Exercise 1:

(a) Which of the following statement(s) is/are correct?

(i) A single ICE curve is a local explanation method.

(ii) Robust local explanation methods should return similar explanations for similar observations.

(iii) In ordinary Gower's distance all feature receive different weight.

(b) Which of the following statement(s) about local surrogate models is/are correct?

(i) Surrogate models produced by LIME should have the same prediction as the model to be explained for the whole training dataset.

(ii) The choice of the sampling process and the definition of locality are important hyperparameters of LIME that have a large impact on the behavior of the method.

(iii) LIME does not require any adaptions to be applicable to deep learning models for image data.

(iv) LIME requires the surrogate model to use all available features - a selection of features is not allowed.

(v) If the kernel width for the exponential kernel is set to infinity, all observations receive a proximity measure/weight of 1 independent of their distance to $\mathbf{x}$.

Exercise 2:

An insurance company wants to calculate monthly premiums for a disability insurance ("Berufsunfähigkeitsversicherung") with an ML model based on the pension, age, job type and marital status. An appropriate model $\hat{f}$ was already fitted using a large customer dataset but in order to be launched it needs approval by regulators. The regulators put the model to test by using LIME (with exponential kernel) to provide an explanation for typical and critical customers. One of these test instances $\mathbf{x}$ is a 21 year old woman displayed in the first row of the following table.

a) The regulators already generated new instances to fit the surrogate model. For simplicity, we assume that three instances are enough to fit the model. Fill out the missing fields in the following table, where $\phi_\sigma(\mathbf{z}) = \exp(-d(\mathbf{x},\mathbf{z})^2/\sigma^2)$ is the exponential kernel similarity measure for a specific σ and $d(\cdot)$ as the Gower's distance.

	pension	age	job type	marital status	$\hat{f}$	$d(\mathbf{x},\mathbf{z}.)$	$\phi_{\sigma=0.15}(\mathbf{z}.)$	$\phi_{\sigma=0.5}(\mathbf{z}.)$
$\mathbf{x}$	1800	21	sedentary	single	30.6	-	-	-
$\mathbf{z}_1$	1600	21	sedentary	married	25.8	0.25	0.06	
$\mathbf{z}_3$	2200	32	sedentary	married	85.2	0.32	0.01	
$\mathbf{z}_2$	1200	23	physically	single	74.9	0.49	0.00	0.38

How does the kernel width σ influence the proximity measure? What would happen if the kernel width is set too small?

b) The regulatories fit two different local surrogate models (g_1 and g_2) on the re-weighted data (here, three observations). The following table compares the prediction of $\hat{f}$ to the ones of the two surrogate models for the three instances.

	$\hat{f}$	g_1	g_2
$\mathbf{x}$	30.6	34.8	31.1
$\mathbf{z}_1$	25.8	28	26.1
$\mathbf{z}_3$	85.2	105	92.7
$\mathbf{z}_2$	74.9	90	68.9

Which of the two surrogate models do you prefer? Compute the local faithfulness for both surrogate models using the weights from a) with $\sigma = 0.15$.
Hint: consider $L(\hat{f}, g, \phi_{\mathbf{x}})$ from the lecture.

c) The surrogate model g_1 corresponds to a linear model using all three features, while g_2 corresponds to a random forest with 500 trees. Would you still prefer the model you chose in b)?

d) Discuss whether for the faithfulness assessment in b) it makes sense to use a new sampled dataset instead of the one the local surrogate model was fitted on.