Name: Salvador, Louis Abraham	Date: 10-04-25
Section CIT 401A IT31S1	Instructor: Richard Kho

I. BUSINESS UNDERSTANDINGRichard

In this CRISP-DM exercise, your task is to build a classification model that achieves the highest possible accuracy by experimenting with different combinations of data preparation techniques. You'll be expected to apply sanitation, feature selection, and feature extraction. methods to improve model performance and compare their effects systematically. As you move through the phases — from Business Understanding to Modeling—consider how each preprocessing choice influences the final outcome. Your goal is not just to build a model, but to demonstrate the value of thoughtful data preparation in predictive analytics.

Reflection (10 pts):

a. Based on your understanding of model creation combined with data preparation task(s), do you see any benefit in performing this task? Please elaborate your answer. (10 pts)

- if the model doesn't have enough data it will be noisy,incomplete or unorganized, then the algorithm will perform poorly. Applying data preparation methods like feautre selection, feature extraction will the model can learn patterns more effectively
- Example if the dataset have many missing values and attributes. ignoring these issues, the model may give inaccurate predictions. if we handle missing data properly, encode the categories and remove the irelevant attributes, the quality of the dataset improve and the model as well

II. DATA UNDERSTANDING

Extract the dataset from our Git repository located in the '20251004-Lab-Wk11' directory, and perform your own assessment and evaluation of the file's contents.

Reflection (40 pts):

1) Based on your assessment, what do you think the dataset is about? (5 pts)

- Each row links an order with metadata: Channel(inbound/outbound), issues category/sub-category, order/product details, agent/manager handling the case, timestamps for reporting/responding, and a customer satisfaction score(CSAT). the goal is to predict the CSAT from the interaction and order features.

2) What are the details of the dataset:

a) List the field names along with their corresponding data types. (15 pts)

- Unique id is a string or identifier (object) that is unique to each record.
- channel_name — string or category (object) – channel of contact (inbound, outbound, etc.).
- category — string / categorical (object) — the top-level issue category (for example, Order, Product Query, Return).
- Sub-category — text or categorical (object) — indicates a more specific issue kind.
- There are around 28,756 non-null customer remarks, with approximately 66% of them missing. Good for NLP, but scarce.

Name: Salvador, Louis Abraham	Date: 10-04-25
Section CIT 401A IT31S1	Instructor: Richard Kho

- Order_id — string / identifier (object) — numerous missing (about 67,675 non-null).
- The order_date_time string/timestamp (object) has a considerable number of missing timestamps (about 17,214 non-null values). Needs to be parsed to datetime.
- Issue_reported at — string / timestamp (object) — the time when the customer reported the problem; this should be converted to datetime.
- issue_responded — string / timestamp (object) — the response timestamp, which should be converted to datetime.
- Survey_response_Date — string / date (object) — the date the CSAT was provided; parseable to date.
- Customer_City — string/categorical (object) — numerous missing (about 17,079 non-null).
- Product_category: string / categorical (object) — many lacking.
- Item_price — numeric (float) — item price; there are many missing values and some formatting errors in the raw file.
- The connected_handling_time numeric (float) variable is highly sparse, with only a few non-null values compared to the earlier observation of about 242.
- Agent_name — string or categorical (high cardinality) – thousands of distinct names.
- Supervisor — string / categorical (with high cardinality).
- Manager — string / categorical (with high cardinality).
- Tenure Bucket — string / category — agent tenure bracket (for example, 0-30, >90, on-the-job training).
- Agent Shift — string / category — Morning, Evening, Night.
- CSAT Score — integer or numeric (goal) — customer satisfaction score (1–5). This is numeric in the raw data, but it should be considered as the class label (nominal) for classification purposes.

b) Indicate the number of records. (2 pts)

Total rows (instances): 85,907
Total attributes (columns): 20

c) Identify the attribute in the dataset that is suitable to serve as the class label. (3 pts)

- CSAT Score is a good choice for the class name because it symbolizes the outcome (customer satisfaction) that we wish to predict.
- Practical note: Because CSAT is extremely skewed (most ratings = 5), it is frequently beneficial to convert it to a binary goal (e.g., Satisfied vs Unsatisfied) or utilize class-balancing procedures before modeling.

3) Provide a thorough analysis of the dataset in terms of its readiness for model evaluation. Do you observe any anomalies in the dataset? (15 pts)

Name: Salvador, Louis Abraham	Date: 10-04-25
Section CIT 401A IT31S1	Instructor: Richard Kho

1. Severe class imbalance (CSAT).
    - Most buyers give the highest rating (5). In modeling studies, the classifier frequently predicts class 5, whereas other classes have zero true positives => Weka shows? For precision/F-measurement. This is the most significant modeling issue.

