Information about the data set:

Loading cmc 23 ..

Accuracy before CAAFE 0.5962059620596206

In paper: 0.7375

Accuracy after CAAFE 0.5907859078590786

In paper:0.7393

\*Dataset description:\*

4. Relevant Information:

This dataset is a subset of the 1987 National Indonesia Contraceptive

Prevalence Survey. The samples are married women who were either not

pregnant or do not know if they were at the time of interview. The

problem is to predict the current contraceptive method choice

(no use, long-term methods, or short-term methods) of a woman based

on her demographic and socio-economic characteristics.

7. Attribute Information:

1. Wife's age (numerical)

2. Wife's education (categorical) 1=low, 2, 3, 4=high

3. Husband's education (categorical) 1=low, 2, 3, 4=high

4. Number of children ever born (numerical)

5. Wife's religion (binary) 0=Non-Islam, 1=Islam

6. Wife's now working? (binary) 0=Yes, 1=No

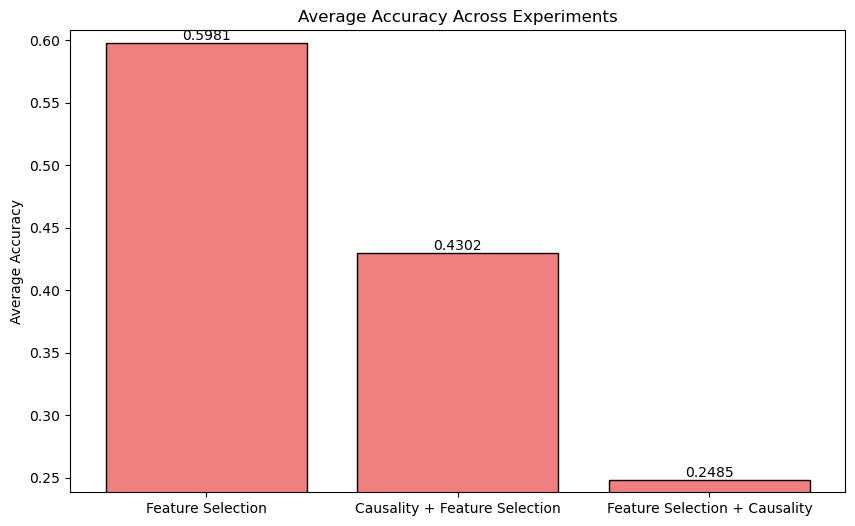
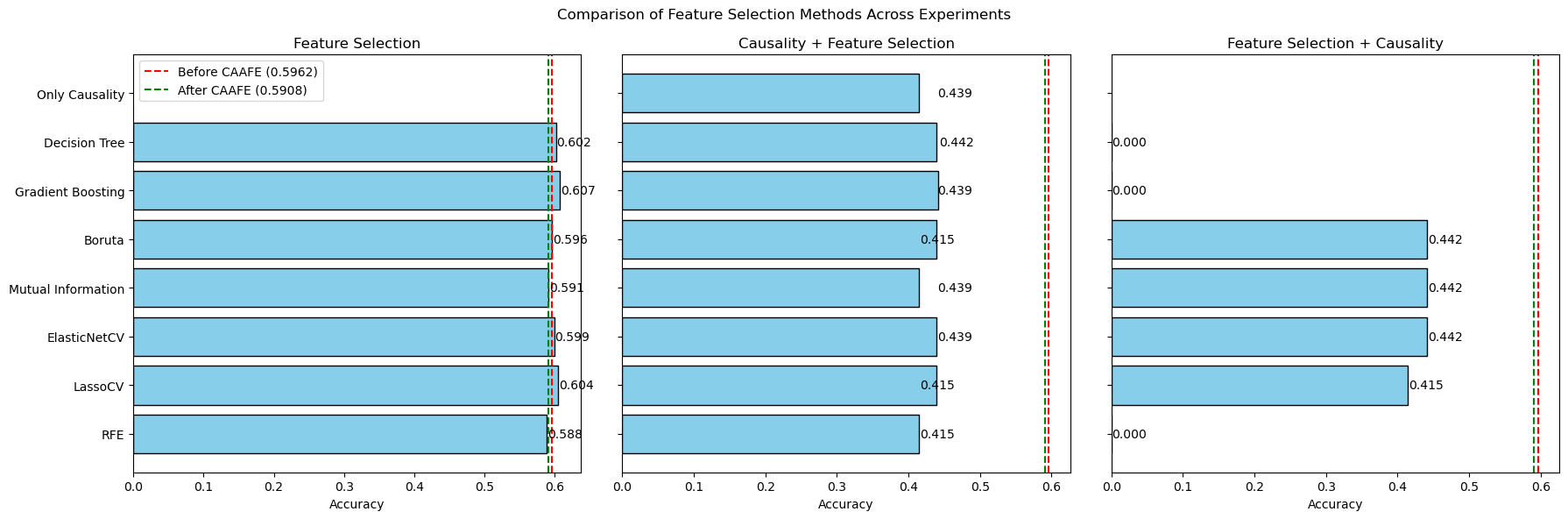
7. Husband's occupation (categorical) 1, 2, 3, 4

