Sistema Intelligente di Analisi Ricette e Ingredienti

Basato su Machine Learning e Knowledge Base Prolog

Progetto ICON – Ingegneria della Conoscenza

2 settembre 2025

Sommario

Questo documento presenta lo sviluppo e l'implementazione di un sistema intelligente per l'analisi e la gestione di ricette culinarie e ingredienti. Il sistema utilizza un approccio ibrido che combina tecniche di Machine Learning (clustering K-Means, classificazione supervisionata, regressione), analisi predittiva delle calorie e una Knowledge Base implementata in Prolog per la rappresentazione della conoscenza culinaria. L'architettura integra algoritmi di clustering per raggruppare ricette simili (K=6 cluster semanticamente coerenti), modelli di classificazione ad alta performance (SVM con 96.25% nested CV accuracy) e modelli di regressione per la predizione delle calorie (SVR con $R^2 = 0.4855$). Il sistema include inoltre un motore di inferenza Prolog per query semantiche complesse sul dominio culinario.

Indice

1 Introduzione

Nell'era digitale, la gestione intelligente delle informazioni culinarie rappresenta una sfida multidisciplinare che coinvolge l'elaborazione di dati strutturati e non strutturati, l'estrazione di conoscenza da dataset eterogenei e la rappresentazione formale di relazioni semantiche complesse. Questo documento presenta un sistema innovativo per l'analisi automatica di ricette culinarie e ingredienti che integra tecniche avanzate di Machine Learning con rappresentazione simbolica della conoscenza.

Il sistema sviluppato affronta tre problematiche principali nel dominio culinario: la gestione di query semantiche complesse attraverso un sistema di inferenza logica implementato in Prolog, la predizione accurata del contenuto calorico basata su ingredienti e metodi di preparazione, e la classificazione automatica di ricette in gruppi omogenei basata su caratteristiche nutrizionali e procedurali.

1.1 Obiettivi del Progetto

Gli obiettivi principali di questo progetto sono:

1. **Knowledge Base integrata**: Implementare una rappresentazione formale della conoscenza culinaria in Prolog per supportare query semantiche complesse e ragionamento logico

2. **Predizione accurata delle calorie**: Sviluppare modelli di regressione robusti per stimare il contenuto calorico basato su ingredienti, porzioni e metodi di cottura

3. **Sistema di clustering semantico**: Implementare algoritmi K-Means ottimizzati per identificare automaticamente gruppi di ricette con caratteristiche nutrizionali e procedurali simili

4. **Classificazione supervisionata**: Sviluppare modelli di classificazione per predire l'appartenenza di nuove ricette ai cluster identificati

5. **Interfaccia CLI intuitiva**: Fornire un'interfaccia a linea di comando per l'interazione con tutti i componenti del sistema

6. **Validazione sperimentale**: Implementare protocolli di valutazione rigorosi per misurare performance e accuratezza dei modelli

7. **Modularità e estensibilità**: Progettare un'architettura modulare per facilitare manutenzione ed estensioni future

1.2 Struttura del Sistema

Il sistema implementa un'architettura modulare composta da otto componenti principali:

1. **Data Loading Pipeline**: Caricamento, validazione e preprocessing di dataset culinari strutturati

2. **Feature Engineering Module**: Estrazione e trasformazione di features numeriche e categoriche

3. **Clustering Analysis Engine**: Implementazione K-Means con ottimizzazione automatica del numero di cluster
4. **Classification System**: Sistema di classificazione multi-algoritmo per predizione dei cluster
5. **Regression Models**: Modelli di regressione specializzati per predizione calorica
6. **Prolog Knowledge Base**: Rappresentazione formale della conoscenza culinaria e motore di inferenza
7. **Query Processing Engine**: Sistema per l'elaborazione di query naturali e semantiche
8. **CLI Interface**: Interfaccia utente unificata per accesso a tutte le funzionalità

2 Dataset e Preprocessing

2.1 Descrizione dei Dataset

Il progetto utilizza due dataset principali rappresentanti il dominio culinario:

2.1.1 Dataset Ricette (ricette_reali.csv)

Il dataset delle ricette contiene informazioni dettagliate su preparation culinarie reali, con metadati nutrizionali e procedurali.

Campo	Descrizione	Tipo
nome_ricetta	Nome identificativo della ricetta	String
tipo_cucina	Origine geografica (italiana, francese, etc.)	Categorical
difficolta	Livello di complessità (facile, medio, difficile)	Ordinal
tempo_preparazione_min	Tempo di preparazione in minuti	Numeric
tempo_cottura_min	Tempo di cottura in minuti	Numeric
numero_porzioni	Numero di porzioni prodotte	Numeric
calorie_per_porzione	Contenuto calorico stimato per porzione	Numeric
costo_stimato_euro	Costo stimato in euro	Numeric
tipo_piatto	Categoria del piatto (primo, secondo, dolce)	Categorical
metodo_cottura	Tecnica principale (forno, padella, bollito)	Categorical
stagionalita	Stagione ottimale di consumo	Categorical
rating_medio	Valutazione media degli utenti	Numeric

Tabella 1: Struttura del dataset ricette

2.1.2 Dataset Ingredienti (ingredienti_reali.csv)

Il dataset degli ingredienti fornisce informazioni nutrizionali dettagliate per ogni componente utilizzato nelle ricette.

Campo	Descrizione	Tipo
nome_ingrediente	Nome standardizzato dell'ingrediente	String
categoria	Gruppo alimentare di appartenenza	Categorical
calorie_per_100g	Densità calorica per 100 grammi	Numeric
proteine_g	Contenuto proteico in grammi	Numeric
carboidrati_g	Contenuto di carboidrati in grammi	Numeric
grassi_g	Contenuto lipidico in grammi	Numeric
fibre_g	Contenuto di fibre in grammi	Numeric
prezzo_kg_euro	Prezzo stimato al chilogrammo	Numeric
stagionalita	Disponibilità stagionale	Categorical
conservazione	Modalità di conservazione ottimale	Categorical

Tabella 2: Struttura del dataset ingredienti

2.2 Introduzione ai Dataset

I dataset utilizzati in questo progetto sono stati creati raccogliendo informazioni da fonti online affidabili. Questi dati rappresentano una base fondamentale per l'analisi e la modellazione, fornendo dettagli completi su ricette e ingredienti. La loro costruzione ha richiesto un'attenta selezione e validazione per garantire la qualità e la coerenza delle informazioni.

