1. Machine learning fundamentals
    a. Unsupervised: The narrator is the machine learning algorithm. Just as an unsupervised machine learning algorithm would identify patterns and give labels on its own for an unlabelled data set, the narrator identifies creatures by their characteristics, group them into categories, and give them unique names. All this work is done by the narrator alone, with no outside assistance.
    b. Supervised: As the narrator classifies entpeople by gender based on their features, their entfriend goes through a list of people and helps the narrator label them. The narrator then uses this labeled list to improve their classification model. This is akin to a machine learning algorithm being given a labeled data set, and using it to improve its model. Just as a researcher would provide a completely labeled datasets for supervised learning, the entfriend provided a labeled list of all their entneighbors for the narrator to use.
    c. Semi-supervised: While the narrator now has a consistent level of classification, they occasionally require an additional effort when they come upon an unfamiliar creature. The narrator uses their classification method, as well as labels acquired from different species, to make an educated guess. This is similar to semi-supervised machine learning algorithms, where they'll use their model as well as a given labeled data set. What separates this from supervised learning is that the labels the narrator possesses is not a complete labeled list of every creature in Middle Earth. Similarly, with semi-supervised machine learning, the algorithm only has access to a partially labeled data set.

2. Independent or not independent

    a. If events A and B are independent, then then for all values of A and B, the following must be true: P(A) = P(A|B). To determine independence, all values of A and B must be tested using that equation. Thus:

P(A)	**P(B)**	**P(A|B)**	**P(A) = P(A|B)**	**Independent**
P(A=red) = 0.25	P(B=hat) = 0.3	0.075 / 0.3 = 0.25	0.25 = 0.25	Yes
P(A=red) = 0.25	P(B=T-shirt) = 0.3	0.075 / 0.3 = 0.25	0.25 = 0.25	Yes
P(A=red) = 0.25	P(B=skirt) = 0.2	0.05 / 0.2 = 0.25	0.25 = 0.25	Yes
P(A=red) = 0.25	P(B=shoes) = 0.2	0.05 / 0.2 = 0.25	0.25 = 0.25	Yes
P(A=blue) = 0.25	P(B=hat) = 0.3	0.075 / 0.3 = 0.25	0.25 = 0.25	Yes
P(A=blue) = 0.25	P(B=T-shirt) = 0.3	0.075 / 0.3 = 0.25	0.25 = 0.25	Yes
P(A=blue) = 0.25	P(B=skirt) = 0.2	0.05 / 0.2 = 0.25	0.25 = 0.25	Yes
P(A=blue) = 0.25	P(B=shoes) = 0.2	0.05 / 0.2 = 0.25	0.25 = 0.25	Yes
P(A=green) = 0.5	P(B=hat) = 0.3	0.15 / 0.3 = 0.5	0.5 = 0.5	Yes
P(A=green) = 0.5	P(B=T-shirt) = 0.3	0.15 / 0.3 = 0.5	0.5 = 0.5	Yes
P(A=green) = 0.5	P(B=skirt) = 0.2	0.1 / 0.2 = 0.5	0.5 = 0.5	Yes
P(A=green) = 0.5	P(B=shoes) = 0.2	0.1 / 0.2 = 0.5	0.5 = 0.5	Yes

Note that the values of the numerator in the middle **P(A|B)** column are retrieved from the joint distribution table of A and B. Based on these calculations, all values of A and B check the independence condition so A and B are indeed independent.

b. The uniform distributions of A and B are continuous, indicated by the ranges of a and b. The definition of continuous distributions states that two continuous distributions are independent if the product of their marginal distributions is equal to their joint distribution, or $f_{A,B}(a,b) = f_A(a) * f_B(b)$. Continuous distributions can be marginalized by solving for its integral based on the derivative of its joint distribution. As a result, the marginal distribution of $f_A(a)$ is $\int_{-1}^{0} \frac{4}{3} db$ or $\frac{4b}{3}$ bound from -1 to 0 and the marginal distribution of $f_B(b)$ is $\int_{0}^{1} \frac{4}{3} da$ or $\frac{4a}{3}$ bound from 0 to 1. When the two are multiplied, their product is $\frac{16ab}{9}$, not matching the joint distribution,$f_{A,B}(a,b)$. This indicates that A and B are not independent random variables.

c. If A and B are independent, then they must fit the following condition:

P(A, B) = P(A)P(B). By the law of total probability, that equation can be expanded to support the inclusion of event C, resulting in the following:

$\sum_{c = \{-1, 1\}} P(A,\ B \mid C)\ P(C)$. At this point, the probabilities of event B are required.

