CSE 535 Mobile Computing
Assignment 2, Models Report
Group 22: Pranay Jagtap, Evan Lam, Julia Liu, Andrew Wingersky

Introduction:

The machine learning models chosen for the purpose of developing an online application service capable of accepting human pose skeletal key points as an input and generating a sign language classifier as an output were as follows: k-nearest neighbors (KNN), decision tree (DT), support vector machine (SVM), and random forest (RF). In order to evaluate these chosen models, metrics such as precision, recall, F1-score, and accuracy are used to provide clear insight into the chosen models' performance.

Before beginning an analysis of the chosen models, it may be worth noting that in order to choose which models to pursue for this assignment, a variety of supervised learning classification techniques were chosen (due to the nature of the problem we set out to solve). Of those explored (including neural networks), KNN, DT, SVM, and RF were chosen due to improved classification outputs in addition to ease of implementation and parameter tuning.

Metric Analysis:

KNN	Precision	Recall	F1-Score	Accuracy
BUY	0.92	0.92	0.92	0.97
COMMUNICATE	0.85	1	0.92	0.97
FUN	0.67	1	0.8	0.91
HOPE	1	1	1	1
MOTHER	0.71	0.36	0.48	0.84
REALLY	0.58	0.54	0.56	0.84

Table 1: KNN Metrics

From the above table, we can see that the KNN model does very well at predicting gestures BUY, COMMUNICATE, FUN, and HOPE. Two gestures, MOTHER and REALLY, while having decent accuracy ratings, have poorer F1-Scores (representing the harmonic mean between precision and recall). "MOTHER" suffers from a lower accuracy due to a higher amount of false negatives (with the number of false positives being fairly low), while gesture "REALLY" suffers from a moderate increase to false positive and false negative counts (with both roughly rising equally). Observing the nature of the signs, we would expect to see more confusion between these two signs since the location of the hands during the motion are in very similar places relative to the nose. We also note an average accuracy of 0.9216.

DT	Precision	Recall	F1-Score	Accuracy
BUY	0.9	0.75	0.82	0.93
COMMUNICATE	0.73	1	0.84	0.93
FUN	0.8	1	0.89	0.95
HOPE	1	0.85	0.92	0.97
MOTHER	0.83	0.36	0.5	0.85
REALLY	0.44	0.62	0.51	0.79

Table 2: DT Metrics

From the above table, we can see that the DT model does very well at predicting gestures BUY, COMMUNICATE, FUN, and HOPE. Two gestures, MOTHER and REALLY, while having decent accuracy ratings, have poorer F1-Scores. MOTHER suffers from a lower accuracy due to a higher amount of false negatives, while REALLY suffers an increase in the amount of false positives. We also note an average accuracy of 0.9033, lower than that of KNN, which was to be expected due to the nature of decision trees being significantly less complex.

SVM	Precision	Recall	F1-Score	Accuracy
BUY	0.75	1	0.86	0.94
COMMUNICATE	0.85	1	0.92	0.97
FUN	0.92	1	0.96	0.98
HOPE	1	0.85	0.92	0.97

MOTHER	0.8	0.57	0.67	0.89
REALLY	0.67	0.62	0.64	0.87

Table 3: SVM Metrics

From the above table, we can see that the SVM model does very well at predicting gestures BUY, COMMUNICATE, FUN, and HOPE. Two gestures, MOTHER and REALLY, while having decent accuracy ratings, have poorer F1-Scores. MOTHER suffers from a lower accuracy due to a higher amount of false negatives, while REALLY suffers from a moderate increase to false positive and false negative counts (with more false negatives than false positives). We also note an average accuracy of 0.9366, higher than both KNN and DT.

RF	Precision	Recall	F1-Score	Accuracy
BUY	0.71	0.83	0.77	0.91
COMMUNICATE	0.79	1	0.88	0.95
FUN	0.8	1	0.89	0.95
HOPE	1	0.85	0.92	0.97
MOTHER	0.88	0.5	0.64	0.88
REALLY	0.62	0.62	0.62	0.86

Table 4: RF Metrics

From the above table, we can see that the RF model does very well at predicting gestures BUY, COMMUNICATE, FUN, and HOPE. Two gestures, MOTHER and REALLY, while having decent accuracy ratings, have poorer F1-Scores. MOTHER suffers from a lower accuracy due to a higher amount of false negatives, while REALLY suffers from a moderate increase to false positive and false negative counts (with an apparent equal number of both). We also note an average accuracy of 0.92, higher than DT only (though very comparable to KNN). Again, knowing how the different supervised machine learning models work, we anticipated the performance to compare this way. Random forest works as an enhancement on decision trees, so we expect the performance to increase; however, the complexity of the machine still does not reach that of KNearest Neighbors.

Feature Engineering:

In order to combat the curse of dimensionality suffered by many machine learning models, we decided to limit the feature set to only thought necessary. We discard features which do not provide relevant information required for classification. The given gestures (buy, fun, hope, really, communicate, mother) don't have any movements from parts like knees and such features are discarded. We discovered that removing the elbow data sets, we limit computation time while minimally impacting our model metrics. Additionally, in the models discussed, we created our own feature set, the distance of skeletal points from the nose (serving to center all points to a fixed and unmoving bodily location).

