Module 3 term project

All cosmetic companies in California are required to report their products (since 2005) if they could cause cancer, developmental birth defects, or harm the reproductive system. This dataset contains information about them: company names, chemical names, cosmetic types, date product was registered, date product was discontinued (if it was) etc...

Here, I have applied the lazy learning nearest neighbours method to predict whether certain cosmetics were discontinued or not. My results are on the cross table below. 100% of the predictions were correct!!

I used k=20 though my data had 88954 examples, and the textbook recommended to take the square root of n as k. This was because when I took that number(298), R complained that kNN had too many ties meaning that it wasn't being able to process that for some reason. My only possible question if any, would be how do you know what a suitable training and testing number is? I roughly took a 8:2 training:testing ratio like in the textbook but I don't know if that is ideal.

next comment:

I did get the same thing with just about any number so yes I was actually using the answer as part of the prediction which makes a lot more sense because when I explored the data it didn't look like the dates (most of the predictor variables) had any connection with whether the product was discontinued since products from all years were discontinued. The only real predictor of this would be ChemicalCount but the majority of the data has 1 or 2 chemicals, with very few having more, so it's not really a strong distinguishing point.

So my new k=230 which was about the biggest k I could use before the knn function ran into ties (ideally I would use 298, the square root of n=88955 otherwise). My results are in the picture below and the knn function has basically predicted most of the isDiscontinued to be false regardless of whether the actual answer was true or false. I think it is doing that because most of the products in the dataset have not been discontinued. Since the predictors don't have any real correlation with the target variable knn has simply taken on the bias in the data.

