Data Science and Engineering (DSE) Record, Volume 5, issue 1.

Modeling to Predict Cyprinid Herpes Virus 2 (CyHV-2) in Goldfish

Siriyaporn Rattana [1] and Pree Thiengburanathum[2]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] College of Arts, Media and Technology, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
siriyaporn_r@cmu.ac.th

Abstract. Goldfish are popular ornamental fish, especially in Thailand, which is one of the top importers of goldfish globally. Simultaneously, the industry for exporting fish is also among the top in the world. However, the exchange of fish from various sources makes them susceptible to diseases such as Cyprinid Herpesvirus 2, a significant disease in goldfish with a mortality rate of 50% to 100%. The purpose of this research is to design a predictive model to identify Cyprinid Herpesvirus 2 infections in goldfish, utilizing data gathered from 13 ornamental fish shops across 5 districts in Chiang Mai province. The dataset was imbalanced, with the number of non-infected samples (PCR=0) being higher than the infected samples (PCR=1). Therefore, bootstrapping was used to increase the number of infected samples by 54 to balance the dataset. Subsequently, the Mutual Information method was employed to determine the relationship scores between features and the infection variable (MI Score). The Fixed Threshold method was employed to select features most relevant to the infection variable from a total of 46 features based on MI Scores ranging from 0 to 0.24. Using the K-Nearest Neighbors model with n_neighbors=2, it was found that an MI Score of 0.19 was most suitable for this dataset. The features with an MI Score greater than or equal to 0.19 were the pH value of water, The water temperature, Total length of the fish (CM), and Length of the fish tank (CM). These variables were then used to train several models, including Decision Tree Classification model, Random Forest Classification model, Logistic Regression model, and K-Nearest Neighbors model. The Random Forest Classifier emerged as the most effective model, with training data results of {Accuracy: 99.992757%, Recall: 99.993054%, Precision: 99.992757%, F1 Score: 99.992748%}, and test data results of {Accuracy: 93.333333%, Recall: 91.666667%, Precision: 93.333333%, F1 Score: 93.055556%}.

Keywords: Machine-learning, Prediction, Fish Diseases

1 Introduction

The goldfish (Carassius auratus) is a freshwater species from the Cyprinidae family, making it a smaller relative of the carp family. It originates from China and Japan (Research Institute of Ornamental Aquatic Animals and Aquatic Plants, 2015). Goldfish, commonly raised in aquariums and ponds, typically have an average lifespan of 10 to

30 years. In contrast, wild goldfish can live up to an average of 41 years. They usually reach a size ranging from 4.7 to 16.1 inches (Mohr, K., n.d.). Renowned for their ornamental appeal, goldfish have been popular pets for generations, maintaining their popularity from past to present. These captivating fish hold a significant value and are often regarded as top-ranking aquatic pets.

Figure 1: The picture shows an example of a Pearlscale goldfish.

Currently, the cultivation of ornamental fish is highly popular, particularly goldfish. Thailand is among the world's major exporters of aquatic animals and aquatic animal products. The country's production and export value of aquatic animals have consistently seen continuous growth. Thailand holds the top position in goldfish imports, ranking 6th globally, with a value of 92,178 baht according to freshwater aquarium import statistics. Additionally, Thailand ranks 7th in the world for goldfish exports, with an export value of 46,287,180 baht based on freshwater aquarium export statistics (Department of Fisheries, 2021).

Table 1: Statistics on the Top 10 Freshwater Ornamental Fish Imports via Suvarnabhumi Airport in June 2021 (Department of Fisheries, 2021)

Aquatic species	Quantity (Unit)	Price (Baht)
1.Cardinal Tetra	52,500	94,252
2.Rummy-nose tetra	24,100	45,941
3.Ember tetra	14,660	31,715
4.Nerite snail	9,900	31,267
5.Neon tetra	8,700	23,972
6.Goldfish	7,773	92,178
7.Dwarf cichlid	6,750	38,344
8.Flagtail prochilodus	6,600	20,845
9.Chivellia fish	6,400	68,973
10.Cherry barb	6,395	11,556

Table 2: Statistics on the Top 10 Freshwater Ornamental Fish Exports via Suvarnabhumi Airport in June 2021 (Department of Fisheries, 2021)

Types of aquatic animals	Quantity (Unit)	Price (Baht)
1. Siamese fighting fish	1,651,217	17,679,719
2. Guppy	948,111	2,730,333
3. Swordtail fish	403,509	1,456,651
4. Flowerhorn cichlid	349,576	4,140,774
5. Bumblebee goby	287,878	773,720
6. Goldfish	235,364	4,316,312
7. Flying fox fish	216,657	765,613
8. German blue ram	176,673	967,568
9. Red-tailed catfish	166,420	429,549
10. Cherry shrimp	165,841	990,265

The exchange of fish from different breeding sources or environments significantly impacts disease outbreaks in fish. The most common diseases are bacterial (54.9%), viral (22.6%), parasitic (19.4%), and fungal (3.1%). Notably, viral diseases affecting aquatic animals include iridoviruses, reoviruses, rhabdoviruses, nodaviruses, and herpesviruses, with Cyprinid herpesvirus-2 being particularly detrimental to goldfish. A study by Waltzek et al. (2009) highlighted goldfish's high susceptibility to CyHV-2 infection, emphasizing the need for effective disease prevention and management strategies in aquaculture to mitigate economic losses.

