FINAL PROJECT

INSTRUCTOR:
AIGUL B. MIMENBAYEVA

MACHINE LEARNING-BASED PREDICTION OF DEMOGRAPHIC FOOD CONSUMPTION TRENDS

SE - 2327

MARDEN ARUZHAN
SAILAUOVA ULDANA
SHAMIL NARTAY

CONTENTS OF THE THESIS

1. RELEVANCE
2. RESEARCH AIM
3. OBJECTIVES
4. LITERATURE REVIEW
5. PROBLEM STATEMENT: INPUT/OUTPUT
6. COLLECTION, ANALYSIS AND PROCESSING OF INPUT DATA
7. MODELS AND METHODS FOR SOLVING THE PROBLEM
8. RESULTS
9. CONCLUSION

RELEVANCE

IMPACT ON SCIENCE AND SOCIETY

PRACTICAL SIGNIFICANCE

THEORETICAL VALUE

PRACTICAL VALUE

IMPACT ON SCIENCE AND SOCIETY

- PREDICTING FOOD CONSUMPTION TRENDS IS CRUCIAL IN THE CONTEXT OF GLOBAL CHANGES IN LIFESTYLE, DIETARY HABITS, AND DEMOGRAPHIC STRUCTURES.
- MACHINE LEARNING METHODS ALLOW FOR MODELING AND ANALYZING COMPLEX RELATIONSHIPS BETWEEN DEMOGRAPHIC FACTORS (AGE, GENDER, SOCIO-ECONOMIC STATUS) AND FOOD PREFERENCES, OPENING NEW AVENUES FOR RESEARCH AND PRACTICAL APPLICATIONS.

PRACTICAL SIGNIFICANCE

THEORETICAL VALUE

PRACTICAL VALUE

PRACTICAL SIGNIFICANCE

- PREDICTIONS OF FOOD CONSUMPTION TRENDS HELP GOVERNMENTS AND ORGANIZATIONS DEVELOP MORE EFFECTIVE POLICIES IN AREAS LIKE FOOD SECURITY, SUSTAINABLE AGRICULTURE, AND POPULATION NUTRITION.
- THE APPLICATION OF THESE PREDICTIONS IN BUSINESS ALLOWS COMPANIES TO OPTIMIZE PRODUCTION AND DISTRIBUTION OF FOOD PRODUCTS, AS WELL AS TAILOR MARKETING AND PRODUCT ASSORTMENT TO CONSUMER PREFERENCES.

THEORETICAL VALUE

- USING MACHINE LEARNING TO ANALYZE FOOD CONSUMPTION TRENDS CONTRIBUTES TO THE DEVELOPMENT OF NEW STATISTICAL AND MODELING METHODS IN FIELDS LIKE SOCIOLOGY, ECONOMICS, AND FOOD SCIENCE.
- UNDERSTANDING THE RELATIONSHIPS BETWEEN DEMOGRAPHIC CHARACTERISTICS AND CONSUMER BEHAVIOR HELPS DEVELOP NEW THEORIES IN SOCIO-ECONOMIC RESEARCH.

IMPACT ON SCIENCE AND SOCIETY

PRACTICAL SIGNIFICANCE

THEORETICAL VALUE

PRACTICAL VALUE

- IMPROVING PREDICTIONS FOR PRODUCTION PLANNING AND LOGISTICS IN THE FOOD INDUSTRY.
- OPTIMIZING MARKETING STRATEGIES BY CONSIDERING FORECASTED CHANGES IN CONSUMER PREFERENCES AND BEHAVIOR.
- REDUCING RISKS OF FOOD SHORTAGES OR SURPLUSES, CONTRIBUTING TO IMPROVED FOOD SECURITY.

RESEARCH AIM

The aim of this research is to develop a machine learning-based framework for predicting food consumption trends by analyzing the impact of demographic factors, enhancing predictive accuracy, and providing actionable insights for stakeholders in the food industry.

