Use of machine-learning algorithms to predict urine culture results in cats

Materials and methods

Feature preprocessing

Patient data, including sex, age, color, clarity, USG, and ph, were utilized as features. Numeric features: protein, glucose, ketones, hemoprotein, bilirubin, WBC, and RBC were accounted for to target positive culture. Ignore_features (visit number; BW) were excluded from training as its values contain NaN values and strings. A positive culture was encoded using the one-hot encoding (0 and 1). Cat sex was splitted into 4 different groups.

Machine-learning model training

Two Feline UTI prediction models were trained. The ExtraTreesClassifier (ET) model was trained with all patient features excluding Sed Bac Quantity (SBQ) and Sed Bac Type (SBT). The second model, ExtraTreesClassifier (ET) trained with excluded features SBQ; SBT. The first model excluded SBQ and SBT due to limited resources to obtain data from the veterinary lab. The second model was trained to analyze AUC performance by just training the excluded data.

	predict neg (0)	predict pos (1)
true neg (0)	962	31
true pos (1)	100	123

First model (Data_limted)

	predict neg (0)	predict pos (1)
true neg (0)	954	39
true pos (1)	105	118

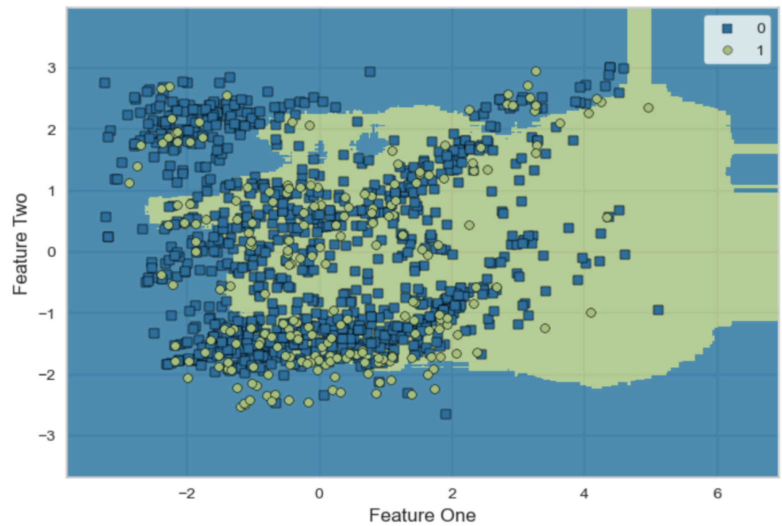
Second model (SBQ_SBT)

Figure 1. Data table illustrating population from two different models.

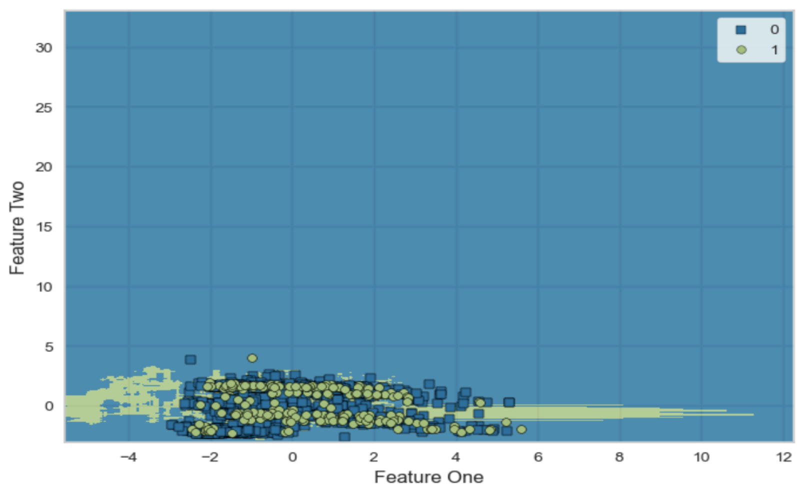
The ET model (Fig.1) utilized the training set data, allowing us to determine how numerical features affect the prediction performance for Feline UTI. The model trained using a repeated 10-folds for each of the 10 candidates, totaling 100 times to produce the average confusion matrix to tune the hyperparameters, using the built-in function:

Tuned_et = tune_model(et) (Python3; ipykernel).

The boundary plot (Fig. 2) visualizes how plots spread for trained set data. Through visualization, we can analyze that all trained data sets were spread out for the first model (Data_limited), and the decision boundary will be in the negative linear slope form. While the second model (SBQ, SBT) reflects skewed trained data sets on the right-hand side of the plot, it will include the decision boundary in the form of a curve hyper-plane to fit the dimension of the space.