Cyprinid herpesvirus-2 (CyHV-2), also known as Goldfish Hematopoietic Necrosis Virus, is a significant pathogen affecting goldfish. It belongs to the family of Herpes Viral Hematopoietic Necrosis Disease (HVHND) viruses and exhibits a mortality rate ranging from 50% to 100%. Infected fish display various symptoms, including reduced appetite, fatigue, gasping behavior, loss of equilibrium, and sometimes resting at the tank bottom before succumbing. Visible signs of infection encompass anemia, pale skin, and gums, protruding eyes on both sides, and erosion and necrosis of the gums. Timely treatment is crucial, as delayed intervention might lead to missed opportunities for effective treatment. Fish diseases are also prone to rapid transmission, particularly due to the common practice of housing multiple collected fish together in ponds. This often prevents the comprehensive testing of all fish and consequently leads to extensive damage caused by the spread of infection (Groff et al., 1998).

2 Literature Review

2.1 Using machine learning for infection classification

From the literature review, there are those who are interested and have studied models used for predicting various diseases or infections, such as SJ Divinely et al. (2019) aimed to classify fish diseases and identify significant factors influencing their

occurrence, primarily using fish images. The study found that the bio14 variable, representing precipitation in the driest month, significantly impacted disease confirmation, with an AUC of 0.761. Additionally, the Probabilistic Neural Network (PNN) algorithms demonstrated the highest accuracy in diagnosing fish diseases. This work enhances the understanding of fish disease classification and could improve aquatic animal health management in the future. (Divinely, S., & et al.,2019).

Next, In the research conducted by Golden et al. (2019), a comprehensive study was undertaken to investigate the application of machine learning techniques for predicting the prevalence of Listeria spp. in the environment of pasture-raised poultry farms. Specifically, this study aimed to compare the effectiveness of two popular models: Random Forest (RF) and Gradient Boosting Machine (GBM). The researchers collected soil and fecal samples from 11 pasture-based poultry farms over a period from 2014 to 2017, analyzing the data to assess the spread of Listeria spp. within this specific agricultural context. Although the study did not specify particular survey factors, the results demonstrated the strong predictive capabilities of both models. The RF model achieved an impressive Area Under the Curve (AUC) value of 0.905, indicating high accuracy in predictions, while the GBM model followed closely with an AUC of 0.855. However, when focusing on soil samples, the GBM model outperformed the RF model, reaching an AUC of 0.873 compared to RF's AUC of 0.700. (Golden C.E., & et al., 2019).

The research conducted by Liang et al. (2019) aimed to forecast the outbreak of African Swine Fever (ASF) and identify significant factors influencing the occurrence of this disease. The study extensively analyzed various meteorological factors, including annual mean temperature, daily temperature variations, temperature uniformity, seasonal temperature fluctuations, and total rainfall during different periods, such as the driest and wettest months. Data on ASF outbreaks were obtained from the Global Animal Disease Information System maintained by the Food and Agriculture Organization (FAO) and the World Organization for Animal Health (OIE). This dataset included crucial details such as the timing of outbreaks, geographic coordinates of affected areas, and the number of animals impacted. In addition to the outbreak data, climate-related information encompassing nineteen global variables was sourced from the WorldClim climate database. For data analysis, the researchers employed various machine learning methods, including Naïve Bayes, Random Forest, and Support Vector Machine (SVM). The findings revealed that the Random Forest algorithm achieved the highest accuracy, reaching 98.29% for the training set and 98.19% for the test set. Notably, the most significant factor associated with the confirmation of ASF outbreaks was identified as "rainfall during the driest month." This underscores the important role that meteorological factors play in the spread of ASF. The insights derived from this study could inform the development of effective management and prevention strategies for ASF in the future, based on the established relationships between weather conditions and disease outbreaks. (Liang, R., &et al., 2019).

Next, the study conducted by Malki et al. (2020) focuses on forecasting the number of confirmed COVID-19 cases and examining the relationship between weather variables and the transmission of the virus in Italy, although the sample size used in the

research is not specified. The primary objective of the study is to analyze various factors, including hours of sunlight, temperature, median age, population density, infection ratio, humidity, intensive care unit (ICU) beds per 1,000 people, urban percentage, fertility rate, and wind speed. To analyze the data, the research team employed a wide range of machine learning methods, including Decision Tree, K Neighbors Regressor, Extra Trees Regressor, Support Vector Machine, and Random Forest, among others, to assess the performance of each model. The analysis revealed that the KNN regressor achieved the highest performance regarding mean squared error (MSE) and root mean squared error (RMSE), with values of 1.49381e+07 and 3782.07, respectively. In contrast, the Extra Trees algorithm and Random Forest performed best in terms of mean absolute error (MAE) and root mean squared logarithmic error (RMSLE), yielding values of 365.563 and 1.3027, respectively. These findings underscore the significance of temperature and humidity as important factors in predicting COVID-19 death rates. (Malki, Z., & et al., 2020).