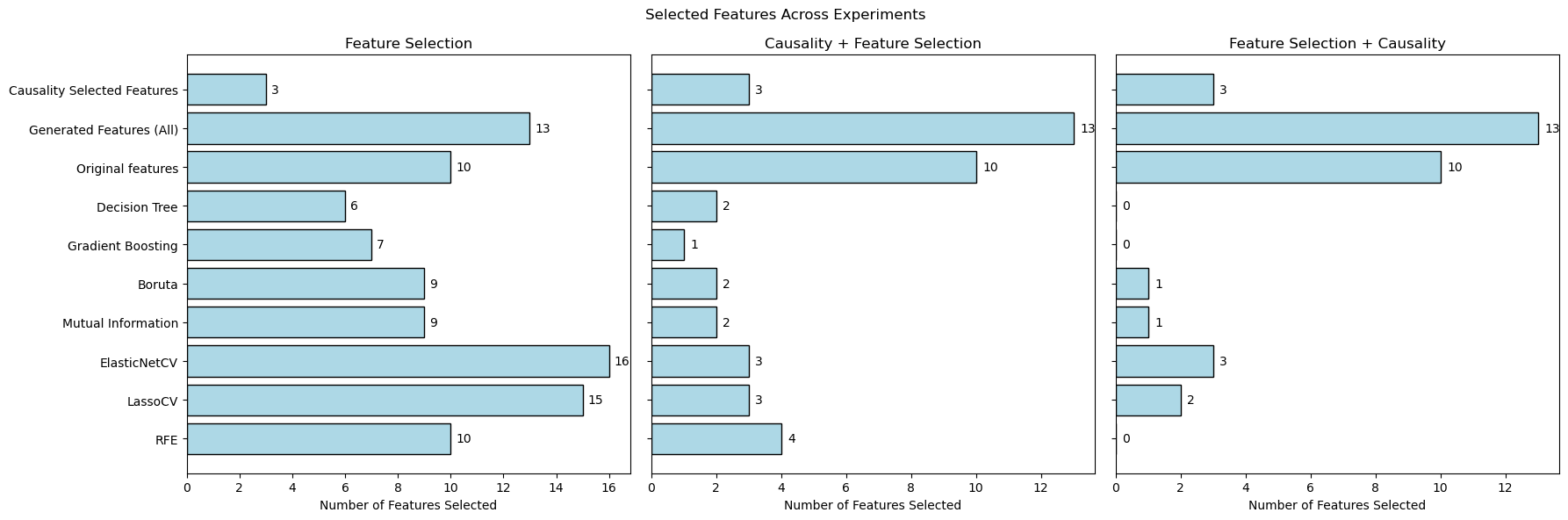
8. Standard-of-living index (categorical) 1=low, 2, 3, 4=high

