Question 1:

In order to evaluate how well our model performs, we prefer to use model performance rather than model accuracy that is prone to biases. We looked up for the current prevalence of T1D patients around the world, to understand how rare the condition is. The prevalence on T1D is ~9.5% around the world (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7146037/). For this purpose, we want to show that accuracy score might be misleading. We assume that the dataset we acquired is a representative sample of the population, and the sensitivity is around 50% (meaning the model is bad). We assume that we collected 10,000 samples.

		Prediction	
		No	Yes
Truth	No	9000	0
	Yes	500	500

We assume as well that the specificity is 100%. Now, let us calculate the model's accuracy:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} = \frac{9000 + 500}{10000} = 95\%$$

We can see that the accuracy score is not bad although the sensitivity is low, thus, using this model we can misdiagnose half of the T1D patients and still get a high accuracy score. In other words, the accuracy score is indicative/relevant when the condition we want so survey is prevalent.

Question 2:

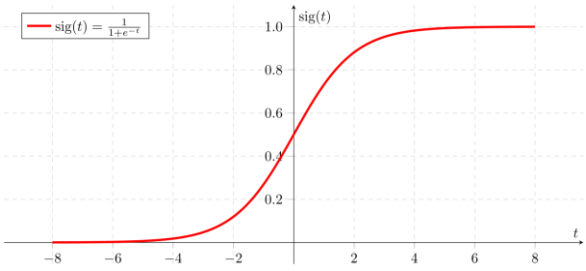
There isn't a correct answer and it depends on many parameters. On the first hand, using multiple features might result with better precision because of the additive value of extra information, on the other hand, data visualization is more complicated, because as we humans can only observe and imagine 3D. Moreover, adding a lot of features might be computationally expensive and requires feature scaling and approaches of features selection which might add to the complexity of the model. Another aspect that we should consider, is that the more the features the more time and effort it takes to collect the data (assuming they aren't available online☺)

Question 3:

The problem described can be pertained by many aspects and angles, on one hand, if there are a lot of features, if we apply nonlinear model it can be computationally expensive and the fact that the researcher reports similar observation based on her subjective assessment, doesn't mean that the biopsies' results are not linearly separable (multi-dimensional linear classifier). On the other hand, if the differences between the classes are not apparent to the histologist, it might indicate that there is a complex boundary separating between the two classes, and projecting the features on the multi-dimensional non-linear space might assist in finding a precise classifier.

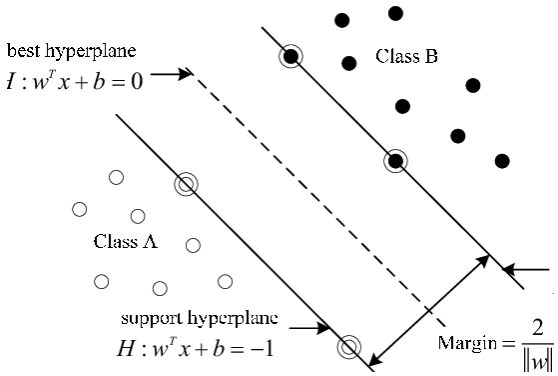
Question 4:

Different ML algorithms have different characteristics and may be implemented on several interesting problems. Logistic regression (LR) is a common and relatively simple classifier results with a probability value (0 to 1) for each sample that represents how well it belongs to a certain class. LR is based on a statistical approach which splits the predictions to the classes using a sigmoid function and a defined threshold.



https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

Linear SVM (LSVM) is considered a more complex algorithm which results with a linear (hyper-plane) boundary between the different classes (n-dimensional clusters). The algorithm tries to maximize the margins' width between the two classes, thus increasing the certainty of the separation for the future samples to be classified.



https://www.researchgate.net/figure/The-hyperplane-of-SVM-The-use-of-linear-SVM-to-construct-the-hyperplane-may-sometimes_fig1_257723607

Discussing the hyper-parameters' topic, in LR, we use a single hyper-parameter λ that is responsible for regularization of the weights by increasing the loss for increasing the weights, and by that penalty, overfitting is reduced.

In LSVM, we also use λ hyper-parameter, that is regularizing the weights, but the main difference is that in the SVM context, the λ parameter is used to indicate how much we can tolerate misclassification (or under-fitting, as for LR), and how soft is our margin.