DS3002 Data Mining Project Report FurGenius

Ameer Abdullah
Department of Data Science
FAST-NUCES, Lahore, Pakistan
Submitted To: Ms. Esha Tur Razia Babar

***Abstract*—Effective nutrition is critical for the health, recovery, and well-being of animals across species, yet current dietary planning methods rely on generalized guidelines that fail to address individualized needs. This project introduces will be designed to efficiently predict animal nutrition by integrating machine learning (ML), natural language processing (NLP), and generative AI. The system analyzes structured data (e.g., weight, medical history) and unstructured inputs (e.g., veterinary notes, prescriptions) to specify dietary recommendations. By leveraging predictive analytics and NLP to interpret complex medical records, the platform supports veterinarians and caregivers in making data-driven decisions, improving disease management, and optimizing recovery outcomes. The system emphasizes transparency through explainable AI (XAI), sustainability by prioritizing cost-effective and ecofriendly ingredients, and scalability for applications in shelters, farms, and zoos. This approach bridges the gap between veterinary expertise and technological innovation, aiming to revolutionize animal healthcare by enhancing efficiency, reducing errors, and fostering individualized care.**

I. Motivation

Nutrition plays a pivotal role in maintaining the health and vitality of animals, whether they are pets, livestock, or wildlife. However, conventional dietary planning methods often rely on broad, species-generic guidelines that overlook critical individual variations such as medical conditions, genetic predispositions, and environmental factors. For instance, animals with chronic illnesses or specific allergies require meticulously curated diets, yet caregivers frequently lack tools to systematically address these needs. Current practices also struggle to synthesize unstructured data (e.g., handwritten veterinary notes) and dynamic health metrics, leading to suboptimal care and delayed recovery. Recent advancements in artificial intelligence (AI) have demonstrated transformative potential in healthcare, including animal nutrition. Studies by Azzimani et al. (2022) and Zhang et al. (2023) highlight how AI can automate nutrient analysis, predict dietary deficiencies, and interpret behavioral cues to optimize feeding strategies [1][3]. Meanwhile, research by Deka et al. (2024) underscores the importance of precision feeding and sustainable practices in aligning nutritional science with realworld applications [2]. These findings underscore a critical opportunity: integrating AI into veterinary nutrition to create personalized, evidence-based solutions.

II. Dataset Description

For this project the dataset was not readily available online and not even on Kaggle and UCIML. We need tabular and textual (structured and unstructured data) for this project. We did research and we found out medical conditions common in animals and then we found dietary recommendations that are suitable for that animal to fight against that disease. Then we were able to find out veterinary notes too prescribing which diet category is most suitable. So, we gather the information related to this and then we synthetically generated textual data of almost 6000 instances and 4 features such as veterinary notes, prescriptions and dietary recommendations. Now for tabular dataset, we purely generated it synthetically due to unavailability of the dataset. However, the dataset was available for human's diet plan so we got a picture of how tabular data might look like. So, we created almost 6000 instances dataset with 17 features such as age, weight, nutritional requirements, etc. However, for scaling this project and for deployment and for production level, it is recommended and strictly advised to have extensive research and to consult caregivers and veterinarians for the authenticity of the data.

III. Dietary Recommendations

Using the dataset, our task is to specify each animal into a specific dietary category based on medical conditions and nutritional requirements. The evaluation of the model is based on classification metrics including accuracy, precision, recall, and Fscore. In the case of our project, accuracy takes precedence as our primary goal is to accurately categorize animals into dietary recommendations that are most suitable for them.

IV. Explainable AI

With the purpose of enhancing transparency and building trust in the model's performance, the use of XAI reveals the reasoning behind predictions including the impact of each feature that aids in model improvement.

V. Initial Data Assessment & Preprocessing

Clean and prepare the data for analysis by handling missing values, outliers, scaling, encoding, and multicollinearity.

A. *Key Actions:*

The datasets were merged by combining the `tabular (synthetic_tabular_dataset.csv)` and `textual (synthetic_textual_dataset.csv)` data using the `Animal_ID`. Missing values in numerical columns were filled with the median, while categorical columns were filled with the mode. Outliers were detected using the IQR method, and outlier values were capped to the IQR boundaries. Min-Max scaling was applied to numerical features to standardize their range. Categorical variables, such as Species and Gender, were encoded using one-hot encoding. To address multicollinearity, the Variance Inflation Factor (VIF) was calculated, and features with a VIF greater than 10 were removed. I have visualized missing values before and after so below is visual representation.

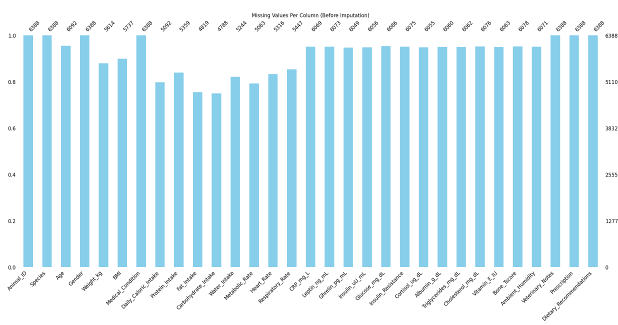

Fig. 1: Before Missing Values

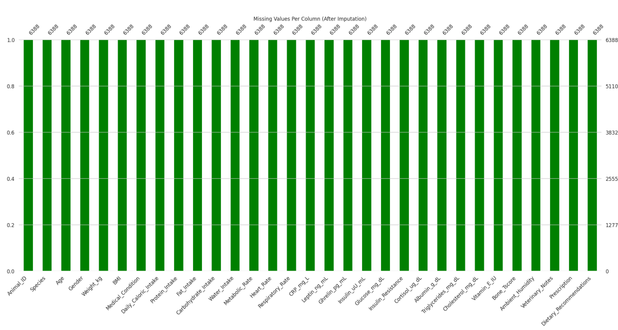

Fig. 2: After Missing Values

After this I have visualized the outliers in the dataset so here is a visual representation.

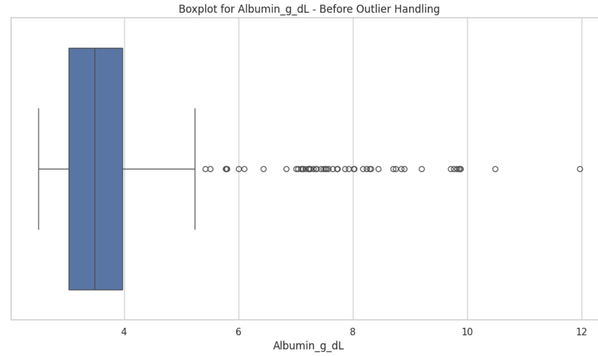

Fig. 3: Before Outliers BoxPlot

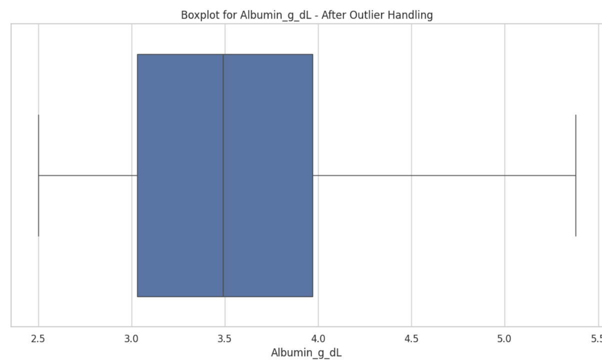

Fig. 4: After Outliers BoxPlot

Following these numerical features were scaled and here is a visual representation of them.

