final_project

Start Assignment

- Due May 27 by 11:59p.m.
- Points 100
- Submitting a file upload
- File Types html, pdf, ipynb, and py
- Available Apr 29 at 12p.m. - Jun 23 at 11:59p.m.

Final Project

Students will use Python programming and Python libraries such as `pandas`, `altair`, `sklearn` and `numpy` to wrangle and transform the data provided and answer a predictive question about the dataset below.

Students should be creating a full analysis from the beginning (importing data into the Jupyter notebook) to end (communicating their methods and conclusions in a Jupyter notebook).

In the Jupyter notebook, the code cells will read in the dataset, transform the data, create an estimator and visualize the data. Markdown cells will be used throughout the document to narrate the analysis to communicate the question asked, methods used and the conclusion reached.

Data description and important things to know

For this project, students will be using a dataset regarding the different types of Canadian cheeses.

The original data was found on the Government of Canada's Open Government Portal but has unfortunately been taken down.

We have done a bit of wrangling and cleaning for you already and have provided you with a modified version of the dataset.

The data is stored in `cheese_data.csv` which is located in a data folder of the `final_project` directory.

The data were obtain from Kaggle and follows an Open Government Licence (Canada).

The following columns have been included but you are **NOT required to use all of them!** We are including some instructions and suggestions on how to approach certain columns.

Column Name	Instructions
`CheeseId`	Drop this column however, you may also set it as your index if you so desire. This does not contribute to the prediction.
`ManufacturerProvCode`	Apply appropriate transformation.
`ManufacturingTypeEn`	Apply appropriate transformation.
`MoisturePercent`	Apply appropriate transformation.
`FlavourEn`	This column could be transformed with CountVectorizer, however, we acknowledge that this may be above your capabilities. You can drop this column or if you want to challenge yourself, you can leave it in and transform it accordingly.
`CharacteristicsEn`	This column could be transformed with CountVectorizer, however, we acknowledge that this may be above your capabilities. You can drop this column or if you want to challenge yourself, you can leave it in and transform it accordingly.
`Organic`	Apply appropriate transformation.
`CategoryTypeEn`	Apply appropriate transformation.
`MilkTypeEn`	Apply appropriate transformation.
`MilkTreatmentTypeEn`	Apply appropriate transformation.
`RindTypeEn`	We did not use this column in our analysis for convenience.
`CheeseName`	This column could be transformed with CountVectorizer, however, we acknowledge that this may be above your capabilities. You can drop this column or if you want to challenge yourself, you can leave it in and transform it accordingly.
`FatLevel`	A suitable column to use as a prediction target.

Specific requirements for submission

You are required to submit your solution as a jupyter Notebook (.ipynb file) along with its PDF or HTML version.

Imagine that you are interning at a company and have been tasked with working on this problem, with the expectation of sharing your analysis widely. As a result, your .ipynb should include both the code and a narrative to ensure that people unfamiliar with the project can comprehend your analysis. In particular, it should include the following sections.

- **Title** (1 point)
- **Introduction** (5 points)
  - Clearly state the goal of the project, the question you aim to address through your project, and the expected outcome.
  - Briefly provide a context and rationale for the significance of the question you aim to answer.
  - Formulate the question to be solved within the framework of supervised machine learning. Is this problem best suited for classification or regression?
- **Exploratory Data Analysis** (10 points)
  - Read in the data and split it into the train and test sets.
  - Create at least two visualizations to better understand the characteristics of the data.
  - Describe the dataset in your own words. Here are some example questions you may want to address.
    - Discuss any challenges or peculiarities of the dataset.
    - Are there any inconsistencies or errors in the data that need to be addressed?
    - How much data is missing in the dataset? Are there any patterns in the missing data?
    - Are there any obvious relationships between features or between the target and any of the features?
    - What is the distribution of the data? Are the datapoints skewed in some way?
    - Do you need to deal with class imbalance? Explain your answer in your own words.
  - Specify the metrics that will be used to evaluate the success of your project (accuracy, precision, recall, F1-score, ROC-AUC, etc.). Discuss why these metrics are chosen and how they relate to the project's goal.
- **Preprocessing (8 points)**
  - Take the necessary steps to clean and preprocess the data.

