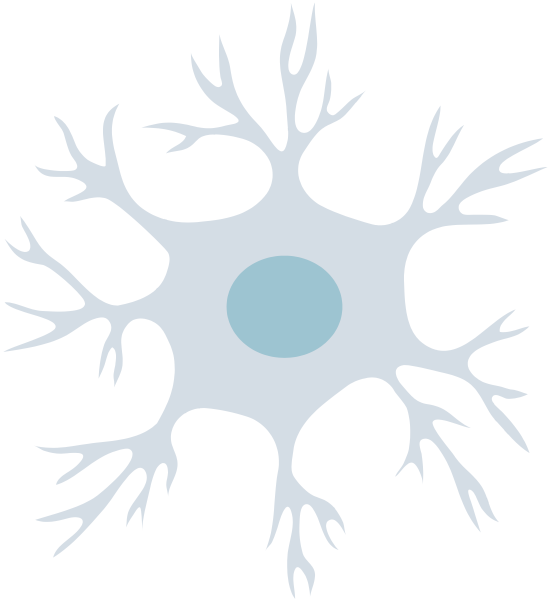
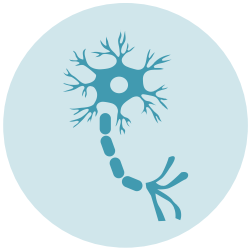


# Parkinson's Disease Predictive Machine Learning Model

Luna Pérez Troncoso

# Introduction



The **early detection of PD** is a **growing priority** within both clinical practice and research.

# Introduction



The **early detection of PD** is a **growing priority** within both clinical practice and research.

**Many individuals remain undiagnosed until the disease has already progressed**, limiting the effectiveness of available therapeutic interventions.

# Introduction



The **early detection of PD** is a **growing priority** within both clinical practice and research.

**Many individuals remain undiagnosed until the disease has already progressed**, limiting the effectiveness of available therapeutic interventions. Consequently, **timely diagnosis** can significantly **influence patient outcomes and long-term quality of life**.

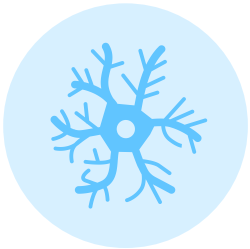
# Introduction



The **early detection of PD** is a **growing priority** within both clinical practice and research.



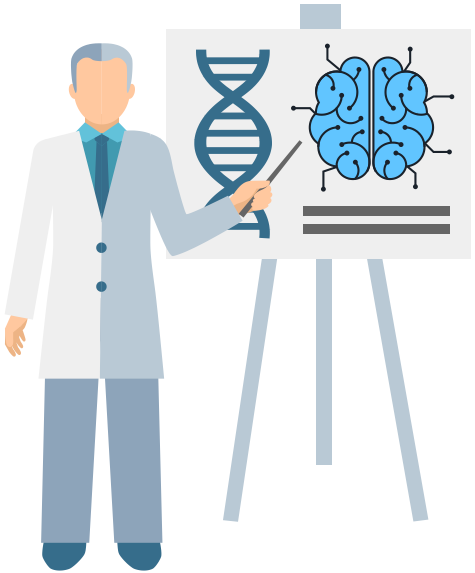
**Many individuals remain undiagnosed until the disease has already progressed**, limiting the effectiveness of available therapeutic interventions. Consequently, **timely diagnosis** can significantly **influence patient outcomes and long-term quality of life**.

