3a. Output file: knn1 test.csv and knn3 test.csv for k = 1 and k = 3 respectively. Classification accuracy for testing set when data is normalized as printed on the terminal for both k = 1 and k = 3 is 94.4%.

For knn1test.csv

|  |  |  |
| --- | --- | --- |
| y | predicted\_y | distance1 |
| 1 | 1 | 0.435637 |
| 1 | 1 | 0.502393 |
| 3 | 3 | 0.307184 |
| 1 | 1 | 0.749164 |
| 2 | 2 | 0.505433 |
| 1 | 1 | 0.736888 |
| 2 | 2 | 0.551148 |
| 3 | 3 | 0.386098 |
| 2 | 2 | 0.736976 |
| 3 | 3 | 0.455962 |
| 1 | 1 | 0.412799 |
| 3 | 3 | 0.794681 |
| 1 | 1 | 0.464415 |
| 2 | 3 | 0.657739 |
| 1 | 1 | 0.708059 |
| 2 | 2 | 0.359972 |
| 2 | 2 | 0.526015 |
| 2 | 2 | 0.445299 |
| 1 | 1 | 0.530518 |
| 2 | 2 | 0.42713 |
| 1 | 1 | 0.701313 |
| 2 | 1 | 0.837492 |
| 2 | 2 | 0.688625 |
| 3 | 3 | 0.843975 |
| 3 | 3 | 0.457966 |
| 3 | 3 | 0.470893 |
| 2 | 2 | 0.645909 |
| 2 | 2 | 0.69649 |
| 2 | 2 | 0.468595 |
| 1 | 1 | 0.737747 |
| 1 | 1 | 0.430784 |
| 2 | 2 | 0.595971 |
| 3 | 3 | 0.480854 |
| 1 | 1 | 0.474846 |
| 1 | 1 | 0.424035 |
| 1 | 1 | 0.584431 |
|  |  |  |

For knn3 test.csv:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| y | predicted\_y | distance1 | distance2 | distance3 |  |
| 1 | 1 | 0.435637 | 0.693688 | 0.737051 |  |
| 1 | 1 | 0.502393 | 0.589404 | 0.763428 |  |
| 3 | 3 | 0.307184 | 0.359242 | 0.465533 |  |
| 1 | 1 | 0.749164 | 0.783611 | 0.784655 |  |
| 2 | 2 | 0.505433 | 0.534214 | 0.536515 |  |
| 1 | 1 | 0.736888 | 0.791617 | 0.795099 |  |
| 2 | 2 | 0.551148 | 0.624061 | 0.64659 |  |
| 3 | 3 | 0.386098 | 0.473519 | 0.497897 |  |
| 2 | 2 | 0.736976 | 0.741784 | 0.793114 |  |
| 3 | 3 | 0.455962 | 0.484584 | 0.524093 |  |
| 1 | 1 | 0.412799 | 0.480835 | 0.51219 |  |
| 3 | 3 | 0.794681 | 0.820842 | 0.897338 |  |
| 1 | 1 | 0.464415 | 0.468269 | 0.541311 |  |
| 2 | 3 | 0.657739 | 0.668826 | 0.677064 |  |
| 1 | 1 | 0.708059 | 0.712277 | 0.741843 |  |
| 2 | 2 | 0.359972 | 0.423217 | 0.470589 |  |
| 2 | 2 | 0.526015 | 0.556992 | 0.600778 |  |
| 2 | 2 | 0.445299 | 0.570793 | 0.611902 |  |
| 1 | 1 | 0.530518 | 0.540582 | 0.540658 |  |
| 2 | 2 | 0.42713 | 0.494316 | 0.500936 |  |
| 1 | 1 | 0.701313 | 0.812321 | 0.824176 |  |
| 2 | 2 | 0.837492 | 0.891375 | 0.893752 |  |
| 2 | 2 | 0.688625 | 0.690918 | 0.698305 |  |
| 3 | 3 | 0.843975 | 0.907923 | 0.948622 |  |
| 3 | 3 | 0.457966 | 0.607727 | 0.60944 |  |
| 3 | 3 | 0.470893 | 0.482999 | 0.534787 |  |
| 2 | 2 | 0.645909 | 0.67654 | 0.683388 |  |
| 2 | 1 | 0.69649 | 0.731552 | 0.743939 |  |
| 2 | 2 | 0.468595 | 0.495035 | 0.506395 |  |
| 1 | 1 | 0.737747 | 0.740751 | 0.742296 |  |
| 1 | 1 | 0.430784 | 0.437383 | 0.484356 |  |
| 2 | 2 | 0.595971 | 0.600341 | 0.646582 |  |
| 3 | 3 | 0.480854 | 0.511923 | 0.540721 |  |
| 1 | 1 | 0.474846 | 0.489784 | 0.553789 |  |
| 1 | 1 | 0.424035 | 0.437725 | 0.471186 |  |
| 1 | 1 | 0.584431 | 0.635513 | 0.652496 |  |

