**Data 622: Machine Learning and Big Data**
**Assignment 3**
*Dirk Hartog*

Support Vector Machines (SVM) are one of many supervised machine learning techniques used to predict outcomes of a target variable. As a classification technique SVMs work by maximizing the separation between the classes in the target variable. These models work well with non linear data by using a kernel trick and are considered to be an efficient “out of the box” classifier. SVM’s have been researched for their use in predictive modeling in many domains including predicting the presence or absence of diseases. Guhathakurata S, Kundu S, Chakraborty A, et al. evaluated the performance of SVM’s to predict COVID - 19 and when compared to other classification models (K-nearest neighbors, Näive Bayes, Random Forest, AdaBoost and Decision Trees), SVM out performed all of them (1).

For this project using a SVM model to predict Parkinson’s Disease from the features of the data set is indicated. First, the data set was not very large, about 2,105 observations and from previous assignments it was thought that the relationships between the features were not linear. Another strength of using an SVM is that they can handle data with high dimensionality. While the number of dimensions did not exceed the number of observations the initial model built uses 32 features. The primary aims of this assignment were to build a Support Vector Machines to predict the presence or absence of Parkinson’s Disease and to compare the performance to previous models built (Logistic Regression, Decision Trees, and Random Forest) with this particular data set. Three total SMV models were built using Python’s library Scikit Learn, using the radial basis function (RBF) kernel. The radial basis function was used for its application when data is non linear data. For all models data was centered as the RBF kernel assumes a normal distribution and categorical variables were converted to dummy variables.

The first model (SVM_All) was built using all features (listed in the section “Data Key” in the Jupyter notebook) and default parameters (C = 1, gamma = “scale”, kernel = “rbf’). In the second model (SVM_CV) cross-validation using the Gridsearch function was used to find the best parameters for C (regularization) and gamma which were then used to build another model with all of the features. Lastly a model (SVM_IMP) was built with a narrowed down selection of features based on importance from the second model. The thought was in reducing the number of features we might eliminate any unnecessary noise by features that were not contributing much to the prediction. For SVMs with nonlinear kernels, permutation importance can be used to estimate feature importance by measuring the contribution of each feature to a fitted model's statistical performance. Cross Validation was again deployed and a model was fitted using the recommended parameters for gamma and C (0.01 and 100 respectively). The results of all three SVM models are listed in the table below along with the confusion matrix for the third model.

| Metric | SVM_All | SVM_CV | SVM_IMP |
|---|---|---|---|
| F1_score | 0.836186 | 0.845105 | 0.890000 |
| Sensitivity | 0.865823 | 0.863291 | 0.901266 |

| | | | |
|---|---|---|---|
| Specificity | 0.658228 | 0.658228 | 0.658228 |
| Precision | 0.808511 | 0.827670 | 0.879012 |
| Accuracy | 0.787975 | 0.802215 | 0.860759 |
| AUC | 0.762025 | 0.781857 | 0.847257 |

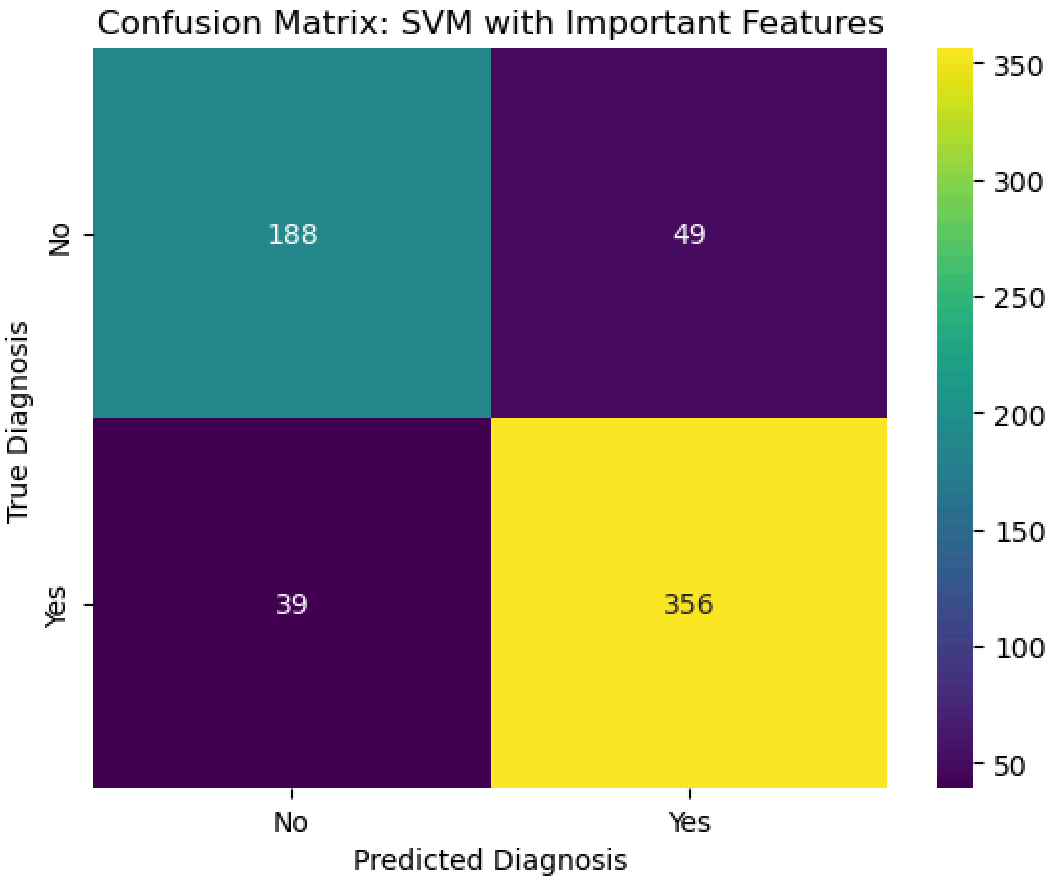

***Fig. 3***

When comparing the performance of all the SVM models, the final model built SVM_IMP, using selected features by importance and cross validation, had the highest F1 score, sensitivity, precision, and accuracy. Specificity for all SVM models were the same. The second model improved upon the first using cross validation, underpinning the importance of other parameters like regularization and gamma.