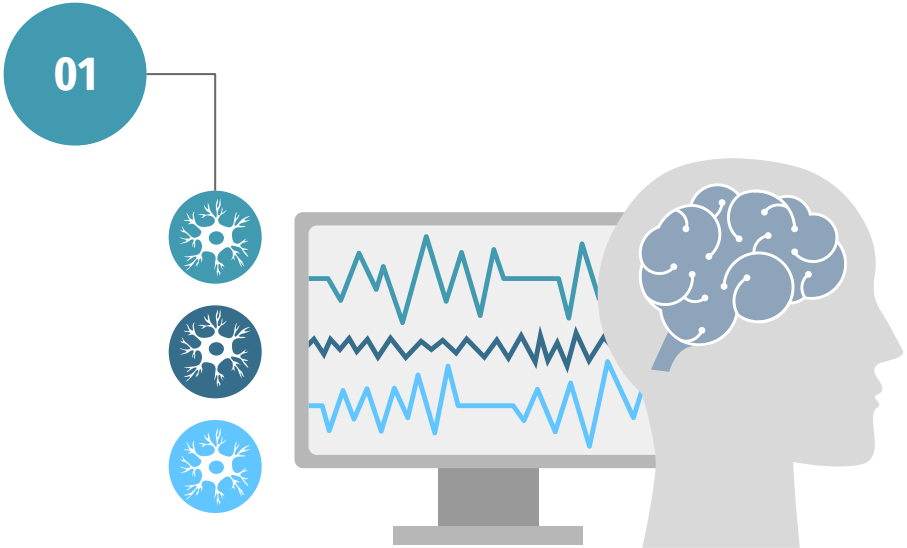


Early identification could allow for **timelier monitoring, lifestyle adjustments, and targeted therapeutic strategies** that may slow **disease progression or improve quality of life**.

# Why develop a predictive model?

A reliable predictive system has the potential to **support clinicians in recognizing subtle signs that might otherwise go unnoticed**.





# Why develop a predictive model?

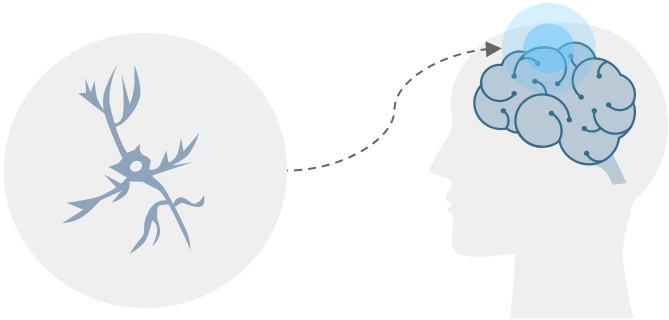
A reliable predictive system has the potential to **support clinicians in recognizing subtle signs that might otherwise go unnoticed**.

01

# Why develop a predictive model?

A reliable predictive system has the potential to **support clinicians in recognizing subtle signs that might otherwise go unnoticed**.

Predictive modeling can help researchers **gain deeper insight into the complex interactions that contribute to the onset of neurodegenerative disorders**



# Why develop a predictive model?

A reliable predictive system has the potential to **support clinicians in recognizing subtle signs that might otherwise go unnoticed**.

Predictive modeling can help researchers **gain deeper insight into the complex interactions that contribute to the onset of neurodegenerative disorders**

# Why develop a predictive model?

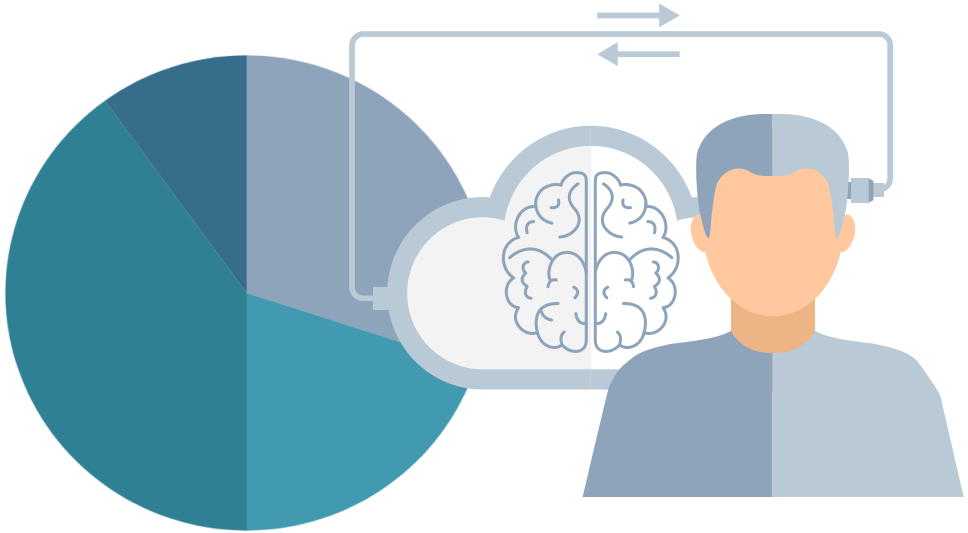
A reliable predictive system has the potential to **support clinicians in recognizing subtle signs that might otherwise go unnoticed**.