OBJECTIVES

1	Analyze the impact of demographic factors on food consumption trends
2	Develop machine learning models for predicting food consumption
3	Evaluate and compare the performance of the models
4	Identify actionable insights for stakeholders
5	Provide recommendations for sustainable food consumption practices

LITERATURE REVIEW

Publication	Authors and year	Data acquisition method	Dataset size	Model(s) used	Accuracy
• Machine learning prediction of the degree of food processing	• Menichetti G & al., 2023	• Converting legacy data – used existing USDA FNDDS (2009–2010) and SR Legacy databases	• 7,253 foods (FNDDS), 7,793 foods (SR Legacy)	• Random Forest (ensemble × 5)	• AUC: NOVA1 0.9804; NOVA2 0.9632; NOVA3 0.9696; NOVA4 0.9789
• Predicting Unreported Micronutrients From Food Labels: Machine Learning Approach	• Razavi R & Xue G., 2023	• Sharing/exchanging data – obtained USDA FNDDS food label data from public nutrition sources	• 5,624 foods	• RF, Gradient Boosting (GB)	• Classification accuracy ≥0.80; vitamin B12=0.94; phosphorus=0.94; selenium=0.83; regression R^2 from 0.28 to 0.92
• Predicting nutrient composition using deep learning	• Han Y. et al., 2022	• Collecting new data – laboratory-measured nutritional values for Asian food samples	• 2,174 samples (145 food types)	• CNN model	• Accuracy = 91.3%

PROBLEM STATEMENT

The goal of this project is to develop and evaluate predictive models that estimate the Total_Mean value based on demographic, product, and consumption data.
The dataset contains records of food consumption in various countries and years, with multiple attributes describing food type, gender, number of consumers, and related statistical indicators.

The machine learning models are applied to analyze how different socio-demographic and product features affect the average consumption level (Total_Mean), and to predict this value for new combinations of input variables.

Input Data

The input variables correspond to tabular data entries describing food consumption characteristics.
Each record (row) includes the following types of information:

- Country, Year – geographical and temporal context of consumption.
- FoodCode, FoodName, Application – identifiers and description of the food product.
- Gender – demographic attribute of the consumers.
- Number_of_consumers, Consumers_Mean, Consumers_Median, Consumers_SD, Consumers_P95, etc. – statistical measures of consumption distribution.
- Number_of_subjects, Total_Median, Total_SD, Total_P95 – aggregated indicators of total food intake.

All features are numerical or categorical and are preprocessed using:

- imputation of missing values,
- scaling of numeric features,
- ordinal encoding of categorical variables.

Output Data

The output variable of the predictive models is:

- Total_Mean — a continuous numerical target representing the mean total consumption of the given food item under specified demographic and contextual conditions.

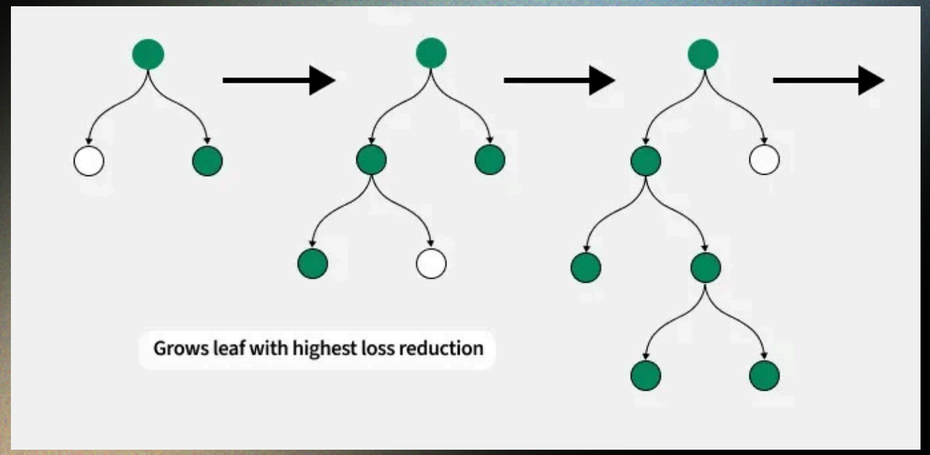
Thus, the machine learning task is supervised regression, where the model learns to predict Total_Mean from the other input attributes.