    - Identify different types of features from the dataset and define a column transformer to carry out the necessary preprocessing.
    - Briefly justify your choices.
- **Methods & Results (20 points)**
    - Code
        - Begin by training a baseline model.
        - Proceed to train a basic linear model.
        - Explore and select additional suitable supervised machine learning models appropriate for the problem.
        - Conduct feature engineering (creating new features relevant for the prediction task) and feature selection (if applicable) and hyperparameter optimization for one or two promising models.
        - Choose the final model or models based on your chosen evaluation criteria and report the final results on the test set with clarity and detail. Include results for each chosen metric. In case of a classification problem, show confusion matrix and interpret true positives, true negatives, false positives, and false negatives.
    - Writing
        - Explain the rationale behind your feature engineering and feature selection techniques used to choose relevant features.
        - Explain the rationale behind the choice of classification algorithms/models.
        - If multiple models were evaluated, provide a concise comparison of their performance.
        - Discuss why the chosen model outperformed others or met the project's goals better.
- **Discussion** (10 points)
    - Write concluding remarks.
        - Interpret results in the context of project's goals.
        - Relate the findings back to the initial problem statement.
        - If your model provides feature importance scores, present and discuss them. Explain which features had the most influence on predictions.
        - Discuss any limitations of the model or the approach taken.
        - Discuss potential sources of bias, data quality issues, or other factors that might affect the results.
        - Discuss other ideas that you did not try but could potentially improve the performance/interpretability.
- **References (3 points)**
    - Include well-cited and appropriate references for any code or content used from the module or external sources. Each reference should be accurately formatted, providing complete details for readers to locate the original sources easily

Notes:

- *The methods and results sections should be clear, well-organized, and supported with appropriate visualizations and explanations.*
- *All tables and figures should have figure/table numbers and titles.*
- *If you realize that you are repeating a lot of code try to organize it in functions. Clear presentation of your code, experiments, and results is the key to be successful in this project. You may use code from lecture notes or previous lab solutions with appropriate attributions.*

How to Submit?

In order to start your final project, you need to start your own Jupyter server following the instructions below, and submit the end product, with outputs shown, to Canvas.

-- *To make new files* --

1. Go to CANVAS and click on the **JupyterHubN (https://canvas.ubc.ca/courses/143629/external_tools/40165?display=borderless)** tab on the left (or on the given link).
2. A new browser tab will open. Click on "Start My Server"
3. Locate yourself to the "final_project" directory where the data is stored.
4. On the Launcher's tab, to start a new notebook, click Python under the 'Notebooks' section. To write the scripts, choose Text File on the 'Other' section.
5. After completing your work, ensure that you've executed your notebook sequentially from beginning to end, displaying all the output within the notebook. Export your notebook as an ipynb file and as a pdf file.
    1. To save as a pdf, click on File -> Export Notebook As... -> Export Notebook to PDF
    2. To save as an html, click on File -> Export Notebook As... -> Export Notebook to HTML
    3. To save as ipynb, just click on File -> Download

-- *To submit* --

1. Go to Canvas. Click on Assignments and choose 'Final Project'.
2. Upload your files.
3. Click on the submit button.

Note: You can only submit files with .ipynb, .html and .pdf extensions.

Project rubric (revised)

