ML4QS Assignment 3

Akhil Chandran(2732963), Oromia G. Sero(2720857), Vidhya Narayanasamy(2733527)
Group 44

1. Introduction:

The invention of smart gadgets is one of the most amazing technological advances where smart watches, wearables and smartphones, in particular, are highly crucial for self-tracking. These gadgets' sensors can keep track of activities and can give the user insight into their way of life. Such information would increase user awareness and support living a healthier life.

In this experiment, we'll make use of one such dataset that was compiled using sensors that recorded various human movements. With the help of movement patterns, we have tried to determine if a person has parkinson's disease. The dataset is preprocessed prior to prediction, and following feature engineering techniques, it is subjected to several machine learning models. The methods for preprocessing data, feature engineering, and machine learning models are covered in detail in the sections that follow.

2. Dataset Description:

In order to develop an algorithm that can identify and analyze the movement singularities of degenerative diseases like Alzheimer's and Parkinson's as well as gait problems in senior people, these signals have been collected. The upper of a right slipper that is worn on the lower limb has been fitted with inertial sensors (a triaxial accelerometer and a triaxial gyroscope). This makes it possible to gather information in order to track people's daily activities less obtrusively[1].

Ten volunteers performed the following seven actions to replicate the signals: (1) Bradykinesia (2) Sitting with tremor (3) Ataxic gait (4) Myopathic gait (5) Muscle atrophy (6) Not pathological gait (7). The recording protocol stipulates that each movement for each subject be performed ten times, with a 20-second recording taking place between each repetition. The signals were recorded at a sampling rate of 100 Hz.

3. Data Preprocessing:

The dataset was treated to exploratory data analysis in order to incorporate all the attributes and get precise predictions. This requires that the dataset be free of missing values and outliers, and that any such values be handled appropriately.

3.1 Missing Values:

First, any missing values in the data were verified. Multiple characteristics lacked certain values. All of the features gathered from the sensor gyroscope resulted in a large percentage of NULL values. In the initial dataset, gyroscope measurements are significantly outnumbered by accelerometer measurements. This indicates that the gyroscope data has a number of missing values when the measurements are combined into a single dataset to build instances. To get the dataset ready for modeling, these missing variables must be filled up. There are various methods for filling in the missing values, such as utilizing the mean, median, or linear interpolation.

3.2 Outliers

For the detection of potential outliers, five different methods are considered: Chauvenet's criterion, mixture models, distance-based methods, the Kalman Filter and the Local Outlier Factor (LOF) method. According to Hoogendoorn and Funk, Chauvenet's criterion assumes a normal distribution of the data [2]. It was observed that this assumption is not always valid. Therefore, Chauvenet's criterion is not regarded to be a suitable outlier detection method for the problem at hand. Since there was no clear segregation of the data points over different distributions, mixture models were also not regarded to be suitable. The Kalman filter compares the observed values with the anticipated values to create a model for the expected values based on prior data and to determine how noisy a fresh measurement is[2]. Lastly, the LOF algorithm is extremely computationally expensive and is therefore disregarded for outlier detection [3].

In this study, we used the Kalman filter was used as an integrated approach to identify outliers and impute missing values on each individual attribute, i.e., accelerometer and gyroscope measurements in the x, y, and z-planes. Based on historical data, the Kalman filter generates a model for expected values. It calculates how noisy a new measurement is by comparing observed and predicted values[2]. In the presence of noise, this will aid in determining the best estimate by combining measurements. Measured values that deviate from expectations were replaced with a more reliable expected value based on past values, and unknown values were imputed with predicted values using this method.

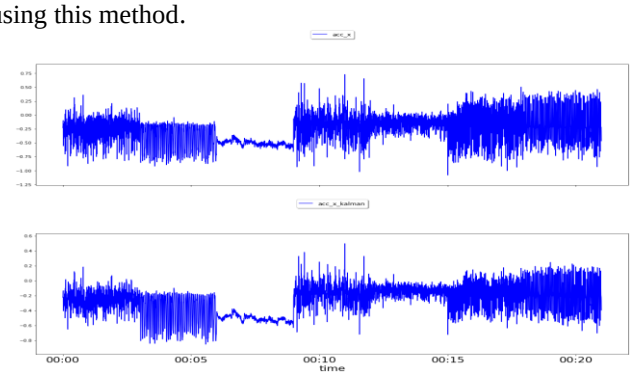

Figure1: Kalman filter applied to acc_x of inertia sensors of the first individual (first plot -original values and the second plot - values after applying the Kalman filter)

The Kalman filter was set up with trivial transition and observation matrices, which were then used to find other best parameters for the data at hand. Figure 1 shows the Kalman filter applied to one of the accelerometer measurements, namely, acc_x, of the first individual in our dataset. The difference in scale indicates the removal of some outlier values at both ends.