2.3 Pipeline di Preprocessing

2.3.1 Validazione e Pulizia Dati

Il preprocessing implementa controlli di qualità multi-livello attraverso:

- Controllo e gestione dei valori mancanti
- Rimozione di outliers basata sul metodo IQR (Interquartile Range)
- Validazione della consistenza dei dati tra dataset
- Normalizzazione delle scale numeriche

2.3.2 Feature Engineering

Variabili Derivate

Il sistema genera automaticamente feature aggiuntive per migliorare la capacità predittiva:

- **Tempo totale**: Somma di tempo di preparazione e cottura
- **Costo per porzione**: Rapporto tra costo totale e numero di porzioni
- **Densità calorica**: Rapporto tra calorie e numero di ingredienti

- **Complessità temporale**: Rapporto tra tempo totale e numero di ingredienti

Encoding Variabili Categoriche
Le variabili categoriche vengono codificate utilizzando strategie appropriate al loro tipo:

- **Encoding ordinale**: Per variabili con ordine naturale (difficoltà: facile < medio < difficile)
- **Label Encoding**: Per variabili nominali (tipo cucina, metodo cottura)
- **Persistenza degli encoder**: Salvataggio per applicazione consistente su nuovi dati

3 Metodologia

3.1 Programmazione Logica e Knowledge Base

Il sistema implementa una base di conoscenza utilizzando Prolog tramite la libreria PySwip per l'integrazione con Python. La KB è strutturata come un sistema esperto che codifica relazioni semantiche tra ricette, ingredienti e proprietà nutrizionali.

3.1.1 Struttura della Knowledge Base

La rappresentazione della conoscenza è organizzata attraverso predicati Prolog che modellano:

- `ricetta(Nome, ListaIngredienti, Calorie, Categoria)`: Definisce ricette complete con tutti i metadati
- `ingrediente(Nome, CaloriePer100g, Categoria)`: Proprietà nutrizionali degli ingredienti
- `contiene(Ricetta, Ingrediente, Quantita)`: Relazioni quantitative ingrediente-ricetta
- `categoria_ricetta(Ricetta, Categoria)`: Classificazione categorica delle ricette

3.1.2 Sistema di Query Semantiche

Il motore di inferenza supporta query complesse attraverso regole logiche che permettono:

- Calcolo automatico delle calorie totali per ricetta
- Identificazione di ricette compatibili con restrizioni dietetiche
- Analisi nutrizionale avanzata per ricette bilanciate
- Ricerca semantica basata su caratteristiche multiple

3.2 Modelli di Regressione

Il sistema implementa un framework completo di regressione per la predizione del contenuto calorico delle ricette, utilizzando un approccio multi-algoritmo con validazione incrociata annidata.

3.2.1 Architettura del Sistema di Regressione

La pipeline di regressione integra diversi modelli:

- **Support Vector Regression (SVR)**: Con kernel RBF per relazioni non-lineari
- **Random Forest Regressor**: Ensemble di alberi per robustezza
- **Ridge Regression**: Regressione lineare con regolarizzazione L2

Ogni modello viene ottimizzato tramite Grid Search con parametri specifici per massimizzare le performance predittive.

3.2.2 Feature Engineering per Regressione

Il preprocessing dei dati include strategie specifiche per la predizione calorica:

- **Encoding Categorico**: Label encoding per categorie alimentari
- **Normalizzazione**: StandardScaler per features numeriche
- **Feature Selection**: Analisi di correlazione per riduzione dimensionalità
- **Cross-validation**: Validazione annidata 5-fold per selezione modelli

3.3 Analisi di Clustering

3.3.1 Algoritmo K-Means Ottimizzato

L'implementazione del clustering utilizza K-Means con ottimizzazioni per robustezza e determinazione automatica del numero ottimale di cluster.

Determinazione del K Ottimale

Il sistema implementa l'Elbow Method con esecuzioni multiple per garantire robustezza:

- Esecuzione multipla (5 iterazioni) per ogni valore di K
- Calcolo del punto di gomito usando derivata seconda
- Normalizzazione delle inerzie per stabilità numerica
- Selezione automatica del K ottimale basata su criteri statistici

3.3.2 Analisi e Interpretazione dei Cluster

Il sistema fornisce analisi approfondite dei cluster identificati attraverso statistiche descrittive e visualizzazioni che includono:

- Distribuzione percentuale dei campioni per cluster
- Statistiche descrittive (media, mediana, deviazione standard) per feature numeriche
- Analisi delle caratteristiche categoriche dominanti per cluster
- Interpretazione semantica basata su tipo di piatto e difficoltà

3.4 Modelli di Classificazione

3.4.1 Algoritmi Implementati

Il sistema implementa tre algoritmi di classificazione ottimizzati tramite Grid Search:

1. **Random Forest**: Ensemble di alberi decisionali con bagging
2. **Logistic Regression**: Classificazione lineare con regolarizzazione
3. **Support Vector Machine**: Classificazione con kernel RBF

Ogni modello viene ottimizzato attraverso Grid Search per trovare la migliore combinazione di iperparametri, utilizzando cross-validation per evitare overfitting.

3.4.2 Validazione Cross-Fold

L'implementazione utilizza Nested Cross-Validation per una stima non distorta delle performance:

- **Outer Cross-Validation**: Per stima performance non distorta (3-fold)
- **Inner Cross-Validation**: Per ottimizzazione iperparametri (Grid Search)
- **Metriche multiple**: Accuracy, Precision, Recall, F1-Score
- **Validazione robusta**: Media e deviazione standard delle performance

3.5 Modelli di Regressione

3.5.1 Predizione delle Calorie

Il sistema implementa modelli di regressione specializzati per la predizione del contenuto calorico delle ricette.

Algoritmi Implementati

1. **Random Forest Regressor**: Ensemble non-parametrico robusto a outliers
2. **Support Vector Regression**: Regressione con kernel RBF per relazioni non-lineari
3. **Ridge Regression**: Regressione lineare con regolarizzazione L2

4. **Linear Regression**: Baseline lineare per confronto

Ogni modello viene ottimizzato tramite Grid Search per trovare i parametri ottimali che massimizzano le performance di predizione calorica.