Niu et al. (2021) focuses on predicting the incidence of Peste des Petits Ruminants (PPR), a contagious disease affecting small ruminants such as sheep and goats. The study analyzes a dataset comprising 2,977 cases, aiming to forecast the occurrence of PPR and identify significant factors contributing to its emergence. Key factors examined in the study include the average temperature during the warmest quarter, seasonal variation in precipitation, and specific rainfall measurements recorded in July, August, and December, as well as altitude. The data utilized in this research spans outbreaks from 2008 to 2018, sourced from the Food and Agriculture Organization (FAO) for China, Bangladesh, and Morocco. Various machine learning methods were employed in the analysis, including BayesNet, Naive Bayes, and Random Forest. The results revealed that the Random Forest (RF) algorithm achieved an impressive accuracy rate of 99.10% in both training and testing datasets. Additionally, the study highlighted critical influencing factors, notably "precipitation during the warmest quarter," "rainfall recorded in July," and "rainfall recorded in August," which are crucial for developing effective strategies for disease management and prevention of PPR outbreaks in the future. (Niu, B. & et al., 2015).

L. J. Muhammad et al. (2021) focused on developing predictive models for COVID-19 infection using data from 263,007 recorded cases in Mexico. The primary objective of this research was to create a model capable of accurately forecasting COVID-19 infections by considering various significant factors, including age, pneumonia, asthma, cardiovascular diseases (CVDs), obesity, diabetes, chronic kidney diseases (CKDs), sex, hypertension, and tobacco use, which has been identified as a critical risk factor for increased infection likelihood. Additionally, the study utilized RT-PCR test results to confirm COVID-19 infections within the sample population. The researchers employed a range of machine learning methodologies, such as Support Vector Machine (SVM), Decision Tree, Naive Bayes, Artificial Neural Network (ANN), and Logistic Regression, to analyze the data and evaluate the predictive accuracy of each model. The analysis revealed that the Decision Tree model achieved the highest accuracy at 94.99%, which is a promising result for predicting COVID-19 infections. Furthermore, the SVM and Naive Bayes models demonstrated exceptional performance in terms of

sensitivity and specificity, with scores of 93.34% and 94.30%, respectively. (L.J. Muhammad & et al, 2021).

Next, Richard Du et al. (2021) focuses on the application of machine learning technology to predict SARS-CoV-2 infection based on blood test results and chest radiographs from 5,148 patients treated across 24 hospitals in Hong Kong. The primary objective was to develop an accurate prediction model that leverages various critical factors, including other related diseases and clinical infections such as bacterial pneumonia (Bacterial PNA) and viral pneumonia (Viral PNA), alongside COVID-19. In this research, the team employed multiple analytical approaches, including a machine learning model (ML model), a clinical model, radiologist consensus, and a combination of the radiologist and ML model. The results indicated that the machine learning model, utilizing commonly accessible laboratory markers, demonstrated excellent precision in predicting SARS-CoV-2 infection, with an area under the curve (AUC) ranging from 89.9% to 95.8%, sensitivity between 55.5% and 77.8%, and specificity from 91.5% to 98.3%. This highlights the potential of machine learning technology to effectively predict COVID-19 infections. (Richard, D. & et al, 2021).

The study conducted by Ibrahim et al. (2021) focuses on predicting COVID-19 infections using data from 114 patients treated at Taizhou Hospital in Zhejiang Province, China, between January 17 and February 1, 2020. The primary objective was to analyze and forecast COVID-19 infection based on various significant medical factors, including neutrophil-lymphocyte ratio, neutrophil count, hemoglobin levels, lymphocyte count, lymphocyte/monocyte ratio, basophils, mean corpuscular volume (MCV), platelet count, monocytes, white blood cell count, hematocrit, procalcitonin levels, eosinophil percentage (E%), mean red blood cell volume, thrombocytocrit, and the results from RT-PCR tests for COVID-19.In this study, the research team utilized a range of statistical analyses and machine learning techniques, including a meta-classifier (Classification via Regression), logistic regression, lazy classifier (IBk), rule-learner (PART), decision-tree (J48), and Bayes classifier to create highly accurate predictive models. Among these, the CR meta-classifier demonstrated the highest accuracy rate of 84.21%, establishing it as the most effective model for predicting COVID-19 infection based on the analyzed data. (Ibrahim, A. & et al, 2021).

Finally, the authors of this study are Safavi et al. (2022), with the primary objective of forecasting Lumpy Skin Disease Virus (LSDV) infections based on meteorological and geological characteristics. While the specific sample size is not disclosed, the research examines several critical factors, including meteorological data, animal population density, land cover data, and elevation information. The outbreak data for Lumpy Skin Disease was obtained from the Climatic Research Unit (CRU TS4.04), covering the period from January 2011 to December 2019. Additionally, data regarding the density of cattle and buffalo populations were sourced from the GLW 3 database on global livestock distribution. Land cover information was gathered from the GLC-SHARE Version 1.0 (Beta Release), and elevation data were retrieved from Version 2.1.0 of the Global Geospatial Elevation Dataset (GRAY_50M_SR) available through the Natural Earth database. For data analysis, the researchers employed various modeling techniques, including Random Forest, AdaBoost, Logistic Regression, Bagging, Support Vector Machine, Decision Tree, Artificial Neural Networks (ANN), and XGBoost. The

results indicated that the ANN model demonstrated the highest performance in terms of area under the curve (AUC), achieving a value of 0.97. This finding highlights the significance of meteorological variables as influential factors in the predictive analysis conducted in this study. Overall, the research underscores the importance of integrating meteorological and geographical data to analyze and predict disease occurrences in livestock, which can aid in effective disease management and mitigate the risk of pathogen spread in the future. (Safavi, E., 2022).Data and Methodology

2.2 Data

The data obtained from the research on 'The survey of Cyprinid herpes virus 2 (CyHV-2) in goldfish from ornamental fish shops in Chiang Mai Province' by Chayanit and Witit, which includes a total of 62 variables and 102 samples.