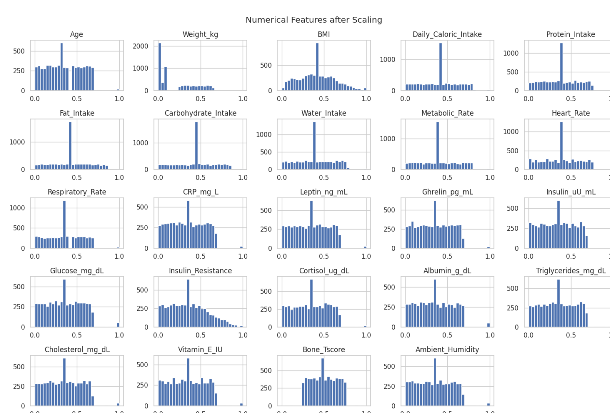

Fig. 5: Numerical Features After Scaling

Just after this correlation heatmap was introduced to interpret the relationships between multiple variables in a dataset.

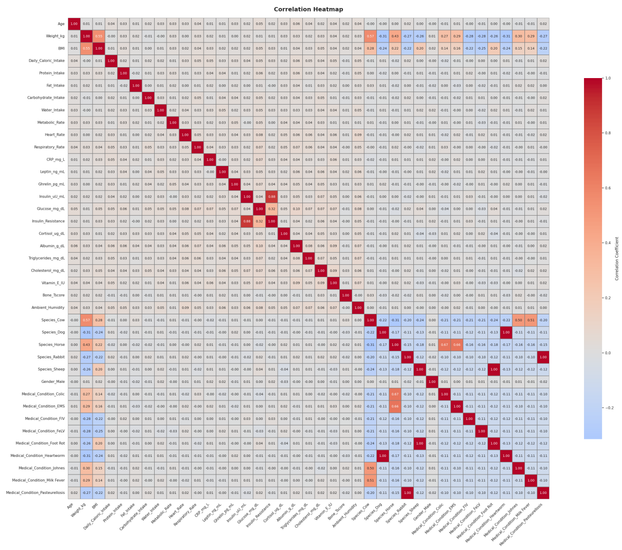

Fig. 6: Correlation Heatmap

And by completing the data preprocessing part, we finally visualize the multicollinearity.

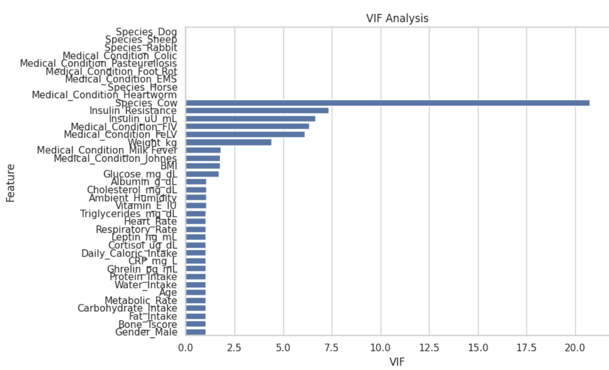

Fig. 7: VIF Analysis

Now, it is turned to preprocess some textual dataset. Use of NLP proved to be beneficial for cleaning the unstructured data.

B. Textual Data Preprocessing

Now, it is turned to preprocess some textual dataset. Use of NLP proved to be beneficial for cleaning the unstructured data.

- Word Count Distribution for Veterinary Notes (Histogram) A histogram with kernel density estimate visualizes the word count distribution of processed Veterinary Notes. It reveals the length variability of notes, informing text feature engineering.

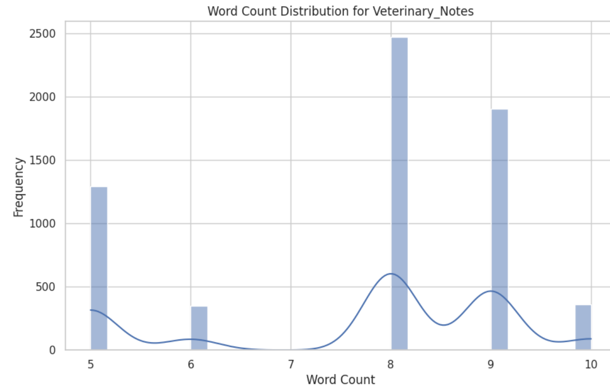

Fig. 8: Word Count Distribution for Veterinary Notes

- Word Count Distribution for Prescriptions (Histogram) A histogram with kernel density estimate shows the word count distribution of processed Prescriptions. It highlights prescription text length patterns, aiding NLP feature selection.

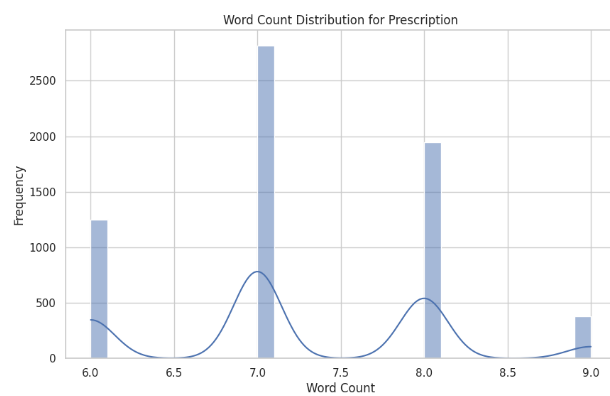

Fig. 9: Word Count Distribution for Prescriptions

- Sample TF-IDF Feature Distributions for Veterinary Notes (Histograms) Histograms display the distribution of the first five TF-IDF features from Veterinary Notes. They illustrate the sparsity and range of TF-IDF values, guiding feature scaling decisions.

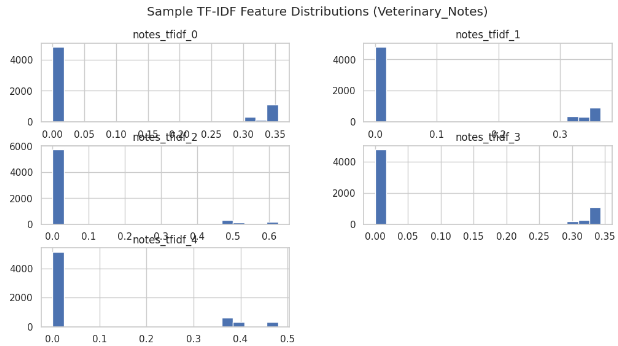

Fig. 10: Sample TF-IDF Feature Distributions for Veterinary Notes

- Sample TF-IDF Feature Distributions for Prescriptions (Histograms) Histograms visualize the distribution of the first five TF-IDF features from Prescriptions. They show feature value distributions, supporting text feature analysis for modeling.