3b. Output file: knn1 train.csv and knn3 train.csv. Classification accuracy for training set when data is normalized as printed on the terminal is 100% for k = 1 and 97.8% for k = 3.

Output of knn1.train.csv

|  |  |  |  |
| --- | --- | --- | --- |
| y | predicted\_y | distance1 |  |
| 3 | 3 | 0 |  |
| 3 | 3 | 0 |  |
| 2 | 2 | 0 |  |
| 3 | 3 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 3 | 3 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 3 | 3 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 1 | 1 | 0 |  |
| 3 | 3 | 0 |  |
| 3 | 3 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 3 | 3 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 3 | 3 | 0 |  |
| 1 | 1 | 0 |  |
| 1 | 1 | 0 |  |
| 1 | 1 | 0 |  |
| 3 | 3 | 0 |  |
| 1 | 1 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 3 | 3 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 3 | 3 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 3 | 3 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 1 | 1 | 0 |  |
| 3 | 3 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 3 | 3 | 0 |  |
| 3 | 3 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 3 | 3 | 0 |  |
| 3 | 3 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 3 | 3 | 0 |  |
| 3 | 3 | 0 |  |
| 2 | 2 | 0 |  |
| 3 | 3 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 1 | 1 | 0 |  |
| 3 | 3 | 0 |  |
| 1 | 1 | 0 |  |
| 3 | 3 | 0 |  |
| 1 | 1 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 1 | 1 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 3 | 3 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 3 | 3 | 0 |  |
| 3 | 3 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 3 | 3 | 0 |  |
| 3 | 3 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 3 | 3 | 0 |  |
| 3 | 3 | 0 |  |
| 3 | 3 | 0 |  |
| 3 | 3 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 3 | 3 | 0 |  |
| 1 | 1 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 1 | 1 | 0 |  |
| 3 | 3 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 3 | 3 | 0 |  |
| 3 | 3 | 0 |  |
| 1 | 1 | 0 |  |
| 1 | 1 | 0 |  |
| 3 | 3 | 0 |  |
| 3 | 3 | 0 |  |
| 3 | 3 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 3 | 3 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |
| 1 | 1 | 0 |  |
| 2 | 2 | 0 |  |
| 2 | 2 | 0 |  |