1) **Explained variable:** Y: PCR Test result (0: Not infected, 1: Infected)

2) **Explanatory variable:**

Table 3: shows the meaning of each variable as follows.

Column Name	Data Description	Unit	Measurement level	Data Type
date	Date of the survey	-	ordinal	date
codf	Fish sample code	-	nominal	text
DO	Dissolved oxygen	milligrams per liter	ratio	float
temp	Water temperature	degrees Celsius	interval	float
ammo	Total ammonia value in water	milligrams per liter	ratio	float
nit	Nitrite value in water	milligrams per liter	ratio	float
ph	Water's acidity or alkalinity value (pH)	-	interval	float
tw	Fish tank width	centimeter	ratio	float
tl	Length of the fish tank	centimeter	ratio	float
th	Height of the fish tank	centimeter	ratio	float
wh	Height of the water level in the fish tank	centimeter	ratio	float
vol	Volume of water in the fish tank	liters	ratio	float
den	fish stocking density	-	ordinal	integer
den1	low stocking density	-	nominal	binary
den2	moderate stocking density	-	nominal	binary
den3	high stocking density	-	nominal	binary
den4	Specify the number of fish in the tank	fish count	ratio	integer
den5	Categorize the number of fish in the tank	-	ordinal	integer
sw1	Fish Swimming Behavior: Normal	-	nominal	binary
sw2	Fish Swimming Behavior: Swimming beneath the water surface	-	nominal	binary
sw3	Fish Swimming Behavior: Gasping	-	nominal	binary
sw4	Fish Swimming Behavior: flashing	-	nominal	binary
sw5	Fish Swimming Behavior: floating	-	nominal	binary
sw6	Fish Swimming Behavior: bottom sitting	-	nominal	binary
sw7	Fish Swimming Behavior: ataxia	-	nominal	binary
sw8	Fish Swimming Behavior: erratic	-	nominal	binary
bf	Fish Behavior	-	nominal	integer
bf1	Fish Behavior: Lethargy	-	nominal	binary
bf2	Fish Behavior: Alert response	-	nominal	binary
rr	Respiratory rate	times per minute	ratio	integer

bw	Fish weight	gram	ratio	float
tlf	Total length of the fish	centimeter	ratio	float
slf	Standard length of the fish	centimeter	ratio	float
exl	External signs of diseases	-	nominal	text
exlg1	Grouping External Signs Based on Systems: normal	-	nominal	binary
exlg2	Grouping External Disease Signs by System: Abnormalities in gills	-	nominal	binary
exlg3	Grouping External Disease Signs by System: Abnormalities in skin and fins	-	nominal	binary
exlg4	Grouping External Disease Signs by System: Systemic symptoms such as abnormalities in internal organs, fluid in the abdominal cavity, and body swelling	-	nominal	binary
exlg5	Grouping External Disease Signs: Other Abnormalities	-	nominal	binary
ects	External Parasites on Skin	-	nominal	binary
ectst	Type of parasites found on the skin	-	nominal	text
ectst2	Group of parasites found on the skin	-	nominal	integer
ectsn4	Number of parasites found on the skin at 40x magnification	-	nominal	text
sectsn4	Number of parasites on the skin at 40x magnification	-	ordinal	integer
ectg	External parasites near gills	-	nominal	binary
ectgt	Type of parasites found near gills	-	nominal	text
ectgt2	Group of parasites found near gills	-	nominal	integer
ectgn4	Number of parasites found near gills at 40x magnification	-	nominal	text
sectgn4	Total number of parasites near gills at 40x magnification	-	ordinal	integer
level	Evaluate the severity of parasite infestation	-	ordinal	integer
qua1	Goldfish disease quarantine	-	nominal	binary
qua2	Goldfish Disease Quarantine Method	-	nominal	text
qua3	Group the methods of goldfish disease quarantine	-	nominal	integer
dura	Duration of releasing fish into the tank	-	ordinal	integer
lot	Separation of containers by lot	-	nominal	binary
size	Separation of containers by fish size	-	nominal	binary
sepa1	Separation of raising new fish batches from old fish batches.	-	nominal	binary
sepa2	Raising old and new fish batches together	-	nominal	binary
ndna	Quantity of DNA	A260/A280	ratio	float
dsdna	Quantity of DNA	dsDNA(μg/ml)	ratio	float
pcr1	First-round PCR test results	-	nominal	integer
pcr2	Second-round PCR test results	-	nominal	integer

