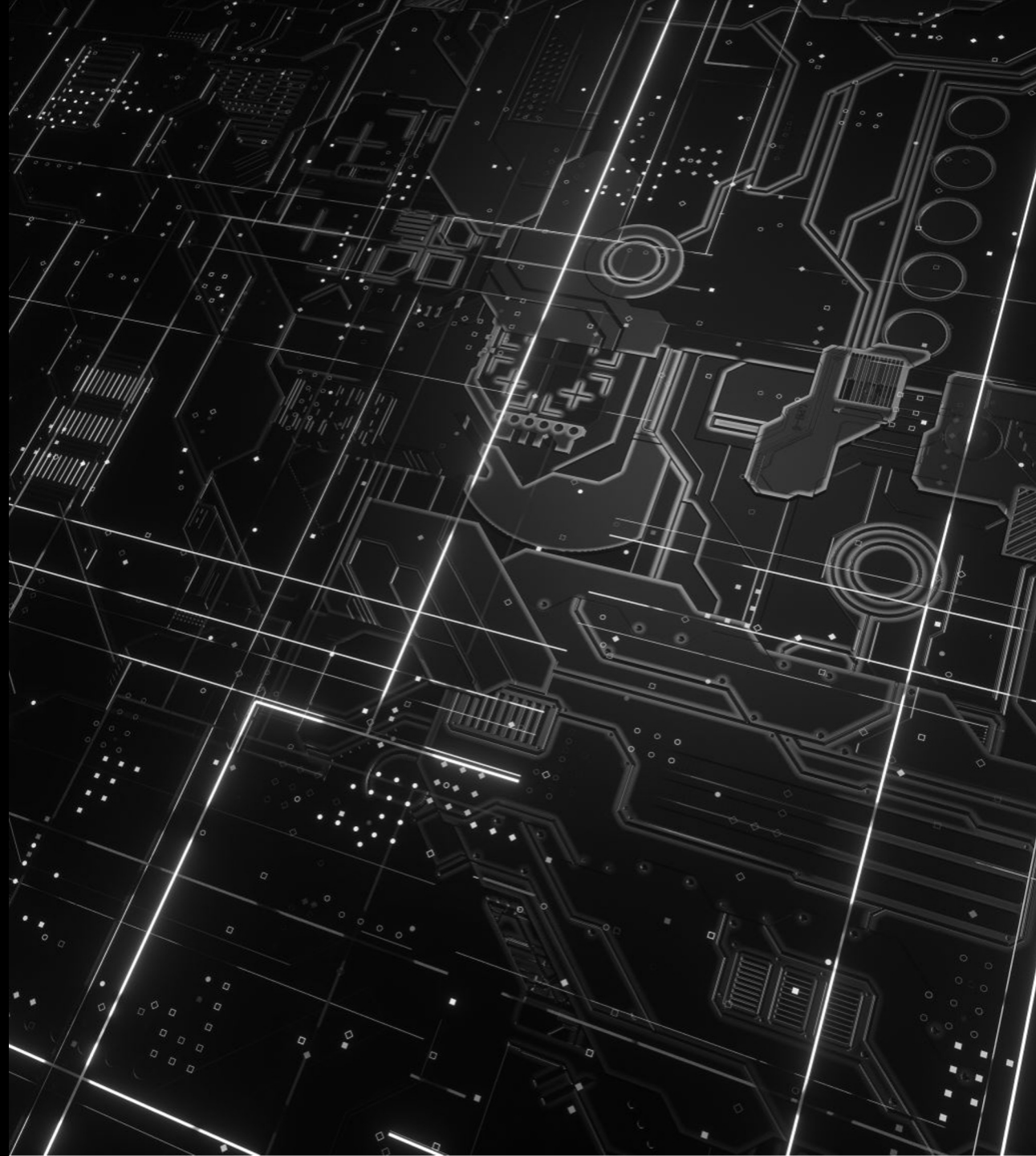
RANDOM FOREST AND NEURAL NETWORK COMPARISON – MUSHROOM CLASSIFICATION

Progress Report

Caroline Smith

Computer Science Senior

Seminar



INTRODUCTION/ LITERATURE REVIEW

- Mushrooms can serve a wide range of purposes – anti-inflammatory, nutritional, etc.
- Different factors of varying importance help determine whether a mushroom is poisonous or edible – important to determine whether they can be used for the above-mentioned purposes in medicine and food
- The data set used is neither large nor complex, so this project was suitable for a beginner machine learning project
- In similar studies, Random Forest and other algorithms were more commonly used than Neural Network algorithms, which contributed to my decision on which two algorithms to implement

METHODOLOGY

Language/ Libraries

- Python 3.10
- PyCharm IDE
- Pandas
- Scikit-Learn
- Random Forest Model Artificial Neural Network Model
- ReLU (Rectified Linear Unit) as activation for NN
- Adam – optimization algorithm

Evaluation Metrics

- F1 Score – average of precision and recall
- Precision – true positives to false positives
- Recall – true positives to false negatives
- Accuracy – correct predictions to total predictions
- Confusion Matrix – true positives, true negatives, false positives, false negatives