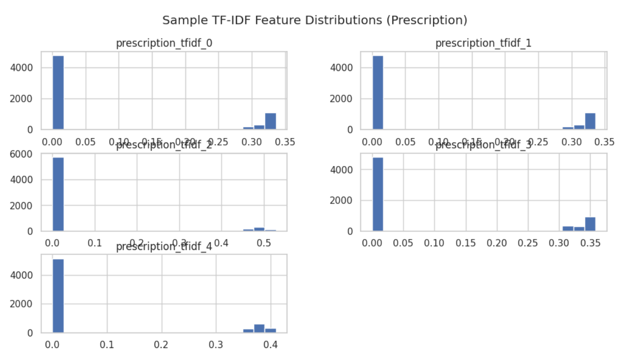

Fig. 11: Sample TF-IDF Feature Distributions for Prescriptions

- Explained Variance by SVD Components (Line Plot) A line plot shows the cumulative explained variance ratio for Truncated SVD components applied to TF-IDF features of Veterinary Notes. It indicates the number of components needed to capture significant variance, optimizing dimensionality reduction.

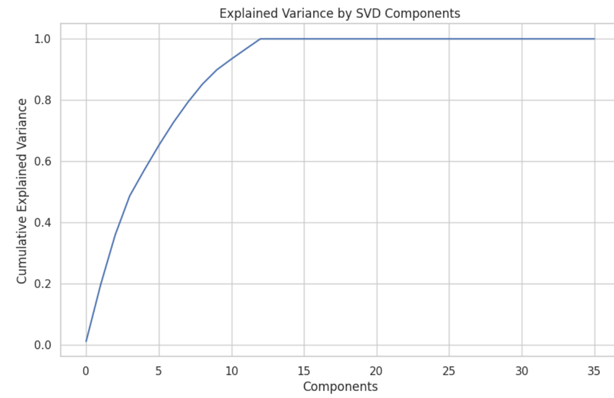

Fig. 12: Explained Variance by SVD Components

VI. Exploratory Data Analysis

Analyze data distributions, relationships, and patterns to inform modeling decisions. Key actions included analyzing the target variable by visualizing the class distribution of Dietary Recommendations through bar charts. Statistical summary was generated, providing descriptive statistics (mean, median, standard deviation, etc.) for numerical features. Correlation analysis was performed by creating a heatmap to identify correlations between numerical features. Visualizations were created, including bar charts to show the relationship between species and dietary plans, box plots to examine weight distribution by species, a word cloud to highlight frequent terms in veterinary notes, and bubble charts to illustrate the relationship between age and weight, categorized by species and diet. The overall results are plotted in the following figures.

- Medical Conditions vs. Dietary Recommendations (Bubble Chart) A bubble chart visualizes the relationship between Medical Conditions and Dietary Recommendations, with bubble sizes scaled by condition frequency. It highlights which diets are commonly associated with specific conditions, aiding in understanding dietary patterns.

Fig. 13: Medical Conditions vs. Dietary Recommendations

- Correlation Analysis (Heatmap) A correlation heatmap displays pairwise correlations among numerical features (e.g., Age, Weight, Glucose). It identifies strong relationships (e.g., Weight and BMI) to guide feature selection and detect multicollinearity.

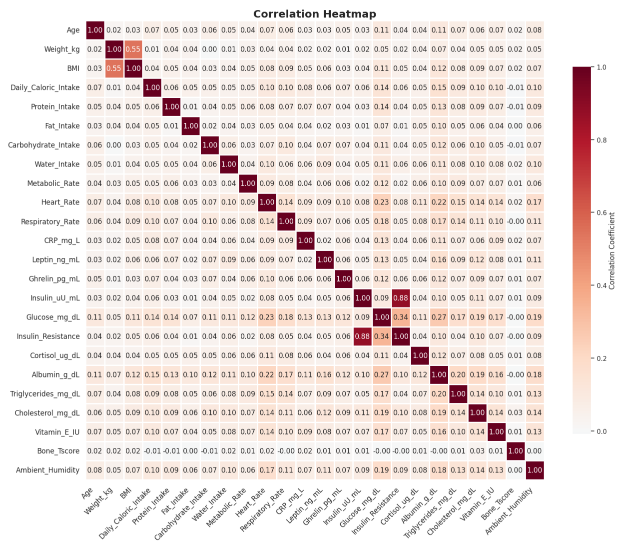

Fig. 14: Correlation Analysis

- Feature Distribution by Dietary Recommendation (Boxplot) Boxplots show the distribution of numerical features (e.g., Weight, Glucose) across Dietary Recommendations. They reveal variations and outliers, informing feature relevance for dietary classification.

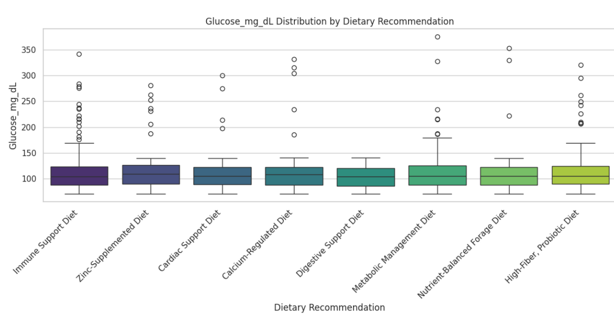

Fig. 15: Feature Distribution by Dietary Recommendation (Boxplot)

- Numeric Feature Distributions (Histograms) Histograms with kernel density estimates visualize the distribution of numerical features (e.g., Age, Weight). They highlight feature skewness and spread, aiding in preprocessing decisions.

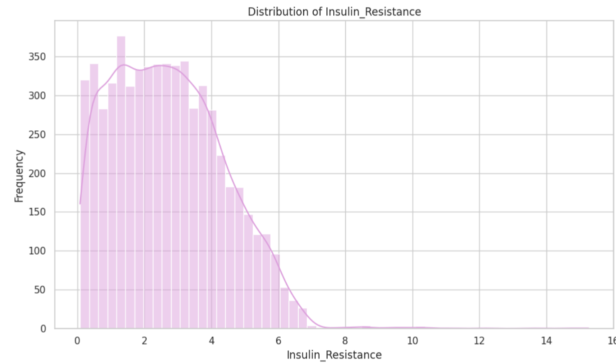

Fig. 16: Numeric Feature Distributions

- Average Glucose Level by Medical Condition and Dietary Recommendation (c) A grouped bar chart shows average Glucose levels by Medical Condition and Dietary Recommendation. It identifies condition-specific glucose patterns, relevant for dietary planning.

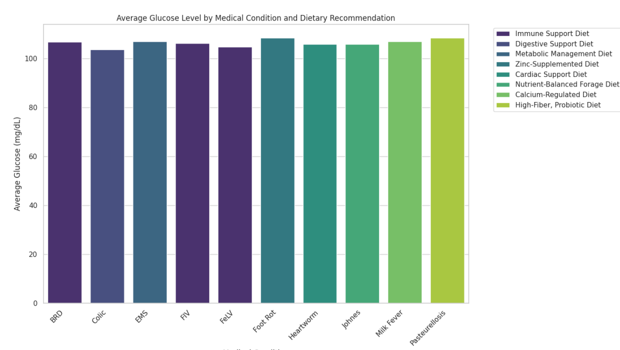

Fig. 17: Grouped Bar Chart

- 3D PCA of Tabular Features (3D Scatter Plot) A 3D scatter plot visualizes the first three PCA components of numerical features, colored by Dietary Recommendations. It reveals class separability and data structure in reduced dimensions.