Predictive modeling can help researchers **gain deeper insight into the complex interactions that contribute to the onset of neurodegenerative disorders**

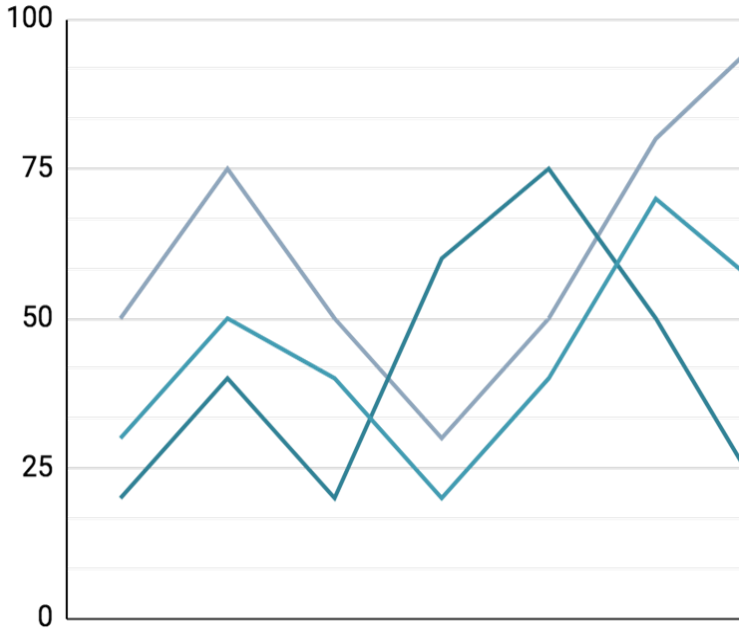
Beyond clinical impact, creating a predictive model encourages the integration of modern **data-driven approaches** into neurological healthcare, which stands out as a promising path to **more personalized and proactive patient care**.

# Objective

**Integrating diverse variables** (demographic, lifestyle, clinical, cognitive, and symptom-related variables) into a unified predictive framework, the project seeks to **evaluate multiple machine learning algorithms and determine their capability to accurately identify patients at risk.**