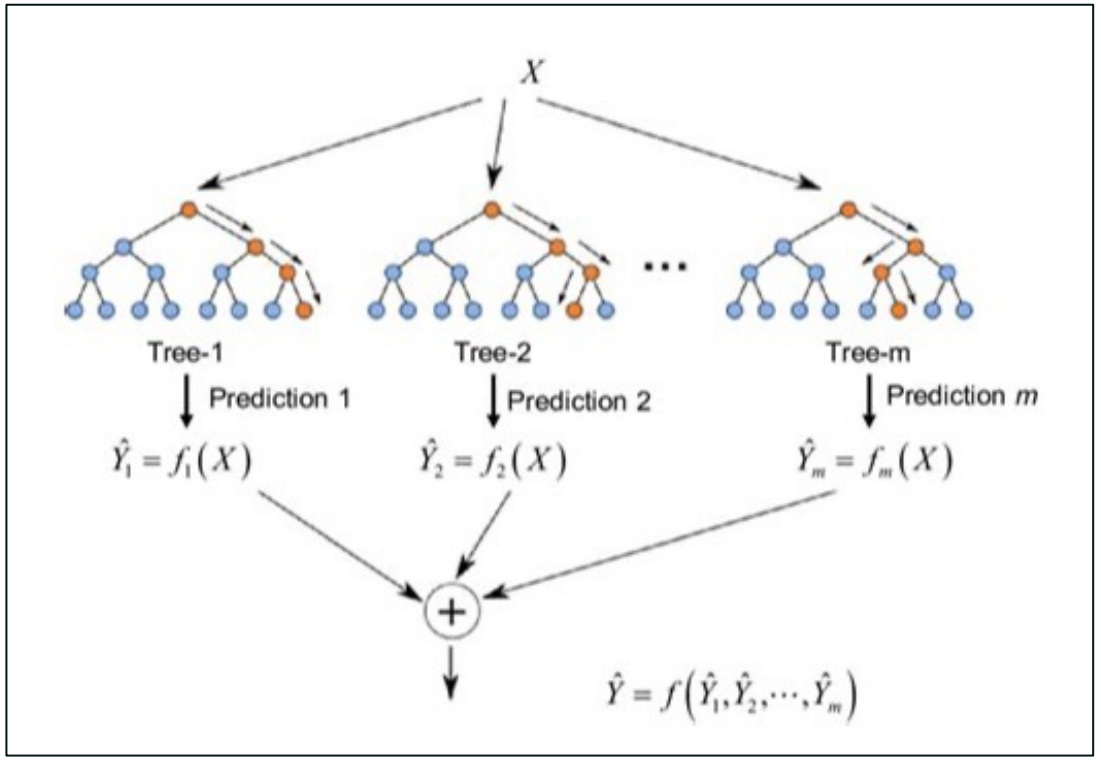
Dataset

- Source – Kaggle
- 8124 rows
- 22 mushroom characteristics to help with determination
- First column of Y-axis – determination of edible vs. poisonous – this is the column we are predicting on
- “Unknown” is placed in “Poisonous” category
- “p” (poisonous) or “e” (edible) – edible is represented by 0, poisonous is represented by 1
- Distribution – 52% edible, 48% poisonous – relatively even distribution

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	stalk-shape	stalk-root	stalk-surface-above-ring	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring	veil-type	veil-color	ring-number	ring-type	spore-print-color
2	p	x	s	n	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	k
3	e	x	s	y	t	a	f	c	b	k	e	c	s	s	w	w	p	w	o	p	n
4	e	b	s	w	t	l	f	c	b	n	e	c	s	s	w	w	p	w	o	p	n
5	p	x	y	w	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	k
6	e	x	s	g	f	n	f	w	b	k	t	e	s	s	w	w	p	w	o	e	n
7	e	x	y	y	t	a	f	c	b	n	e	c	s	s	w	w	p	w	o	p	k
8	e	b	s	w	t	a	f	c	b	g	e	c	s	s	w	w	p	w	o	p	k
9	e	b	y	w	t	l	f	c	b	n	e	c	s	s	w	w	p	w	o	p	n
10	p	x	y	w	t	p	f	c	n	p	e	e	s	s	w	w	p	w	o	p	k