3.3 Data Transformations

Due to the relevance of higher frequency measurements in explaining different types of movements in the dataset, the low-pass filtering method was not applied.

4. Feature Engineering:

A variety of other features can be inferred in addition to those that were mentioned in Section 2. When modeling the data, these features may be useful. The extra characteristics include clustering, frequency-based, time-based, and principal component analysis features.

Principal component analysis (PCA) is an unsupervised algorithm with the goal to extract useful information from a dataset, in order to create new orthogonal variables (principal components) [4]. The number of principal components is determined using the amount of explained variance.

After dealing with noises and missing values, Principal component analysis (PCA) was used to identify the attributes that best explain the variance. PCA was applied to all Kalman filtered attributes. Figure 2 shows that the explained variance by the principal components begins to decline after 2. As a result, two components were chosen and added to the dataset going forward.

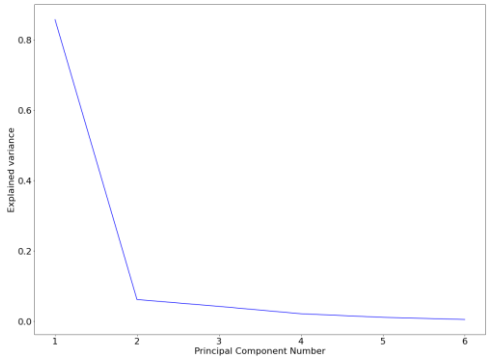

Figure2: Explained variance by principal components ranked on importance.

Furthermore, time domain and frequency (Fourier) transformations were used to enhance our predictive performance.

Time Domain Transformation: Utilizing the time domain is another method for producing features. A descriptive value, such as the mean, minimum, maximum, or standard deviation, is produced and utilized to explain each numerical attribute using a time period of length[2].New time domain features were derived in an attempt to assess patterns and fluctuations in measurements that would be useful for prediction.

For comparison, window sizes of 5 seconds (20 instances), 10 seconds (40 instances), and 20 seconds (80 instances) were chosen. Figure 3 depicts the first individual's Kalman filtered attribute acc_x. The 5-second window appears to make a small difference in the amount of noise, whereas the 20-second window appears to smooth out a lot of variances. As a result, a window size of 10 seconds (40 instances) was chosen as a good balance between these two. Mean imputation was used for any NAN value that appeared in the newly generated attributes.

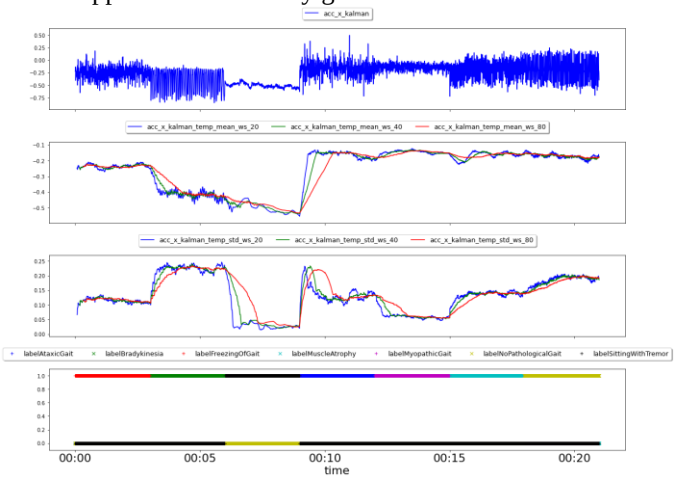

Figure3: Numerical temporal aggregation with different window sizes

Frequency Domain Transformation: In addition to time-based features, frequency-based features can also be utilized to condense the data across a period of time. In order to observe the number of frequencies per time frame, a Fourier transformation is performed. As a result, attributes such as the greatest amplitude frequency, the average of the frequency-weighted signals, and the power spectrum entropy are produced [2].

The amplitudes of the frequencies were computed using a Fourier Transformation with a window size of 10 seconds(40 instances). This window size will assist us in identifying interesting frequencies without introducing too many features. The maximum frequency, frequency signal weighted average, and power spectral entropy

are all displayed in Figure 4. The frequency signal weighted average provides little visual information here, whereas the power spectral entropy has the most noise.

We added the three aggregate features to the dataset for all sets of features generated using Kalman filtering for all individuals.