	date	codf	DO	temp	ammo	nit	ph	tw	tl	th	...	qua3	dura	lot	size	sepa1	sepa2	ndna	dsdna	pcr1	pcr2
0	9-9-2564	AD1	6.30	24.7	1.0	0.50	7.80	25.0	73.0	38.0	...	4	NaN	1	1	1	1	1.907	751.31	0	0
1	9-9-2564	AD2	7.30	24.6	0.2	0.01	7.89	43.0	66.5	46.0	...	4	NaN	1	1	1	1	1.987	1579.00	1	0
2	8-10-2564	AD3	5.09	25.0	0.2	0.05	7.45	42.0	67.0	43.0	...	1	1.0	1	1	1	1	1.835	226.00	1	1
3	8-10-2564	AD4	4.47	27.5	0.2	0.05	7.55	39.0	75.0	38.0	...	1	1.0	1	1	1	1	1.880	334.46	0	0
4	8-10-2564	AD5	4.19	24.8	0.4	0.20	7.40	75.0	90.0	46.0	...	1	1.0	1	1	1	1	1.865	468.38	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
97	10-16-2564	MJ4	5.73	21.9	2.0	1.00	6.32	39.0	76.0	39.0	...	1	2.0	1	1	1	1	2.023	1933.20	0	1
98	10-16-2564	MJ5	6.22	21.6	3.0	0.02	4.93	39.0	76.0	39.0	...	1	2.0	1	1	1	1	2.110	1863.10	0	0
99	12-11-2564	MJ6	6.08	23.5	1.0	0.05	7.82	91.0	111.0	29.0	...	1	2.0	1	1	1	1	1.883	2417.60	1	1
100	12-11-2564	MJ7	6.22	23.0	1.0	0.50	7.23	39.0	76.5	39.0	...	1	2.0	1	1	1	1	1.879	3462.80	0	0
101	12-11-2564	MJ8	5.31	23.7	0.6	0.20	7.46	39.0	76.5	39.0	...	1	2.0	1	1	1	1	1.842	2757.20	1	1

102 rows × 62 columns

Figure 2: Load related datasets.

2.3 Methodology

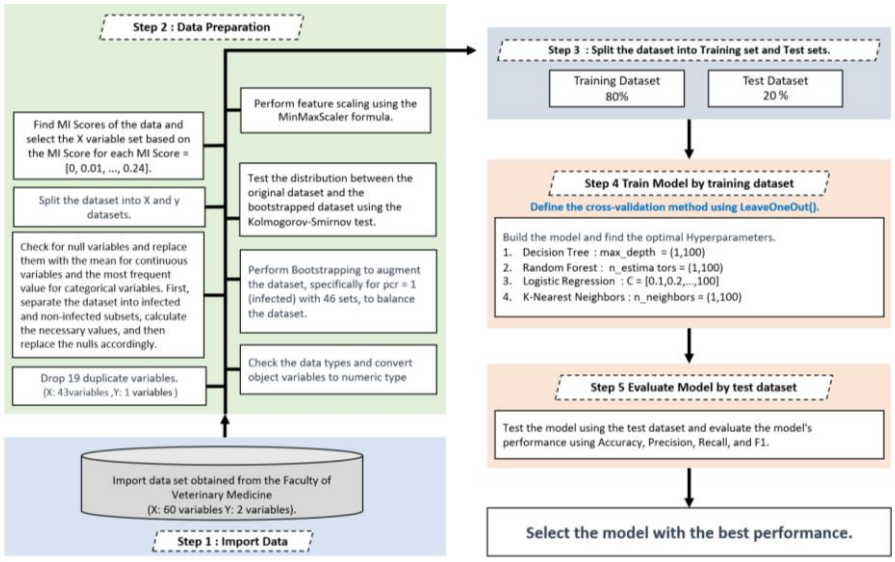

Figure 3: The research framework overall design.

1) Import Data

The research framework shown in Figure 3 is structured in five phases. The first phase involves importing the entire dataset with 62 variables from the Faculty of Veterinary Science's Excel file into Google Sheets.

2) Data Preparation

The second step involves data preparation, consisting of eight sub-steps. The first sub-step is removing duplicate columns by discarding 19 unnecessary columns. The second sub-step involves checking data types and converting object variables to

numeric types. Handling missing values is the focus of the third sub-step, where missing entries are replaced with the mean for continuous variables and the mode for categorical variables. This requires first separating the dataset into infected and non-infected groups, calculating necessary values, and then replacing missing values accordingly. The fourth sub-step deals with imbalanced datasets, where non-infected samples (PCR = 0) outnumber infected ones (PCR = 1). To resolve this, Bootstrapping is applied to increase the infected dataset to 46 samples. The fifth sub-step tests the distribution between the original and bootstrapped datasets using the Kolmogorov-Smirnov test. The sixth sub-step consists of partitioning the dataset into Dataset X and Dataset Y. The seventh sub-step focuses on finding correlations and Mutual Information (MI) scores to evaluate the relationship between X and y and selecting features suitable for the K-Nearest Neighbors model (n_neighbors=2) using the Fixed Threshold method within the range of 0.00 to 0.24. Model performance is gauged through accuracy metrics on training and test datasets, identifying the MI scores that achieve the most effective results. Features with MI scores meeting or exceeding the selected thresholds are considered relevant to infection and will be used for developing prediction models in subsequent experiments.After applying the Fixed Threshold method to select MI Scores, we found that the optimal MI Score is 0.19. This consists of 4 features with MI Scores greater than or equal to 0.19, namely the pH value of water, the water temperature, total length of the fish (CM) and length of the fish tank (CM). The highest accuracy values obtained for the training dataset and the test dataset are 97.4576 and 93.3333, respectively.The eighth sub-step involves scaling continuous datasets using Min Max Scaler.