INITIAL DATASET

fullcifocoss

BW	Country	Year	FoodCode	FoodName	AgeClass	SourceAgeClass	Gender	Number_of_consumers	Consumers_Mean	Consumers_Median	Consumers_P05	Consumers_P90	Consumers_P95	Consumers_P975	Consumers_Standard_deviation	Number_of_subjects	Total_Mean	Total_Median	Total_P05	Total_P90	Total_P95	To
0	China	2002	A000G	Oat grain	All	All	All	1157	60.623	0	8.3333	116.6667	150	166.6667		66172	1.06	0	0		0	
0	China	2002	A000G	Oat grain	All	All	female	608	55.8676	0	8.3333	116.6667	133.3333	158.3333		33953	1.0004	0	0		0	
0	China	2002	A000G	Oat grain	All	All	male	549	65.8895	0	10	133.3333	158.3333	166.6667		32219	1.1227	0	0		0	
0	China	2002	A000N	Buckwheat	All	All	All	167	55.2645	0	8.3333	100	116.6667	158.3333		66172	0.1395	0	0		0	
0	China	2002	A000N	Buckwheat	All	All	female	82	54.7053	0	8.3333	83.3333	116.6667	183.3333		33953	0.1321	0	0		0	
0	China	2002	A000N	Buckwheat	All	All	male	85	55.8039	0	8.3333	100	116.6667	133.3333		32219	0.1472	0	0		0	
0	China	2002	A000P	Barley grains	All	All	All	61	38.5792	0	6.6667	66.6667	83.3333	116.6667		66172		0	0		0	
0	China	2002	A000P	Barley grains	All	All	female	28	37.5	0	3.3333	66.6667	83.3333	100		33953		0	0		0	
0	China	2002	A000P	Barley grains	All	All	male	33	39.4949	0	8.3333	83.3333	116.6667	166.6667		32219		0	0		0	
0	China	2002	A000T	Maize grain	All	All	All	2422	86.9653	0	8.3333	183.3333	358.3333	500		66172	3.1831	0	0		0	
0	China	2002	A000T	Maize grain	All	All	female	1355	81.9742	0	8.3333	166.6667	346.6667	500		33953	3.2714	0	0		0	
0	China	2002	A000T	Maize grain	All	All	male	1067	93.3037	0	11.6667	200	400	533.3333		32219	3.0899	0	0		0	
0	China	2002	A001B	Common millet grain	All	All	All	9069	73.2742	0	10	166.6667	216.6667	250		66172	10.0424	0	0		66.6667	
0	China	2002	A001B	Common millet grain	All	All	female	4740	71.5673	0	10	158.3333	200	243.3333		33953	9.9911	0	0		66.6667	
0	China	2002	A001B	Common millet grain	All	All	male	4329	75.143	0	10	166.6667	216.6667	266.6667		32219	10.0963	0	0		66.6667	
0	China	2002	A001D	Rice grain	All	All	All	57489	238.1073	183.3333	33.3333	450	533.3333	600		66172	206.8632	183.3333	0		516.6667	
0	China	2002	A001D	Rice grain	All	All	female	29564	222.065	166.6667	33.3333	433.3333	500	550		33953	193.3594	166.6667	0		475	
0	China	2002	A001D	Rice grain	All	All	male	27925	255.0912	200	33.3333	490	550	616.6667		32219	221.0938	200	0		550	
0	China	2002	A001K	Rye grain	All	All	All	52	139.5192	0	3.3333	333.3333	750	1000		66172	0.1096	0	0		0	
0	China	2002	A001K	Rye grain	All	All	female	25	148.3333	0	5	500	666.6667	1000		33953	0.1092	0	0		0	
0	China	2002	A001K	Rye grain	All	All	male	27	131.358	0	3.3333	116.6667	750	1833.3333		32219	0.1101	0	0		0	
0	China	2002	A001L	Sorghum grain	All	All	All	258	87.1253	0	16.6667	200	216.6667	266.6667		66172	0.3397	0	0		0	
0	China	2002	A001L	Sorghum grain	All	All	female	141	80.5201	0	16.6667	166.6667	216.6667	258.3333		33953	0.3344	0	0		0	
0	China	2002	A001L	Sorghum grain	All	All	male	117	95.0855	0	16.6667	200	241.6667	350		32219	0.3453	0	0		0	
0	China	2002	A001N	Common wheat grain	All	All	All	365	92.3425	0	8.3333	216.6667	266.6667	300		66172	0.5094	0	0		0	
0	China	2002	A001N	Common wheat grain	All	All	female	197	94.8942	0	8.3333	233.3333	281.6667	308.3333		33953	0.5506	0	0		0	
0	China	2002	A001N	Common wheat grain	All	All	male	168	89.3502	0	8.3333	200	250	266.6667		32219	0.4659	0	0		0	
0	China	2002	A002A	Job's tears grain	All	All	All	74	38.9414	0	6.6667	66.6667	83.3333	166.6667		66172		0	0		0	
0	China	2002	A002A	Job's tears grain	All	All	female	42	36.746	0	3.3333	66.6667	83.3333	83.3333		33953		0	0		0	
0	China	2002	A002A	Job's tears grain	All	All	male	32	41.8229	0	8.3333	66.6667	83.3333	250		32219		0	0		0	
0	China	2002	A002G	Buckwheat flour	All	All	All	85	48.1569	0	16.6667	100	116.6667	125		66172		0	0		0	