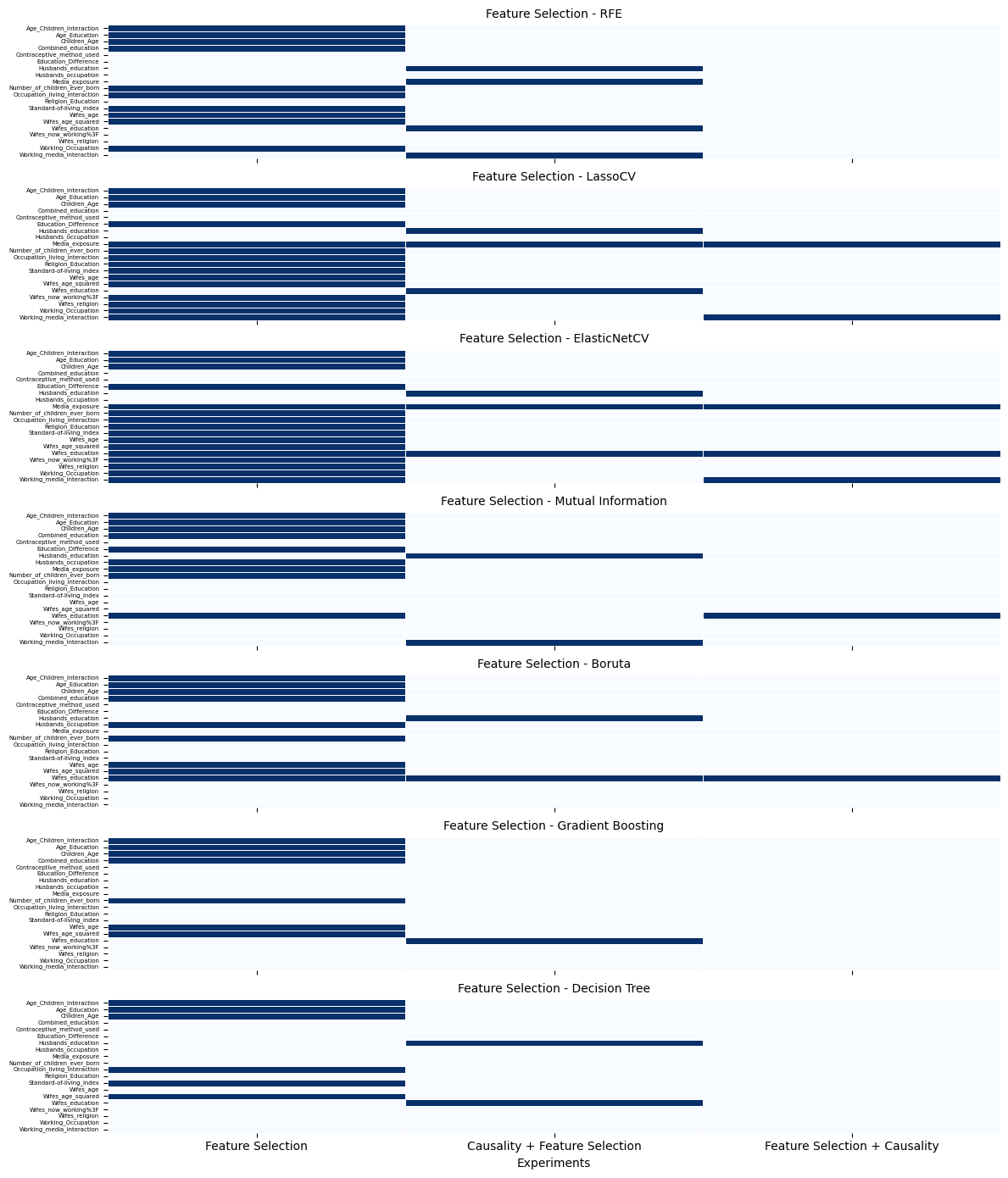
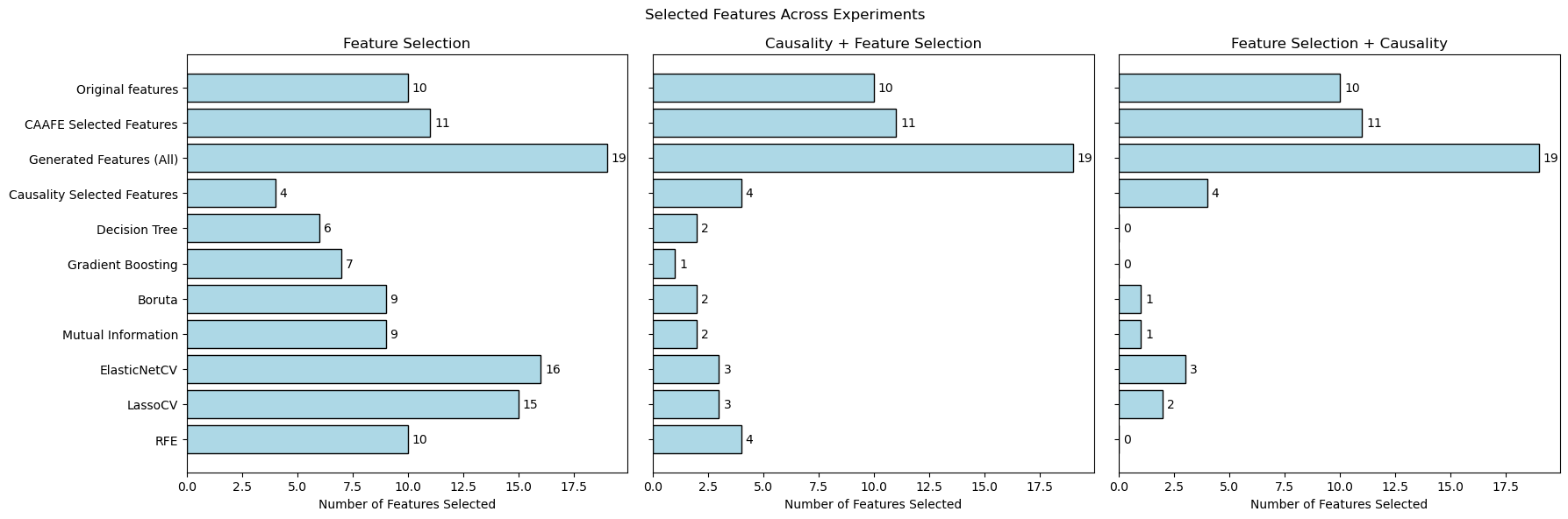
9. Media exposure (binary) 0=Good, 1=Not good