Because it was given that B = A * C, the probability mass function of B can be acquired by finding the joint probability mass function of A and C. Or, P(B) = P(B | A)P(A) where B is replaced by A * C to get

P(B) = P(A * C | A) P(A). The likelihood can be interpreted as the probability of event A and C occurring given that event A occurs, so the equation can be further reduced to simply P(B) = P(C)P(A).

A	**C**	**A * C**	**P(A)**	**P(C)**	**P(A * C)**		**P(B = 1)**
1	1	1 * 1 = 1	0.5	0.9	0.45		0.45
1	-1	1 * -1 = -1	0.5	0.1	0.05		
-1	1	-1 * 1 = -1	0.5	0.9	0.45		
-1	-1	-1 * -1 = 1	0.5	0.1	0.05		

As seen in the table, the probability that event B results in a value of 1 when event A results in a value of 1 can be calculated with B = A * C, giving 0.45. This can be compared to the substitution of the equation,

P(A, B) = $\sum_{c = \{-1, 1\}} P(A,\ B \mid C)\ P(C)$, from earlier.

$\sum_{c = \{-1, 1\}} P(A = 1,\ B = 1 \mid C = 1)\ P(C = 1)$ can be reduced to

P(A=1) * P(C=1), making P(A, B) equal to 0.45. Now, the original independence condition, P(A, B) = P(A)P(B) can be solved to get the following:

0.45 = 0.5 * 0.45. This equation is not true, therefore A and B are not independent

d. For gaussian distributions, correlation of two random variables also indicates independence between the two and vice versa. This means if it is proven that two gaussian distributions are uncorrelated, it is also proven that they are independent. Uncorrelation between two gaussian distributions can be proved using its covariance matrix, E[(A-E[A]) * (B-E[B])]. Based on the values given in the problem, the variables in this equation can be substituted in to result in the following, E[(A-0) * (B-1)] which can be further reduced to E[A * (B-1)]. It is also given that E[A * (B-1)] is equal to -1. Based on properties of gaussian distributions, when the covariance matrix of two gaussian distributions is something other than 0, then those two distributions are uncorrelated and independent. Therefore, A and B are uncorrelated and independent.

3. K-nearest neighbors classification
    a. Distance function
        i. `print(get_dist(Xtrain,Xtrain[0,:])[0])` prints 0.0
        ii. `print(get_dist(Xtrain,Xtest[0,:])[10])` prints 6069462.0
        iii. `print(get_dist(Xtrain,Xtest[10,:])[50])` prints 5661744.0
    b. Prediction
        i. `print(ytest_pred[:20])` prints [7 2 1 0 4 1 9 4 6 7 0 9 9 0 1 3 7 7 3 4]
        ii. `print(ytest[:20])` prints [7 2 1 0 4 1 4 9 5 9 0 6 9 0 1 5 9 7 3 4]
    c. Classification Accuracy: 0.6694
    d. Hypermeter tuning

m	*K*	accuracy	runtime (sec)
100	1	0.6794	10.69760513305664
100	3	0.6694	12.126908302307129
100	5	0.6267	12.694426774978638
1000	1	0.869	103.1755998134613
1000	3	0.872	105.43249988555908
1000	5	0.8585	105.63866019248962
2500	1	0.9136	256.8390715122223
2500	3	0.9187	259.2170822620392
2500	5	0.9107	260.18890261650085

        i. While it is technically feasible in the accuracy the model provides, it is not advisable because of how the runtime increases as the sample size

increases. My computer processes training data at a rate of about 10 objects per second. Ignoring additional processing time, the time to process all 60000 training datasets would be estimated to be around 6000 seconds, or 100 minutes.

ii. The three higher values of *m* resulted in three higher levels of accuracy. But for each of the three levels of accuracy, the greater the *K* value, the lesser the prediction accuracy.