Final Project Part 2 Report
Calvin Schaul
IDSN542 Machine Intelligence

Part 2 Report Begins with "Correlations (part 2)"

Domain

My grandpa has Parkinson's disease. Watching the disease progress in one of the smartest and hardworking people I know has been incredibly challenging. Further, the strain on my family to be present for the seemingly endless amount of testing and assessment has been a constant source of stress. What if there was a better way to identify if a patient has Parkinson's?

For our final machine intelligence project, I selected a dataset containing synthetic data on Parkinson's patients (real data not publicly available due to Health Insurance Portability and Accountability Act HIPAA [1]). This data (described in depth in the sections below) contains a variety of categorical and numeric data across clinical measurements, standardized Parkinson's test scores, and other lifestyle and symptom data.

My goal is to use this data to accurately predict if a patient has Parkinson's based on their medical and lifestyle data.

Dataset

This dataset was obtained from Kaggle (link) and specifically noted it was using synthetic data, not real personal data. There are 2,105 rows of data with 35 attributes (included a patient ID number as well as a diagnosis "target" attribute). Throughout the many attributes in this dataset, I can work towards my goal of a better way of diagnosing Parkinson's by demonstrating that machine learning models can ingest normal clinical and personal data to help predict a diagnosis.

Depending on the scope of my final project and how much more image processing is covered throughout our coursework, I identified a second dataset that uses patient drawings of lines and spirals and groups them into healthy patients and Parkinson's patients. This study came from RMIT University and Dandenong Neurology in Australia. (link)

Problem Type

The problem I am aiming to solve with this dataset is a classification problem. The two categories that I ultimately care about are if a patient is healthy or if a patient has Parkinson's. Since this dataset contains a diagnosis attribute, I will use that as my target value to classify a patient based on their other health and personal data.

Attributes

Below is a table of the attributes available within this dataset. Diagnosis will be used as the target attribute to classify based on all of the other data for a patient.

By creating a simple function to check for null values across the dataset, all data was verified to be present. There were no missing values.

Attribute	Type	Description
PatientID	Numeric	Unique identifier assigned to each patient (3058–5162).
Age	Numeric	Age of the patient (50–90 years).
Gender	Categorical	Gender of the patient: 0 = Male, 1 = Female.
Ethnicity	Categorical	Ethnicity of the patient: 0 = Caucasian, 1 = African American, 2 = Asian, 3 = Other.
EducationLevel	Categorical	Education level: 0 = None, 1 = High School, 2 = Bachelor's, 3 = Higher.
BMI	Numeric	Body Mass Index (15–40).
Smoking	Categorical	Smoking status: 0 = No, 1 = Yes.
AlcoholConsumption	Numeric	Weekly alcohol consumption in units (0–20).
PhysicalActivity	Numeric	Weekly physical activity in hours (0–10).
DietQuality	Numeric	Diet quality score (0–10).
SleepQuality	Numeric	Sleep quality score (4–10).
FamilyHistoryParkinsons	Categorical	Family history of Parkinson's Disease: 0 = No, 1 = Yes.
TraumaticBrainInjury	Categorical	History of traumatic brain injury: 0 = No, 1 = Yes.
Hypertension	Categorical	Presence of hypertension: 0 = No, 1 = Yes.

Diabetes	Categorical	Presence of diabetes: 0 = No, 1 = Yes.
Depression	Categorical	Presence of depression: 0 = No, 1 = Yes.
Stroke	Categorical	History of stroke: 0 = No, 1 = Yes.
SystolicBP	Numeric	Systolic blood pressure (90–180 mmHg).
DiastolicBP	Numeric	Diastolic blood pressure (60–120 mmHg).
CholesterolTotal	Numeric	Total cholesterol levels (150–300 mg/dL).
CholesterolLDL	Numeric	LDL cholesterol levels (50–200 mg/dL).
CholesterolHDL	Numeric	HDL cholesterol levels (20–100 mg/dL).
CholesterolTriglycerides	Numeric	Triglyceride levels (50–400 mg/dL).
UPDRS	Numeric	Unified Parkinson's Disease Rating Scale score (0–199). Higher = greater severity.
MoCA	Numeric	Montreal Cognitive Assessment score (0–30). Lower = greater cognitive impairment.
FunctionalAssessment	Numeric	Functional assessment score (0–10). Lower = greater impairment.
Tremor	Categorical	Presence of tremor: 0 = No, 1 = Yes.
Rigidity	Categorical	Presence of muscle rigidity: 0 = No, 1 = Yes.
Bradykinesia	Categorical	Presence of bradykinesia: 0 = No, 1 = Yes.
PosturalInstability	Categorical	Presence of postural instability: 0 = No, 1 = Yes.
SpeechProblems	Categorical	Presence of speech problems: 0 = No, 1 = Yes.
SleepDisorders	Categorical	Presence of sleep disorders: 0 = No, 1 = Yes.
Constipation	Categorical	Presence of constipation: 0 = No, 1 = Yes.
Diagnosis	Categorical	Parkinson's Disease diagnosis status: 0 = No, 1 = Yes.
DoctorInCharge	Text	A redacted ID of the doctor present for the patient