11	e	b	s	y	t	a	f	c	b	g	e	c	s	s	w	w	p	w	o	p	k
12	e	x	y	y	t	l	f	c	b	g	e	c	s	s	w	w	p	w	o	p	n
13	e	x	y	y	t	a	f	c	b	n	e	c	s	s	w	w	p	w	o	p	k
14	e	b	s	y	t	a	f	c	b	w	e	c	s	s	w	w	p	w	o	p	n
15	p	x	y	w	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	n
16	e	x	f	n	f	n	f	w	b	n	t	e	s	f	w	w	p	w	o	e	k
17	e	s	f	g	f	n	f	c	n	k	e	e	s	s	w	w	p	w	o	p	n
18	e	f	f	w	f	n	f	w	b	k	t	e	s	s	w	w	p	w	o	e	n
19	p	x	s	n	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	k
20	p	x	y	w	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	n
21	p	x	s	n	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	n
22	e	b	s	y	t	a	f	c	b	k	e	c	s	s	w	w	p	w	o	p	n
23	p	x	y	n	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	n
24	e	b	y	y	t	l	f	c	b	k	e	c	s	s	w	w	p	w	o	p	n
25	e	b	y	w	t	a	f	c	b	w	e	c	s	s	w	w	p	w	o	p	n
26	e	b	s	w	t	l	f	c	b	g	e	c	s	s	w	w	p	w	o	p	k
27	p	f	s	w	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	n
28	e	x	y	y	t	a	f	c	b	n	e	c	s	s	w	w	p	w	o	p	n
29	e	x	y	w	t	l	f	c	b	w	e	c	s	s	w	w	p	w	o	p	n
30	e	f	f	n	f	n	f	c	n	k	e	e	s	s	w	w	p	w	o	p	k
31	e	x	s	y	t	a	f	w	n	n	t	b	s	s	w	w	p	w	o	p	n