Criteria	Ratings				Pts
Title & Author	**1 pts** **Excellent** Included appropriate title and author name	**0.5 pts** **Good** The title could be clearer and concise or no author name provided.	**0 pts** **Poor** Missing or inappropriate title and missing author name		1 pts
Introduction	**5 pts** **Excellent** The introduction provides a crystal-clear articulation of the project's goal, the specific question to be answered, and the expected outcomes. It offers a comprehensive and compelling context for the question's significance, showcasing a deep understanding of the problem's real-world relevance. It excellently frames the question within the context of supervised machine learning, clearly indicating whether it's a classification or regression problem, and justifying the choice effectively. It is well-written, well-structured, free from jargon or unnecessary technical details and it engages the reader's interest and curiosity, making them eager to continue reading and learn more about the problem and its relevance.	**4 pts** **Good (4 or below)** The introduction effectively states the project's goal and the question to be addressed. It provides a clear context and rationale, effectively explaining why the question is significant and worth addressing. It effectively formulates the question as a supervised machine learning problem, indicating the problem type and providing a reasonable justification. While the introduction is well-structured, there is room for further depth and specificity.	**3 pts** **Satisfactory (3 or below)** The introduction states the project's goal and the question, but it lacks some clarity or conciseness. The expected outcomes are mentioned but may not be fully elaborated. It provides some context and rationale but may lack depth or comprehensive coverage of the problem's significance. It formulates the question as a supervised machine learning problem, but the explanation of problem type and justification may be somewhat vague or lacking in depth. There are some issues with writing and structure.	**2 pts** **Poor (2 or below)** The introduction lacks clarity in stating the project's goal and the question it seeks to answer. The expected outcomes are unclear or missing. It fails to provide meaningful context or rationale for the significance of the question. It does not effectively frame the question within the context of supervised machine learning, or it fails to specify whether it's a classification or regression problem.	5 pts
EDA Exploratory data analysis	**10 pts** **Excellent** The student's EDA is exemplary. They read in the data and successfully split it into train and test sets. Their visualizations are insightful, providing a deep understanding of the data's characteristics. The dataset description is thorough, addressing all provided example questions with exceptional clarity. They identify challenges, inconsistencies, and missing data patterns effectively. The discussion of relationships between features and the target variable is comprehensive, and they assess data distribution and class imbalance expertly. They specify appropriate evaluation metrics, justifying their choices with a clear connection to the project's goals.	**8 pts** **Good (8 or below)** The student's EDA is well-executed. They read in the data and split it into train and test sets appropriately. Their visualizations offer valuable insights, though there may be room for improvement in terms of depth. The dataset description is comprehensive, addressing most example questions with clarity. They identify challenges, inconsistencies, and missing data patterns effectively. The discussion of relationships between features and the target variable is reasonably clear, and they assess data distribution and class imbalance adequately. They specify evaluation metrics that are suitable, with a justified connection to the project's goals.	**6 pts** **Satisfactory (6 or below)** The student's EDA is acceptable. They read in the data and split it into train and test sets as required. Their visualizations provide some understanding of the data, though they may lack depth or context. The dataset description covers most example questions, but some answers may be brief or lack clarity. They identify challenges, inconsistencies, and missing data patterns but may not delve deeply into the analysis. The discussion of relationships between features and the target variable is present but may be somewhat lacking in detail. They assess data distribution and class imbalance to some extent. They specify evaluation metrics, but the justification may be limited or somewhat unclear.	**4 pts** **Poor (4 or below)** The student's EDA falls short of expectations. They may not have properly read in the data or split it into train and test sets. Their visualizations lack meaningful insights or may be missing altogether. The dataset description is incomplete, failing to address several example questions. They may overlook challenges, inconsistencies, or missing data patterns. The discussion of relationships between features and the target variable is lacking or unclear. Data distribution and class imbalance are not effectively assessed. Evaluation metrics may be missing or poorly justified, with no clear connection to the project's goals.	10 pts
Preprocessing Preprocessing of the features	**8 pts** **Excellent** he student's preprocessing is outstanding. They effectively clean and preprocess the data, addressing any inconsistencies, missing values, or outliers. They demonstrate a deep understanding of feature types, correctly identify them, and define a comprehensive column transformer for preprocessing. Their	**6 pts** **Good (6 or below)** The student's preprocessing is well-executed. They successfully clean and preprocess the data, addressing most inconsistencies and missing values. They identify feature types accurately and define a column transformer for preprocessing that covers the essential steps. Their justification for preprocessing choices is	**4 pts** **Satisfactory (4 or below)** The student's preprocessing is acceptable. They make an effort to clean and preprocess the data, but there may be some inconsistencies or missing value issues left unaddressed. They identify feature types, but there could be minor inaccuracies or omissions in their classification. The defined column transformer includes	**2 pts** **Poor (2 or below)** The student's preprocessing falls short of expectations. They may not adequately clean or preprocess the data, leaving significant inconsistencies or missing value problems unhandled. Their identification of feature types may be inaccurate or incomplete. The column transformer definition may be missing or insufficient,	8 pts

