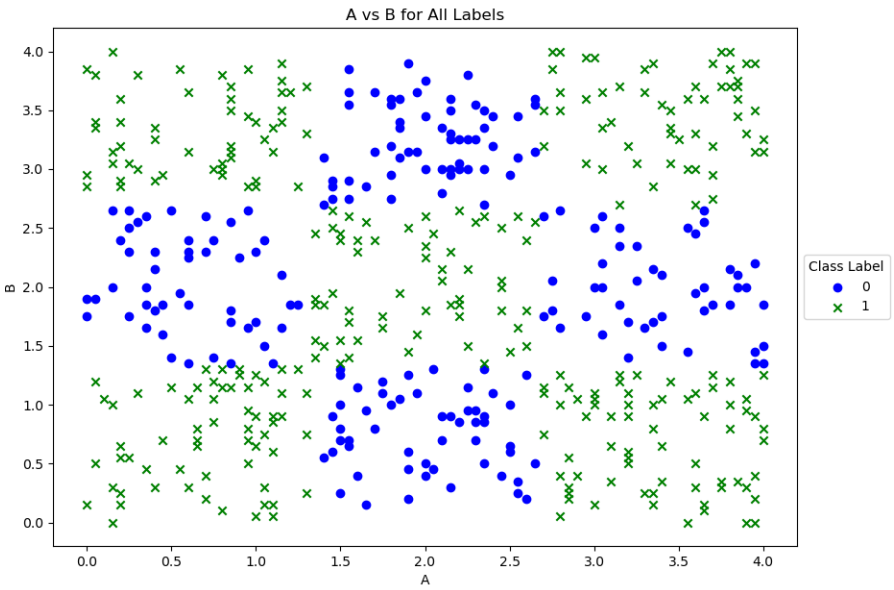
Artificial Intellience, HW 5 Coding Report

Alexandra Koletsos, ak4749

14 December 2023

Part I: Classification

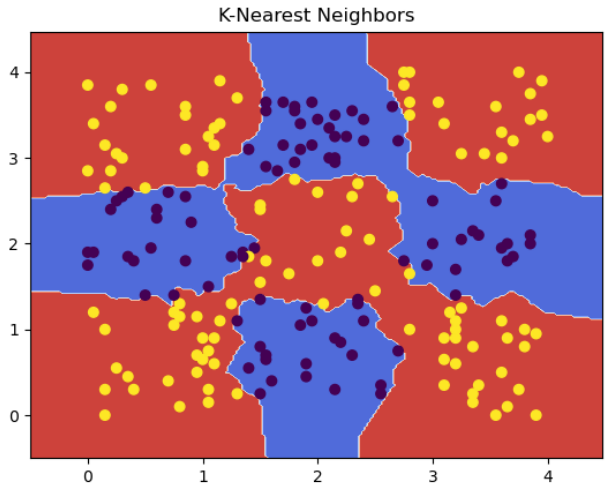
A scatterplot of the dataset showing the two classes with two different patterns.



(i)

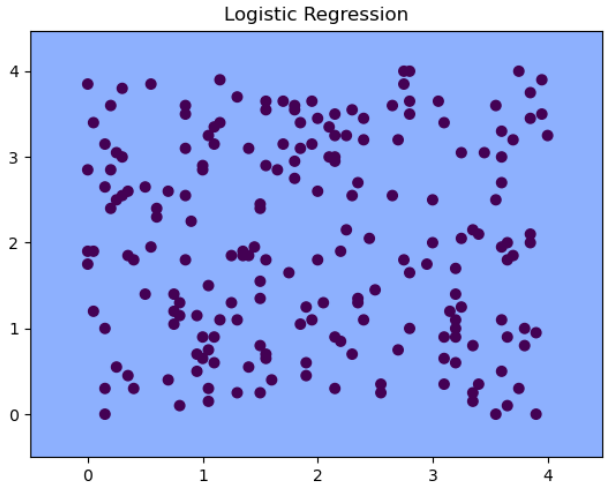
The K-Nearest Neighbors classifier achieves a training and testing accuracy of 93% with parameters $leaf\_size = 5$ and $n\_neighbors = 3$. We can see from the scatterplot of the dataset that the data for each class is already conveniently grouped into clusters, so the KNN decision boundary has high accuracy and does not overfit the training data even when

using a small number of neighbors.