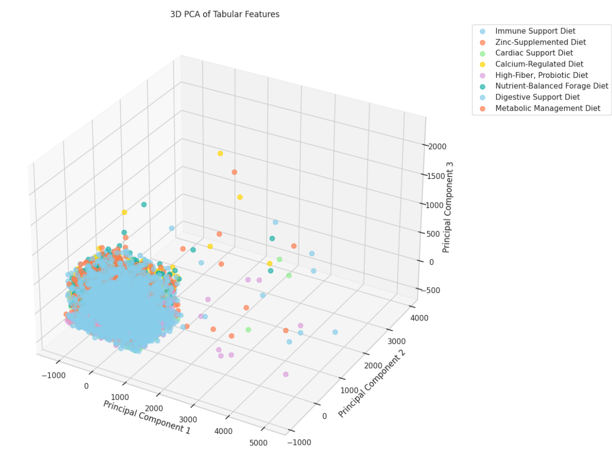

Fig. 18: 3D PCA of Tabular Features

- t-SNE Plot of Text Data (Scatter Plot) A t-SNE plot projects TF-IDF features from Veterinary Notes into 2D, colored by Dietary Recommendations. It shows clustering of text data, indicating dietary-related patterns in notes.

Fig. 19: t-SNE Plot of Text Data

- UMAP Plot of Text Data (Scatter Plot) A UMAP plot visualizes TF-IDF features from Veterinary Notes in 2D, colored by Dietary Recommendations. It highlights text data structure and class separation for NLP insights.

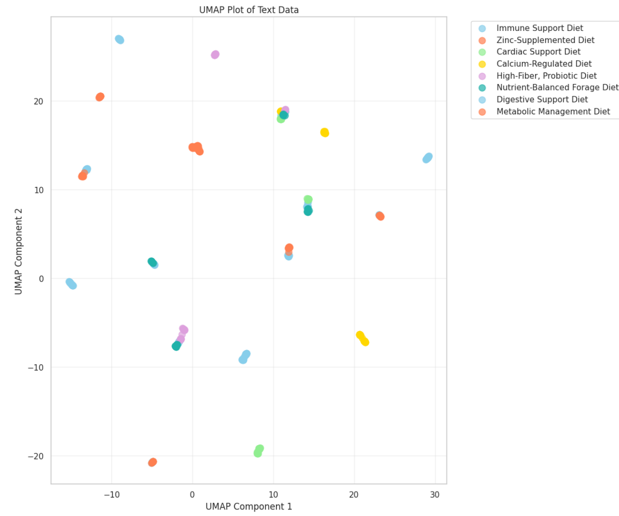

Fig. 20: UMAP Plot of Text Data

- Hierarchical Clustering (Dendrogram) A dendrogram displays hierarchical clustering of numerical features using Ward's method. It groups samples by similarity, revealing underlying data structures for dietary classification.

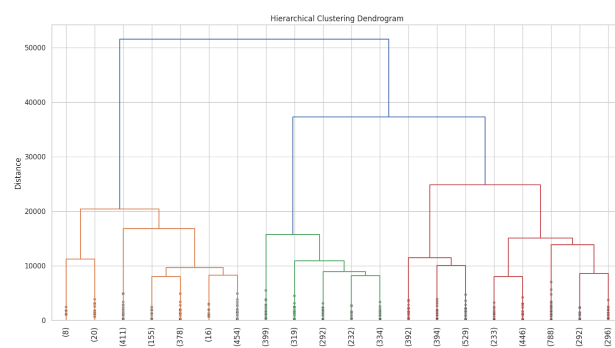

Fig. 21: Dendrogram

- 3D Scatter Plot of Age, Weight, and Daily Caloric Intake A 3D scatter plot visualizes Age, Weight, and Daily Caloric Intake, colored by Dietary Recommendations. It explores multi-feature relationships critical for nutrition planning.

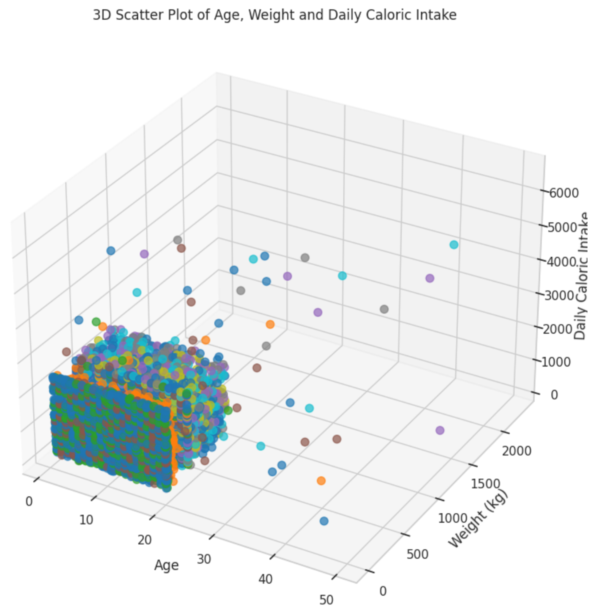

Fig. 22: 3D Scatter Plot

- Silhouette Plot for Clustering A silhouette plot evaluates clustering quality for three clusters of scaled numerical features. It assesses cluster cohesion and separation, informing potential dietary groupings

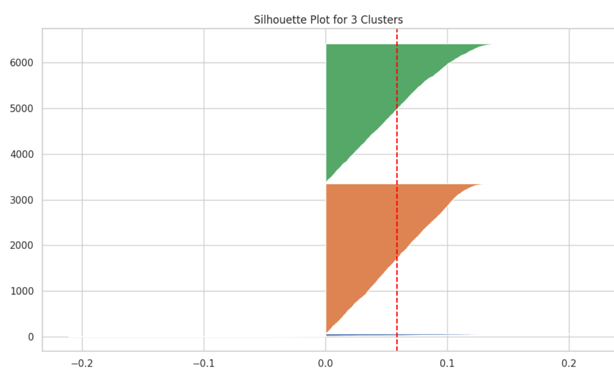

Fig. 23: Silhouette Plot

- Multi-Class Precision-Recall Curve Precision-recall curves for each Dietary Recommendation class are plotted using a Logistic Regression model. They evaluate classification performance, highlighting model strengths for imbalanced classes.