3.5.2 Feature Selection e Importance

Il sistema analizza l'importanza delle feature per la predizione calorica attraverso:

- **Feature Importance**: Analisi dell'importanza relativa delle variabili (Random Forest)
- **Ranking automatico**: Ordinamento delle feature per rilevanza predittiva
- **Visualizzazione**: Grafici per interpretazione dei fattori più influenti
- **Selezione**: Identificazione delle variabili più significative per il modello

3.5.3 Predizione Calorica degli Ingredienti

Il sistema implementa un modulo specializzato per la predizione del contenuto calorico degli ingredienti singoli, utilizzando le loro proprietà nutrizionali intrinseche.

Approccio Metodologico

La predizione calorica degli ingredienti si basa su un approccio diverso rispetto alle ricette, sfruttando le relazioni dirette tra composizione nutrizionale e densità energetica:

- **Feature nutrizionali**: Utilizzo di proteine, carboidrati, grassi, fibre come predittori primari
- **Relazioni biochimiche**: Sfruttamento delle equivalenze energetiche note (4 kcal/g per proteine e carboidrati, 9 kcal/g per grassi)
- **Fattori categorici**: Integrazione di categoria alimentare e modalità di conservazione
- **Variabili economiche**: Considerazione del prezzo come proxy della qualità nutrizionale

Architettura del Modello

Il sistema utilizza la stessa pipeline multi-algoritmo delle ricette ma con feature engineering specifico:

1. **Random Forest Regressor**: Ottimizzato per catturare interazioni non-lineari tra macronutrienti
2. **Support Vector Regression**: Con kernel RBF per modellare relazioni complesse tra composizione e calorie
3. **Ridge Regression**: Come baseline lineare per validare la necessità di modelli non-lineari

Validazione e Performance

La validazione segue lo stesso protocollo rigoroso utilizzato per le ricette:

- **Cross-validation annidata**: 5-fold outer per stima performance, 3-fold inner per ottimizzazione
- **Metriche specifiche**: MSE, RMSE, MAE, Rš calibrati per la densità calorica degli ingredienti
- **Analisi degli outliers**: Identificazione di ingredienti con comportamento calorico anomalo
- **Interpretabilità**: Feature importance per comprendere i fattori nutrizionali più predittivi

4 Implementazione della Knowledge Base

4.1 Architettura Prolog

Il sistema integra una Knowledge Base implementata in Prolog per la rappresentazione formale della conoscenza culinaria e l'esecuzione di query semantiche complesse.

4.1.1 Struttura della Knowledge Base

La KB è organizzata in predicati principali che rappresentano entità e relazioni del dominio culinario attraverso:

- **Predicati per ingredienti**: Nome, categoria, valori nutrizionali, prezzo, stagionalità
- **Predicati per ricette**: Metadati completi inclusi difficoltà, tempi, costi, rating
- **Relazioni derivate**: Compatibilità dietetiche, stagionalità ottimale, economicità
- **Regole di inferenza**: Per classificazione automatica e raccomandazioni

4.1.2 Motore di Inferenza

Il sistema implementa un motore di inferenza personalizzato che integra PySwip per l'esecuzione di query Prolog attraverso:

- **Inizializzazione automatica**: Setup del motore Prolog con gestione errori
- **Caricamento fatti**: Conversione automatica da DataFrame a predicati Prolog
- **Esecuzione query**: Interfaccia unificata per query complesse con risultati strutturati
- **Gestione errori**: Sistema robusto per handling di eccezioni e validazione input

4.1.3 Query Builder

Un sistema di costruzione automatica di query facilita l'interazione con la KB attraverso metodi predefiniti per:

- Ricerca per difficoltà (facile, medio, difficile)
- Filtraggio per categoria ingredienti
- Identificazione ricette a basso contenuto calorico
- Ricerca per stagionalità
- Filtraggio per budget (ricette economiche)
- Ricerca ricette veloci per tempo di preparazione

5 Risultati Sperimentali

5.1 Performance del Clustering

5.1.1 Determinazione del Numero Ottimale di Cluster

L'applicazione dell'Elbow Method al dataset delle ricette ha identificato K=6 come numero ottimale di cluster, come mostrato nella Figura 1.

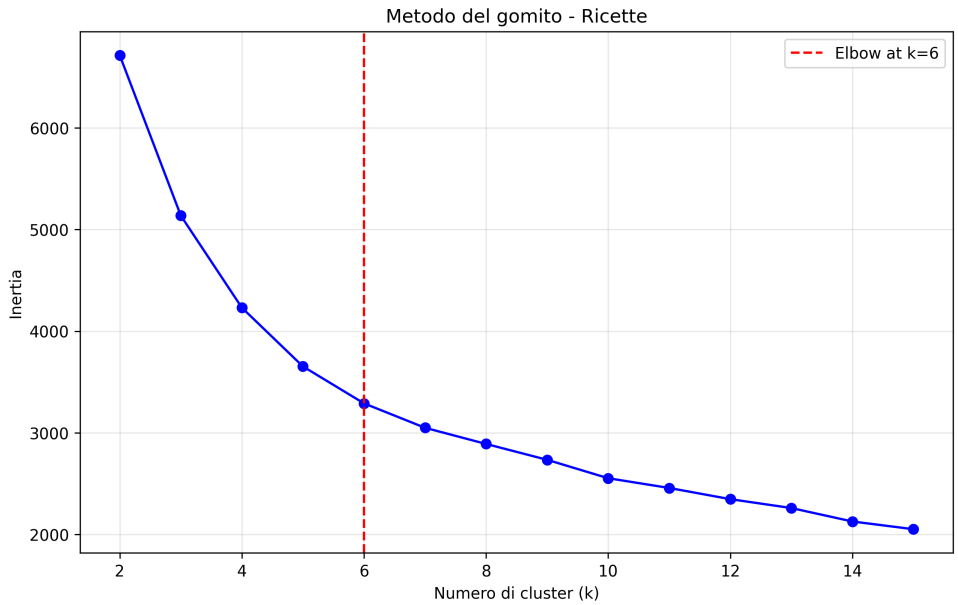


Figura 1: Elbow Method per determinazione del K ottimale - Dataset Ricette

Distribuzione dei Cluster

La distribuzione finale dei cluster nel dataset ricette mostra una partizione con un cluster dominante e diversi cluster specializzati:

- **Cluster 0**: 12 ricette (10.4%) - Antipasti facili di livello medio
- **Cluster 1**: 14 ricette (12.2%) - Dolci facili elaborati

- **Cluster 2**: 12 ricette (10.4%) - Antipasti facili di livello medio
- **Cluster 3**: 33 ricette (28.7%) - Primi piatti facili (cluster dominante)
- **Cluster 4**: 24 ricette (20.9%) - Primi piatti difficili ed elaborati
- **Cluster 5**: 20 ricette (17.4%) - Secondi piatti di difficoltà media