Table 4 : The accuracy values of the training dataset and the test dataset from the KNeighborsClassifier model (n_neighbors=2) at each MI Scores value

MI scores	Accuracy Training	Accuracy Test	Number of Variables	Independent Variables (X)
0.00	97.4576	90.0000	42	ph,temp,tlf,tl,DO,rr,slf,ectst2,vol,wh,th,ectgt2,ammo,sectgn4,sepa1,qua3,exlg2,bw,tw,den5,level,sectsn4,lot,den,nit,sw2,dura,sw3,exlg1,sepa2,sw5,sw6,sw7,exlg4,sw8,exlg5,exlg3,bf1,bf2,sw1,sw4,size
0.01	97.4576	90.0000	30	ph,temp,tlf,tl,DO,rr,slf,ectst2,vol,wh,th,ectgt2,ammo,sectgn4,sepa1,qua3,exlg2,bw,tw,den5,level,sectsn4,lot,den,nit,sw2,dura,sw3,exlg1,sepa2
0.02	97.4576	90.0000	24	ph,temp,tlf,tl,DO,rr,slf,ectst2,vol,wh,th,ectgt2,ammo,sectgn4,sepa1,qua3,exlg2,bw,tw,den5,level,sectsn4,lot,den
0.03	97.4576	90.0000	24	ph,temp,tlf,tl,DO,rr,slf,ectst2,vol,wh,th,ectgt2,ammo,sectgn4,sepa1,qua3,exlg2,bw,tw,den5,level,sectsn4,lot,den

0.04	97.4576	90.0000	24	ph,temp,tlf,tl,DO,rr,slf,ectst2,vol,wh,th,ectg t2,ammo,sectgn4,sepa1,qua3,exlg2,bw,tw,den5, level,sectsn4,lot,den
0.05	97.4576	90.0000	21	ph,temp,tlf,tl,DO,rr,slf,ectst2,vol,wh,th,ectg t2,ammo,sectgn4,sepa1,qua3,exlg2,bw,tw,den5, level
0.06	97.4576	80.0000	17	ph,temp,tlf,tl,DO,rr,slf,ectst2,vol,wh,th,ectg t2,ammo,sectgn4,sepa1,qua3,exlg2
0.07	97.4576	80.0000	16	ph,temp,tlf,tl,DO,rr,slf,ectst2,vol,wh,th,ectg t2,ammo,sectgn4,sepa1,qua3
0.08	97.4576	80.0000	15	ph,temp,tlf,tl,DO,rr,slf,ectst2,vol,wh,th,ectg t2,ammo,sectgn4,sepa1
0.09	97.4576	80.0000	14	ph,temp,tlf,tl,DO,rr,slf,ectst2,vol,wh,th,ectg t2,ammo,sectgn4
0.10	97.4576	80.0000	14	ph,temp,tlf,tl,DO,rr,slf,ectst2,vol,wh,th,ectg t2,ammo,sectgn4
0.11	97.4576	80.0000	13	ph,temp,tlf,tl,DO,rr,slf,ectst2,vol,wh,th,ectg t2,ammo
0.12	97.4576	83.3333	12	ph,temp,tlf,tl,DO,rr,slf,ectst2,vol,wh,th,ectg t2
0.13	97.4576	83.3333	11	ph,temp,tlf,tl,DO,rr,slf,ectst2,vol,wh,th
0.14	98.3051	90.0000	9	ph,temp,tlf,tl,DO,rr,slf,ectst2,vol
0.15	98.3051	80.0000	8	ph,temp,tlf,tl,DO,rr,slf,ectst2
0.16	97.4576	86.6667	7	ph,temp,tlf,tl,DO,rr,slf
0.17	98.3051	83.3333	6	ph,temp,tlf,tl,DO,rr
0.18	96.6102	90.0000	5	ph,temp,tlf,tl,DO
0.19	**97.4576**	**93.3333**	**4**	**ph,temp,tlf,tl**
0.20	96.6102	83.3333	3	ph,temp,tlf
0.21	98.3051	80.0000	2	ph,temp
0.22	98.3051	80.0000	2	ph,temp

3**) Split the dataset into Training set and Test sets.**

Following the completion of all eight data preparation steps, the third step involves dividing the data into 80% for training and 20% for testing, utilizing the train_test_split function from the sklearn module in Python.

4**) Modeling**

In the fourth step, we will train the models and find the best hyperparameters for each model. We will consider Decision Tree Classifier, Random Forest Classifier, Logistic Regression, and K-Nearest Neighbors (KNN) Classification. We will use Leave-One-Out Cross-Validation for this purpose.

In the final step, we will evaluate the performance of all four models by comparing metrics such as accuracy, precision, recall, and F1 score to identify the most suitable model for this dataset. Following that, a website will be developed using Python's Streamlit, enabling public access to the model.