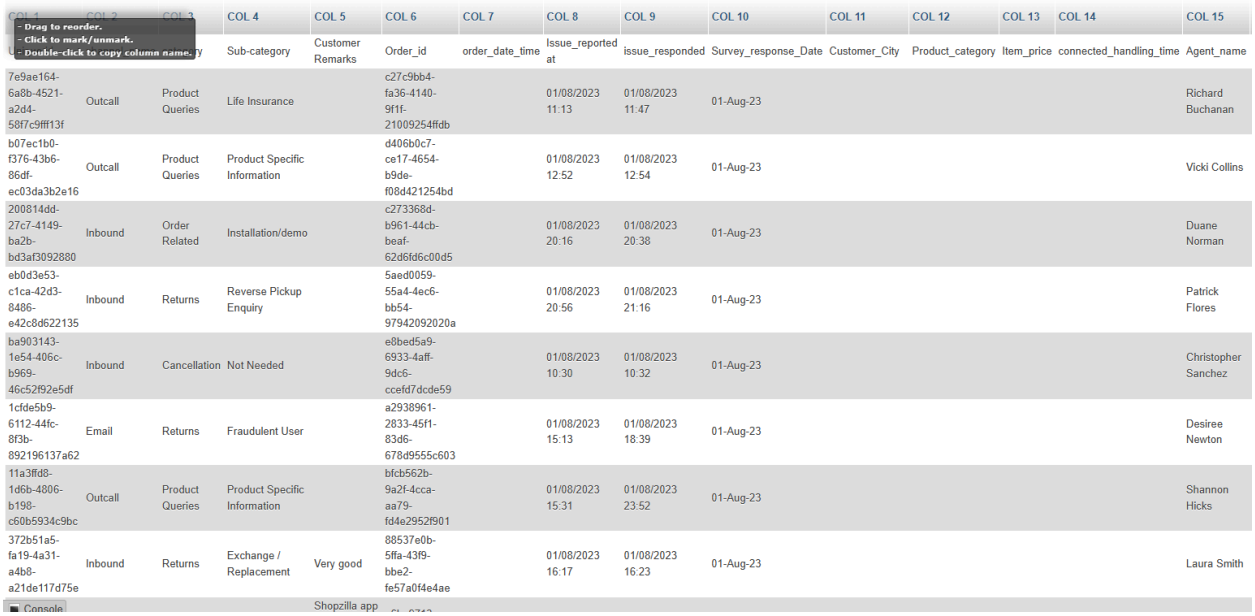
2.There is a lot of missing data in critical fields.
    - order_date_time and Customer_City contain a substantial fraction of missing data (about 17k non-null out of 85k).
    - The Customer Remarks section is generally empty (~28k non-null vs. ~57k absent).
    - There are several missing values in item_price and product_category.
    - connected_handling_time has significantly few non-null rows. These render certain features useless or necessitate active imputation.

3.Small sample sizes after sampling.
    - When balanced incorrectly (or with aggressive sampling), you may wind up with a small training set (e.g., 20 instances), resulting in unstable metrics and... For certain classes. Be sure to keep enough samples per class.

4.Possible mismatched labels or whitespace
    - If CSAT values contain stray whitespace or are typed inconsistently (for example, "5"), Weka may treat them as separate categories — check and normalize.

III. DATA PREPARATION

Your task is to import the dataset into a MySQL DB table.



COL 1	COL 2	COL 3	COL 4	COL 5	COL 6	COL 7	COL 8	COL 9	COL 10	COL 11	COL 12	COL 13	COL 14	COL 15
- Drag to reorder. - Click to mark/unmark. - Double-click to copy column name.		y	Sub-category	Customer Remarks	Order_id	order_date_time	Issue_reported at	issue_responded	Survey_response_Date	Customer_City	Product_category	Item_price	connected_handling_time	Agent_name
7e9ae164-6a8b-4521-a2d4-58f7c9fff13f	Outcall	Product Queries	Life Insurance		c27c9bb4-fa36-4140-9f1f-21009254ffdb		01/08/2023 11:13	01/08/2023 11:47	01-Aug-23					Richard Buchanan
b07ec1b0-f376-43b6-86df-ec03da3b2e16	Outcall	Product Queries	Product Specific Information		d406b0c7-ce17-4654-b9de-f08d421254bd		01/08/2023 12:52	01/08/2023 12:54	01-Aug-23					Vicki Collins
200814dd-27c7-4149-ba2b-bd3af3092880	Inbound	Order Related	Installation/demo		c273368d-b961-44cb-beaf-62d6fd6c00d5		01/08/2023 20:16	01/08/2023 20:38	01-Aug-23					Duane Norman
eb0d3e53-c1ca-42d3-8486-e42c8d622135	Inbound	Returns	Reverse Pickup Enquiry		5aed0059-55a4-4ec6-bb54-97942092020a		01/08/2023 20:56	01/08/2023 21:16	01-Aug-23					Patrick Flores
ba903143-1e54-406c-b969-46c52f92e5df	Inbound	Cancellation	Not Needed		e8bed5a9-6933-4aff-9dc6-ccefd7dcde59		01/08/2023 10:30	01/08/2023 10:32	01-Aug-23					Christopher Sanchez
1cfde5b9-6112-44fc-8f3b-892196137a62	Email	Returns	Fraudulent User		a2938961-2833-45f1-83d6-678d9555c603		01/08/2023 15:13	01/08/2023 18:39	01-Aug-23					Desiree Newton
11a3ffd8-1d6b-4806-b198-c60b5934c9bc	Outcall	Product Queries	Product Specific Information		bfcb562b-9a2f-4cca-aa79-fd4e2952f901		01/08/2023 15:31	01/08/2023 23:52	01-Aug-23					Shannon Hicks
372b51a5-fa19-4a31-a4b8-a21de117d75e	Inbound	Returns	Exchange / Replacement	Very good	88537e0b-5ffa-43f9-bbe2-fe57a0f4e4ae		01/08/2023 16:17	01/08/2023 16:23	01-Aug-23					Laura Smith
Console				Shopzilla app										