(ii)

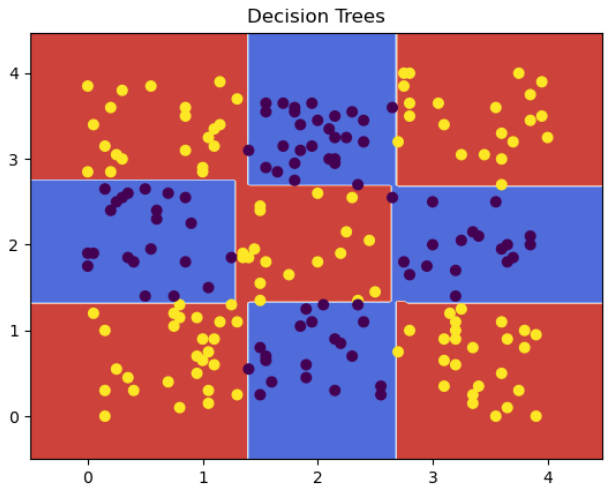
The Logistic Regression classifier achieves a training accuracy of 61% and a testing accuracy of 57% with parameter $C = 0.1$. We can see from the scatterplot of the dataset that the dataset is not linearly separable, so logistic regression has low accuracy and is not a suitable classifier for our data.



(iii)

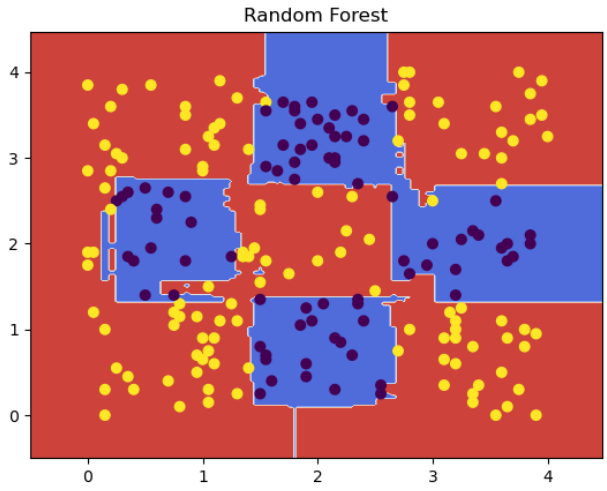
The Decision Trees classifier achieves a training accuracy of 97% and a testing accuracy of 99% with parameter $max\_depth = 7$. Since the data for each class is grouped into clusters, the decision tree classifier creates a decision boundary that resembles a 3×3 grid. The

algorithm tries to separate the dataset such that all leaf nodes (nodes that don't split the data further) belong to a single class. The scatterplot of the dataset can be easily split into 9 sections, so the decision tree classifier has high accuracy and is suitable for our data.



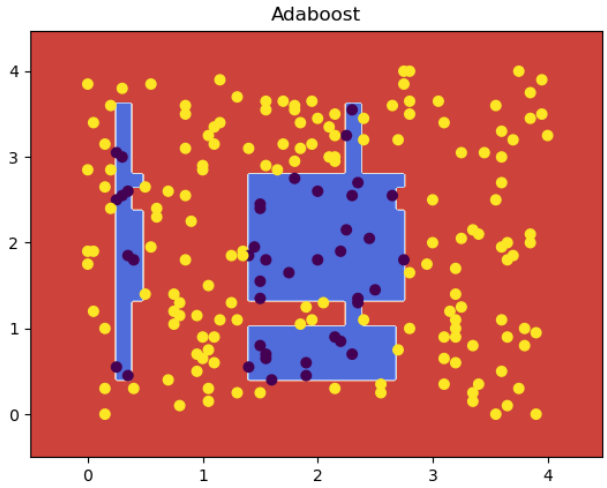
(iv)

The Random Forest classifier achieves a training accuracy of 94% and a testing accuracy of 96% with parameters $max\_depth = 5$ and $min\_samples\_split = 3$. Random forest uses a collection of decision trees to make predictions, and averages the predictions of all the trees to produce the final prediction. Employing similar logic to (iii), random forest achieves high accuracy because the scatterplot of the dataset can be easily split into 9 sections. However, random forest performs slightly worse than the decision trees classifer on this dataset. This may be due to the small dataset and clear cluster patterns in the data, so a single decision tree is enough to accurately predict the data without overfitting.