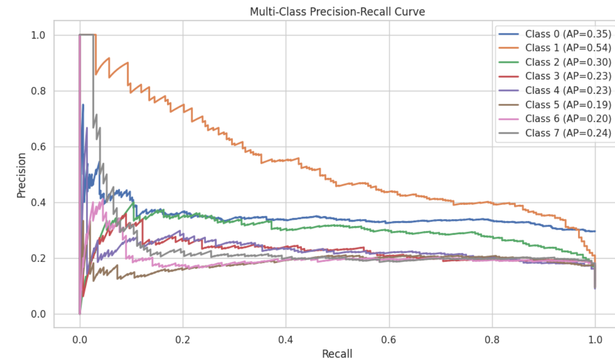

Fig. 24: Multi-Class Precision-Recall Curve

VII. Feature Selection

The goal is to identify the most relevant features for predicting Dietary Recommendations using mutual information, reduce dimensionality, and ensure a robust feature set for the FurGenius project's machine learning models. The code also generates EDA visualizations to analyze feature importance and correlations, providing insights into the data's structure and suitability for dietary classification in veterinary nutrition. Selected 14 features out of 29. Selected features:

- `Weight_kg`
- `Medical_Condition_Johnes`
- `Medical_Condition_Milk_Fever`
- `BMI`
- `Medical_Condition_FIV`
- `Medical_Condition_FeLV`
- `Leptin_ng_mL`
- `Insulin_uU_mL`
- `Ambient_Humidity`
- `Triglycerides_mg_dL`
- `Respiratory_Rate`
- `Gender_Male`
- `Water_Intake`
- `Carbohydrate_Intake`
  - Tabular Features Score.

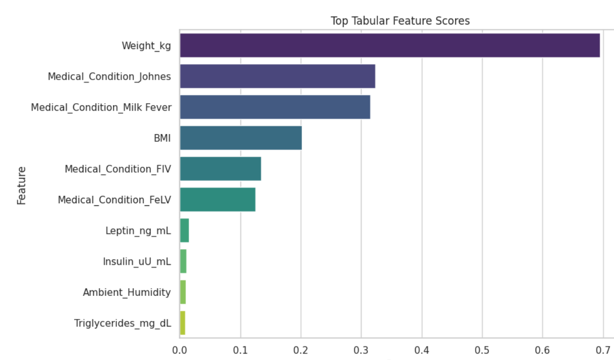

Fig. 25: Top Tabular Features Score

  - And the heatmap of selected features.

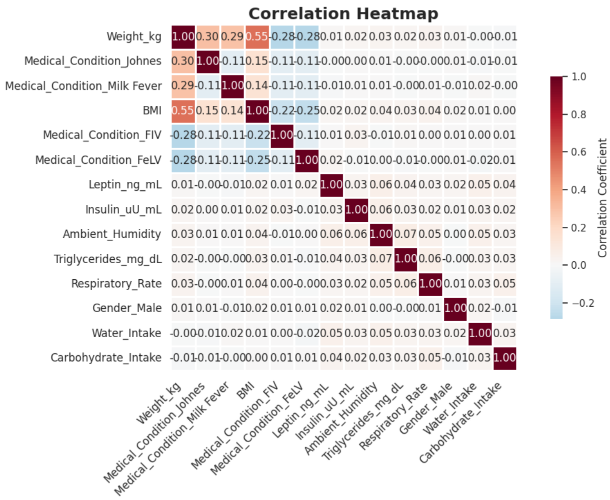

Fig. 26: Selected Features Heatmap

VIII. Multi-Modal Integration

Combine tabular and text data into a unified feature set for modeling. Key actions included generating fixed length embeddings for text data, such as Veterinary Notes, to convert it into numerical representations. The tabular features were then merged with the text embeddings into a single dataset. Finally, validation was performed to ensure consistency and alignment between the modalities, ensuring there were no missing values and that the data was properly Integrated. Shape of selected tabular features: (6388, 14) Final integrated feature set shape: (6388, 34)

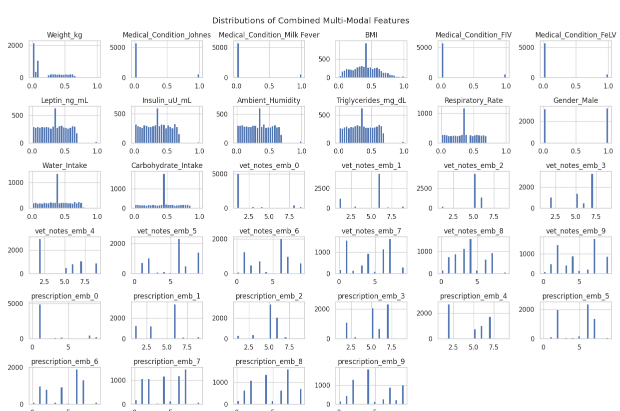

Fig. 27: Distribution of combined multi-modal features

IX. Class Imbalance

Mitigate class imbalance in the target variable (Dietary Recommendations). Key actions included analyzing the class distribution to assess the counts of minority and majority classes. SMOTE oversampling was applied to the training set to oversample the minority classes. Class weight adjustment was used, incorporating class weights (e.g., class weight='balanced' in logistic regression) to penalize misclassifications of underrepresented classes. Models were then trained on the resampled data while validating performance on the original test set. Here is overview of before vs after of class distribution after applying SMOTE.

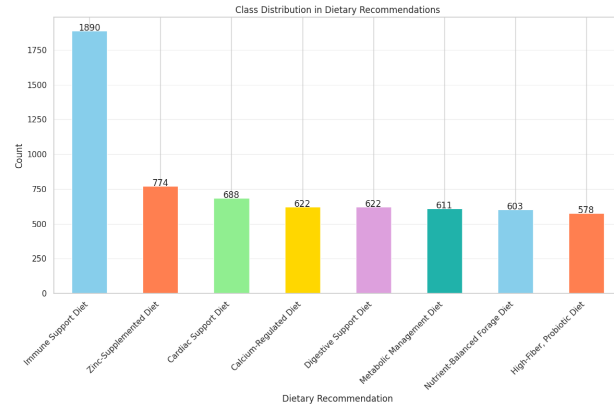

Fig. 28: Before Class Imbalance Solution

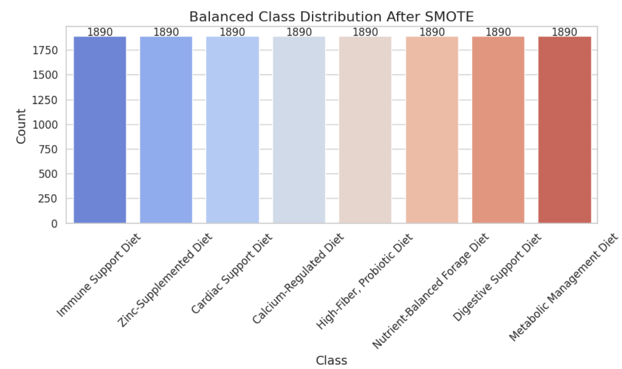

Fig. 29: After applying SMOTE

X. Model Implementation

The model implementation phase, as executed in the `implement_models` function, evaluates multiple machine learning models to predict dietary recommendations. The code splits the balanced dataset (`X_balanced`, `y_balanced`) into training and testing sets using an 80:20 ratio, encodes the target variable using `LabelEncoder`, and trains nine models: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM, SVM, k-NN, and Neural Network. Each model undergoes 5-fold cross-validation on the training set to compute mean accuracy and standard deviation, followed by full training and testing to obtain test accuracy (Random Forest: [0.9177], XGBoost: [0.9018). A bar plot visualizes test accuracies highlighting model performance. The best model Random Forest is selected based on test accuracy, and its classification report and confusion matrix are generated, with the latter visualized as a heatmap. A soft-voting ensemble of the top three models are created, achieving an accuracy of 0.9107, and replaces the best model if superior. This step ensures reliable dietary classification essential for veterinary nutrition by comparing a range of algorithms and using ensemble techniques to boost accuracy, with clear visualizations that make the results easy to understand.