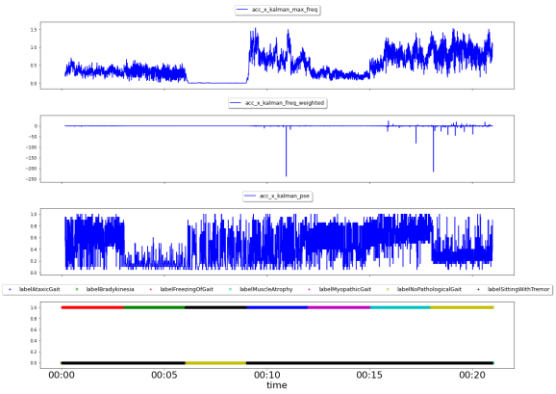

Figure4: Frequencies with the aggregated features for the Fourier transformation

Clustering: Using the unsupervised machine learning technique called clustering, it is possible to put related cases together. In the context of this study, the use of clustering may provide more context to groupings of features, which may improve the predictive performance of models. K-means can be used as one method of clustering. Due to the iterative nature of the technique, K-means enables the clustering of vast volumes of data [5].

To maximize the value of the groups of observations, the right number of clusters must be identified. Several different measures can be used to determine the right number of clusters. The silhouette coefficient, which assesses resemblance both within and between clusters of examples, is one of these metrics. A result of +1 implies high intra-cluster similarity and low inter-cluster similarity, with the silhouette score ranging from 1 to +1[6].

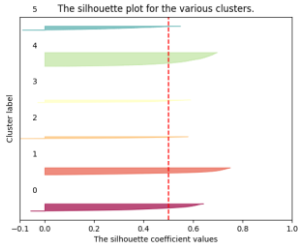

Figure5: The Silhouette plot for the various clusters

In this experiment, the various sensor features are subjected to K-means clustering in order to provide context for the sensor measurements. For the accelerometer and gyroscope data, this leads to separate clusterings. Maximizing the silhouette score for each sensor leads to the identification of the most suitable number of clusters. In this case, the maximum silhouette score of various clusters is 0.5 as shown in the figure5.

5. Feature Selection:

After the feature engineering process, we moved forward to find the most impactful features from the set of all features. Going through Kalman Filtering, Principal Component Analysis, temporal and frequency transformation, has created new features which may have strong predictive power. This was done using the forward selection method of feature classification.

Performing an individual person-wise feature selection procedure seemed wise and proved effective in finding the best features out of the classification process. Running the forward selection operation to find 10 feature outputs over all features produced a set of features most accurate in predicting across a 70-30 train to test split. This was performed across datasets for each individual—ten in all—to find the best set of features repeatedly present in the set of best predicting features from each of the results. With a cutoff for accuracy score of 0.95, a set of five features out of them were segregated. From this, the features present in all ten sets of results were narrowed down to features as presented in Figure6 these scores were averaged across features that most frequently appeared among the ten people. These were then chosen as the “selected features” as part of the predictive model analysis.

In addition, performing the feature selection procedure on the combined dataset to infer optimized performance conditions yielded Figure7. This shows that there was a relative leveling-out of the accuracy curve past eight features. We took this under consideration for efficiency and performance reasons.

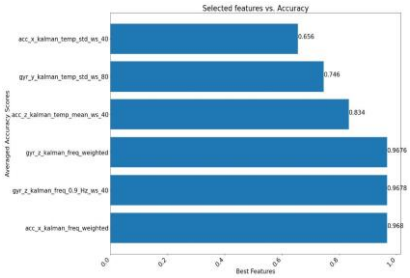

Figure6: Accuracy of the features selected

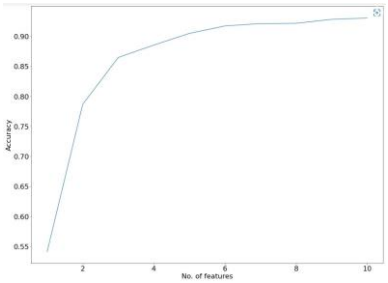

Figure7: Accuracy for the no. of features

6. Predictive Models:

The aim of this paper is to predict Parkinson's disease based on the corresponding movement types. This can be translated to a binary classification problem, where one instance is classified to be one of the given eighteen activities. Since there are eighteen activity labels, this is a multi-class classification problem. The most suitable models for binary-class classification, neural network (NN), k-nearest neighbors (kNN), and naive Bayes (NB) [7]. These were used for feature evaluation for the selected sets of features. To determine how good these models perform in relation to each other, several evaluation metrics can be used in combination with visual evaluation methods. The accuracy shows the number of correctly-classified instances with respect to all of the instances in the data. Since the data contains no extreme class imbalances, the accuracy seems to be an adequate measure for the evaluation of model performance.

7. Evaluation:

We decided to evaluate the performances on one non-deterministic predictive model—Feed Forward Neural Network; and two deterministic models—k-nearest nearest neighbors and Naives Bayes; without the notion of time. Time was not considered since date-time information was not part of the initial dataset, and our date-time assignment based on the data description from our source, was only fairly justifiable enough to perform speculative time series studies and feature engineering.