Output of knn3.train.csv

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| y | predicted\_y | distance1 | distance2 | distance3 |  |
| 3 | 3 | 0 | 0.346414 | 0.787925 |  |
| 3 | 3 | 0 | 0.459637 | 0.490447 |  |
| 2 | 2 | 0 | 0.298161 | 0.510193 |  |
| 3 | 3 | 0 | 0.346414 | 0.594293 |  |
| 1 | 1 | 0 | 0.29204 | 0.303615 |  |
| 2 | 2 | 0 | 0.371996 | 0.392936 |  |
| 2 | 2 | 0 | 0.338092 | 0.365893 |  |
| 2 | 2 | 0 | 0.498078 | 0.618737 |  |
| 3 | 3 | 0 | 0.419341 | 0.426766 |  |
| 1 | 1 | 0 | 0.385706 | 0.412489 |  |
| 2 | 2 | 0 | 0.395534 | 0.399022 |  |
| 2 | 2 | 0 | 0.446224 | 0.510164 |  |
| 3 | 3 | 0 | 0.391748 | 0.411644 |  |
| 1 | 1 | 0 | 0.3611 | 0.368565 |  |
| 2 | 2 | 0 | 0.41781 | 0.432402 |  |
| 1 | 1 | 0 | 0.265017 | 0.284269 |  |
| 1 | 1 | 0 | 0.27819 | 0.389363 |  |
| 3 | 3 | 0 | 0.40157 | 0.427158 |  |
| 3 | 3 | 0 | 0.312933 | 0.36868 |  |
| 2 | 2 | 0 | 0.297347 | 0.324728 |  |
| 2 | 2 | 0 | 0.534483 | 0.585101 |  |
| 1 | 1 | 0 | 0.315111 | 0.34513 |  |
| 2 | 2 | 0 | 0.434612 | 0.450154 |  |
| 1 | 1 | 0 | 0.328143 | 0.334877 |  |
| 3 | 3 | 0 | 0.458575 | 0.564628 |  |
| 2 | 2 | 0 | 0.325448 | 0.365893 |  |
| 2 | 2 | 0 | 0.585101 | 0.65587 |  |
| 3 | 3 | 0 | 0.349959 | 0.433286 |  |
| 1 | 1 | 0 | 0.308035 | 0.369988 |  |
| 1 | 1 | 0 | 0.3611 | 0.390823 |  |
| 1 | 1 | 0 | 0.29204 | 0.308035 |  |
| 3 | 3 | 0 | 0.444361 | 0.466236 |  |
| 1 | 1 | 0 | 0.385587 | 0.396039 |  |
| 1 | 1 | 0 | 0.372011 | 0.39326 |  |
| 2 | 2 | 0 | 0.534483 | 0.65587 |  |
| 3 | 3 | 0 | 0.295253 | 0.4329 |  |
| 2 | 2 | 0 | 0.469181 | 0.476994 |  |
| 1 | 1 | 0 | 0.368565 | 0.413736 |  |
| 3 | 3 | 0 | 0.3426 | 0.367877 |  |
| 2 | 2 | 0 | 0.569779 | 0.611013 |  |
| 1 | 1 | 0 | 0.286111 | 0.330991 |  |
| 3 | 3 | 0 | 0.341974 | 0.347344 |  |
| 2 | 2 | 0 | 0.345495 | 0.423719 |  |
| 2 | 2 | 0 | 0.362449 | 0.509079 |  |
| 1 | 1 | 0 | 0.301508 | 0.315801 |  |
| 2 | 2 | 0 | 0.65664 | 0.714027 |  |
| 1 | 1 | 0 | 0.384115 | 0.396039 |  |
| 1 | 1 | 0 | 0.295001 | 0.301508 |  |
| 2 | 2 | 0 | 0.526404 | 0.555698 |  |
| 1 | 1 | 0 | 0.385587 | 0.543638 |  |
| 1 | 1 | 0 | 0.328143 | 0.339009 |  |
| 3 | 3 | 0 | 0.28089 | 0.419993 |  |
| 2 | 2 | 0 | 0.324728 | 0.375741 |  |
| 2 | 2 | 0 | 0.446224 | 0.51232 |  |
| 2 | 2 | 0 | 0.528744 | 0.531947 |  |
| 1 | 1 | 0 | 0.474465 | 0.556003 |  |
| 2 | 2 | 0 | 0.351502 | 0.389146 |  |
| 2 | 2 | 0 | 0.362449 | 0.36422 |  |
| 2 | 2 | 0 | 0.380395 | 0.388799 |  |
| 3 | 3 | 0 | 0.322745 | 0.368798 |  |
| 3 | 3 | 0 | 0.308334 | 0.347344 |  |
| 1 | 1 | 0 | 0.265017 | 0.315801 |  |
| 2 | 2 | 0 | 0.443085 | 0.450154 |  |
| 3 | 3 | 0 | 0.359411 | 0.391748 |  |
| 3 | 3 | 0 | 0.353321 | 0.382641 |  |
| 2 | 2 | 0 | 0.440517 | 0.498078 |  |
| 2 | 1 | 0 | 0.744417 | 0.758562 |  |
| 1 | 1 | 0 | 0.371987 | 0.397836 |  |
| 2 | 2 | 0 | 0.273964 | 0.345495 |  |