- Here is a visual representation of test accuracies and confusion matrix.

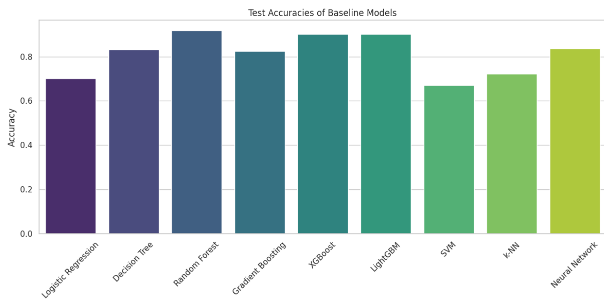

Fig. 30: Test accuracies

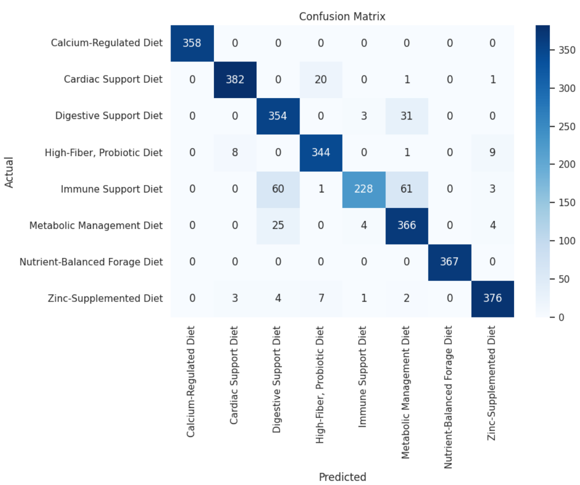

Fig. 31: confusion matrix

XI. Hyperparameter Tuning

The hyperparameter tuning phase, executed via the `tune_hyperparameters` function, optimizes the best performing model from the initial evaluation for predicting dietary recommendations. The code splits the balanced `dataset (X_balanced, y_balanced)` into training and testing sets using an 80:20 ratio and applies RandomizedSearchCV with 5-fold cross-validation to evaluate 20 parameter combinations from a predefined grid `(e.g., n_estimators, max_depth for Random Fore` A line plot visualizes mean test accuracy versus `n_estimators`, highlighting performance trends. The best parameters achieve a cross-validation accuracy of 0.9131 and a test accuracy of 0.9190, ensuring optimized model performance for accurate dietary classification in veterinary nutrition. Best Parameters: `{'n_estimators': 300, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': 30}`

- `Mean test score vs n_estimators visualizatio`

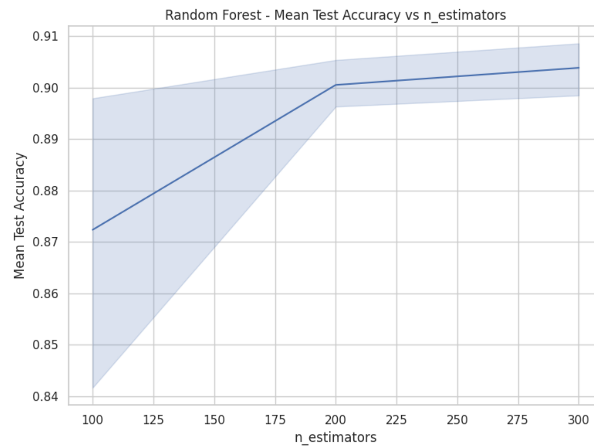

Fig. 32: Mean test score vs `n_estimators` visualization

XII. Model Evaluation

Models are evaluated using accuracy, precision, recall, F1-score, and ROC-AUC, with 5-fold cross-validation for reliability. The code computes metrics on the test set, generates a classification report and visualizes a confusion matrix heatmap and ROC curves for each class. Outputs highlight per-class performance, identifying misclassifications. In veterinary nutrition, high precision and recall are crucial to avoid harmful dietary assignments, ensuring the model meets clinical standards for safe and effective recommendations.

- Here is a visual representation of Final confusion matrix.

Fig. 33: Final confusion matrix

- Here is a visual representation of F1 Score.

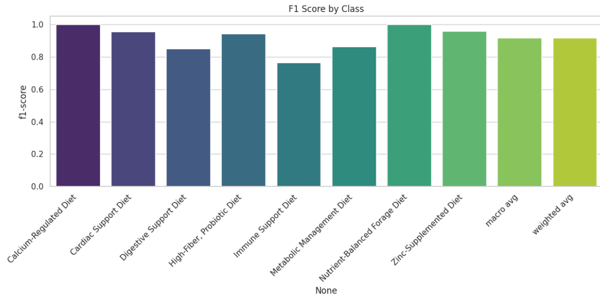

Fig. 34: F1 Score

XIII. Model Explainability

LIME and SHAP provide interpretability for model predictions, explaining feature contributions. The code uses LIME to explain a single test instance and SHAP to compute global feature importance. Outputs include a LIME explanation plot (highlighting feature impacts) and SHAP summary plots (bar and dot formats) showing features like Weight and Veterinary Notes terms as key drivers. LIME and SHAP provided interpretability. LIME explained individual predictions, highlighting the top five contributing features. SHAP summary plots visualized global feature importance, and force plots detailed local predictions. This step builds trust among veterinarians by making AI predictions transparent, aligning with clinical decisionmaking processes in animal healthcare.

- Here is a visual representation of LIME Explanation for Test Instance.

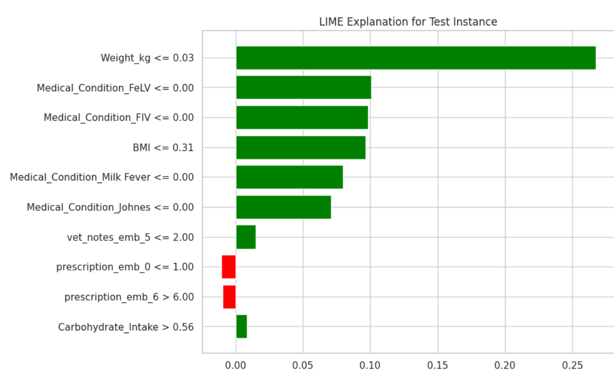

Fig. 35: LIME Explanation for Test Instance

- Here is a visual representation of Shap Summary.