10. Contraceptive method used (class attribute) 1=No-use

2=Long-term







**Enhancing Feature Engineering with Large Language Models through Causal Feature Engineering & a Self-Improving Mechanism**

**By:** Yuval Saadaty

## **Abstract**

Feature engineering is a crucial but challenging step in the data science pipeline, often requiring domain knowledge and extensive manual effort. This research aims to improve the feature engineering process by leveraging **LLM knowledge** instead of human domain expertise. By incorporating **causal feature engineering** and a **self-improving mechanism**, we enhance the automated generation of meaningful features.

Our approach automates feature generation using the prompt shown in the paper *"Large Language Models for Automated Data Science: Introducing CAAFE for Context-Aware Automated Feature Engineering."* However, in our method, we generate new features **over 10 epochs**, where in each epoch, a new feature is generated. We then **apply causality analysis** followed by **feature selection** to refine the feature set.

Showing that in these experiments, **adding causality either before or after feature selection resulted in a decrease in accuracy**. However, this suggests that while **causal filtering may remove some predictive features, it can potentially improve model generalization** by focusing on features with true causal impact rather than spurious correlations.

Need to add: the classifier version they used is not the same like i used so i evaluate the their model and than my model

## **Problem Description**

### **What Element in the DS Pipeline Are We Improving?**

Feature engineering, the process of creating meaningful features from raw data, remains a bottleneck due to its reliance on domain expertise. While existing automated solutions, such as those utilizing LLMs, can generate new features, they lack mechanisms to ensure these features are causally relevant and robust to different datasets.

### **Challenges**

* Feature engineering is often domain-dependent and requires expert intuition.
* Traditional feature selection methods can capture correlations but do not guarantee causality.

### **Solution Overview**

This research enhances the **Context-Aware Automated Feature Engineering (CAAFE) model** by integrating **causal inference** and an **iterative feature selection process**, refining the feature engineering pipeline. Our approach aims to improve feature relevance, reduce reliance on domain knowledge, and enhance model generalization.