5) Results

Table 5 : comparing the performance of all 4 models on the training dataset

Model	Hyperparameter	%Accuracy	%Recall	%Precision	%F1 score
Decision Tree	max_depth=13	99.601622	99.578885	99.601622	99.600035
Random Forest	**n_estimators=18**	**99.992757**	**99.993054**	**99.992757**	**99.992748**
Logistic Regression	C=100	65.848182	65.670461	65.848182	65.679836
K-Nearest Neighbors	n_neighbors =2	97.341735	97.469391	97.341735	97.339979

Table 6 : comparing the performance of all 4 models on the test dataset

Model	Hyperparameter	%Accuracy	%Recall	%Precision	%F1 score
Decision Tree	max_depth=13	86.666667	83.333333	86.666667	86.111111
Random Forest	**n_estimators=18**	**93.333333**	**91.666667**	**93.333333**	**93.055556**
Logistic Regression	C=100	70.000000	75.000000	70.000000	69.696970
K-Nearest Neighbors	n_neighbors =2	80.000000	75.000000	80.000000	79.166667

From the study and development of suitable models for predicting CyHV-2 infection in goldfish from 13 ornamental fish shops in Chiang Mai province, it was found that the data was imbalanced and there were fewer data samples, with non-infected fish samples being 72.55% more than infected ones. Therefore, the Bootstrapping technique was used to increase the number of infected fish sam-ples by adding another 46 samples to balance the dataset. After that, the Mutual Information technique was used to find the MI Score, which indicates the rela-tionship between features and the infection

variable. The Fixed Threshold method was then employed to select features from all 42 features related to the infec-tion variable, with MI Scores ranging from 0 to 0.24. Using the K-Nearest Neighbors model with n_neighbors=2, it was found that an MI Score of 0.19 was most suitable for this dataset. The features with an MI Score greater than or equal to 0.19 were the pH value of water, The water temperature, Total length of the fish (CM), and Length of the fish tank (CM). These four features were then used to develop models, including Decision Tree Classifier, Random Forest Classifier, Logistic Regression, and K-Nearest Neighbors. The most suitable model was found to be the Random Forest Classifier, with {Accuracy, Recall, Precision, F1 score} values of {99.992757%, 99.993054%, 99.992757%, 99.992748%} for the training data and {93.333333%, 91.666667%, 93.333333%, 93.055556%} for the test data, respectively.

To enhance clarity and detail, the study on Cyprinid herpesvirus 2 (CyHV-2) infection in goldfish (Carassius auratus) presents an in-depth analysis of the factors influencing viral infection. It combines statistical and biological insights using machine learning modeling and environmental data. This research employs a Decision Tree Classifier to examine the relationship between environmental factors such as water pH, temperature, Length of the fish tank, and Total length of the fish and CyHV-2 infection in goldfish. The model determined that a hyperparameter of max_depth=13 optimally increased prediction accuracy. Furthermore, unscaled data provided results that better reflected real-world conditions, with Figure 4 highlighting key infection-related factors.

- Impact of Water Temperature

The findings indicate that temperatures between 15°C and 25°C significantly facilitate CyHV-2 transmission and disease development. This insight aligns with Thangaraj et al., 2021, which reported that positive water samples were mostly within 25.2°C to 25.9°C. However, the study acknowledges the limitation of single-time-point water quality data collection, which may constrain long-term infection analysis (Thangaraj, R.S. & et al, 2021).

- Total length of the fish

The study confirms that goldfish measuring 4.05-6.45 cm in length are significantly more susceptible to CyHV-2 infection. This aligns with Jeffery et al., 2007, which found that fish under 7 cm in length experienced higher mortality rates, suggesting that smaller fish are more vulnerable to outbreaks (Jeffery, K.R. & et al, 2007).

- Water pH

The research reveals that pH levels above 7.805 significantly increase CyHV-2 infection risk. This finding supports Amna Marium et al., 2023, which identified an optimal pH range of 6.5-9.0 for reducing fish stress and enhancing immunity. Deviations from this range may increase stress and weaken immune defenses, heightening viral susceptibility (Marium, A. & et al, 2023).

- Length of the fish tank

Data suggest that tanks measuring less than or equal to 149.75 cm effectively reduce CyHV-2 infection risk. Larger ponds can mitigate crowding and improve water quality. While research on pond dimensions and CyHV-2 infection is in its infancy, initial results emphasize that appropriate pond size can alleviate stress and bolster goldfish immunity.