LIGHT GRADIENT BOOSTING MACHINE (LIGHTGBM)



$$L(\theta) = \sum_{i=1}^{n} l(y_i, y_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Formula 4. Objective function of gradient boosting

In [...

```
X = df_clean[[
    'Total_P95',
    'Total_P975',
    'Total_Standard_deviation',
    'Total_Median',
    'Consumers_P975',
    'Consumers_P95',
    'Consumers_Standard_deviation',
    'Consumers_Mean'
]]

y = df_clean['Total_Mean']

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

print("Training set shape:", X_train.shape)
print("Testing set shape:", X_test.shape)
```

```
Training set shape: (44398, 8)
Testing set shape: (11100, 8)
```

Model Training

In [...

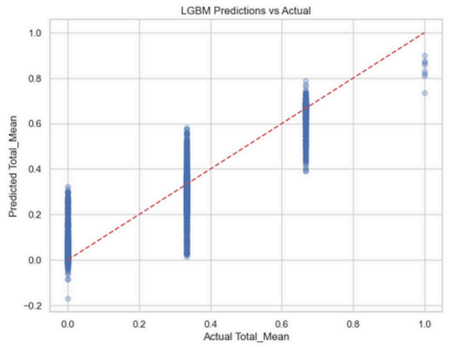
```
lgbm_model = LGBMRegressor(
    n_estimators=500,
    learning_rate=0.05,
    max_depth=7,
    random_state=42
)

lgbm_model.fit(X_train, y_train)
```

```
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing
was 0.001577 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 885
```

In [...

```
y_pred = lgbm_model.predict(X_test)

rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)

print(f"RMSE: {rmse:.4f}")
print(f"R² Score: {r2:.4f}")

import matplotlib.pyplot as plt

plt.figure(figsize=(8,6))
plt.scatter(y_test, y_pred, alpha=0.3)
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'r--')
plt.xlabel("Actual Total_Mean")
plt.ylabel("Predicted Total_Mean")
plt.title("LGBM Predictions vs Actual")
plt.show()
```

```
RMSE: 0.0477
R² Score: 0.9312
```



Save Model

In [...

```
joblib.dump(lgbm_model, "../artifacts/lgbm_model.joblib")
print("Model saved to artifacts/lgbm_model.joblib")
```

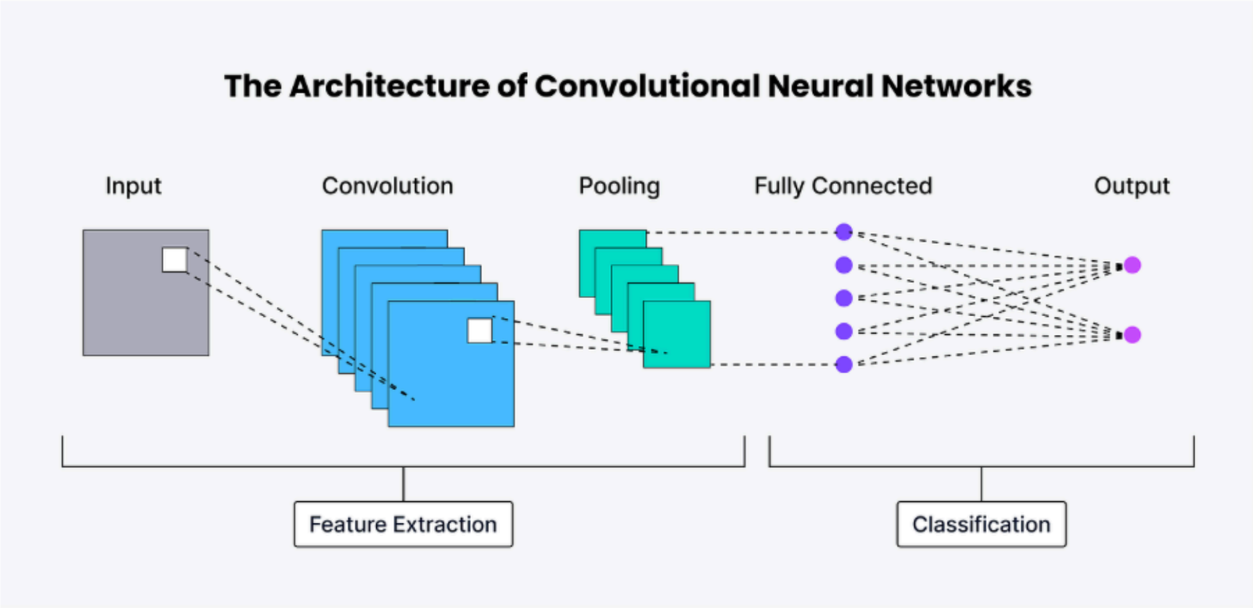
```
Model saved to artifacts/lgbm_model.joblib
```

