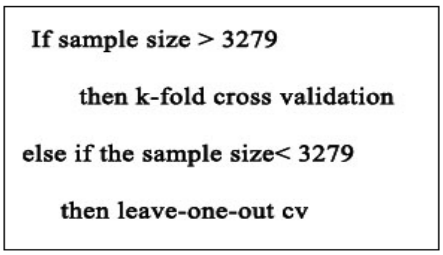

Figure 10: Dataset threshold results

6. Conclusions

In this paper we have presented results from the experiments performed in order to establish the dataset threshold for the performance estimation methods. From the experiments performed, the threshold has been identified when the dataset has 3179 instances where by the difference between the two methods is 0. The establishment of the dataset threshold will help unfamiliar supervised machine learning experimenters such as students studying in the field to categorise datasets based on the number of instances and attributes and then choose appropriate performance estimation method.

7. References

[1] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*: Morgan Kauffman, 2005.

[2] T. Mitchell, *Machine Learning*: MIT Press, 1997.

[3] Alpaydin Ethem, *Introduction to Machine Learning*. Cambridge, Massachusetts, London, England: MIT Press. 2004.

[4] T. Mitchell, "The Discipline of Machine Learning," Carnegie Mellon University, Pittsburgh, PA, USA, 2006.

[5] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, pp. 27-34, 1996.

[6] L. Rokach and O. Maimon. Part C, "Top-down induction of decision trees classifiers - a survey.," *Applications and Reviews, IEEE Transactions on Systems, Man, and Cybernetics*, vol. 35, pp. 476-487., 2005.

[7] T. Caelli and W. F. Bischof, *Machine Learning and Image Interpretation*. York, NY, USA: Plenum Press, 1997.

[8] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*: Kauffman Press., 2002.

[9] P. Cunningham and S. J. Delany, "K-Nearest Neighbour Classifiers," vol. 2008, 2007.

[10] C. Campbell, *An Introduction to Kernel Methods*, 2000.

[11] M. Berthold and D. J. Hand, "Intelligent Data Analysis," 2003.

[12] L. Breiman, "Random Forests," 2001.

[13] M. W. Craven, "Extracting Comprehensible Models from Trained Neural Networks." 1996.

[14] Y. Bengio and Y. Grandvalet, "No Unbiased Estimator of the Variance of KFold Cross-Validation," *Journal of Machine Learning Research*, vol. 5, pp. 1089–1105, 2004.

[15] S. Kostiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*, vol. 31, pp. 249-268, 2007.

[16] J. Mena, "Data Mining Your Website," 1999.

[17] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*: MIT Press, 1999.

[18] T. Fawcett, "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers," 2004.

[19] J. Winkler, M. Niranjan, and N. Lawrence, *Deterministic and Statistical Methods in Machine Learning*: Birkhauser, 2005.

[20] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithm," pp. 1895-1923, 1998.

[21] B. Everitt, *The Analysis of Contingency Tables*: Chapman & Hall/CRC, 1992.

[22] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," pp. 1137-1143, 1995.

[23] E. Micheli-Tzanakou, *Supervised and Unsupervised Pattern Recognition*: CRC Press, 1999.

[24] O. Nelles, "Nonlinear System Identification," 2001.

[25] I. H. Witten and E. Frank, *Data Mining*: Morgan Kauffman, 2000.

[26] Y. e. a. Tang, "Granular Support Vector Machines for Medical Binary Classification Problems," pp. 73-78., 2004.