| 3 | 3 | 0 | 0.257462 | 0.419993 |  |
| 3 | 3 | 0 | 0.308334 | 0.356814 |  |
| 2 | 2 | 0 | 0.432402 | 0.484048 |  |
| 3 | 3 | 0 | 0.40157 | 0.408437 |  |
| 2 | 2 | 0 | 0.50163 | 0.541787 |  |
| 2 | 2 | 0 | 0.60553 | 0.626984 |  |
| 2 | 2 | 0 | 0.530092 | 0.54228 |  |
| 1 | 1 | 0 | 0.280764 | 0.302644 |  |
| 1 | 1 | 0 | 0.302644 | 0.34779 |  |
| 3 | 3 | 0 | 0.3426 | 0.356374 |  |
| 1 | 1 | 0 | 0.396321 | 0.396756 |  |
| 3 | 3 | 0 | 0.312933 | 0.406732 |  |
| 1 | 1 | 0 | 0.34779 | 0.383162 |  |
| 1 | 1 | 0 | 0.334877 | 0.336705 |  |
| 2 | 2 | 0 | 0.325448 | 0.338092 |  |
| 2 | 2 | 0 | 0.476994 | 0.489178 |  |
| 1 | 1 | 0 | 0.241843 | 0.315111 |  |
| 1 | 1 | 0 | 0.465621 | 0.469829 |  |
| 1 | 1 | 0 | 0.393417 | 0.403138 |  |
| 2 | 2 | 0 | 0.611013 | 0.616668 |  |
| 1 | 1 | 0 | 0.393893 | 0.541504 |  |
| 2 | 2 | 0 | 0.537505 | 0.558348 |  |
| 3 | 3 | 0 | 0.459637 | 0.502021 |  |
| 2 | 2 | 0 | 0.616668 | 0.763791 |  |
| 2 | 2 | 0 | 0.293975 | 0.333749 |  |
| 2 | 3 | 0 | 0.41675 | 0.430724 |  |
| 3 | 3 | 0 | 0.314014 | 0.394333 |  |
| 3 | 3 | 0 | 0.341974 | 0.356401 |  |
| 2 | 2 | 0 | 0.37386 | 0.406967 |  |
| 1 | 1 | 0 | 0.390656 | 0.413216 |  |
| 1 | 1 | 0 | 0.486759 | 0.496547 |  |
| 2 | 2 | 0 | 0.665331 | 0.699409 |  |
| 3 | 3 | 0 | 0.531966 | 0.545082 |  |
| 3 | 3 | 0 | 0.257462 | 0.28089 |  |
| 1 | 1 | 0 | 0.317632 | 0.364673 |  |
| 2 | 2 | 0 | 0.484048 | 0.48695 |  |
| 3 | 3 | 0 | 0.322745 | 0.424359 |  |
| 3 | 3 | 0 | 0.359411 | 0.411644 |  |
| 3 | 3 | 0 | 0.314014 | 0.324335 |  |
| 3 | 3 | 0 | 0.295253 | 0.389734 |  |
| 2 | 2 | 0 | 0.298161 | 0.443085 |  |
| 1 | 1 | 0 | 0.307983 | 0.339104 |  |
| 2 | 2 | 0 | 0.272841 | 0.293975 |  |
| 1 | 1 | 0 | 0.508059 | 0.565363 |  |
| 3 | 3 | 0 | 0.380631 | 0.382641 |  |
| 1 | 1 | 0 | 0.386065 | 0.392299 |  |
| 1 | 1 | 0 | 0.374937 | 0.383162 |  |
| 2 | 2 | 0 | 0.333749 | 0.37386 |  |
| 1 | 1 | 0 | 0.393893 | 0.505194 |  |
| 1 | 1 | 0 | 0.27819 | 0.280764 |  |
| 3 | 3 | 0 | 0.324335 | 0.397534 |  |
| 2 | 2 | 0 | 0.510164 | 0.528232 |  |
| 1 | 1 | 0 | 0.241843 | 0.303615 |  |
| 3 | 3 | 0 | 0.488834 | 0.514471 |  |
| 3 | 3 | 0 | 0.353321 | 0.397534 |  |
| 1 | 1 | 0 | 0.384452 | 0.425691 |  |
| 1 | 1 | 0 | 0.364183 | 0.392299 |  |
| 3 | 3 | 0 | 0.370984 | 0.380631 |  |
| 3 | 3 | 0 | 0.440569 | 0.514797 |  |
| 3 | 3 | 0 | 0.370984 | 0.531966 |  |
| 2 | 2 | 0 | 0.371996 | 0.414534 |  |
| 2 | 2 | 0 | 0.51232 | 0.586785 |  |
| 2 | 2 | 0 | 0.297347 | 0.307795 |  |
| 2 | 2 | 0 | 0.418662 | 0.456776 |  |
| 2 | 2 | 0 | 0.541612 | 0.549724 |  |
| 2 | 1 | 0 | 0.818194 | 0.91742 |  |
| 3 | 3 | 0 | 0.394333 | 0.443882 |  |
| 1 | 1 | 0 | 0.284269 | 0.286111 |  |
| 2 | 2 | 0 | 0.626449 | 0.634961 |  |
| 2 | 2 | 0 | 0.307795 | 0.327675 |  |
| 1 | 1 | 0 | 0.42577 | 0.45635 |  |
| 2 | 2 | 0 | 0.272841 | 0.360687 |  |
| 2 | 2 | 0 | 0.273964 | 0.380395 |  |