LightGBM (Light Gradient Boosting Machine) is an ensemble learning algorithm based on decision tree boosting. It builds multiple weak learners (decision trees) sequentially, where each tree attempts to correct the errors made by the previous ones. To stabilize variance and handle skewed data distributions, a logarithmic transformation was applied to the target variable (Total_Mean), which improved regression accuracy for highly variable consumption values.

CONVOLUTIONAL NEURAL NETWORK(CNN)

$$(I * K)(x) = \sum_m I(x - m)K(m)$$

Formula 6. One-dimensional convolution operation



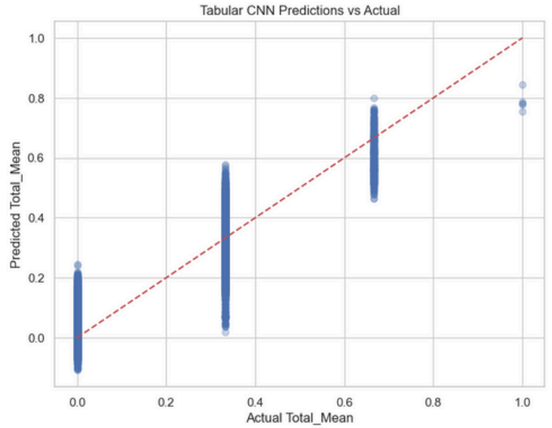
The Convolutional Neural Network (CNN) was adapted for regression on structured tabular data.

The convolutional layer extracts hierarchical feature representations by applying kernels (filters) over the input feature vector.

In [...

```
cnn_model.eval()
y_preds = []

with torch.no_grad():
    for xb, yb in test_loader:
        preds = cnn_model(xb)
        y_preds.append(preds)

y_pred = torch.cat(y_preds, dim=0).numpy()
y_true = y_test.numpy()

rmse = np.sqrt(mean_squared_error(y_true, y_pred))
r2 = r2_score(y_true, y_pred)

print(f"RMSE: {rmse:.4f}")
print(f"R² Score: {r2:.4f}")

plt.figure(figsize=(8,6))
plt.scatter(y_true, y_pred, alpha=0.3)
plt.plot([y_true.min(), y_true.max()], [y_true.min(), y_true.max()], 'r--')
plt.xlabel("Actual Total_Mean")
plt.ylabel("Predicted Total_Mean")
plt.title("Tabular CNN Predictions vs Actual")
plt.show()
```

```
RMSE: 0.0691
R² Score: 0.8546
```