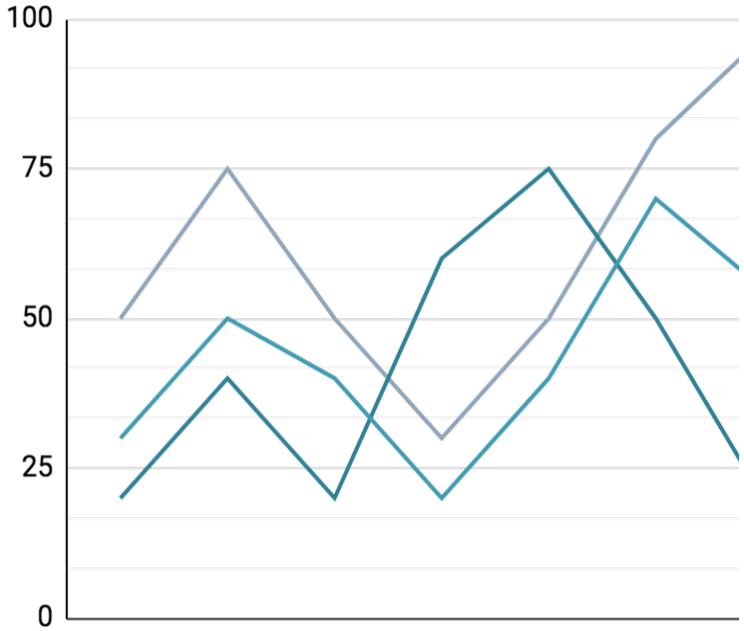


# Data Description and Sources




As part of this project, I selected a synthetic **dataset from Kaggle** generated by Mr. Rabie El Kharoua, to support the development of a predictive model for Parkinson's disease.

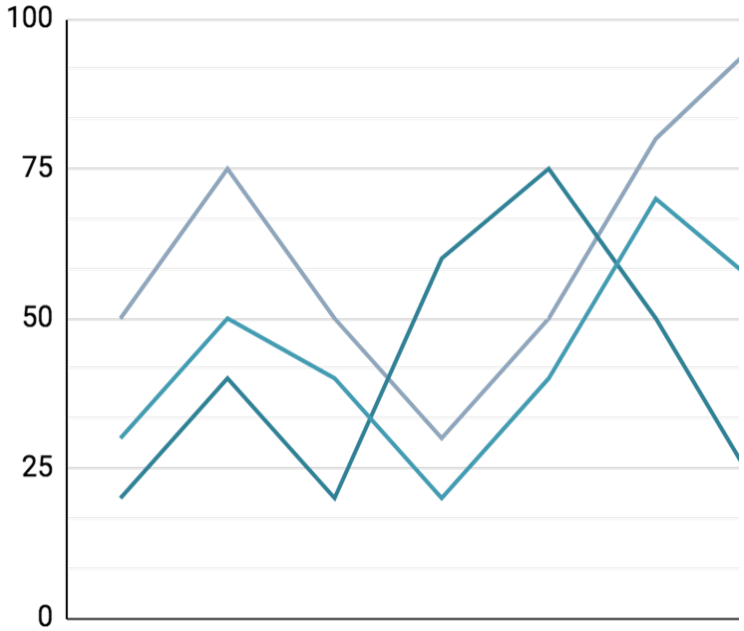
# Data Description and Sources




As part of this project, I selected a synthetic **dataset from Kaggle** generated by Mr. Rabie El Kharoua, to support the development of a predictive model for Parkinson's disease.

This dataset contains diverse features:

# Data Description and Sources

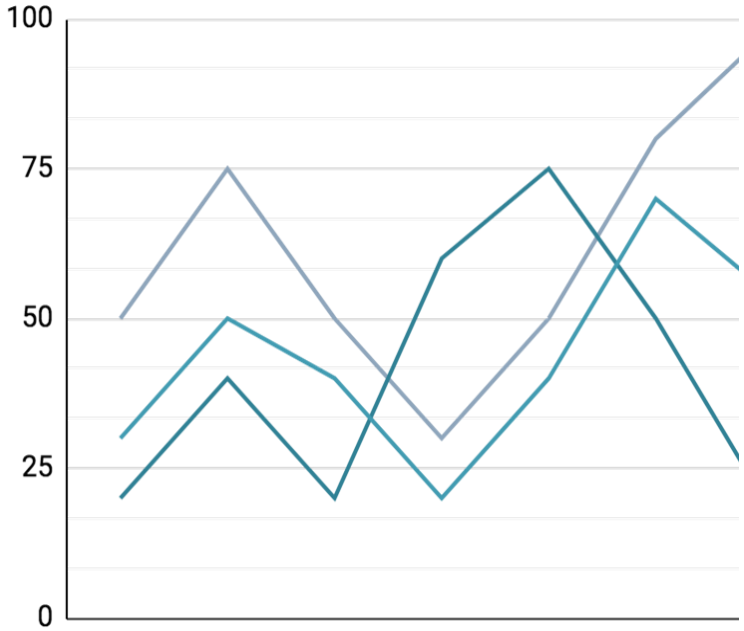



As part of this project, I selected a synthetic **dataset from Kaggle** generated by Mr. Rabie El Kharoua, to support the development of a predictive model for Parkinson's disease.