3ci. When k = 1, it maximizes the accuracy for the training set, which gives accuracy of 100%. However, when k = 3, accuracy becomes 97.8%. k = 1 does not improve accuracy in the test set as it only gives accuracy of 94.4%.

3cii. We train the model on training data during the training phase. If we use testing data during training phase, the model hasn’t seen the data before. However, when we use training data during testing phase, the model has already seen the data. KNN will make predictions based on the nearest neighbours of the data it has been trained on, which is the training set. Therefore, it will always correctly find nearest neighbour when using training data during training phase, leading to maximising the accuracy.

3d. Increasing k can lead to better predictions because there are more neighbours to consider, however, it shouldn’t increase by too much (Reasoning discussed in extreme values of k section). Decreasing k can lead to overfitting, where model learns training data too well, where noise and irrelevant patterns are captured from the training data. This will cause model to fail to generalize to unseen data.

Let’s say I have k = 35. After running the algorithm for testing set, it gives an accuracy of 39%, which is significantly lower than k = 1 or k = 3. If k is too big, underfitting can happen, where the model is too generalized and can fail to capture patterns in the data. This leads to KNN being ineffective.

References:

https://medium.com/@sachinsoni600517/k-nearest-neighbours-introduction-to-machine-learning-algorithms-9dbc9d9fb3b2#:~:text=In%20the%20case%20of%20KNN,underlying%20patterns%20in%20the%20data.