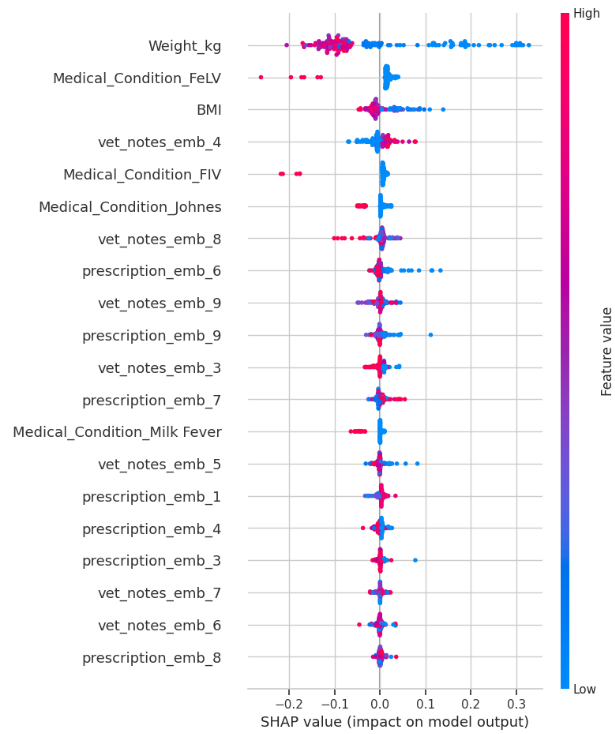

Fig. 36: Shap Summary

- Here is a visual representation of Shap Feature Importance.

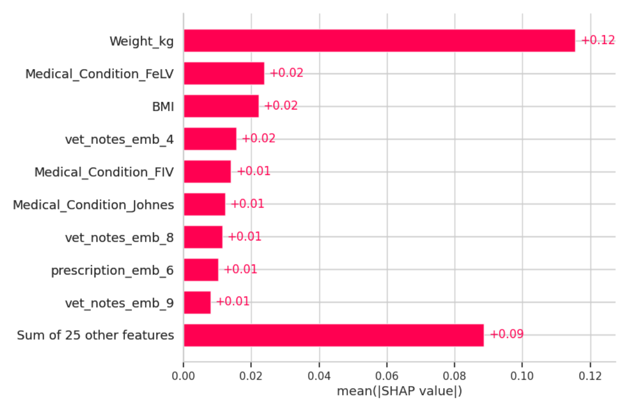

Fig. 37: Shap Feature Importance

XIV. Generative AI Exploration

A GPT-2 model generates synthetic veterinary notes for rare conditions to augment the dataset. The code sets up a text-generation pipeline, prompts GPT-2 with conditionspecific inputs, and produces synthetic samples (e.g., "Monitor symptoms for rare allergy"). Outputs include printed synthetic notes and a word cloud visualizing common terms (e.g., "symptom," "diet"). This step addresses data scarcity for rare veterinary conditions, enhancing model reliability and supporting comprehensive dietary planning for diverse animal health scenarios.

Generated Text

Generate veterinary notes for a Cow diagnosed with Bovine Respiratory Disease:

"An immediate appointment is needed to confirm the diagnosis and determine treatment options. A thorough physical examination will be necessary, including an assessment of respiratory parameters such as oxygen saturation and airway pressure."

Generate veterinary notes for a Cow diagnosed with Bovine Respiratory Disease:

"The veterinarian should determine the cause of illness, including any signs or symptoms which may indicate bacterial pneumonia."

A recent study in *The Journal of Clinical Investigation* found that cow meat is more likely to be contaminated with pathogens than chicken. That finding prompted researchers at Stanford University and Michigan State Health System (MSHS) School – Medical Sciences to perform tests on animals from two locations where they had previously tested feces samples taken after eating beef cattle carcasses.

They uncovered evidence suggesting some animal species could also harbor bacteria resistant as well—a possible sign humans are ingesting infected foods while still alive!

In addition, an experiment conducted by scientists led them toward developing new vaccines against these disease-causing microbes called "Bravia virus neutralizing antibodies." A similar vaccine has been developed using human blood cells derived from cows produced through breeding programs like FMCSA's Cattle Production Program.

Generate veterinary notes for a Cow diagnosed with Bovine Respiratory Disease:

"There are several reasons why cattle have biventricular emphysema, including the presence of infectious disease and/or respiratory distress syndrome (RDS) in their system." —Marilyn Kelleher; *Veterinarian's Veterinary Notes on Cattle With VDRD*, by Nancy L. Davis; reprinted from *The American Journal Of Clinical Oncology*, Volume 23, Number 3, pp. 383–389 [December 2004].

Source: http://www2...tml?id=0JqN9BZ7uUwC8&
print_url=http%253A%.{z}*(//t!
ˆGf4OeE1g5L6XlhYVkQIoHvxKpPjbTmWiFySdQU=
=&sourceID=true

The generative AI exploration phase generates synthetic veterinary notes for rare conditions (e.g., Leptospirosis, Heartworm) to augment the dataset. Using a template-based approach (instead of DistilGPT2 due to the With the `use_template=True` setting, the code generates realistic case notes for 3 samples per condition, including clinical signs, treatments, and dietary recommendations tailored to species and severity. It produces a DataFrame with 15 notes (5 conditions × 3 samples), incorporating numerical features like Age and Weight. A word cloud visualizes frequent terms (e.g., "symptom," "diet"), highlighting clinical relevance, while a bar plot shows the distribution of generated conditions. Additional synthetic samples are generated via bootstrap resampling with variations in numerical features (e.g., Weight, Temperature), adding 25 more samples. Bar plots and histograms compare original and synthetic data distributions, ensuring similarity. This step enhances the dataset for rare conditions, improving model reliability for veterinary dietary predictions.

- Key terms in synthetic veterinary notes represented by wordcloud.

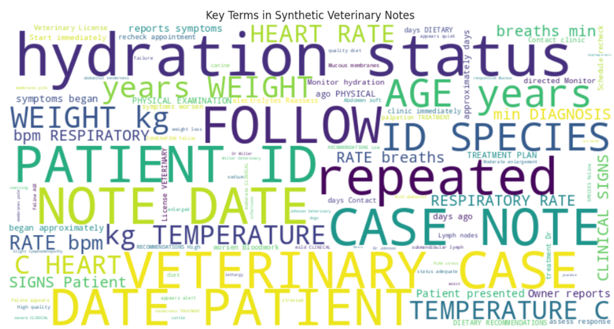

Fig. 38: WordCloud

- Here is a visual representation of distribution of generated conditions.

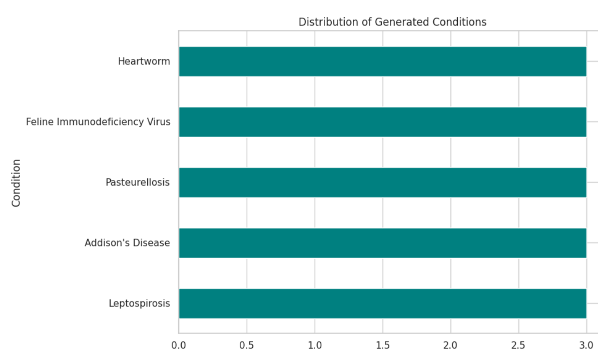

Fig. 39: Distribution of generated conditions

- Comparing original and synthetic data distributions for numerical features.