Correlations

With all of the data available, a correlation matrix was generated to see how each attribute correlated to the “Diagnosis” target attribute. One simple piece of data preperation was done for this step. The DoctorInCharge column was removed since it was listed as “DrXXXConfid” for every row - an unhelpful, redacted piece of information.

Below is the raw pasted output of my correlation matrix:

Correlation of features with 'Diagnosis':

Diagnosis	1.000000
UPDRS	0.398006
Tremor	0.274370
Rigidity	0.185611
Bradykinesia	0.184042
PosturalInstability	0.147519
Age	0.065344
Depression	0.059080
Diabetes	0.057067
AlcoholConsumption	0.036699
BMI	0.030114
Stroke	0.028093
Constipation	0.025327
TraumaticBrainInjury	0.022964
Gender	0.016835
CholesterolTriglycerides	0.015610
CholesterolLDL	0.014707
FamilyHistoryParkinsons	0.013363
PhysicalActivity	0.012940
Smoking	0.005241
EducationLevel	0.004557
SystolicBP	-0.004413
Ethnicity	-0.005068
SleepDisorders	-0.010578
Hypertension	-0.011587
SpeechProblems	-0.012220
CholesterolTotal	-0.019001
CholesterolHDL	-0.019626
DietQuality	-0.022992
DiastolicBP	-0.029074

SleepQuality	-0.043295
PatientID	-0.043508
MoCA	-0.173104
FunctionalAssessment	-0.225036

A brief initial analysis reveals that tremors, a key symptom of Parkinson's, were highly correlated (0.274) with our diagnosis. Further, each of the three Parkinson's assessments had a high absolute value of correlation with the diagnosis The Unified Parkinson's Disease Rating Scale was positively correlated, showing that higher scores on this test were present in patients diagnosed with Parkinson's. The Functional Assessment and Montreal Cognitive Assessment were strongly negatively correlated, showing that patients with low scores on these tests were likely to be diagnosed.

Correlation (part 2)

While certain features, such as the specialized Parkinson's tests, or tremors proved to be highly correlated, further analysis was conducted to see correlations between features. In code, this was done by plotting correlation matrices against each other with a certain threshold for correlation. This way, less-releant features could be excluded.