First model (Data_limted)



Second model (SBQ_SBT)

Figure 2. Data table illustrating AUC performances between two different models.

Machine-learning model performance evaluation

The test set included [types of cats]. The hyperparameters optimized on the training set utilized during ET (with all features excluding sed features) and ET (with only sed features) model testing. Model prediction results reported sensitivity and specificity. Predictions made utilized as a binary classifier, a receiver operating characteristic (ROC) plot was generated, and the area under the curve (AUC) was calculated.

Results

During the study period, 6,732 unique cats were evaluated. Of these cats, 5,850 were tested for Feline UTI prediction: 4,634 in the training set, and 1,216 in the test set. Under the category of “Positive culture”, there were a total of 4,052. Between collected data, 3,310 of the training sets were under 0 while 742 remaining data sets were set to 1; 0 meaning negative and 1 meaning positive.

The (Fig. 3) displays the summary of collected statistics for the training and test set. Numeric variables (mean, standard deviation): protein (67.129, 101.355), glucose (88.968, 265.289), ketones (0.513, 5.761), hemoprotein (136.512, 113.445), bilirubin (0.178, 0.821), WBC (11.320, 23.903), RBC (39.359, 43.832). Categorical variables resulted in the count of 4052.

The data was measured using the built-in function:

.describe() (Python3; ipykernel).

	Positive culture	sex	Age	BW	USG	ph	Protein	Glucose	Ketones	Bilirubin	Hemoprotein	WBC	RBC
count	3310.0	3310.000000	3310.000000	3146.000000	3310.000000	3310.000000	3310.000000	3310.000000	3310.000000	3310.000000	3310.000000	3310.000000	3310.000000
mean	0.0	2.839577	10.024864	4.737307	1.025450	6.321752	65.202417	91.540785	0.530211	0.202417	133.607251	5.409366	38.916918
std	0.0	1.070754	5.215017	1.620530	0.163491	0.836666	100.236144	268.657465	5.678356	0.878950	114.023744	12.100120	43.904503
min	0.0	-9.000000	0.000000	0.000800	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.0	2.000000	5.900000	3.600000	1.012000	6.000000	25.000000	0.000000	0.000000	0.000000	10.000000	2.000000	2.000000
50%	0.0	2.000000	10.500000	4.500000	1.018000	6.000000	25.000000	0.000000	0.000000	0.000000	150.000000	3.000000	12.000000
75%	0.0	4.000000	14.100000	5.600000	1.033000	7.000000	75.000000	0.000000	0.000000	0.000000	250.000000	3.000000	100.000000
max	0.0	4.000000	23.900000	34.000000	10.220000	9.000000	500.000000	1000.000000	150.000000	6.000000	250.000000	100.000000	100.000000

First model (Data_limted) / col = 0

	Positive culture	sex	Age	BW	USG	ph	Protein	Glucose	Ketones	Bilirubin	Hemoprotein	WBC	RBC
count	742.0	742.000000	742.000000	701.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000
mean	1.0	3.218329	11.416173	4.459857	1.020263	6.261456	75.721024	77.493261	0.438005	0.072776	149.467655	37.687332	41.331536
std	0.0	1.178204	5.354555	1.630667	0.011965	0.846743	105.847516	249.558686	6.117677	0.470550	109.971563	40.216644	43.482828
min	1.0	-9.000000	0.000000	0.450000	1.004000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.0	2.000000	7.400000	3.300000	1.012000	6.000000	25.000000	0.000000	0.000000	0.000000	25.000000	3.000000	2.000000
50%	1.0	4.000000	12.800000	4.200000	1.016000	6.000000	30.000000	0.000000	0.000000	0.000000	250.000000	20.000000	18.000000
75%	1.0	4.000000	15.700000	5.400000	1.025000	7.000000	75.000000	0.000000	0.000000	0.000000	250.000000	100.000000	100.000000
max	1.0	4.000000	24.600000	14.300000	1.066000	9.000000	500.000000	1000.000000	150.000000	6.000000	250.000000	100.000000	100.000000

First model (Data_limted) / col = 1

	Positive culture	Sed Bac Quantity	Sed Bac Type
count	3310.0	3310.000000	3310.000000
mean	0.0	0.052266	0.056495
std	0.0	0.330735	0.357299
min	0.0	0.000000	0.000000
25%	0.0	0.000000	0.000000
50%	0.0	0.000000	0.000000
75%	0.0	0.000000	0.000000
max	0.0	3.000000	4.000000