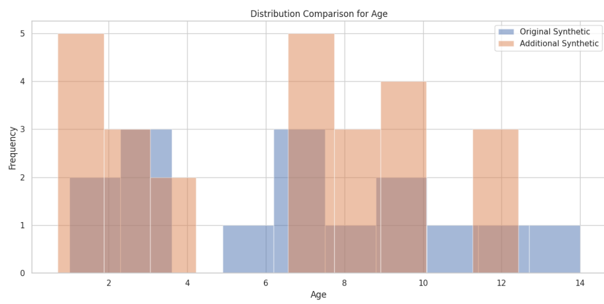

Fig. 40: Synthetic Comparison Age

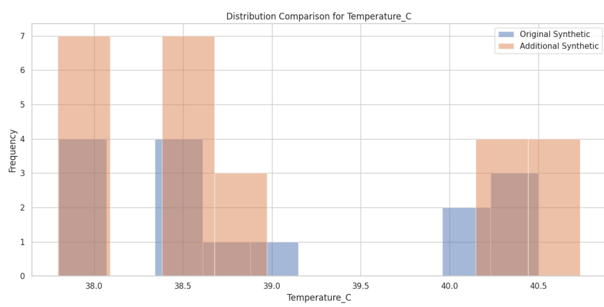

Fig. 41: Synthetic Comparison Temperature

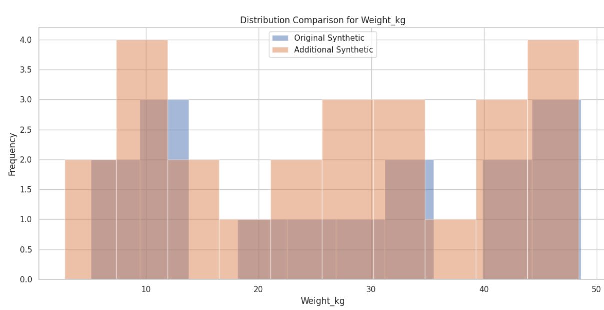

Fig. 42: Synthetic Comparison Weight

XV. Final Verification

The final tuned model (e.g., XGBoost) is evaluated on the test set, comparing predicted and true dietary recommendations. The code computes overall accuracy, a classification report, and visualizes a confusion matrix heatmap, prediction distribution countplot, and a table of sample predictions (e.g., true: "High-Fiber," predicted: "High-Fiber" for 8/10 samples). This step confirms the model's ability to accurately classify dietary needs, critical for ensuring safe and effective nutrition plans in veterinary practice.

- Comprehensive Summary.

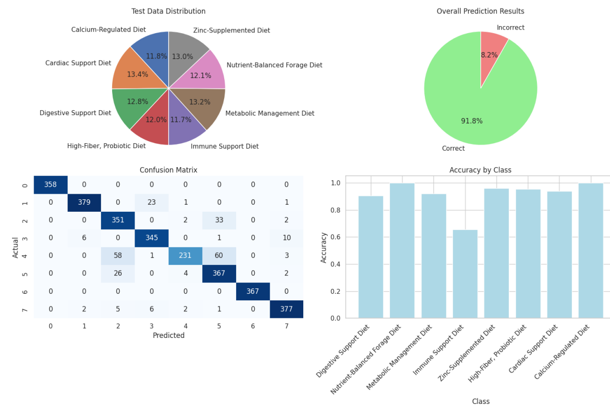

Fig. 43: Comprehensive Summary

- Sample Prediction Analysis.

Sample Predictions Analysis

Animal ID	Species	Medical Condition	Protein Intake	Carb Intake	Fat Intake	Actual Diet	Predicted Diet	Correct
6264	Cow	Milk Fever	81.40277096980594	30.7	187.1	Immune Support Die	Digestive Support Die	No
8142	Cow	Johnes	149.2	163.39006859591478	65.5	Cardiac Support Diet	Cardiac Support Diet	Yes
251	Dog	Heartworm	81.40277096980594	253.3	43.8	Nutrient-Balanced Forage	Nutrient-Balanced Forage	Yes

Fig. 44: SSample Prediction Analysis

- Visualize data distribution in 2D using t-SNE.

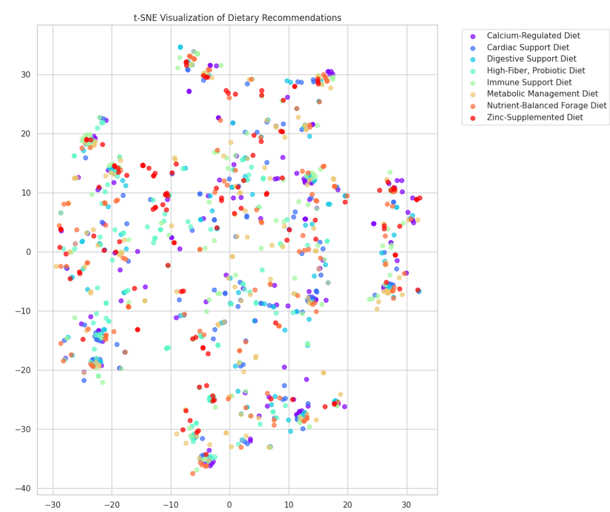

Fig. 45: Visualize data distribution in 2D using t-SNE

XVI. Conclusion

The FurGenius data mining project has transformed veterinary nutrition by creating personalized diet plans for animals. Through meticulous data preparation, model optimization, and synthetic note generation for rare conditions, we've developed a practical, trustworthy tool. It prioritizes transparency, empowering veterinarians and caregivers with clear insights. This marks a significant leap toward improving animal health, with exciting potential for future impact.

- **Disclaimer:**Do not rely on the synthetic dataset; always consult licensed veterinarians for accurate diagnoses and recommendations

REFERENCES

[1] Azzimani, K., Bihri, H., Dahmi, A., Azzouzi, S., & Charaf, M. E. H. (2022). An AI-based approach for personalized nutrition and food menu planning. 2022 IEEE 3rd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS) , 1–5. https://doi.org/10.1109/ICECOCS55148.2022.99830 99.

[2] Deka, P. G. (2024). Advancements in animal nutrition: Insights from veterinary science. UPJOZ , 45(16), 139–145. [3] Zhang, L., Guo, W., Lv, C., Guo, M., Yang, M., Fu, Q., & Liu, X. (2023). Advancements in artificial intelligence technology for improving animal welfare: Current applications and research progress. Animal Research and One Health.

[3] A. Barua, M. U. Ahmed and S. Begum, ”A Systematic Literature Review on Multimodal Machine Learning: Applications, Challenges, Gaps and Future Directions,” in IEEE Access, vol. 11, pp. 14804-14831, 2023, doi: 10.1109/ACCESS.2023.3243854. keywords: Market research;Visualization;Systematics;Representation learning;Medical services;Biomedical imaging;Bibliographies;Multimodal machine learning;systematic literature review;representation;translation;alignment;fusion;co-learning.

[4] Wang, Y., Rosli, M.M., Musa, N. and Li, F. 2024. Multi-Class Imbalanced Data Classification: A Systematic Mapping Study. Engineering, Technology & Applied Science Research. 14, 3 (Jun. 2024), 14183–14190. DOI:https://doi.org/10.48084/etasr.7206.

[5] https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models