Performance metrics across all models built using this data set can be found below. When comparing the SVM models to those from previous assignments, SVM with cross validation and using selected features out performed all models except the Random Forest model (highlighted in the table). Sensitivity rate in the Random Forest model was 97.2% while for the best SVM it was 90.1%.

| Metric | LR | Decision Tree_UPDRS | Decision Tree_Tremor | Random Forest | SVM_All | SVM_CV | SVM_IMP |
|---|---|---|---|---|---|---|---|
| **F1_score** | 0.8564 | 0.852071 | 0.813049 | 0.854260 | 0.836186 | 0.845105 | **0.890000** |
| **Sensitivity** | 0.875 | 0.918367 | 0.826531 | **0.971939** | 0.865823 | 0.863291 | 0.901266 |
| **Specificity** | 0.725 | 0.612500 | 0.662500 | 0.662500 | 0.658228 | 0.700422 | **0.793249** |
| **Precision** | 0.8386 | 0.794702 | 0.800000 | 0.762000 | 0.808511 | 0.827670 | **0.879012** |
| **Accuracy** | 0.818 | 0.802215 | 0.764241 | 0.794304 | 0.787975 | 0.802215 | **0.860759** |

The data had an imbalance between the classes within the target variable. The majority of observations were labeled as having Parkinson's (approx. 60%). Imbalance was not accounted for when building the models and may have affected the model performance in predicting the minority class. Specificity improved along all iterations of the SVM indicating that the imbalance may not have been that strong of an influence, however further evaluation is needed.

The study of Support Vector Machines to predict the presence of Parkinson’s Disease is not new. While no studies were found that used the same data set, other publications have shown the value of SVM’s for the detection of this disease. Govindu A and Palwe S took three approaches, comparing Random Forest, KNN, SVM and Logistic Regression in each approach. SVM yielded the highest accuracy with 5/22 features being used after conducting a Principal Component Analysis to identify those attributes. In a second approach when accounting for data imbalance, KNN performed the best. A third approach in the study found that the Random Forest classifier produced the highest accuracy when using all 22 features, with only adjustment was scaling the data. The authors concluded that both Random Forest and SVM are suited well to handle outliers and are robust models (2). I agree with this as these models can be used with minimal data engineering which make them easy to use. Another study that looked at using vocal changes to detect Parkinson’s Disease found favorable performance using SVM versus decision trees with and without hyperparameter tuning and feature selection (3). We can also see the superiority of SVM over decision Trees in other studies that predict the presence and absence of disease. Chen S, Jian T, Chi C et. al. looked at the detection of prostate cancer and concluded that while the decision tree did not perform as well, the output of the DT model was similar to the clinical pathway (4). Decision trees are known to be interpretable and friendly to non-technical readers making them conceptually more intuitive to an audience like doctors.

The data used in this study was complete, the target outcome was binary, and the underlying distribution of features was unknown making SVM a good choice for this data set.

It is clear from evaluating research that compares many supervised classification techniques that depending on the data, feature engineering/selection techniques, certain models may perform better. It is also reasonable to think that exploring the use of different models and comparing them can serve as an extension of the exploratory phase. When evaluating models using appropriate metrics can also provide confidence in how they performed (5).  Overall, It is imperative to thoroughly explore the data that is being used to identify how the features and target variable may affect the model performance. This in turn can assist in understanding the correct metrics that will provide reliable conclusions on the model performance.

**Resources:**

1. Guhathakurata S, Kundu S, Chakraborty A, Banerjee JS. A novel approach to predict COVID-19 Using Support Vector Machine. Data Science for COVID-19, Academic Press. 2021, Pages 351-364. https://doi.org/10.1016/B978-0-12-824536-1.00014-9.
2. Govindu A, Palwe S. Early detection of Parkinson's disease using machine learning. Procedia Computer Science. Volume 218, 2023, Pages 249-261. https://doi.org/10.1016/j.procs.2023.01.007.
3. Alshammri R, Alharbi G, Alharbi E, Almubark I (2023) Machine learning approaches to identify Parkinson’s disease using voice signal features. *Front. Artif. Intell.* 6:1084001. doi: 10.3389/frai.2023.1084001
4. Chen S, Jian T, Chi C, Liang Y, Liang X, Yu Y, Jiang F and Lu J (2022) Machine Learning-Based Models Enhance the Prediction of Prostate Cancer. Front. Oncol. 12:941349. doi: 10.3389/fonc.2022.941349
5. Ahmad A, Safi O, Malebary S, Alesawi S, Alkayal E. Decision Tree Ensembles to Predict Coronavirus Disease 2019 Infection: A Comparative Study. 2021 https://doi.org/10.1155/2021/5550344