(v)

The AdaBoost classifier achieves a training accuracy of 62% and a testing accuracy of 57% with parameter $n\_estimators = 30$. The AdaBoost algorithm is harder to train on complex, non-linear data because the weak learners will most likely misclassify the data which will affect the subsequent weak learners, and thus will affect the algorithm's final classification. From the scatterplot of the dataset, we can see that the data is non-linear with a complex decision boundary, so this is most likely why AdaBoost performs as badly as linear regression (weak learners only perform slightly better than random, and the data may be too complex for the algorithm to learn enough information at each iteration to significantly reduce the training error).

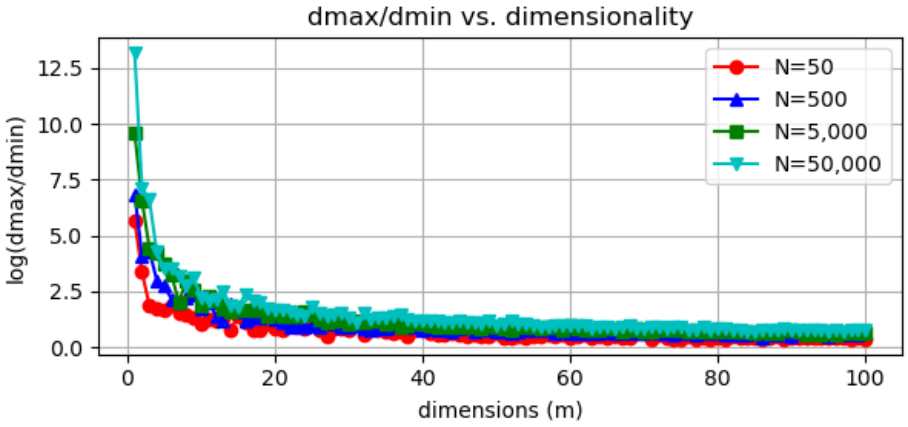


Conclusion. Decision Trees performed best and Logistic Regrssion performed worst. The reasoning for this can be found in the analysis of (ii) and (iii).

Part II - High Dimensionality: Curse or Blessing?

(a)

A graph that visualizes the curse of dimensionality. As the dimension increases, the dmax/dmin ratio (the ratio of the distance of the furthest point from the query to the distance of the nearest point to the query) decreases exponentially.



(b) My Research

***Blessing of dimensionality: mathematical foundations of the statistical physics of data* by A. N. Gorban and I. Y. Tyukin**

In 1900, German Mathematician David Hilbert presented his 10 problems to the International Congress of Mathematicians, among which he presented the **sixth problem**. A.N. Gorban describes the sixth problem as that which "proclaimed expansion of the axiomatic method beyond existing mathematical disciplines, into physics and further on." How does the sixth problem relate to the *blessing of dimensionality*? Well, the blessing of dimensionality initially described the statistical mechanics phenomenon where the laws of a system become simpler as the number of its weakly interacting subsystems approaches infinity (the thermodynamic limit). In accordance with the sixth problem, Soviet Mathematician Khinchin criticized this point of view for lacking a basis of "analytical rigor," and thus began the stream of research relating to ergodicity, limit theorems, and measure concentration phenomena. This led to further research of high-dimensional data analysis, data mining and machine learning.

In 2000, American Statistician David Donoho described the *blessing of dimensionality* as the phenomenon where 'statements about very high-dimensional settings may be made where moderate dimensions would be too complicated'. In high-dimensional space, the squared distance of a random finite set of points to a selected point are likely close to the median squared distance, thus simplifying the extremely complex geometry of the data — thus, a blessing. However, the similarity search in high dimensions becomes difficult and useless for these same reasons — thus, a curse.

Blessing of dimensionality: mathematical foundations of the statistical physics of data by A.N. Gorban and I.Y. Tyukin focuses on research surrounding linear separability, measure concentration theorems, and stochastic separation theorems, and how they relate to the blessing of dimensionality. Notably, the researchers discovered that "random points are all linearly separable from the rest of the set even for exponentially large random sets" and provide examples of probability distributions for concentration and separation theorems with non-vanishing variance. They propose the AI correction method based on stochastic separation theorems, which aims to improve AI systems by increasing the dimensionality of the data. It forms a single measurement vector $\bar{x}$ at a given time t based on a combination of signals from inputs, outputs, and the internal state of the legacy AI system. Each vector $\bar{x}$ is labeled 'correct' or 'incorrect', where the 'incorrect' labels are then corrected by a subsystem. Thus, higher dimensionality of data may improve the future of AI by simplifying the correction process. Although this process has not been heavily tested, it may be a start to showing that high dimensionality is not a curse after all!