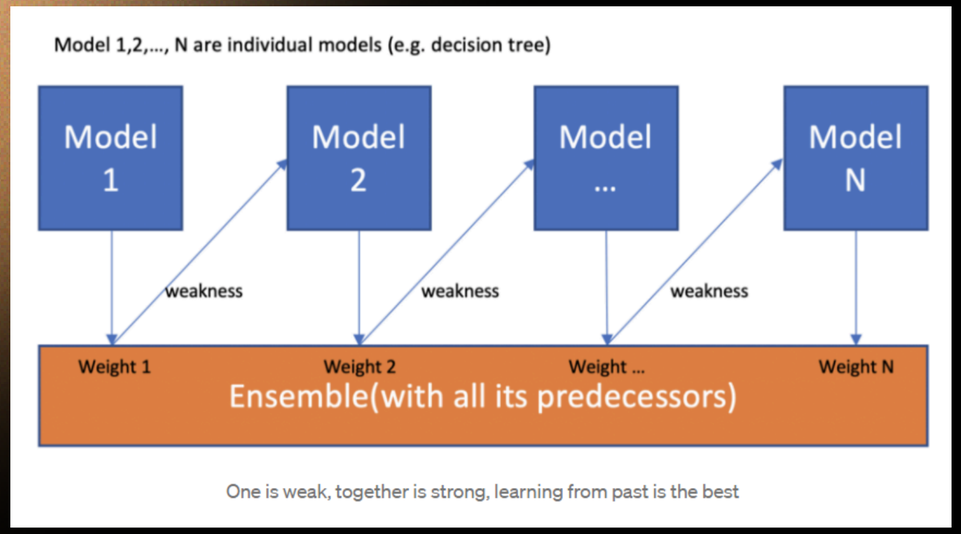
HIST GRADIENT BOOSTING REGRESSOR (HGBR)

$$g_i = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i}$$

Formula 5. Gradient computation in boosting

$$\hat{y}_i^{(m)} = \sum_{t=1}^{m} \eta f_t(x_i)$$

Formula 6. Aggregated prediction of boosted trees

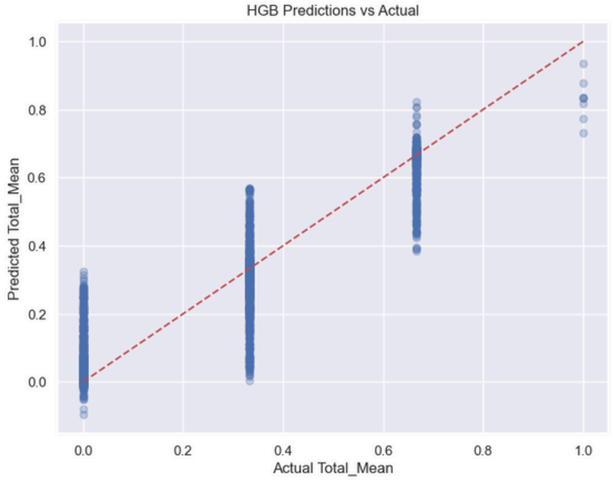


The HistGradientBoostingRegressor (HGBR) accelerates gradient boosting by grouping continuous features into discrete bins (histograms). It reduces computation time and memory usage while maintaining model accuracy.

In [...

```
y_pred = hgb_model.predict(X_test)

rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)

print(f"RMSE: {rmse:.4f}")
print(f"R² Score: {r2:.4f}")

plt.figure(figsize=(8,6))
plt.scatter(y_test, y_pred, alpha=0.3)
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'r--')
plt.xlabel("Actual Total_Mean")
plt.ylabel("Predicted Total_Mean")
plt.title("HGB Predictions vs Actual")
plt.show()
```

```
RMSE: 0.0480
R² Score: 0.9303
```