Reflection (40 pts):

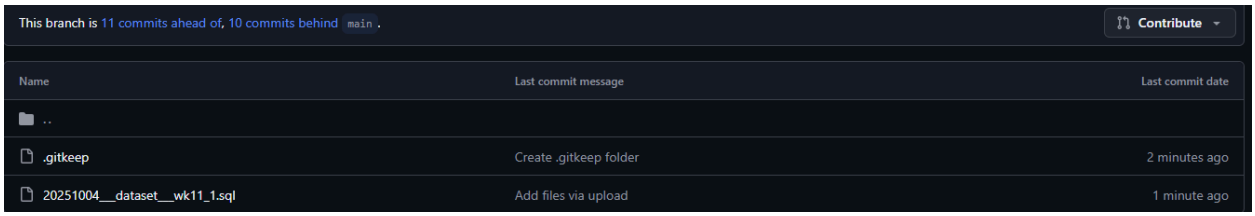
Name: Salvador, Louis Abraham	Date: 10-04-25
Section CIT 401A IT31S1	Instructor: Richard Kho

1) Provide a detailed description of the process you performed on the dataset during the data preparation stage? Was this necessary based on your observations from the “Data Understanding” stage? (30 pts)

- Preprocessing steps were necessary based on my observations in this Data Understanding stage.
- The dataset contained unnecessary attributes, missing or imbalanced values, and inconsistent class distributions that would negatively effect the model performance. Cleaning,Balancing and simplifying the data was it made to be ready for model evaluations and ensure the classifiers to produced valid and interpretable resuts

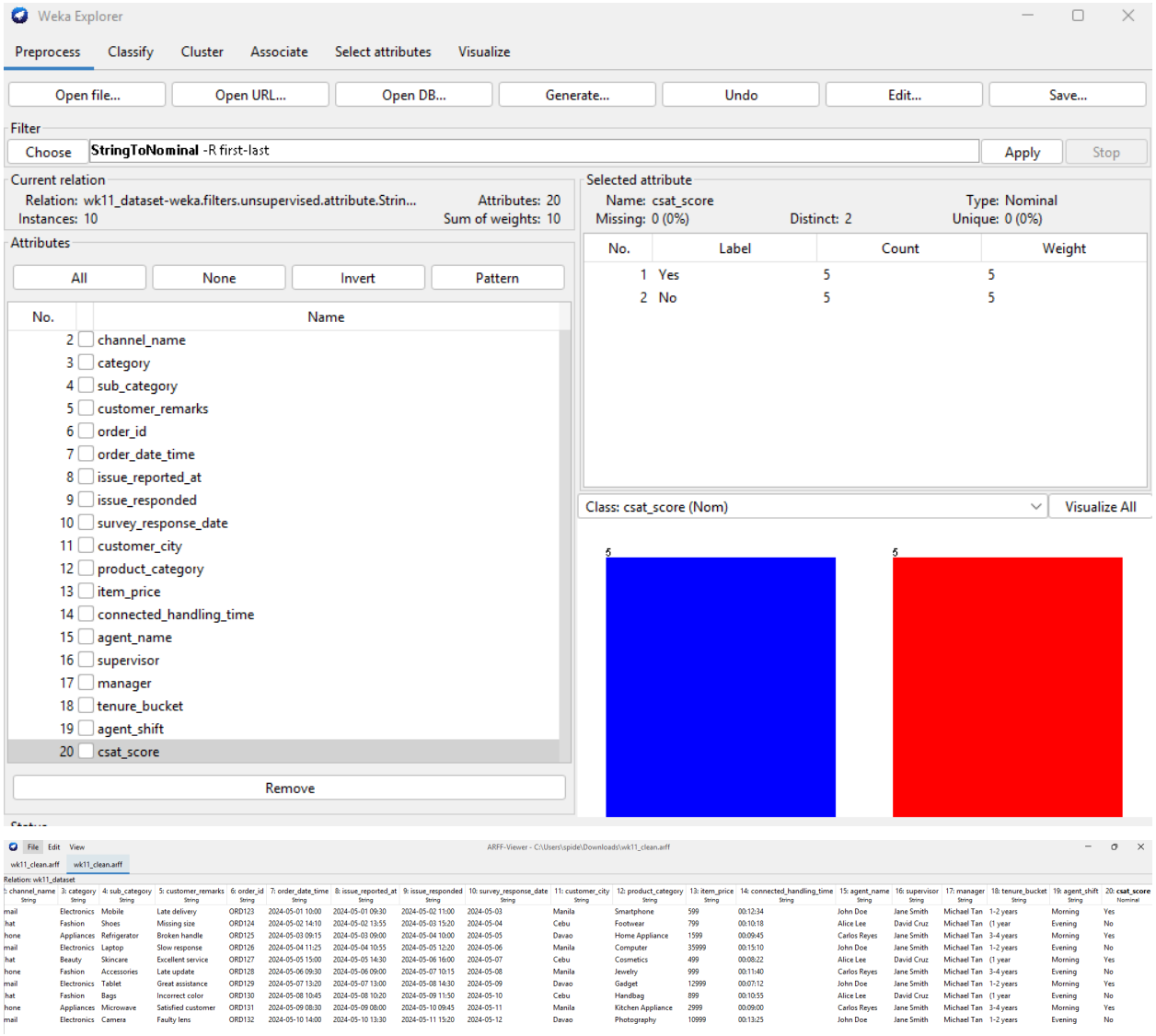
2) Upload your MySQL DB table dump to your own Git branch. (10 pts)