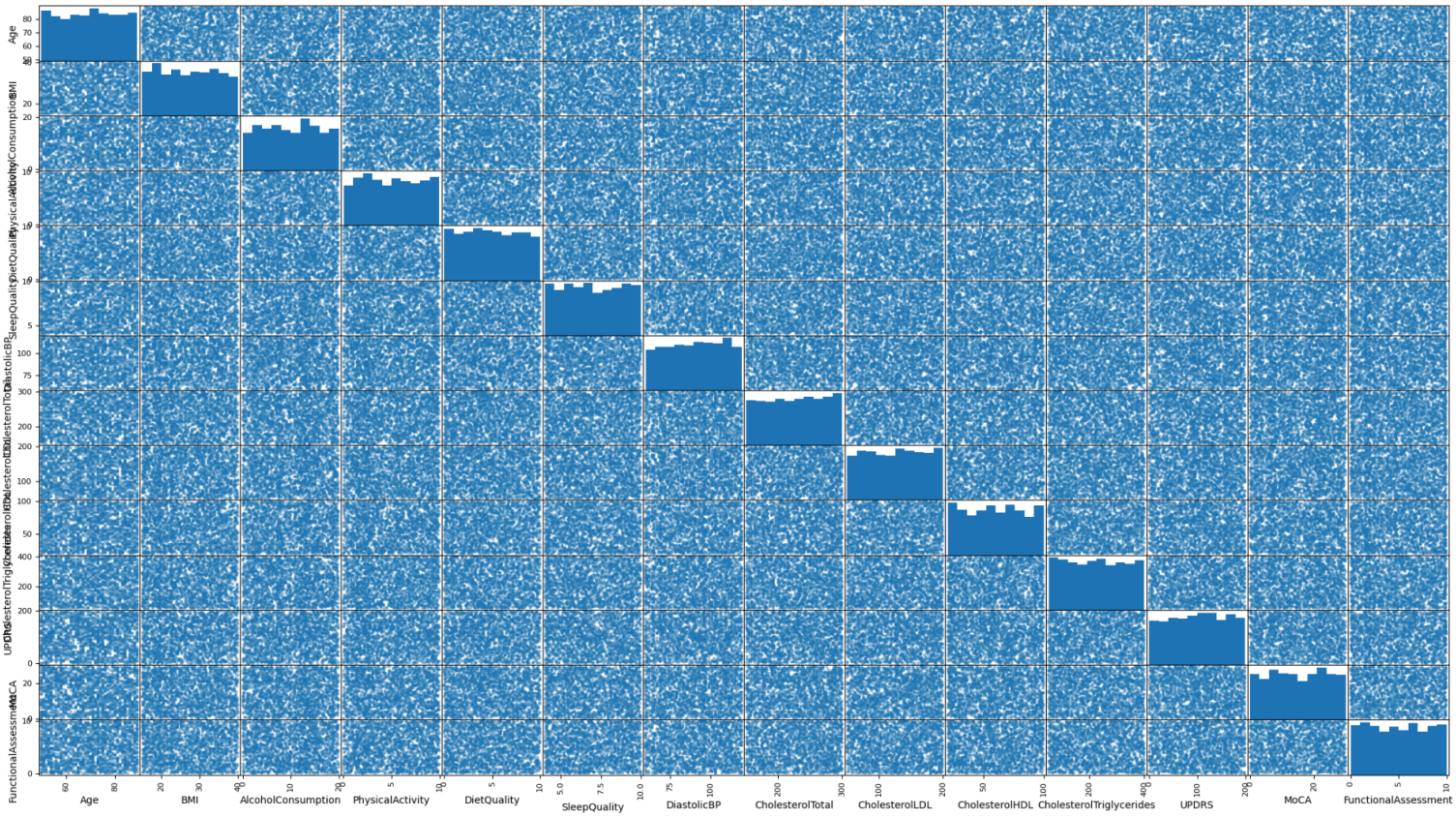


Figure 1 (above): Plot of features with >.01 correlation

Although satisfying and somewhat artistic, this plot did not convey any linear or clear relationships between correlated features.

Models

The three models chosen to analyze this dataset were DecisionTreeClassifier, RandomForestClassifier, and Support Vector Classifier. In general, classification models were needed due to the “0” or “1” nature of Parkinson's diagnosis. The goal of these models is to process the input features about the patient and classify their diagnosis.

DecisionTree and RandomForest were chosen due to my experience with these models for regression in previous assignments and labs. I was curious to see the relationship between the two models with a larger, more complicated dataset and now with the classification task. Support Vector Classifier was chosen due to its inherent Binary classification use. While investigating multi-class classification on image data in a previous lab, I noticed Geron mentioned Support Vector classification for binary classification [2]. Further, the kernel trick used to do higher-degree separation seemed very interesting.

Transformers

Due to the polished nature of this dataset, no custom transformers were used. However, ColumnTransformer and individual pipelines were used to process numerical and categorical data. First, a numeric pipeline was used to apply a standard scalar. Specifically for the Support Vector Classifier, SciKitLearn highly recommends scaling data from their “Practical Use” documentation [3]. This also helped with the wide range of numeric values seen, including age (50-90 among patients in this dataset) and the various clinical measurements (i.e systolic blood pressure or cholesterol). A second pipeline was used to apply One-Hot encoding to categorical data. The categorical data was presented as 0’s and 1’s for Yes/No patient information, such as smoking status. Other categorical information such as ethnicity was listed as 0,1,2,3. This fits well into One Hot encodings structure and didn’t require the rank-based approach of ordinal encoding.

Both of these pipelines were applied together in a full pipeline and applied as a ColumnTransformer to the dataset

Training - Hyperparameters & ROC Results

The dataset was split into training data using a stratified shuffle split. Testing data was set aside for this portion of the lab. With the finalized, transformed training data each of the three models were initialized. Next, hyperparameters were identified for each model and optimized using a cross-validated grid search.

For the Decision Tree, “Criterion”, “Splitter”, “Max Depth”, “Min Samples Split”, “Min Samples Leaf”, “Max Features”, and “Class Weight” were put into the grid search.

For Random Forest, previous class examples form random forest optimization were used to do a grid search on the following hyperparameters: “# of estimators”, “Max Features”, and“Bootstrap”.