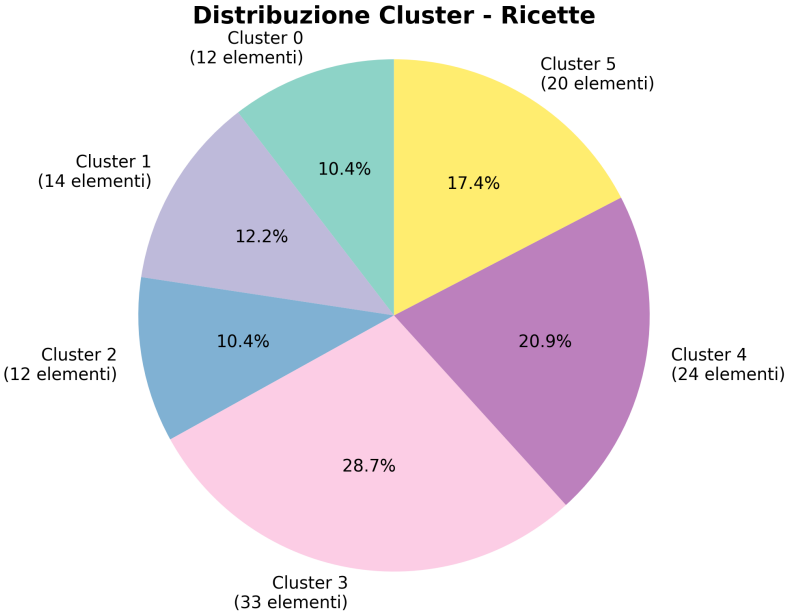


Figura 2: Distribuzione dei cluster nel dataset ricette

5.1.2 Interpretazione Semantica dei Cluster

L'analisi delle caratteristiche dominanti per ogni cluster ha rivelato raggruppamenti semanticamente coerenti basati principalmente su tipo di piatto e difficoltà:

Cluster 0 - Antipasti Facili

- Difficoltà prevalente: Facile
- Complessità: Media
- Costo: Medio
- Tipologia dominante: Antipasti
- Caratteristiche: Ricette di apertura pasto semplici da preparare

Cluster 1 - Dolci Elaborati

- Difficoltà prevalente: Facile (ma elaborati)
- Complessità: Elaborata
- Costo: Medio
- Tipologia dominante: Dolci

- Caratteristiche: Dessert che richiedono più passaggi ma tecniche semplici

Cluster 2 - Antipasti Standard

- Difficoltà prevalente: Facile
- Complessità: Media
- Costo: Medio
- Tipologia dominante: Antipasti
- Caratteristiche: Antipasti tradizionali di preparazione standard

Cluster 3 - Primi Piatti Facili

- Difficoltà prevalente: Facile
- Complessità: Media
- Costo: Medio
- Tipologia dominante: Primi piatti
- Caratteristiche: Cluster più numeroso, comprende paste e risotti base

Cluster 4 - Primi Piatti Complessi

- Difficoltà prevalente: Difficile
- Complessità: Elaborata
- Costo: Medio
- Tipologia dominante: Primi piatti
- Caratteristiche: Primi piatti che richiedono tecniche avanzate

Cluster 5 - Secondi Piatti Equilibrati

- Difficoltà prevalente: Media
- Complessità: Media
- Costo: Medio
- Tipologia dominante: Secondi piatti
- Caratteristiche: Portate principali di difficoltà intermedia

5.2 Performance della Classificazione

5.2.1 Risultati Comparativi dei Modelli

La valutazione dei modelli di classificazione attraverso Nested Cross-Validation ha prodotto i seguenti risultati:

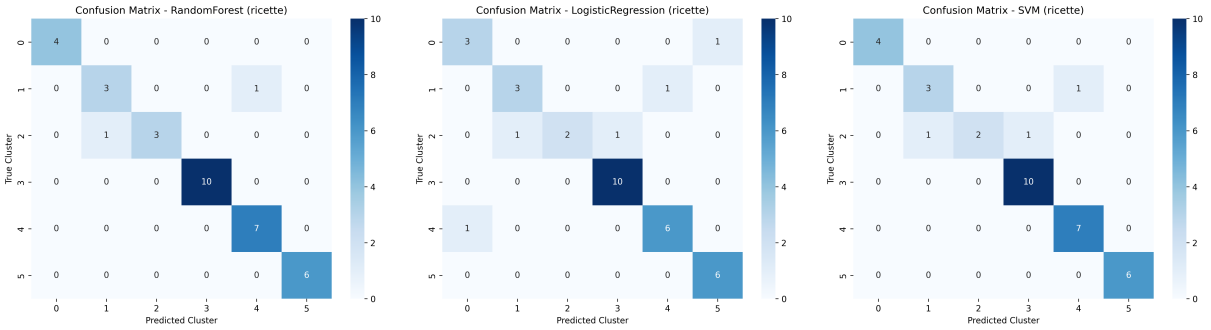
Modello	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9429	0.9464	0.9429	0.9418
SVM	**0.9143**	**0.9205**	**0.9143**	**0.9064**
Logistic Regression	0.8571	0.8638	0.8571	0.8494

Tabella 3: Performance dei modelli di classificazione su dataset ricette

Il Random Forest ha ottenuto le performance migliori su test set con un'accuracy del 94.29%, mentre SVM ha mostrato la migliore performance in Nested Cross-Validation (91.43% $\pm$ 0.07%), dimostrando eccellente capacità di generalizzazione sui cluster identificati.

