Ref:

1. Buda, M., Maki, A. and Mazurowski, M. A. (2017) ‘A systematic study of the class imbalance problem in convolutional neural networks.’ *Neural networks.*106, pp.249-259.
2. Chawla, N. V., Bowyer, K., Hall, L. O. and Kegelmeyer, W. P. (2002) ‘SMOTE: synthetic minority over-sampling technique’ *Journal of Artificial Intelligence Research.* 16(1), pp.321-357.
3. He, H., Bai, Y., Garcia, E. A. and Li, S. (2008) ‘ADASYN: Adaptive synthetic sampling approach for imbalanced learning’ *2008 IEEE International Joint Conference on Neural Networks (IJCNN).* Pp.1322-1328.
4. Thabtah, F., Hammoud, S., Kamalov. F. and Gonsalves, A. (2020) ‘Data imbalance in classification: Experimental evaluation’ *Information Science.* 513(1), pp.429-441.
5. Zhang, S., Li, X., Zong, M., Zhu, X. and Cheng, D (2017). ‘Learning k for kNN Classification’ *ACM Transactions on Intelligent Systems and Technology.* 8(3), pp.1-19.

According to the original dataset, a significant imbalance of classes of data is observed, where in the total 6819 cases, 6599 (96.77%) of which are negative while only 220 (3.23%) is positive. Therefore, it is essential to solve the data imbalance issue in case the model is not trained with bias.

Thabtah et al (2020) suggested a framework of data imbalance solutions. From the data’s aspect, there are two notions to rebalance data, which are under sampling and over sampling. Under sampling means to reduce the cases of major class, while over sampling is to generate synthetic data of minor class till the two classes are balanced. Considering the minor class of this dataset only has 220 cases, using the over sampling method is recommended.

Common over sampling methods include random over sampling, SMOTE, ADASYN, etc (Chawal et al, 2002; Buda et al, 2017; He et al, 2008). Random over sampling method replicates the minor cases randomly, which generates same data and causes bias. SMOTE is based on nearest neighbour and generates data that are not exactly same as the neighbour cases. ADASYN, also based on nearest neighbour algorithm, focuses more on difficult examples (points at the border). This is considered a main difference to SMOTE, which could reduce bias from the data. Therefore, in this project we use ADASYN to solve data imbalance.

As mentioned above, ADASYN is based on nearest neighbour algorithm to choose and thus generate data. Therefore, same to KNN classification, the performance of ADASYN is essentially influenced by *k* value. To find the optimal *k* value for this project, we applied simple KNN model to the training set data and compare the performances of model by setting *k* = 1, 2, 3…74. Note that according to Zhang et al (2017), a typical *k* value for KNN model is the square root of sample size. In this case, it equals 73.86, approximately 74. The performance of model is defined by the proportion that the predicted value not equalling the test set data. By setting the other hyperparameters default, the result of error rates to *k* values is shown in Fig 1.

A graph with a line

Description automatically generated

Fig 1: error rates to k values

From fig 1 we can see that the minimum error rates occur when k = 6 and 7, at 0.0293. Therefore, the lowest odd *k* value *k* = 7 is picked for ADASYN. We apply ADASYN with *k* = 7 only to train set and get a training set of 5275 class 1 and 5254 class 2 cases, and the proportion of two classes reaches 99.6%.