Marisa Paone
MET CS677
8/13/2023
Term Project

For my term project I looked at data from shark attacks. I gathered my data from the global shark attack file from https://www.sharkattackfile.net/incidentlog.htm. I wanted to predict/classify if an unprovoked attack would be fatal or not with respect to features such as the country, species of shark, activity being performed, age of the person attacked, and sex (M/F) of the person attacked. The classifiers I looked at were logistic regression, naive bayesian, decision tree, random forest and kNN. I also looked at Random forest with just features like the country of the attack, species of shark, and activity being performed. The data cleaning that I performed included making some assumptions and filtering the data. What I have taken away from this project is to stay out of the water, and to not take up surfing.

Discuss your results:
Random Forest with all of the features was the most accurate at 86.6% followed by Random forest with Age and Sex features dropped at 86.2%. The best TPR and TNR was logistic regression. The least accurate was the decision tree classifier. It was interesting to see that the true negative ratio wasn't good for this dataset/training sets. The random forest did the worst for TNR. The worst TPR was also Random Forest without age and sex features.

Model	TP	FP	TN	FN	Accuracy	TPR	TNR
Logistic Regression	438	33	28	47	85.3%	0.903	0.459
Naive Bayesian	446	25	11	64	83.7%	0.875	0.31
Decision Tree	404	67	36	39	80.6%	0.912	0.35
Random Forest	470	1	2	73	86.6%	0.866	0.6
kNN (k=1)	396	75	23	52	76.7%	0.884	0.235
kNN (k=2)	454	17	6	69	84.2%	0.868	0.261
kNN (k=3)	434	37	10	65	81.3%	0.870	0.21
kNN (k=4)	462	9	3	72	85.2%	0.865	0.25
kNN (k=5)	459	12	4	71	84.8%	0.867	0.25
Random Forest without Age and Sex	469	2	0	75	86.2%	0.862	0.0

Random Forest for Country, Activity, and Sex

Shark Attacks by Activity

Shark Attacks by Species/Size

Fatal Shark Attacks by Species

Shark Attacks by Country