We started out with the non-deterministic predictive model in Neural Network by first running simulations to find the best regularization parameter (α). The value α=0.001 was found to yield best training and test performance across all individuals—Figure7.

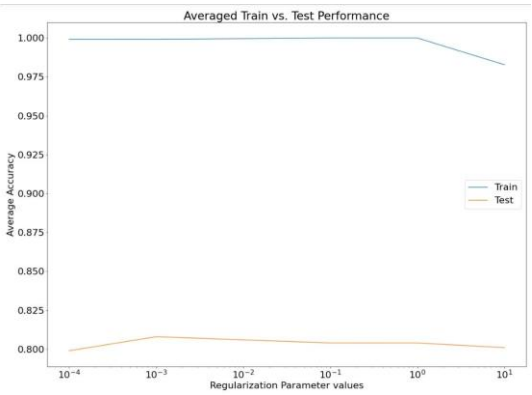

Figure8: Average accuracy for regularization parameters

With regularization value α=0.001, and with gridsearch for all the chosen models, simulations were run across all individuals for certain chosen sets of features—(a) an initial set of all original sensor measurements in addition to Kalman filtered and PCA features; (b) time and (c) frequency (fourier transformed) features; and (d) the chosen features from the feature selection step. This added up to 176 features in total. Performances were measured for the chosen deterministic and non-deterministic predictive models for the sets of features. Figure9 shows the results for individual 1 and individual 10.

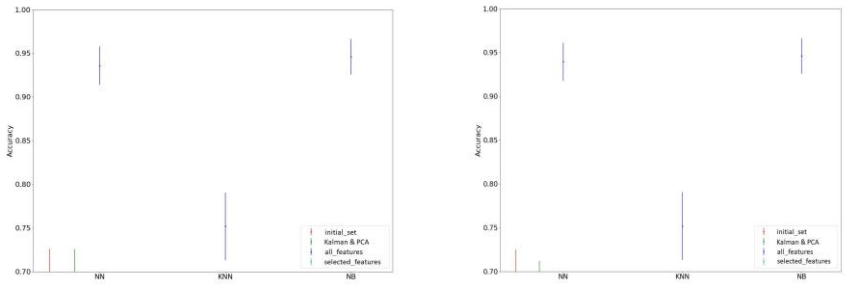

Figure9: Accuracy for individual1(left) and individual 10(right) for different predictive models

8. **Result**:

The initial dataset and the dataset after feature engineering techniques such as Kalman Filtering, PCA, time and frequency domain abstractions produced satisfactory results with an accuracy of ~0.73. The selected features obtained through the feature selection step did not produce desirable results and had a low accuracy. The models Naives Bayes and Neural Networks performed well when all of the features were enabled. Figure8 shows that the accuracy for Naives Bayes is ~0.95 and for Neural networks is ~0.94 for different individuals. The accuracy of the KNN model is relatively low even when more features are included. However, it is obvious from the accuracy of Naives Bayes and Neural Networks model that adding more features improves model performance.

9. **References**:

1. Simulation of Parkinson movement disorders. (2018, 18 mei). Kaggle. Geraadpleegd op 26 juni 2022, van https://www.kaggle.com/datasets/giofra/simulation-of-parkinson-movement-disorders

2. Hoogendoorn, M., & Funk, B. (2018). Machine Learning for the Quantified Self. *Cognitive Systems Monographs*. https://doi.org/10.1007/978-3-319-66308-1

3. Gao, J., Hu, W., Zhang, Z.M., Zhang, X., Wu, O.: Rkof: Robust kernel-based local outlier detection. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg (2011)

4. Abdi, H., Williams, L.J.: Principal component analysis. Wiley interdisciplinary reviews: computational statistics 2(4), 433–459 (2010)

5. Likas, A., Vlassis, N., Verbeek, J.J.: The global k-means clustering algorithm. Pattern recognition 36(2), 451–461 (2003)

6. Thinsungnoena, T., Kaoungkub, N., Durongdumronchaib, P., Kerdprasopb, K., Kerdprasopb, N.: The clustering validity with silhouette and sum of squared errors. learning 3(7) (2015)

7. Ortner, A. (2022, 31 mei). Top 10 Binary Classification Algorithms [a Beginner's Guide]. Medium. Geraadpleegd op 26 juni 2022, van https://towardsdatascience.com/top-10-binary-classification-algorithms-a-beginners-guide-feeacbd7a3e2

8. Visa, S., Ramsay, B., Ralescu, A.L., Van Der Knaap, E.: Confusion matrix-based feature selection. MAICS 710, 120–127 (2011)

9. Juba, B., Le, H.S.: Precision-recall versus accuracy and the role of large data sets. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 4039–4048 (2019)