**Assignment 4**

Authors: Lior Shacohach 206284960 and Moriya Barel 207304353

Question 1

Results:

P = 1:

K = 1: e` = 0.0 | e = 0.3414379084967315 | diff = 0.3414379084967315

K = 3: e` = 0.16875816993464032 | e = 0.28960784313725463 | diff = 0.1208496732026143

K = 5: e` = 0.2016993464052285 | e = 0.2720261437908495 | diff = 0.070326797385621

K = 7: e` = 0.21745098039215668 | e = 0.2638562091503266 | diff = 0.04640522875816991

K = 9: e` = 0.2224183006535946 | e = 0.2577124183006535 | diff = 0.03529411764705889

P = 2:

K = 1: e` = 0.0 | e = 0.3271241830065355 | diff = 0.3271241830065355

K = 3: e` = 0.16660130718954225 | e = 0.2939869281045749 | diff = 0.12738562091503264

K = 5: e` = 0.20581699346405213 | e = 0.27614379084967294 | diff = 0.07032679738562081

K = 7: e` = 0.2167320261437907 | e = 0.2601960784313724 | diff = 0.043464052287581684

K = 9: e` = 0.22431372549019601 | e = 0.2590849673202613 | diff = 0.0347712418300653

P = Infinity:

K = 1: e` = 0.0 | e = 0.33633986928104526 | diff = 0.33633986928104526

K = 3: e` = 0.16849673202614354 | e = 0.2848366013071893 | diff = 0.11633986928104575

K = 5: e` = 0.2031372549019606 | e = 0.26797385620915015 | diff = 0.06483660130718955

K = 7: e` = 0.21666666666666654 | e = 0.26006535947712406 | diff = 0.04339869281045752

K = 9: e` = 0.22307189542483652 | e = 0.25385620915032664 | diff = 0.03078431372549012

Best P, K:

The choice of P does not seem to have too much effect on the results but if we must choose, the best P is infinity, the choice of K, however, seems to have a huge impact on the results and the best K is 9.

Interpretation of the results:

KNN is performing better on the test set with larger values of k. As the value of k increases, the difference between the empirical error and the true error decreases, indicating that larger values of K mean better performance on new data points. Also, in our opinion, the accuracy of KNN is not satisfying and we suspect that the reason might be the imbalance between the classes of the dataset (there are significantly more points labeled 1 than points labeled 2). To our surprise, we ran KNN while applying `under sampling` (sampled 81 points labeled 1 and 81 points labeled 2) and the results were even worse.

Is there Overfitting:

For lower values of K (especially for k=1) there is quite a lot of Overfitting since the empirical error is less than the true error and the difference between them is quite large (good on the sample, bad on the world). The difference between the empirical error and the true error is a lot smaller for larger K values which indicates that our model learned the underlying rule behind the sample and didn’t just learn the sample, hence, there is no, or close to none, Overfitting for larger K values.

Question 2

Results:

P = 1:

K = 1: e` = 0.0 | e = 0.05879999999999997 | diff = 0.05879999999999997

K = 3: e` = 0.02660000000000002 | e = 0.08399999999999995 | diff = 0.05739999999999993

K = 5: e` = 0.04020000000000002 | e = 0.09839999999999995 | diff = 0.058199999999999925

K = 7: e` = 0.05099999999999999 | e = 0.10059999999999988 | diff = 0.049599999999999894

K = 9: e` = 0.06799999999999998 | e = 0.11439999999999988 | diff = 0.0463999999999999

P = 2:

K = 1: e` = 0.0 | e = 0.05579999999999997 | diff = 0.05579999999999997

K = 3: e` = 0.03140000000000002 | e = 0.09599999999999992 | diff = 0.06459999999999991

K = 5: e` = 0.05019999999999998 | e = 0.11099999999999995 | diff = 0.060799999999999965

K = 7: e` = 0.06399999999999996 | e = 0.1207999999999999 | diff = 0.056799999999999934

K = 9: e` = 0.08319999999999998 | e = 0.13119999999999998 | diff = 0.048

P = inf:

K = 1: e` = 0.0 | e = 0.06259999999999998 | diff = 0.06259999999999998

K = 3: e` = 0.03120000000000002 | e = 0.10319999999999993 | diff = 0.07199999999999991

K = 5: e` = 0.05419999999999997 | e = 0.1182 | diff = 0.06400000000000003

K = 7: e` = 0.06919999999999996 | e = 0.1295999999999999 | diff = 0.060399999999999954

K = 9: e` = 0.08999999999999993 | e = 0.14219999999999997 | diff = 0.05220000000000004

Analysis: Overall the results are good, the main two differences between the Haberman Survival dataset results are

1. Here there is no Overfitting.
2. The performance of the model seems to decrease as the values of K increase which is the opposite of the behavior, we saw previously on the Haberman set.

The reason that the accuracy of KNN decreases as the value of K increases might be the density of the data, which refers to the number of instances per area.

Question 3

f does not preserve the area of triangles.

We will show a counter example:

Take d = 2, k = 1 and any combination of u, v, w that forms a triangle with area > 0.

Any function f satisfying the bounds of the JL lemma will reduce the dimension of u, v, and w to 1 and therefore the triangle formed by those points will have area = 0.

Therefore area(< 𝑢, 𝑣, 𝑤 >) ≤ area(< 𝑓(𝑢), 𝑓(𝑣), 𝑓(𝑤) >) is false.