META MODEL

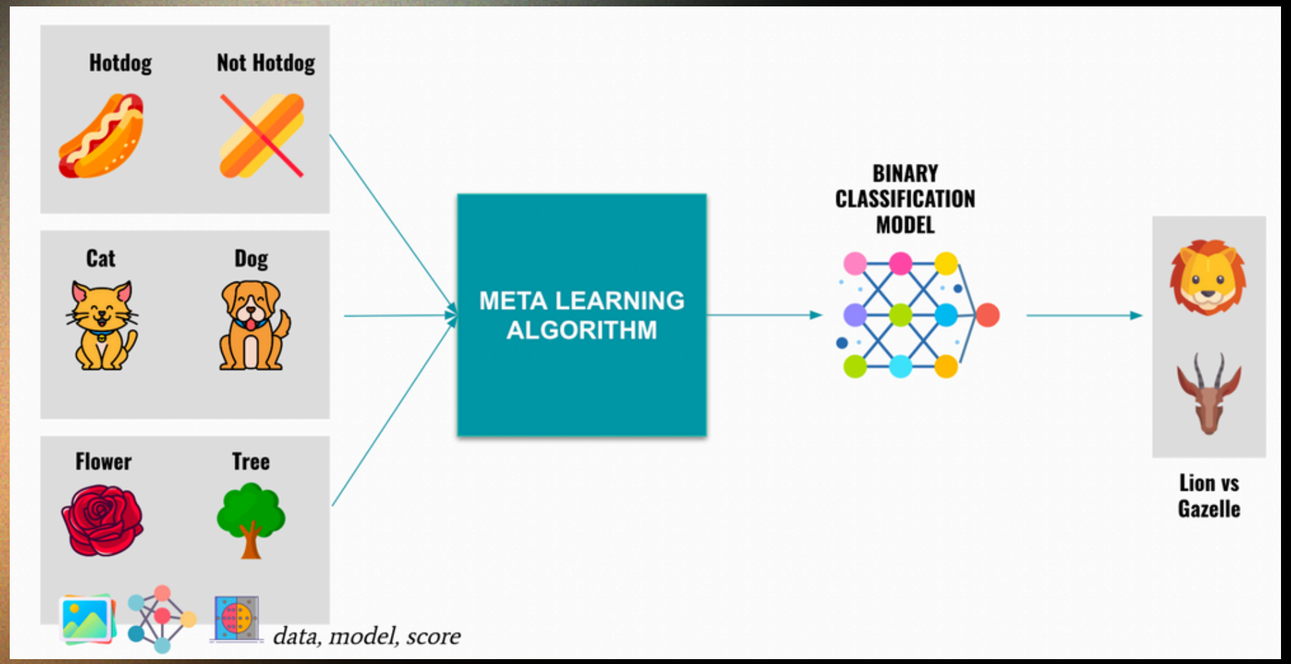
Meta-learning is a higher-level learning paradigm in which a model does not only learn from data but also learns how to learn. Instead of training a single model on one fixed task, meta-learning aims to extract general learning strategies from multiple related tasks so that new tasks can be learned faster and with fewer data.

Meta Learning $L(F) = \sum_{n=1}^{N} l^n$

N tasks

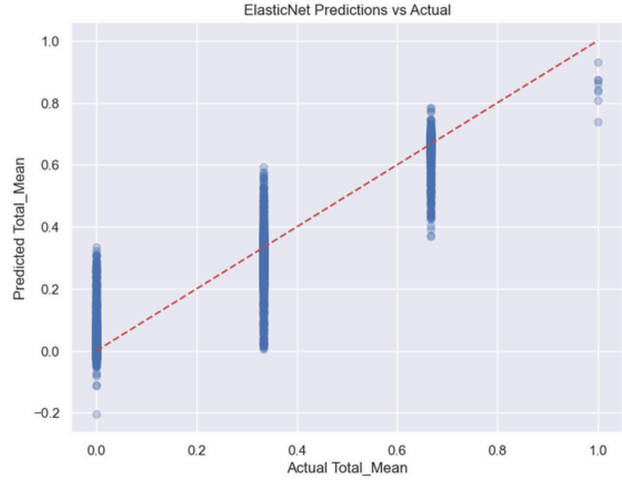
Testing loss for task n after training

- Defining the goodness of a function F




```
In [...
y_pred = meta_model.predict(test_meta)

rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)

print(f"RMSE: {rmse:.4f}")
print(f"R² Score: {r2:.4f}")

plt.figure(figsize=(8,6))
plt.scatter(y_test, y_pred, alpha=0.3)
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'r--')
plt.xlabel("Actual Total_Mean")
plt.ylabel("Predicted Total_Mean")
plt.title("Stacked Model Predictions vs Actual")
plt.show()
```

```
RMSE: 0.0479
R² Score: 0.9304
```



RESULTS

HGB Model:

RMSE: 0.0480
R^2 Score: 0.9303

LGBM Model:

RMSE: 0.0477
R^2 Score: 0.9312

CNN Model:

RMSE: 0.0691
R^2 Score: 0.8546

Meta Model:

RMSE: 0.0479
R^2 Score: 0.9304

CONCLUSION

This study demonstrated that machine-learning methods can effectively predict demographic food-consumption trends using heterogeneous tabular datasets. Among all tested models, LightGBM achieved the highest predictive accuracy, while HistGradientBoostingRegressor showed similarly strong and stable performance. The 1D-CNN captured additional nonlinear relationships but remained less suitable for non-spatial tabular data. The meta-ensemble model further stabilized predictions by integrating outputs from multiple learners. Overall, the models explained approximately 93% of the variability in consumption indicators, confirming that data-driven approaches are highly suitable for forecasting demographic food-consumption patterns and supporting analytical decision-making.

THANK YOU

2025