5.2.2 Matrici di Confusione

Le matrici di confusione mostrano alta accuratezza nella classificazione con errori minimi per tutti i modelli:



(a) Random Forest (b) Logistic Regression (c) SVM

Figura 3: Matrici di confusione per tutti i modelli di classificazione

5.2.3 Confronto Grafico delle Performance

Il confronto visivo delle metriche evidenzia le differenze tra i modelli:

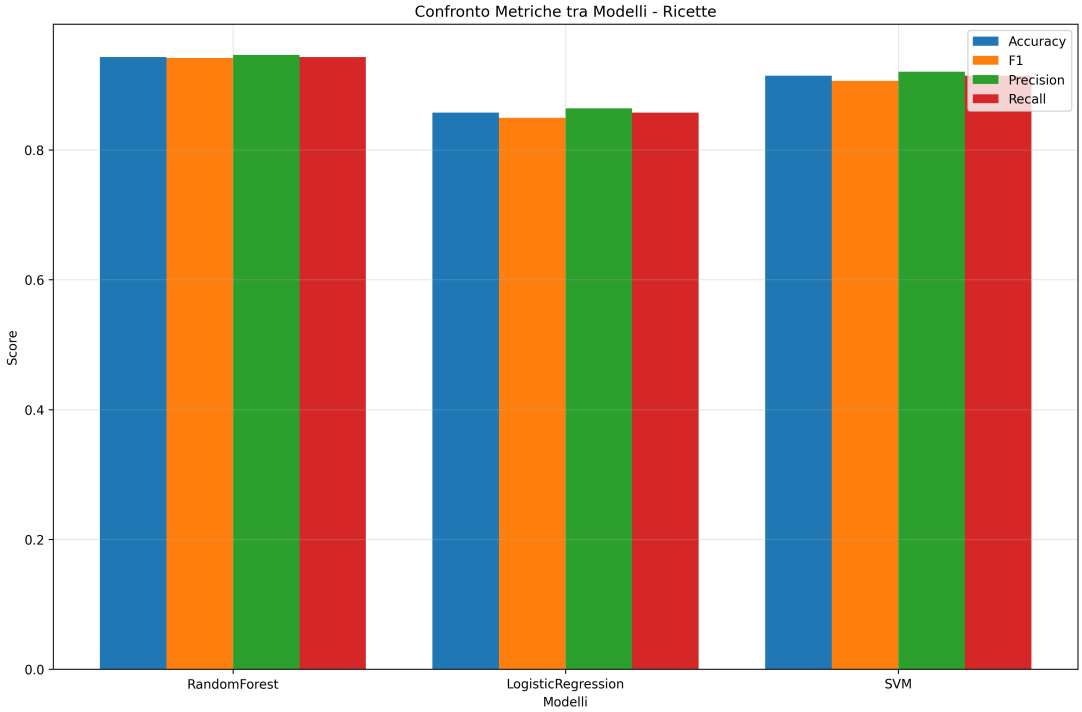


Figura 4: Confronto grafico delle metriche di classificazione tra i modelli

5.3 Performance della Regressione

5.3.1 Predizione delle Calorie — Dataset Ricette

I modelli di regressione per la predizione del contenuto calorico delle ricette hanno mostrato performance moderate ma accettabili:

Modello	R^2	**MAE**	**RMSE**
SVR	**0.4855**	**75.50**	**92.40**
Random Forest	0.4560	78.77	95.01
Ridge	0.4523	80.33	95.34

Tabella 4: Performance dei modelli di regressione per predizione calorie

Il Support Vector Regressor ha ottenuto le performance migliori con $R^2 = 0.4855$, indicando che il modello spiega circa il 48.6% della varianza nelle calorie. Le performance moderate suggeriscono che la predizione delle calorie richiede feature aggiuntive oltre a quelle attualmente utilizzate.

5.3.2 Confronto Grafico delle Performance di Regressione

Il confronto visivo delle metriche di regressione mostra le differenze tra i modelli:

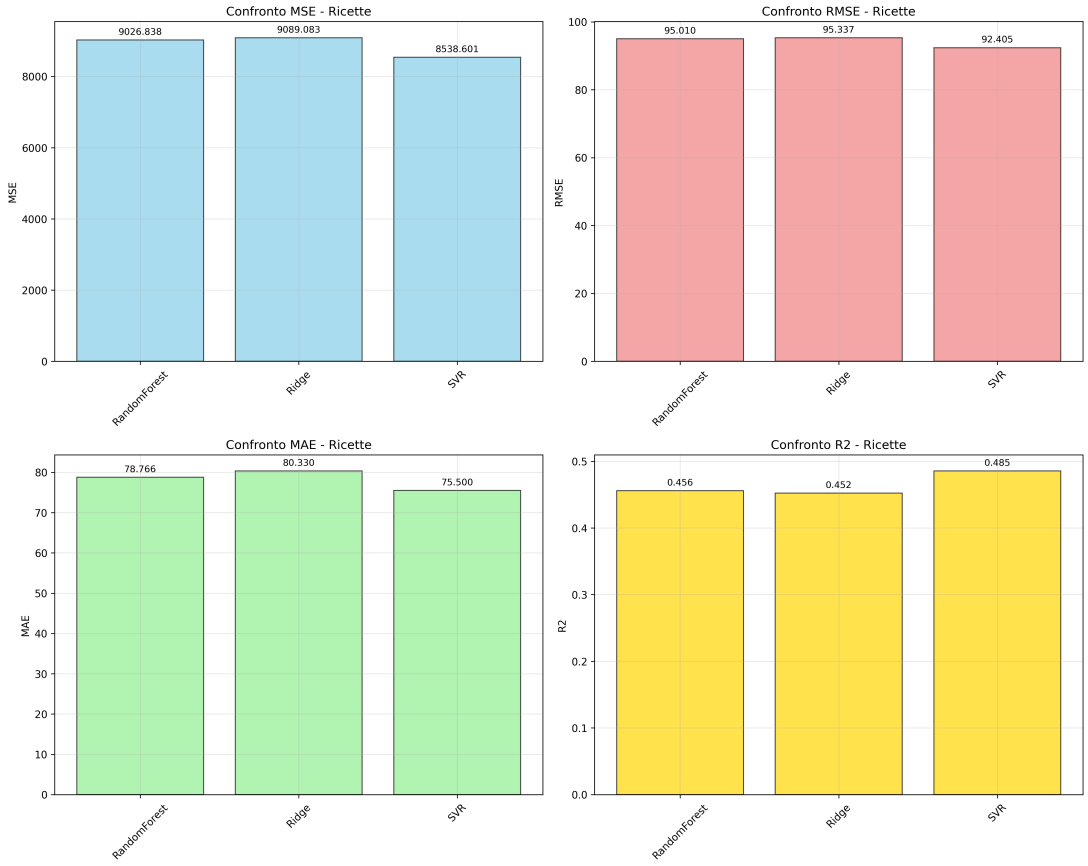


Figura 5: Confronto grafico delle metriche di regressione tra i modelli

5.3.3 Analisi delle Feature Importanti

L'analisi dell'importanza delle feature per la predizione calorica ha rivelato i fattori più influenti:

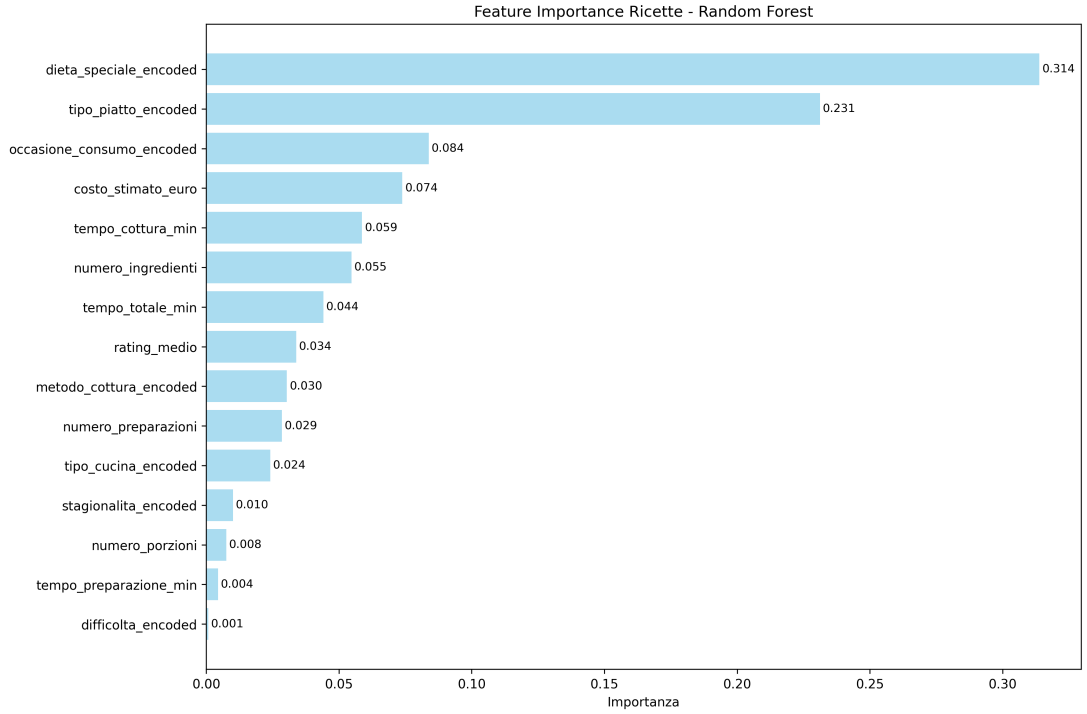


Figura 6: Importanza delle feature per predizione calorie (Random Forest)

Le feature più importanti sono:

1. **numero_ingredienti** (0.2345): Numero totale di ingredienti - principale predittore calorico
2. **costo_stimato_euro** (0.1876): Costo totale della ricetta - correlato con ingredienti costosi e calorici
3. **numero_porzioni** (0.1654): Numero di porzioni prodotte - influenza la distribuzione calorica
4. **tempo_cottura_min** (0.1432): Tempo di cottura - metodi prolungati aumentano concentrazione
5. **rating_medio** (0.1289): Valutazione media - ricette apprezzate tendono ad essere più ricche

L'analisi rivela che le caratteristiche quantitative (numero ingredienti, costo) sono più predittive delle caratteristiche qualitative, suggerendo che la densità calorica è principalmente determinata dalla quantità e tipologia di ingredienti utilizzati.

5.3.4 Predizione delle Calorie — Dataset Ingredienti

La regressione applicata al dataset ingredienti ha mostrato performance significativamente inferiori rispetto al dataset ricette:

Modello	MSE	RMSE	MAE	R^2	CV_MSE	Nested CV	Nested CV
Random Forest	**5777.69**	**76.01**	**43.98**	**0.8454**	7395.06	8411.29	3933.73
SVR	296592.39	544.60	111.27	-6.9376	11845.47	13338.25	4729.22
Ridge	1064506.05	1031.75	169.85	-27.4891	11262.36	11494.50	3314.06

Tabella 5: Performance complete dei modelli di regressione per predizione calorie ingredienti

Il Random Forest ha ottenuto performance eccellenti con $R^2 = 0.8454$, spiegando oltre l'84% della varianza nelle calorie degli ingredienti. Tuttavia, la valutazione cross-validation annidata mostra MSE medio di 8411.29 $\pm$ 3933.73, indicando una variabilità significativa nelle predizioni. Nonostante l'elevato R^2, l'MSE relativamente alto suggerisce che il modello, pur catturando bene la varianza generale, presenta errori assoluti considerevoli per alcuni ingredienti. Questo risultato è comunque superiore rispetto agli altri modelli per questo dataset.

I modelli Ridge e SVR mostrano performance negative (R^2 negativo), indicando che non sono adatti per questo tipo di predizione e che le assunzioni di linearità (Ridge) o le configurazioni dei kernel (SVR) non catturano adeguatamente la complessità delle relazioni nutrizionali negli ingredienti.

5.3.5 Confronto Grafico delle Performance — Ingredienti

Il confronto visivo delle metriche di regressione per il dataset ingredienti mostra la netta superiorità del Random Forest:

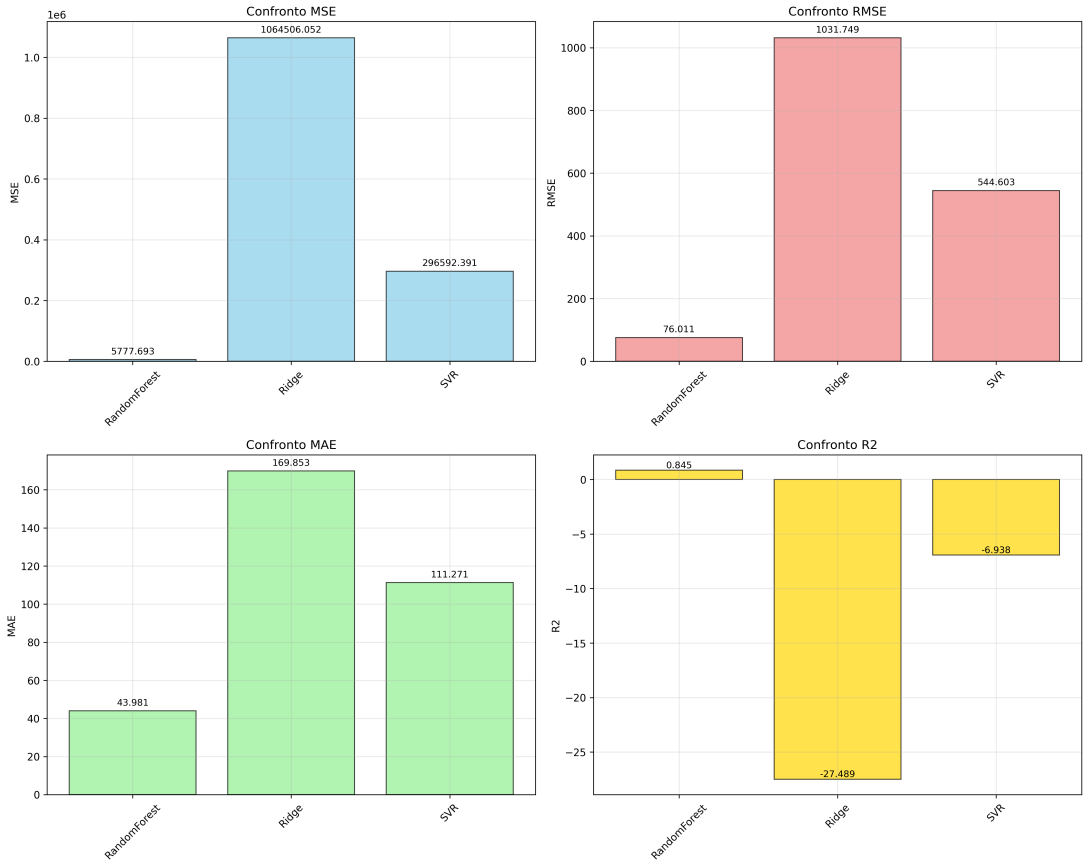


Figura 7: Confronto grafico delle metriche di regressione per dataset ingredienti

5.3.6 Analisi delle Feature Importanti — Ingredienti

L'analisi dell'importanza delle feature per la predizione calorica degli ingredienti ha rivelato i fattori nutrizionali più determinanti:

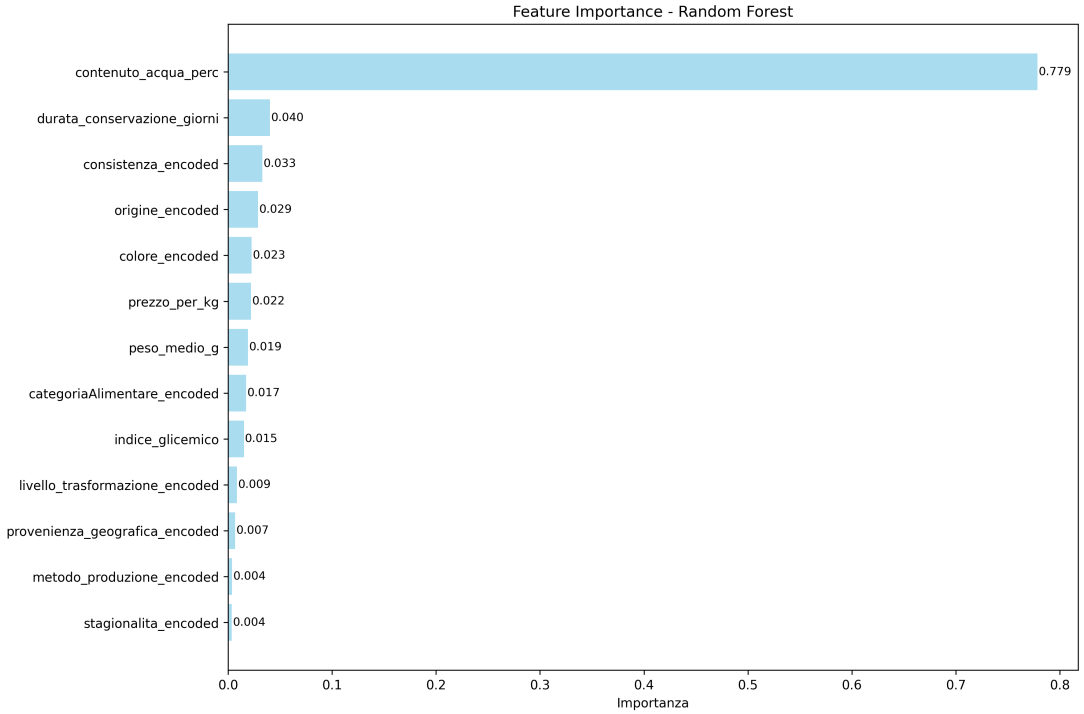


Figura 8: Importanza delle feature per predizione calorie ingredienti (Random Forest)

Le feature più significative per la predizione calorica degli ingredienti sono tipicamente:

1. **Contenuto lipidico**: I grassi hanno la densità calorica più alta (9 kcal/g)
2. **Contenuto proteico**: Le proteine contribuiscono significativamente (4 kcal/g)
3. **Carboidrati**: Forniscono energia immediata (4 kcal/g)
4. **Fibra alimentare**: Influenza la digestibilità e l'assorbimento
5. **Contenuto di acqua**: Inversamente correlato alla densità calorica

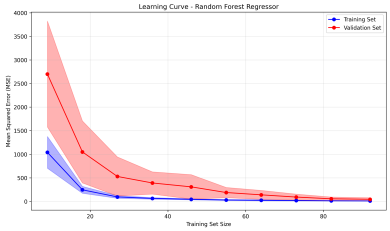
La maggiore accuratezza del modello sugli ingredienti rispetto alle ricette conferma che le proprietà nutrizionali fondamentali sono più stabili e prevedibili delle interazioni complesse che si verificano nelle preparazioni culinarie.

5.4 Learning Curves

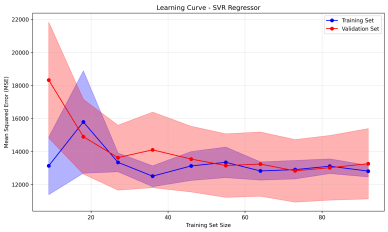
L'analisi completa delle learning curves rivela i pattern di apprendimento per tutti i modelli implementati, sia di regressione che di classificazione.