Criteria	Ratings				Pts
	justification for preprocessing choices is clear, concise, and provides a solid rationale based on best practices and data characteristics.	reasonable, though there may be minor gaps or areas where further detail would be beneficial.	essential preprocessing steps, but there might be room for improvement or further detail. Their justification for preprocessing choices is present, but it may lack depth or clarity in explaining the rationale.	failing to cover essential preprocessing steps. Their justification for preprocessing choices may be missing or unclear, lacking a solid rationale based on data characteristics or best practices.	
Methods and Results	**20 pts** **Excellent** The student's Methods & Results section is exceptional. They meticulously follow the outlined steps, including training a baseline model, a basic linear model, and exploring multiple suitable machine learning models. They conduct feature engineering and selection effectively, showcasing a deep understanding of feature relevance. They optimize hyperparameters for promising models, providing a clear rationale for their choices. The final model selection process is well-justified and based on chosen evaluation criteria, with results reported comprehensively. In the case of classification, they include a confusion matrix and offer a detailed interpretation of true positives, true negatives, false positives, and false negatives. Writing is	**16 pts** **Good (16 or below)** The student's Methods & Results section is well-executed. They follow the outlined steps, including training baseline and linear models, and explore suitable machine learning models. They conduct feature engineering and selection, demonstrating an understanding of feature relevance. They optimize hyperparameters for promising models and provide reasonable justifications. The final model selection process is adequately justified based on evaluation criteria, with results reported clearly. In the case of classification, they include a confusion matrix and offer a basic interpretation. Writing is clear, though there may be minor gaps in rationale and detail.	**10 pts** **Satisfactory (10 or below)** The student's Methods & Results section is acceptable. They generally follow the outlined steps, including training baseline and linear models and exploring machine learning models. They attempt feature engineering and selection, though some steps may be incomplete or lacking detail. They optimize hyperparameters for promising models and provide basic justifications. The final model selection process is somewhat justified based on evaluation criteria, with results reported but possibly lacking depth. In the case of classification, they include a confusion matrix but may provide limited interpretation. Writing is clear but may lack depth in rationale and detail.	**4 pts** **Poor (4 or below)** The student's Methods & Results section falls short of expectations. They may not adequately follow the outlined steps, leading to missing components or key processes. Feature engineering and selection may be minimal or missing. The optimization of hyperparameters is insufficient or missing justifications. The final model selection process lacks clarity or justification based on evaluation criteria. In the case of classification, they may not include a confusion matrix or offer a meaningful interpretation. Writing may be unclear, lacking rationale, and missing crucial details in feature engineering, model selection, and performance comparison.	20 pts

Criteria	Ratings				Pts
	clear, concise, and provides a strong rationale for feature engineering, model selection, and performance comparison.				
Discussion	**10 pts** **Excellent** The student's Discussion section is exceptional. They provide comprehensive concluding remarks that effectively interpret results in the context of the project's goals and relate findings back to the initial problem statement. They present and thoroughly discuss feature importance scores, offering clear insights into which features had the most influence on predictions. They discuss limitations of the model or approach with exceptional clarity, addressing potential sources of bias, data quality issues, or other factors that might affect results in detail. They explore additional ideas not implemented, explaining how they could potentially improve performance or interpretability with clarity and relevance.	**8 pts** **Good (8 or below)** The student's Discussion section is well-executed. They offer concluding remarks that interpret results and relate them to the project's goals and the initial problem statement. They present and discuss feature importance scores, providing insights into influential features. They discuss limitations and potential sources of bias or data quality issues, though there may be minor gaps or areas where further detail would be beneficial. They briefly explore additional ideas for improvement, offering a basic explanation of their potential impact.	**4 pts** **Satisfactory (4 or below)** The student's Discussion section is acceptable. They provide concluding remarks that attempt to interpret results and relate them to the project's goals and the initial problem statement, though clarity may vary. They mention feature importance scores and discuss them to some extent. They acknowledge limitations and potential sources of bias or data quality issues, but the discussion may be somewhat limited or lacking in depth. They briefly mention other ideas for improvement without extensive explanation.	**2 pts** **Poor (2 or below)** The student's Discussion section falls short of expectations. Their concluding remarks may lack clarity or may not effectively interpret results or relate them to the project's goals and problem statement. They may not present or discuss feature importance scores adequately. Limitations and potential sources of bias or data quality issues may be missing or inadequately addressed. The discussion of additional ideas for improvement may be minimal or lacking in explanation. Writing may be unclear, lacking rationale, and missing crucial details in interpreting results and discussing implications.	10 pts
References	**3 pts** **Excellent** The student's References section is exemplary. They	**2 pts** **Good (2 or below)** The student's References section is acceptable. They make an effort to	**1 pts** **Satisfactory (1 or below)** The student's References section falls short of		3 pts

Criteria	Ratings			Pts
	include well-cited and appropriate references for any code or content used from the module or external sources. Each reference is accurately formatted, providing complete details for readers to locate the original sources easily.	include cited references for code, content, or external sources used in the project. References may contain minor formatting inaccuracies or may not be as comprehensive as they could be. While there is an attempt at proper citation, it may lack some details or consistency.	expectations. They may not include properly cited references for code, content, or external sources, or references may contain significant formatting inaccuracies.	
Mechanics	**3 pts** **Excellent** They follow the provided instructions meticulously and in sequence. They also upload the correct files to Canvas and complete the submission process accurately. The output is shown in all the notebooks.	**2 pts** **Acceptable (2 or below)** The student's submission process is acceptable. They generally follow the provided instructions, but there may be some minor deviations or missed steps. Some of the output is not visible in the ipynb or pdf files.	**1 pts** **Poor (1 or below)** The student submitted files fall short of expectations. They may not have followed the provided instructions accurately. The output is missing in the ipynb and pdf files for many cells.	3 pts
				Total Points: 60