The Support Vector Classifier’s hyperparameters “C” (regularization parameter), “Kernel”, “Gamma”, “Class Weight”, and “Probability” were used in grid search. Specifically, Scikit learn mentioned that landing on proper kernel settings was important for this model.

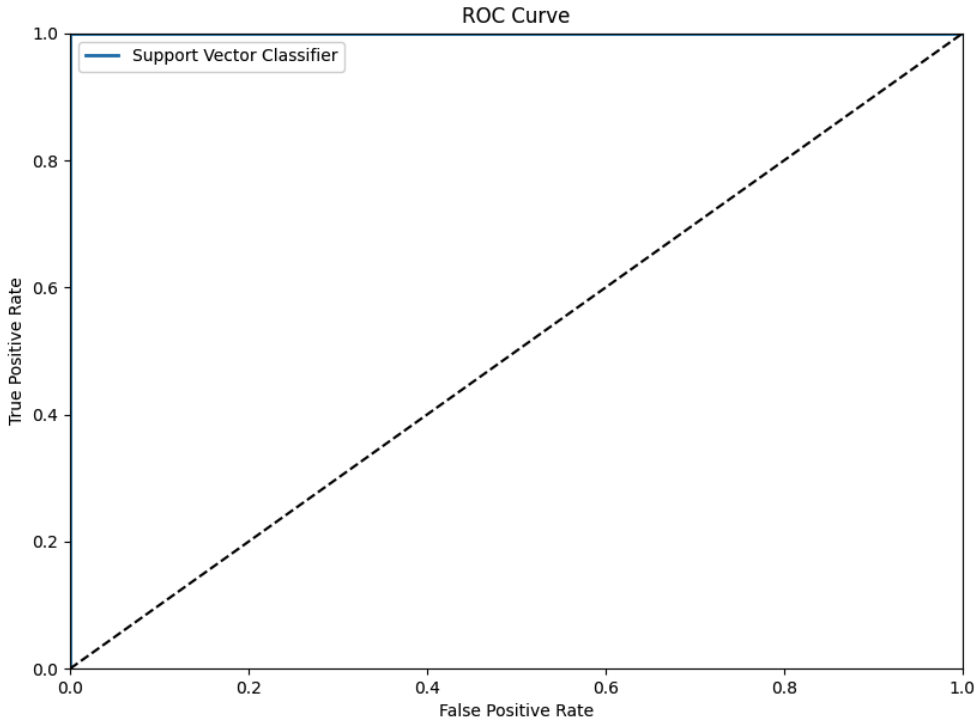
After each model was fit and and the grid search was performed, the following hyperparameters were identified using the best_estimator method.

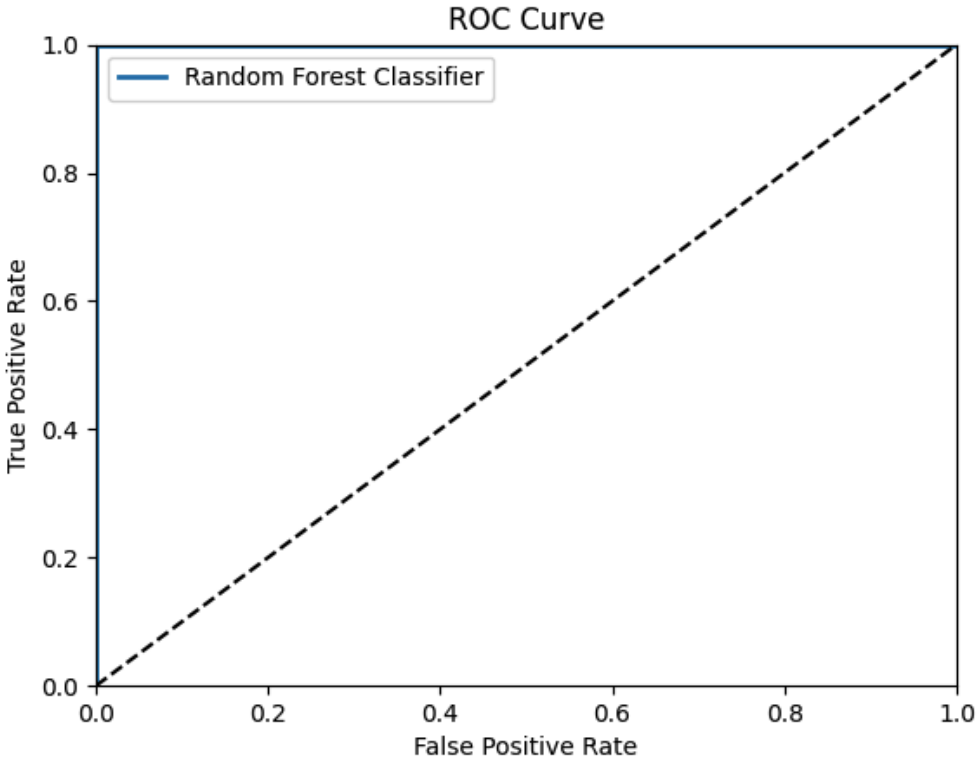
Best parameters for Decision Tree Classifier: {'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'best'}