Second model (SBQ_SBT) / col = 0

	Positive culture	Sed Bac Quantity	Sed Bac Type
count	742.0	742.000000	742.000000
mean	1.0	1.849057	1.376011
std	0.0	1.247450	1.052016
min	1.0	0.000000	0.000000
25%	1.0	0.000000	0.000000
50%	1.0	2.000000	2.000000
75%	1.0	3.000000	2.000000
max	1.0	3.000000	4.000000

Second model (SBQ_SBT) / col = 1

	Positive culture	sex	Age	BW	USG	ph	Protein	Glucose	Ketones	Bilirubin	Hemoprotein	WBC	Pyuria	RBC	Crystals	Epithelial cells transitional
count	3310.0	3310.000000	3310.000000	3146.000000	3310.000000	3310.000000	3310.000000	3310.000000	3310.000000	3310.000000	3310.000000	3310.000000	3310.000000	3310.000000	3310.000000	3310.000000
mean	0.0	2.839577	10.024864	4.737307	1.025450	6.321752	65.202417	91.540785	0.530211	0.202417	133.607251	5.409366	0.206344	38.916918	3.716012	1.110574
std	0.0	1.070754	5.215017	1.620530	0.163491	0.836666	100.236144	268.657465	5.678356	0.878950	114.023744	12.100120	0.404742	43.904503	57.489771	0.673626
min	0.0	-9.000000	0.000000	0.000800	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.0	2.000000	5.900000	3.600000	1.012000	6.000000	25.000000	0.000000	0.000000	0.000000	10.000000	2.000000	0.000000	2.000000	0.000000	1.000000
50%	0.0	2.000000	10.500000	4.500000	1.018000	6.000000	25.000000	0.000000	0.000000	0.000000	150.000000	3.000000	0.000000	12.000000	0.000000	1.000000
75%	0.0	4.000000	14.100000	5.600000	1.033000	7.000000	75.000000	0.000000	0.000000	0.000000	250.000000	3.000000	0.000000	100.000000	0.000000	2.000000
max	0.0	4.000000	23.900000	34.000000	10.220000	9.000000	500.000000	1000.000000	150.000000	6.000000	250.000000	100.000000	1.000000	100.000000	999.000000	3.000000

Original model / col = 0

	Positive culture	sex	Age	BW	USG	ph	Protein	Glucose	Ketones	Bilirubin	Hemoprotein	WBC	Pyuria	RBC	Crystals	Epithelial cells transitional	Epithelial cells squamous
count	742.0	742.000000	742.000000	701.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000
mean	1.0	3.218329	11.416173	4.459857	1.020263	6.261456	75.721024	77.493261	0.438005	0.072776	149.467655	37.687332	0.695418	41.331536	2.958221	1.101078	0.351752
std	0.0	1.178204	5.354555	1.630667	0.011965	0.846743	105.847516	249.558686	6.117677	0.470550	109.971563	40.216644	0.460541	43.482828	51.823115	0.659609	0.568152
min	1.0	-9.000000	0.000000	0.450000	1.004000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.0	2.000000	7.400000	3.300000	1.012000	6.000000	25.000000	0.000000	0.000000	0.000000	25.000000	3.000000	0.000000	2.000000	0.000000	1.000000	0.000000
50%	1.0	4.000000	12.800000	4.200000	1.016000	6.000000	30.000000	0.000000	0.000000	0.000000	250.000000	20.000000	1.000000	18.000000	0.000000	1.000000	0.000000
75%	1.0	4.000000	15.700000	5.400000	1.025000	7.000000	75.000000	0.000000	0.000000	0.000000	250.000000	100.000000	1.000000	100.000000	0.000000	2.000000	1.000000
max	1.0	4.000000	24.600000	14.300000	1.066000	9.000000	500.000000	1000.000000	150.000000	6.000000	250.000000	100.000000	1.000000	100.000000	999.000000	3.000000	3.000000

Original model / col = 1

Figure 3. Data table describing the summary of statistics.

Machine-learning model performance

The ET model (Fig. 1) displays an AUC performance of 80.48% (first model) compared to the second ET model 79.65%. The ET first model showed slightly higher accuracy than the second model by 0.68%. The first ET model performed better on AUC, Accuracy than the second model of ET.