*What is the reason behind the difference in performance between the second and third models?*

In the second and third model different activation functions are used. In order to understand the difference in performance the concept of linear and Rectified Linear Unit (ReLU) needs to be understood. Activation functions are essential in deep learning as it determines the output of the model, as it maps the output values into a certain range. So different activation functions lead to different outputs, which results in different performance of the classification problems

Linear activation is given by a linear equation and thus the function is not restricted to a certain range. The ReLU function is given by:

This implies that all negative values are converted to the value zero and all positive values remain their value. This results in a range from zero to the maximum value. It is found that the ReLU activation function outperforms all other activation functions.1 The great performance of this activation function is due to the sparsity of activations after ReLU.2

<https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6>

[1] Schmidt-Hieber, J. (2018). *Nonparametric regression using deep neural networks with ReLU activation function*. Retrieved from https://arxiv.org/pdf/1708.06633.pdf

[2] Xu, B., Wang, N., Kong, H., Chen, T., & Li, M. (n.d.). *Empirical Evaluation of Rectified Activations in Convolution Network*. Retrieved from https://github.com/