This dataset contains diverse features:

- **Demographic variables** (age, gender, ethnicity, educational level).

# Data Description and Sources

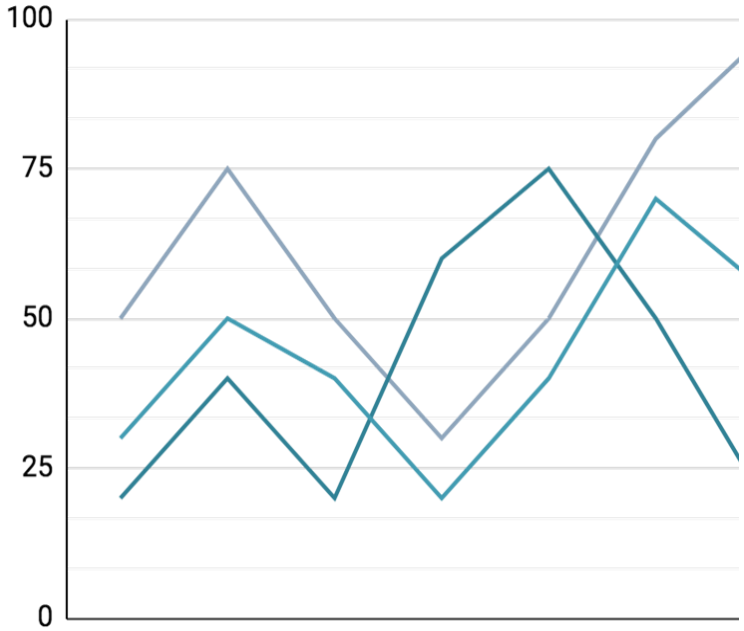



As part of this project, I selected a synthetic **dataset from Kaggle** generated by Mr. Rabie El Kharoua, to support the development of a predictive model for Parkinson's disease.

This dataset contains diverse features:

- **Demographic variables** (age, gender, ethnicity, educational level).
- **Lifestyle Factors** (BMI, smoking, alcohol, physical activity, diet and sleep quality).

# Data Description and Sources

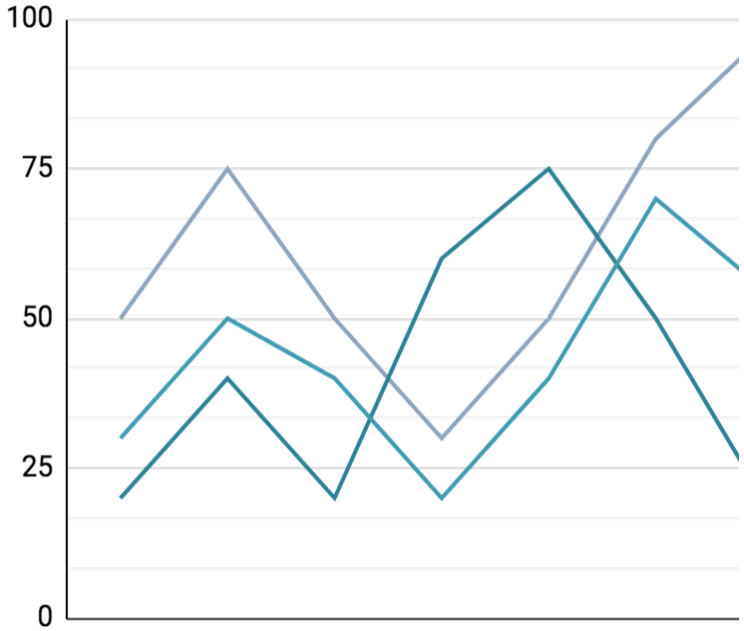



As part of this project, I selected a synthetic **dataset from Kaggle** generated by Mr. Rabie El Kharoua, to support the development of a predictive model for Parkinson's disease.