Figure 4: The export_graphviz output from a Decision Tree Classifier

References

1. Mohr, K. (n.d.). Goldfish aren't the ho-hum fish you thought they were. From: https://www.nationalgeographic.com/animals/fish/facts/goldfish
2. Research Institute of Ornamental Aquatic Animals and Aquatic Plants (2015). Goldfish variety standards and judging criteria in Thailand. From: https://www4.fisheries.go.th/local/file_document/20191004135516_1_file.pdf
3. Department of Fisheries (2021). Statistics of importing and exporting aquatic animals a Suvarnabhumi Airport as of June. From: https://www4.fisheries.go.th/local/index.php/main/view_qr_group/178/1436
4. Groff, J.M., LaPatra, S.E., Munn, R.J., & Zinkl, J.G., (1998). A viral epizootic in cultured populations of juvenile goldfish due to a putative herpesvirus etiology. Journal of Veterinary Diagnostic Investigation 10, 375-378.
5. Thangaraj, R.S., Nithianantham, S.R., Dharmaratnam, A., Kumar, R., Pradhan, P.K., Thangalazhy Gopakumar, S., & Sood, N., (2021). Cyprinid herpesvirus-2 (CyHV-2): a comprehensive review. Reviews in Aquaculture 13, 796-821.
6. Waltzek, T.B., Kurobe, T., Goodwin, A.E., & Hedrick, R.P., (2009). Development of a PolymeraseChain Reaction Assay to Detect Cyprinid Herpesvirus 2 in Goldfish. Journal of Aquatic Animal Health 21, 60-67.
7. Kengsanguansit, C. & Phumsaringkharn, W., (2021). The survey of Cyprinid herpes virus 2 (CyHV-2) in goldfish from ornamental fish shop in Chiang Mai Province.
8. Divinely, S., K.Sivakami, K. & Jayaraj, V, (2019) Fish diseases identification and classification using Machine Learning.
9. Malki, Z., Atlam, E.-S., Hassanien, A.E., Dagnew, G., Elhosseini, M., & Gad, I. (2020). Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. Chaos, Solitons & Fractals
10. Golden, C.E., Rothrock, M.J., & Mishra, A. (2019). Comparison between random forest and gradient boosting machine methods for predicting Listeria spp. prevalence in the environment of pastured poultry farms. Food Research International,122,47-55.
11. Liang, R., Lu, Y., Qu, X., Su, Q., Li, C., Xia, S., Liu, Y., Zhang, Q., Cao, X., Chen, Q., &Niu, B. (2019). Prediction for global African swine fever outbreaks based on a combination of random forest algorithms and meteorological data.
12. Safavi, E. (2022). Assessing machine learning techniques in forecasting lumpy skin disease occurrence based on meteorological and geospatial features.
13. Niu, B., Liang, R., Zhou, G., Zhang, Q., Su, Q., Qu, X., & Chen, Q. (2021). Prediction for Global Peste des Petits Ruminants Outbreaks Based on a Combination of Random Forest Algorithms and Meteorological Data
14. L. J. Muhammad, Ebrahem A., A., Sani S.U., Abdulkadir, A., Chinmay,C., & I. A. Mohammed(2021). Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset.
15. Ibrahim, A., Shigao, H., Mostafa, A.-E., Mohammed N., A.-K., & Minfei, P. (2021). Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms.
16. Silvia, G., Sven M., B., …, & Heike, S.-P. (2015). Herpesviral Hematopoietic Necrosis in Goldfish in Switzerland: Early Lesions in Clinically Normal Goldfish (Carassius auratus) From: https://journals.sagepub.com/doi/full/10.1177/0300985815614974
17. IMB (n.d.). What is a Decision Tree? From:https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,internal%20nodes%20and%20leaf%20nodes

18. Chaiyaphop, J. (n.d.). รู้จักกับ Decision Tree มันคือต้นไม้อะไร ทำงานอย่างไร?
From:https://www.born-todev.com/2022/09/15/%E0%B8%A3%E0%B8%B9%E0%B9%89%E0%B8%88%E0%B8%B1%E0%B8%81%E0%B8%81%E0%B8%B1%E0%B8%9A-decision-tree/
19. Shailey, D. (2022). Decision Trees Explained — Entropy, Information Gain, Gini Index, CCP Pruning.
From: https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c
20. Tony, Y. (2019). Understanding Random Forest.
From: https://towardsdatascience.com/understanding-random-forest-58381e0602d2
21. Onesmus, M. (2022). Introduction to Random Forest in Machine Learning.
From: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/
22. Pimphak, T. (n.d.)., การวิเคราะห์การถดถอยโลจิสติกทวิ, STATISTICS FOR DATA SCIENCE (229711)
23. Plumb, J.A., 1999. Overview of Warmwater Fish Diseases. Journal of Applied Aquaculture 9, 1-10.
24. Jeffery, K.R., Bateman, K., Bayley, A., Feist, S.W., Hulland, J., Longshaw, C., Stone, D.,Woolford, G., Way, K., 2007. Isolation of a cyprinid herpesvirus 2from goldfish, Carassiusauratus (L.), in the UK. Journal of Fish Diseases 30, 649-656.
25. Richard Du, Efstratios D.Tsougenis, JoshuaW. K. Ho, Joyce K.Y. Chan, Keith W. H. Chiu, Benjamin X. H. Fang, MingYen Ng, Siu-Ting Leung, Christine S.Y. Lo,Ho-Yuen F. Wong, Hiu-Yin S. Lam, Long-Fung J. Chiu, TifanyY So, KaTak Wong,Yiu Chung I. Wong, KevinYu, Yiu-CheongYeung, Thomas Chik, JoannaW. K. Pang, Abraham Ka-chungWai, Michael D. Kuo, Tina P. W. Lam, Pek-Lan Khong,Ngai-Tseung Cheung & Varut-Vardhanabhuti,2021. Machine learning application for the prediction of SARS-CoV-2 infection using blood tests and chest radiograph.
26. Data Innovation and Governance Institute, DIGI,2023. สรุปให้!! วิธีการคำนวนด้วย K-Nearest Neighbors
From: https://digi.data.go.th/blog/what-is-k-nearest-neighbors/
27. Two-Sample Kolmogorov-Smirnov Test, Real Statistics
From: https://real-statistics.com/non-parametric-tests/goodness-of-fit-tests/two-sample-kolmogorov-smirnov-test/
28. Marium,A , Chatha, A. M. M.,Naz,S.,FarhanKhan,M., Safdar,W. , Ashraf,L.(2023). Effect of Temperature, pH, Salinity and DissolvedOxygenon Fishes