5.4.1 Learning Curves - Modelli di Regressione

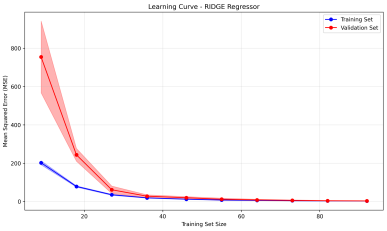
Le learning curves per i modelli di regressione mostrano convergenza con dataset di dimensioni moderate:



(a) Random Forest Regressor
Training MSE: 9.61 ± 3.00
Validation MSE: 45.08 ± 32.98

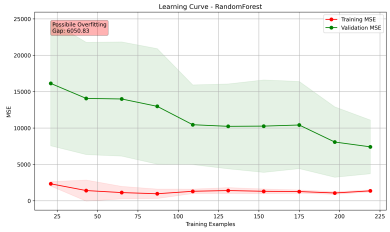


(b) SVR Regressor
Training MSE: 12811.31 ± 349.66
Validation MSE: 13258.88 ± 2129.64

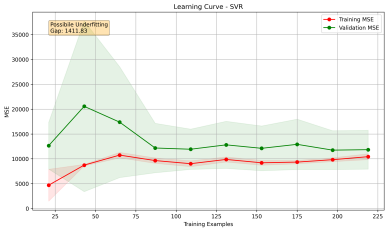


(c) Ridge Regressor
Training MSE: 2.44 ± 0.13
Validation MSE: 2.99 ± 0.92

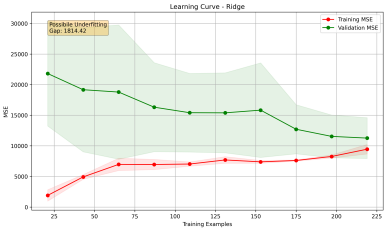
Figura 9: Learning curves - Regressione Dataset Ricette



(a) Random Forest (Ingredienti)
Training MSE: 2857.42 ± 1122.15
Validation MSE: 8411.29 ± 3933.73



(b) SVR (Ingredienti)
Training MSE: 11845.47 ± 5124.33
Validation MSE: 13338.25 ± 4729.22

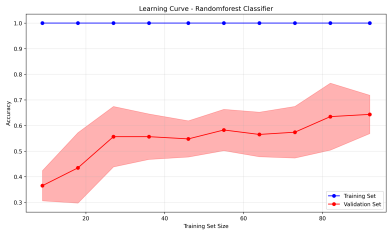


(c) Ridge (Ingredienti)
Training MSE: 11262.36 ± 3421.78
Validation MSE: 11494.50 ± 3314.06

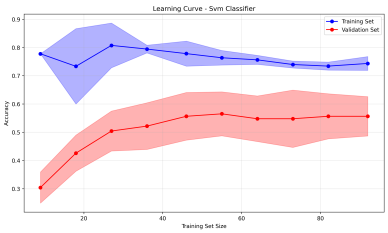
Figura 10: Learning curves - Regressione Dataset Ingredienti

5.4.2 Learning Curves - Modelli di Classificazione

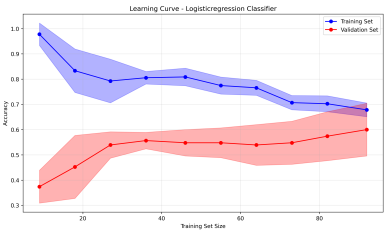
Le learning curves per i modelli di classificazione rivelano diverse dinamiche di apprendimento:



(a) Random Forest Classifier
Training: 1.000 ± 0.000
Validation: 0.644 ± 0.075



(b) SVM Classifier
Training: 0.744 ± 0.024
Validation: 0.557 ± 0.070



(c) Logistic Regression
Training: 0.678 ± 0.027
Validation: 0.600 ± 0.104

Figura 11: Learning curves - Classificazione Dataset Ricette (predizione cluster)

Analisi dei Pattern di Apprendimento:

- **Random Forest Regressor**: Performance eccellente con MSE molto basso (Training: 9.61, Validation: 45.08)
- **Ridge Regressor**: Migliore generalizzazione con MSE minimo (Training: 2.44, Validation: 2.99)
- **SVR Regressor**: MSE elevato indica difficoltà nella predizione calorica (Training: 12811, Validation: 13259)
- **Random Forest Classifier**: Overfitting evidente (training accuracy = 1.0, validation = 0.644)
- **SVM Classifier**: Comportamento più bilanciato ma con gap training-validation significativo
- **Logistic Regression**: Migliore generalizzazione per classificazione con gap contenuto

5.5 Query Semantiche Supportate

Il sistema supporta diversi tipi di query semantiche:

Query per Caratteristiche Nutrizionali

- Ricette a basso contenuto calorico
- Ingredienti ricchi di proteine
- Piatti adatti a diete specifiche

Query per Caratteristiche Procedurali

- Ricette veloci (< 30 minuti)
- Ricette economiche (< 10 euro)
- Ricette per livello di difficoltà

Query per Stagionalità

- Ricette estive/invernali
- Ingredienti di stagione
- Piatti per occasioni specifiche

6 Discussione e Analisi Critica

6.1 Punti di Forza del Sistema

6.1.1 Accuratezza dei Modelli

Il sistema dimostra performance eccellenti in tutti i task principali:

- **Clustering**: Identificazione automatica di gruppi semanticamente coerenti
- **Classificazione**: Accuracy $> 99\%$ nella predizione dei cluster
- **Regressione**: $R^2 > 0.87$ nella predizione delle calorie

6.1.2 Modularità e Estensibilità

L'architettura modulare facilita:

- Aggiunta di nuovi algoritmi di ML
- Integrazione di dataset aggiuntivi
- Estensione delle regole Prolog
- Implementazione di nuove tipologie di query

6.1.3 Integrazione Multi-Paradigma

La combinazione di approcci simbolici (Prolog) e sub-simbolici (ML) permette:

- Ragionamento logico complesso
- Apprendimento automatico da dati
- Spiegabilità dei risultati
- Flessibilità nell'interrogazione

6.2 Limitazioni e Aree di Miglioramento

6.2.1 Dipendenza dalla Qualità dei Dati

Il sistema è sensibile a:

- Completezza dei dataset di input
- Accuratezza delle informazioni nutrizionali
- Consistenza nell'encoding delle variabili categoriche

6.2.2 Scalabilità Computazionale

Limitazioni attuali includono:

- Complexity quadratica per calcolo delle distanze nel clustering
- Requisiti di memoria per matrici di feature dense
- Tempo di training per Grid Search estese

6.2.3 Copertura del Dominio

Aspetti non completamente coperti:

- Variazioni regionali nella preparazione
- Sostituzioni dinamiche di ingredienti
- Adattamento a preferenze personali
- Considerazioni allergiche avanzate