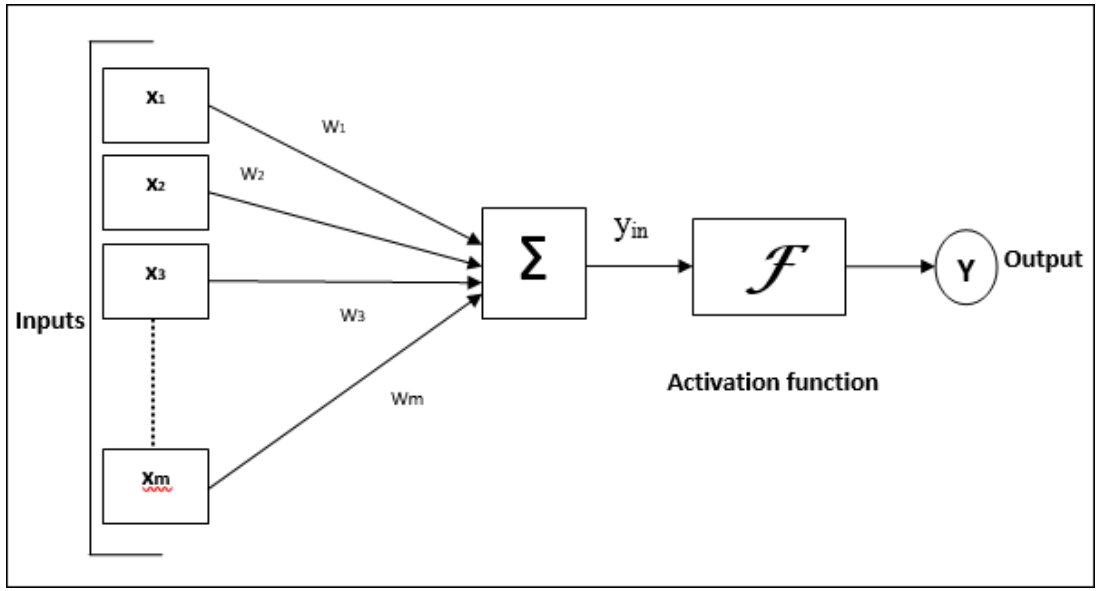
RANDOM FOREST ALGORITHM

- Supervised learning using decision trees
- Bagging method – combination of multiple learning models
- All decision trees are merged to create one large, comprehensive decision tree that increases accuracy
- More straightforward/simple process than Neural Network algorithm



NEURAL NETWORK ALGORITHM

- Large collection of units (neuronodes) connected in a pattern to allow communication between nodes
- Processing elements → Other processing elements
- Arranged in a layer or vector
- Input values are weight-adjusted to produce single input values to each neuronode
- Activation function used – ReLU (Rectified Linear Unit)
- Adam Solver Optimization Algorithm



- Optimization Algorithm and Activation Function taken in as parameters for MLP Classifier from SKLearn
- 3 layers, sizes 8, 10, 16 used for implementation – improved performance from original layers of 8, 8, 8

RESULTS – RANDOM FOREST

- Total score – 1.0
- F1 Score – 1.00
- Precision – 1.00
- Recall – 1.00
- Accuracy – 1.00
- Weighted/Macro Average – 1.00
- Confusion Matrix & Predictions – Demo

RESULTS – NEURAL NETWORK

- Total score – 1.0
- F1 Score – 1.00
- Precision – 1.00
- Recall – 1.00
- Accuracy – 1.00
- Weighted/Macro Average – 1.00
- Confusion Matrix & Predictions – Demo

ANALYSIS

- Random Forest performed better before changes made to Neural Network
- Neural Network had a small number of false positives originally – improved when layer sizes were changed
- According to evaluation metrics, both algorithms performed the same after changes were made

CONCLUSION

- According to my research, if the data set had been smaller, Random Forest would have performed better – if the data set had been larger, Neural Network would have performed better
- Good project to gain experience with ML – not overly complex, training time minimal

... DEMO ...