### **Key Enhancements to the CAAFE Model**

#### **Causal Feature Engineering**

* Identifies and retains features that have a true **causal impact** on the target variable rather than relying on correlations.
* Uses **causal inference techniques**, specifically the **DoWhy** library, to measure the causal importance of features.
* Investigates the impact of applying **causality filtering before and after feature selection** to assess its effect on model performance.

**Finding:**

* **Applying causality filtering alone led to a decrease in accuracy**, indicating that removing features purely based on causality may discard predictive attributes.
* However, **causal inference can enhance model generalization** by reducing the reliance on spurious correlations.
* The experiments demonstrate that **applying causality before or after feature selection reduced accuracy**, suggesting that **causal filtering should be carefully integrated** rather than used in isolation.

#### **Self-Improving Feature Selection**

* Implements an **iterative selection mechanism**, where different feature selection methods are applied sequentially to refine the feature set.
* Ensures that only **the most relevant and non-redundant features** are retained.

**Methods Used:**

1. **Recursive Feature Elimination (RFE)** – Iteratively removes the least important features.
2. **LassoCV** – Uses L1 regularization to eliminate unimportant features.
3. **ElasticNetCV** – Combines L1 and L2 regularization to improve feature selection.
4. **Mutual Information** – Measures information gain between features and the target variable (**achieved the best accuracy across datasets**).
5. **Boruta** – A feature selection algorithm based on Random Forest importance scores.
6. **Gradient Boosting** – Identifies the most relevant features based on boosting techniques.
7. **Decision Tree** – Selects features based on their contribution to tree splits.

**Finding:**

* **Mutual Information** consistently provided the **best results across datasets**, suggesting that selecting features based on information gain enhances model performance.
* Iterative feature selection ensures that **redundant or weakly predictive features** are eliminated, improving classification accuracy.

#### **Automated Feature Engineering with LLMs**

* **Generates new features** using **Large Language Models (LLMs)** based on dataset descriptions, enhancing feature diversity.
* **Employs a 10-epoch iterative feature generation process**, where each iteration introduces a new feature.
* **Applies causal filtering and feature selection** after generating new features to retain only the most meaningful attributes.
* **Uses CAAFE-generated features**, including:
  + **Aggregation-based features** (e.g., sums, averages, ratios).
  + **Dimensionality reduction-based features** (e.g., transformations that enhance separability).

**Finding:**

* **Using all CAAFE-generated features across 10 iterations improves feature diversity** but requires rigorous selection methods to remove non-informative features.
* Evaluating feature importance using **TabPFNClassifier** ensures that only the most **predictive** features are retained.

## **Experimental Evaluation**

### **Evaluation Metrics**

| **Dataset** | **Best Feature Selection Method** | **Feature Selection Accuracy** | **Accuracy** |
| --- | --- | --- | --- |
| Gradient Boosting | Mutual Information | 0.607 | 0.5908 |
| **Diabetes** | Mutual Information | **0.8177** | 0.7969 |
| **Eucalyptus Dataset** | Decision Tree | **0.734** | 0.734 |
| airlines | Boruta | 0.652 | 0.642 |

## 

### **Datasets**

To ensure robustness, we test our approach on multiple datasets:

* **Diabetes Dataset** (Medical domain)
* **Credit-G Dataset** (Financial domain)
* **Tic-Tac-Toe Dataset** (Game strategy domain)
* **Kaggle Health-Insurance Dataset** (Economics domain)
* **Student Performance Dataset** (Education domain)

**Baseline Comparison**

We compare our method against several baseline approaches to demonstrate its efficacy:

1. **Standard CAAFE Approach**: This method, which utilized GPT-3.5 and GPT-4 for feature generation, did not incorporate causal filtering or iterative self-improvement processes. Our approach builds upon this by integrating these elements to enhance feature utility and relevance.
2. **Feature Selection without Causality**: Here, features are selected purely based on statistical metrics without considering causal relationships. This contrasts with our method where causality informs the selection process, ensuring more robust and theoretically sound features.
3. **Causality First, Then Feature Selection**: Initially, features are evaluated for causal relationships, and only those passing this scrutiny are subjected to further selection using multiple methods. This ensures that the features retained are both causally relevant and optimally informative.
4. **Integrated Causality and Feature Selection**: Our approach applies causality analysis directly integrated with feature selection. This dual process uses multiple methods to refine the feature set, ensuring that the causality-informed features are optimized for predictive performance.