-



Name: Salvador, Louis Abraham	Date: 10-04-25
Section CIT 401A IT31S1	Instructor: Richard Kho

IV. MODELING



Reflection (40 pts):

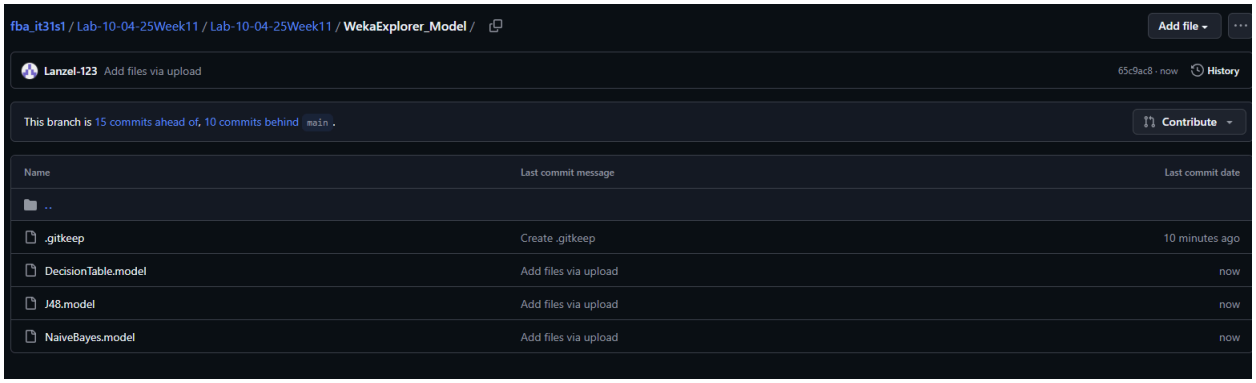
1) Using your top three (3) classification algorithms, complete the table below with your recorded measurements (30 pts):

Classifier	IBK	J48	NaiveBayes
Correctly Classified Instances	6(60%)	10(100%)	8(80%)
Incorrectly	4(40%)	0	2(20%)

Name: Salvador, Louis Abraham	Date: 10-04-25
Section CIT 401A IT31S1	Instructor: Richard Kho

Classified Instances			
Kappa statistics	0.2	1	0.6
Precision (Class: NO)	0.667	1	0.800
Precision (Class: YES)	0.571	1	0.800
Recall (Class: NO)	0.400	1	0.800
Recall (Class: YES)	0.800	1	0.800
F-Measures (Class: NO)	0.500	1	0.800
F-Measures (Class: YES)	0.667	1	0.800

2) Upload the model that achieves the highest measurement to your Git branch. (10 pts)



NOTE:
To build your model, follow these steps:
a. In the “WEKA Explorer” window, under the “Result list” section, select the recorded result with the highest evaluation and right-click.
b. From the options, select “Save model”.
c. In the “Save” window, choose the destination where you want to save the model file (with the .model extension)