: Purpose
  - a list of purchases on credit
- attribute 9: Personal status and sex
  - an arbitrarily ordered list of marital status and sex
- attribute 19: Telephone
  - either have a Telephone (yes) or do not (no)
- attribute 20: Foreign worker
  - either is a foreign worker (yes) or is not (no)

We can use **one-hot encoding** for Attributes 4 and 9 as these features have multiple unique values and **label encoding** for attributes 19 and 20 as these features have only two unique values which makes ordering of label encoding irrelevant.

```
# first we can one-hot encode attribute 4 and 9
german_data_X_encodings = pd.get_dummies(data=german_data_X, columns=['Attribute4','Attribute9']).apply(lambda x: x.apply(lambda x: int(x) if isinstance(x,bool) else x))
display(german_data_X_encodings.loc[:,'Attribute4_A40':].head())
```

✓ 0.0s

	Attribute4_A40	Attribute4_A41	Attribute4_A410	Attribute4_A42	Attribute4_A43	Attribute4_A44	Attribute4_A45	Attribute4_A46	Attribute4_A48	Attribute4_A49	Attribute9_A91	Attribute9_A92	Attribute9_A93	Attribute9_A94
0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
1	0	0	0	0	1	0	0	0	0	0	0	1	0	0
2	0	0	0	0	0	0	0	1	0	0	0	0	1	0
3	0	0	0	1	0	0	0	0	0	0	0	0	1	0
4	1	0	0	0	0	0	0	0	0	0	0	0	1	0

Figure 1: [Attributes 4 and 9 one-hot encoding]

```
# next we apply label encoding to attributes 19 and 20
label_encoder = LabelEncoder()
german_data_X_encodings[['Attribute19', 'Attribute20']] = german_data_X_encodings[['Attribute19', 'Attribute20']].apply(lambda x: label_encoder.fit_transform(x))
display(german_data_X_encodings[['Attribute19', 'Attribute20']].value_counts().reset_index(name='count'))
```

[31] ✓ 0.0s

	Attribute19	Attribute20	count
0	0	0	564
1	1	0	399
2	0	1	32
3	1	1	5

Figure 2: [Attributes 19 and 20 label encoding]

The next set of features appear to have a clear ordinal ranking based on descriptions provided at (https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data) with least credit-worthy values on the left to most credit-worthy values on the right:

- Attribute 1: Status of existing checking account
  - A11 (< 0 DM) → A12 (0 <= & < 200 DM) → A13 (>= 200 DM / salary assignments for at least 1 year) → A14 (no checking account)
- Attribute 3: Credit history
  - A34 (critical account / other credits existing) → A33 (delay in paying off in the past) → A32 (existing credits paid back duly till now) → A31 (all credits at this bank paid back duly) → A30 (no credits taken / all credits paid back duly)
- Attribute 6: Savings account / bonds
  - A61 (< 100 DM) → A62 (100 <= & < 500 DM) → A63 (500 <= & < 1000 DM) → A64 (>= 1000 DM) → A65 (unknown / no savings account)
- Attribute 7: Present employment since
  - A71 (unemployed) → A72 (< 1 year) → A73 (1 <= & < 4 years) → A74 (4 <= & < 7 years) → A75 (>= 7 years)
- Attribute 12: Property
  - A124 (unknown / no property) → A123 (car or other, not in attribute 6) → A122 (building society savings agreement / life insurance) → A121 (real estate)
- Attribute 17: Job
  - A171 (unemployed / unskilled - non-resident) → A172 (unskilled - resident) → A173 (skilled employee / official) → A174 (management / self-employed / highly qualified employee / officer)

We can visualize the value distributions of each feature for each class:

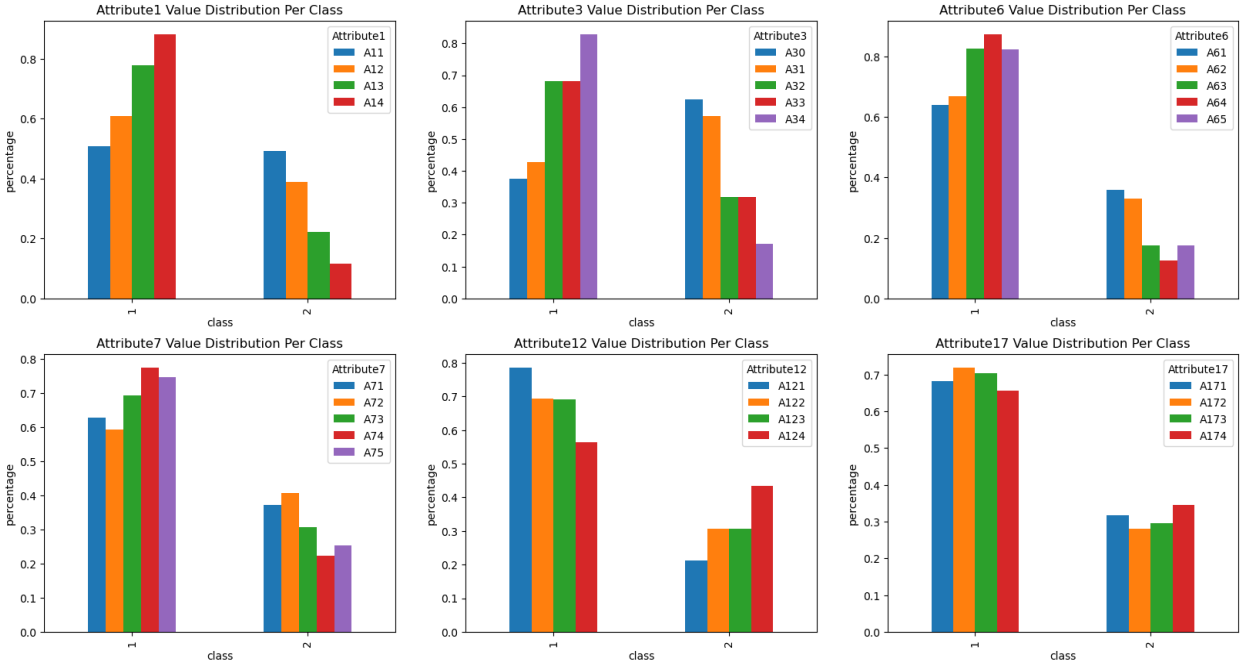

Figure 3: [Ordinal Features Visualization]

It can clearly be observed that the features do seem to closely adhere to their listed orderings and for features that do not, we can re-label them so their lexicographical order better fits the feature's observed order. This results in the following updated class distributions:

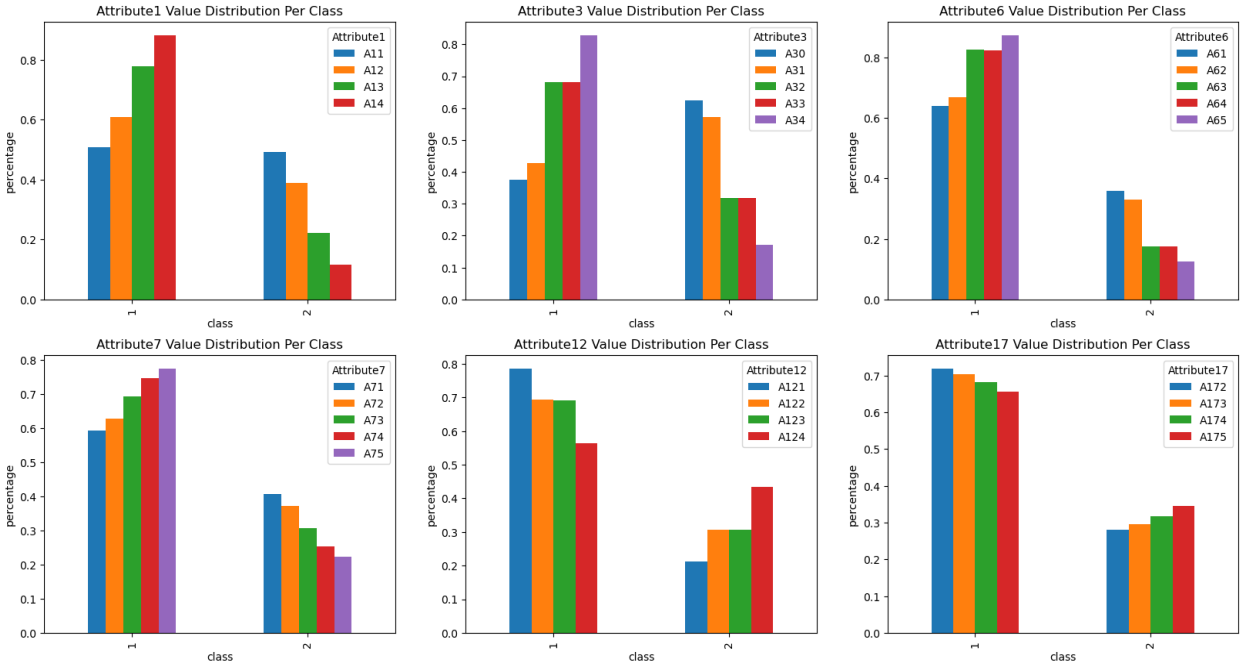

Figure 4: [Ordinal Features Visualization (Re-Labelled)]

Label encoding the above features should now result in labels that better capture feature order, allowing for more meaningful distance measures between objects during k-nearest neighbors classification.