This dataset contains diverse features:

- **Demographic variables** (age, gender, ethnicity, educational level).
- **Lifestyle Factors** (BMI, smoking, alcohol, physical activity, diet and sleep quality).
- **Medical History** (TBI, Depression, Stroke, Diabetes…)

# Data Description and Sources

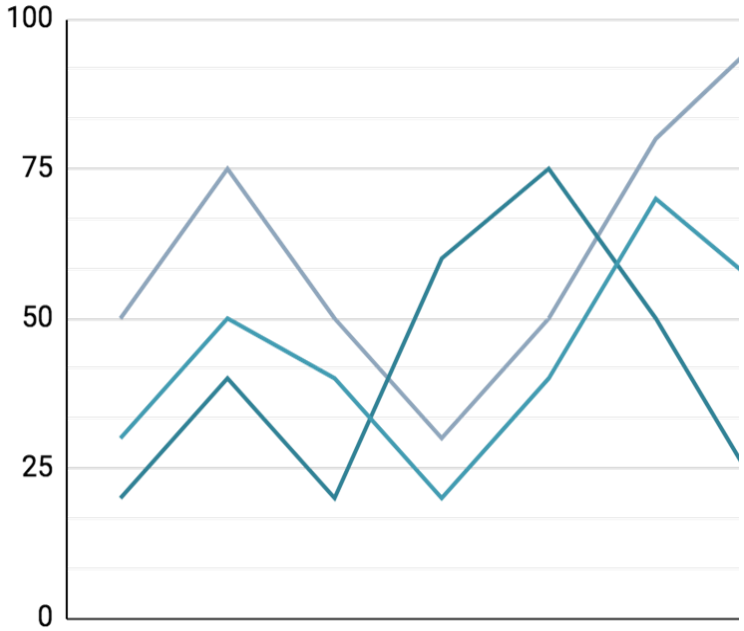



As part of this project, I selected a synthetic **dataset from Kaggle** generated by Mr. Rabie El Kharoua, to support the development of a predictive model for Parkinson's disease.

This dataset contains diverse features:

- **Demographic variables** (age, gender, ethnicity, educational level).
- **Lifestyle Factors** (BMI, smoking, alcohol, physical activity, diet and sleep quality).
- **Medical History** (TBI, Depression, Stroke, Diabetes…)
- **Clinical Meassurements**(Cholesterol, Triglycerides, BP).

# Data Description and Sources

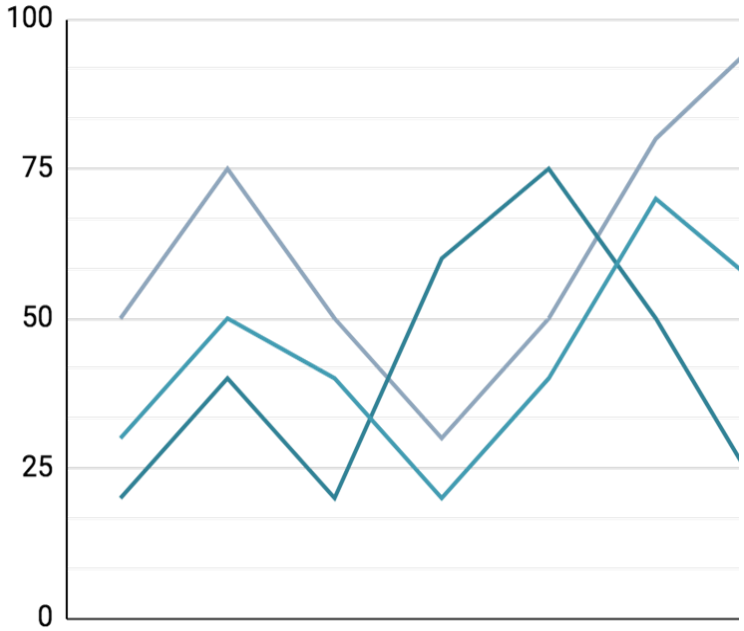



As part of this project, I selected a synthetic **dataset from Kaggle** generated by Mr. Rabie El Kharoua, to support the development of a predictive model for Parkinson's disease.