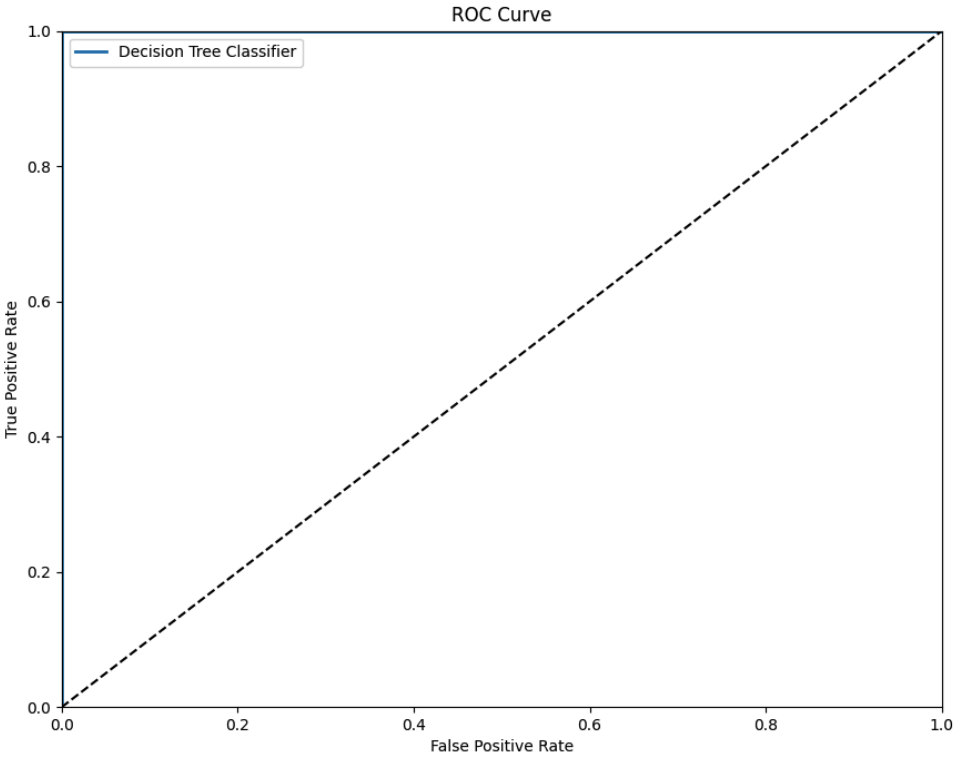
Best parameters for Random Forest Classifier: {'max_features': 2, 'n_estimators': 30}

Best parameters for Support Vector Classifier: {'C': 0.1, 'class_weight': None, 'gamma': 'scale', 'kernel': 'linear', 'probability': True}

With the models trained, performance was identified using ROC plotting to see the true positive vs false positive rate across models. The results suggest severe overfitting and that the models have predicted the patterns of the training data. Further analysis will be conducted in the next part of this lab when looking at test data.


ROC Curve
Support Vector Classifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve
Random Forest Classifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0



To confirm my hypothesis of overfitting, a sample of the training dataset was predicted using the decision tree. The model had 100% probability of its prediction across 10 samples and its mean calculation of the diagnosis exactly matched the actual calculated mean.

```
[[1. 0.]
[0. 1.]
[1. 0.]
[1. 0.]
[0. 1.]
[0. 1.]
[1. 0.]
[0. 1.]
[1. 0.]
[1. 0.]]
Diagnosis mean: 0.6193586698337292
Predictions mean: 0.6193586698337292
```

In conclusion, the models are overfitting at this stage which may cause problems when moving to test data, however exact predictions are not expected when passing in data the model hasn't trained on yet.

References

[1] https://www.hhs.gov/hipaa/index.html
[2] Hands on Learning with Scikit-Learn, Geron - Chapter 3 Classification
[3] https://scikit-learn.org/stable/modules/svm.html#tips-on-practical-use