Tables 5 through 8 serve to show the observed values of evaluation metrics (tables on the left) and the minimal change after removing elbow data from the feature set(tables on the right).

KNN	Precision	Recall	F1-Score	Accuracy
BUY	1	0.92	0.96	0.99
COMMUNICATE	0.92	1	0.96	0.99
FUN	0.95	0.99	0.96	0.96
HOPE	0.92	1	0.96	0.99
MOTHER	0.8	0.57	0.66	0.94
REALLY	0.62	0.62	0.62	0.92

Table 5: KNN after Removing Elbow Data

KNN	Precision	Recall	F1-Score	Accuracy
BUY	-0.08	0	-0.04	-0.02
COMMUNICATE	-0.07	0	-0.04	-0.02
FUN	-0.28	0.01	-0.16	-0.05
HOPE	0.08	0	0.04	0.01
MOTHER	-0.09	-0.21	-0.18	-0.1
REALLY	-0.04	-0.08	-0.06	-0.08

Table 6: DIfferences Noted for KNN

KNN model with elbow data shows an observable accuracy of 0.965, leading to a change of 0.043.

DT	Precision	Recall	F1-Score	Accuracy
BUY	0.54	0.58	0.56	0.91
COMMUNICATE	0.69	0.82	0.74	0.95

DT	Precision	Recall	F1-Score	Accuracy
BUY	0.36	0.17	0.26	0.02
COMMUNICATE	0.04	0.18	0.1	-0.02

FUN	0.95	0.9	0.92	0.91
HOPE	0.85	0.92	0.88	0.97
MOTHER	0.89	0.57	0.7	0.94
REALLY	0.44	0.62	0.52	0.88

Table 7: DT after Removing Elbow Data

FUN	-0.15	0.1	-0.03	0.04
HOPE	0.15	-0.07	0.04	0
MOTHER	-0.06	-0.21	-0.2	-0.09
REALLY	0	0	-0.01	-0.09

Table 8: Differences Noted for DT

DT model with elbow data shows an observable accuracy of 0.9266, leading to a change of 0.023.

SVM	Precision	Recall	F1-Score	Accuracy
BUY	0.91	0.83	0.86	0.98
COMMUNICATE	0.65	1	0.78	0.95
FUN	0.99	0.99	0.98	0.98
HOPE	1	1	1	1
MOTHER	0.88	0.5	0.64	0.94
REALLY	0.64	0.69	0.66	0.93

Table 9: SVM after Removing Elbow Data

SVM	Precision	Recall	F1-Score	Accuracy
BUY	-0.16	0.17	0	-0.04
COMMUNICATE	0.2	0	0.14	0.02
FUN	-0.07	0.01	-0.02	0
HOPE	0	-0.15	-0.08	-0.03
MOTHER	-0.08	0.07	0.03	-0.05
REALLY	0.03	-0.07	-0.02	-0.06

Table 10: Differences Noted for SVM

SVM model with elbow data shows an observable accuracy of 0.9633, leading to a change of 0.026.

RF	Precision	Recall	F1-Score	Accuracy
BUY	0.82	0.75	0.78	0.96
COMMUNICATE	0.9	0.82	0.86	0.97

RF	Precision	Recall	F1-Score	Accuracy
BUY	-0.11	0.08	-0.01	-0.05
COMMUNICATE	-0.11	0.18	0.02	-0.02

FUN	0.9	1	0.94	0.93
HOPE	1	0.92	0.96	0.99
MOTHER	0.88	0.5	0.64	0.93
REALLY	0.57	0.62	0.6	0.91

Table 11: RF after Removing Elbow Data

FUN	-0.1	0	-0.05	0.02
HOPE	0	-0.07	-0.04	-0.02
MOTHER	0	0	0	-0.05
REALLY	0.05	0	0.02	-0.05

Table 12: Differences Noted for RF

RF model with elbow data shows an observable accuracy of 0.9483, leading to a change of 0.0283.

In all cases, we see only marginal changes in accuracy. While there are some enhancements to accuracy in some cases for certain models and certain signs, there was almost an equal degree of decrease in performance for other signs and models. By including elbow, we realized that we were also doubling our feature set from 4 features to 8 features. Another rationale for removing the elbow data was that the variation of the elbow movement was not as drastic between signs and could also be prone to overfitting.

Closing Remarks:

We note that the model which seems to perform the best in terms of accuracy in predicting all gestures was the support vector machine (which we expect to see when considering how this model attempts to create a defining/characterizing function for all gestures). The k-nearest neighbors algorithm also performed fairly well, which is also understandable when considering that, when performed properly, a gesture of one sign should have similar features to other gestures of the same sign. Decision tree and random forest had average results, but were chosen due to their ability to handle large feature sets in a computationally affordable way.

In attempting to reduce the computational price of a web-hosted classification service, we also explored creating and removing feature sets, to create a classification scheme which computed only on relevant and characteristic data (thus reducing training time and allowing for the creation of a more responsive online application service).