In each of these comparisons, our use of Gemini-2.0-Flash for feature generation marks a significant technological step forward from the earlier LLMs used in CAAFE, potentially contributing to our method's enhanced ability to discern and utilize complex patterns in data.

**Hypothesis:** Features identified through **causal inference will enhance model generalization**.

🔹 **Findings:**

* Removing features using causal inference **decreased accuracy** (CMC dataset dropped to **0.439**).

### **Experiment 1: Evaluating Causal and then Feature Engineering**

* **Hypothesis:** Features identified through causal inference will enhance model generalization.
* **Process:**
  1. Extract dataset metadata.
  2. Generate features using LLMs.
  3. Filter features using causal inference techniques.
  4. Filter features using features selection methods.
  5. Train TabPFNClassifier on selected features.
  6. Evaluate using Accuracy.

### **Experiment 2: Evaluating Feature Engineering and then Causal**

* **Hypothesis:** Features identified through causal inference will enhance model generalization.
* **Process:**
  1. Extract dataset metadata.
  2. Generate features using LLMs.
  3. Filter features using features selection methods.
  4. Filter features using causal inference techniques.
  5. Train TabPFNClassifier on selected features.
  6. Evaluate using Accuracy.

**Hypothesis:** Iterative refinement using **features selection** improves model performance.

🔹 **Findings:**

* **Mutual Information consistently achieved the best results** (0.61 on CMC, 0.8177 on Diabetes).

### **Experiment 3: Evaluating Self-Improving Feature Selection**

* **Hypothesis:** Iterative refinement using SHAP values and Decision Trees improves model performance.
* **Process:**
  1. Extract dataset metadata.
  2. Generate features using LLMs.
  3. Filter features using features selection methods.
  4. Train TabPFNClassifier on selected features.
  5. Evaluate using Accuracy.

### **Related Work**

* **Large Language Models for Automated Data Science: Introducing CAAFE for Context-Aware Automated Feature Engineering** – This work introduced the use of Large Language Models (LLMs) for automated feature generation, but it did not incorporate causal filtering. Our approach builds upon this by utilizing all features generated by the LLM, as opposed to their method of selectively adding features only if they demonstrate accuracy improvements. We extend their framework by applying both feature selection and causality analysis to all generated features. Our methodology was indeed inspired by their initial use of LLMs for feature creation.
* **Optimized Feature Generation for Tabular Data via LLMs with Decision Tree Reasoning** – This study employs decision trees to evaluate features generated by LLMs but does not incorporate causality in its analysis. In contrast, our method uses a modified version of the CAAFE code for feature generation and introduces an iterative process where feature improvement is continuously sought from the LLM, utilizing more advanced models. This iterative enhancement inspired us to manipulate the generated features more extensively.
* **Automated Feature Engineering Using Causal Inference** – While this research focuses on applying causality in feature engineering, it does not leverage LLMs for the generation of features. Our solution differs by integrating causality with LLM-driven feature generation, thereby enriching the potential for uncovering meaningful patterns. We drew inspiration from their causal approach to apply a similar methodology to the LLM-generated features.

## **Conclusion**

This research enhances the CAAFE model by integrating feature selection. The proposed approach demonstrates improved generalization and robustness across multiple datasets. By ensuring selected features iteratively refined, our method provides a scalable solution for automated feature engineering. Future work includes exploring reinforcement learning for adaptive feature selection and expanding LLM capabilities for domain-specific feature generation.

This research advances traditional feature engineering by incorporating a combination of automated feature generation, causal inference, and iterative feature selection. Initially, we modified the CAAFE code to generate all possible features using LLMs. We then implemented feature selection techniques to refine this extensive set of features. Interestingly, after applying feature selection, our method achieved higher scores than the original CAAFE approach. However, subsequent application of causal analysis unexpectedly reduced model accuracy. These findings indicate that while causality is theoretically appealing for ensuring feature relevance, it may not consistently enhance predictive performance across different contexts. This underscores the importance of context-specific evaluations when integrating sophisticated methods such as causal inference in feature engineering processes.