REFERENCES

- *4 reasons why deep learning and neural networks aren't always the right choice*. Built In. (n.d.). Retrieved October 19, 2022, from https://builtin.com/data-science/disadvantages-neural-networks
- Al-Masri, A. (2019, January 29). *How does back-propagation in Artificial Neural Networks Work?* Medium. Retrieved October 19, 2022, from https://towardsdatascience.com/how-does-back-propagation-in-artificial-neural-networks-work-c7cad873ea7
- Biau, G., & Scornet, E. (2016, April 19). *A Random Forest Guided Tour - Test*. SpringerLink. Retrieved September 11, 2022, from https://link.springer.com/article/10.1007/s11749-016-0481-7
- *Diagram of multivariate random forest algorithm. - researchgate*. (n.d.). Retrieved September 12, 2022, from https://www.researchgate.net/figure/Diagram-of-multivariate-random-forest-algorithm_fig1_340063319
- Do, T. (2022, February 8). *Types of neural network algorithms in machine learning (+ real-world examples)*. Omdena. Retrieved September 11, 2022, from https://omdena.com/blog/types-of-neural-network-algorithms-in-machine-learning/
- Google. (n.d.). *Python machine learning*. Google Books. Retrieved September 11, 2022, from https://books.google.com/books?hl=en&lr=&id=GOVOCwAAQBAJ&oi=fnd&pg=PP1&dq=python%2Bin%2Bmachine%2Blearning&ots=NddyGcXOXJ&sig=D2pqIiKSTwzDCWi67y7d8XX-alE#v=onepage&q=python%20in%20machine%20learning&f=false
- Korstanje, J. (2021, August 31). *The F1 score*. Medium. Retrieved October 19, 2022, from https://towardsdatascience.com/the-f1-score-bec2bbc38aa6
- *Machine learning random forest algorithm - javatpoint*. www.javatpoint.com. (n.d.). Retrieved October 19, 2022, from https://www.javatpoint.com/machine-learning-random-forest-algorithm
- *Mushroom Classification*. Kaggle. (2016, December 1). Retrieved September 11, 2022, from https://www.kaggle.com/datasets/uciml/mushroom-classification/code?resource=download
- Pankajray. (2020, June 4). *Artificial Neural Network (ANN)*. Medium. Retrieved October 19, 2022, from https://medium.com/ai-knowledge/artificial-neural-network-ann-ed6fa5b9c1b0
- *Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods : American Journal of roentgenology : Vol. 212, no. 1 (AJR)*. American Journal of Roentgenology. (n.d.). Retrieved September 11, 2022, from https://www.ajronline.org/doi/full/10.2214/AJR.18.20224
- Simplilearn. (2022, September 14). *Understanding the machine learning process: Key steps*. Simplilearn.com. Retrieved October 19, 2022, from https://www.simplilearn.com/what-is-machine-learning-process-article
- *Supervised learning algorithms*. Section. (n.d.). Retrieved October 19, 2022, from https://www.section.io/engineering-education/supervised-learning-algorithms/
- Suresh, A. (2021, June 22). *What is a confusion matrix?* Medium. Retrieved October 19, 2022, from https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5
- Tank, K. (2020, September 30). *Mushroom classification using different classifiers*. Medium. Retrieved September 14, 2022, from https://medium.com/analytics-vidhya/mushroom-classification-using-different-classifiers-aa338c1cd0ff
- *Tuning of the structure and parameters of a neural network using an improved genetic algorithm*. IEEE Xplore. (n.d.). Retrieved September 11, 2022, from https://ieeexplore.ieee.org/abstract/document/1176129
- *Underground networking: The amazing connections beneath your feet*. National Forest Foundation. (n.d.). Retrieved October 19, 2022, from https://www.nationalforests.org/blog/underground-mycorrhizal-network
- Verma, U. (2019, November 22). *Data Cleaning and preprocessing*. Medium. Retrieved October 19, 2022, from https://medium.com/analytics-vidhya/data-cleaning-and-preprocessing-a4b751f4066f
- *What are the disadvantages of Random Forest?* Rebellion Research. (2022, April 7). Retrieved October 19, 2022, from https://www.rebellionresearch.com/what-are-the-disadvantages-of-random-forest