This dataset contains diverse features:

- **Demographic variables** (age, gender, ethnicity, educational level).
- **Lifestyle Factors** (BMI, smoking, alcohol, physical activity, diet and sleep quality).
- **Medical History** (TBI, Depression, Stroke, Diabetes…)
- **Clinical Meassurements**(Cholesterol, Triglycerides, BP).
- **Cognitive and Functional Assessments** (UPDRS, MoCA…).

# Data Description and Sources




As part of this project, I selected a synthetic **dataset from Kaggle** generated by Mr. Rabie El Kharoua, to support the development of a predictive model for Parkinson's disease.

This dataset contains diverse features:

- **Demographic variables** (age, gender, ethnicity, educational level).
- **Lifestyle Factors** (BMI, smoking, alcohol, physical activity, diet and sleep quality).
- **Medical History** (TBI, Depression, Stroke, Diabetes…)
- **Clinical Meassurements**(Cholesterol, Triglycerides, BP).
- **Cognitive and Functional Assessments** (UPDRS, MoCA…).
- **Symptom indicators** (Tremor, Constipation, Rigidity).

# Methodology

01

## Standarization/Scaling

StandarScaler
MinMaxScaler or None

# Methodology

01

**Standarization/Scaling**

StandarScaler
MinMaxScaler or None

02

**Imbalance Correction**

SMOTENC or None

# Methodology

01

**Standarization/Scaling**

StandarScaler
MinMaxScaler or None

02

**Imbalance Correction**

SMOTENC or None

03

**Feature Selection or Dimensionality Reduction**

SelectKBest, RFECV, LinearDiscriminantAnalysis or None

# Methodology

01

**Standarization/Scaling**

StandarScaler
MinMaxScaler or None

02

**Imbalance Correction**

SMOTENC or None

03

**Feature Selection or Dimensionality Reduction**

SelectKBest, RFECV, LinearDiscriminantAnalysis or None

04

**Classification Models**

LogisticRegression, QuadraticDiscriminantAnalysis, LinearSVC, SVC, KNeighborsClassifier, XGBClassifier, LGBMClassifier, RandomForestClassifier, ExtraTreesClassifier, StackingClassifier, VotingClassifier

# Methodology

01

## Standarization/Scaling

StandarScaler
MinMaxScaler or None

02

## Imbalance Correction

SMOTENC or None

03

## Feature Selection or Dimensionality Reduction

SelectKBest, RFECV,
LinearDiscriminantAnalysis or None

04

## Classification Models

LogisticRegression, QuadraticDiscriminantAnalysis,
LinearSVC, SVC, KNeighborsClassifier,
XGBClassifier, LGBMClassifier,
RandomForestClassifier, ExtraTreesClassifier,
StackingClassifier, VotingClassifier

05

## Hiperparameter Fine-Tuning and Cross-Validation

GridSearchCV and
RandomizedSearchCV

# Methodology

01

## Standarization/Scaling

StandarScaler
MinMaxScaler or None

02

## Imbalance Correction

SMOTENC or None

03

## Feature Selection or Dimensionality Reduction

SelectKBest, RFECV,
LinearDiscriminantAnalysis or None

04

## Classification Models

LogisticRegression, QuadraticDiscriminantAnalysis,
LinearSVC, SVC, KNeighborsClassifier,
XGBClassifier, LGBMClassifier,
RandomForestClassifier, ExtraTreesClassifier,
StackingClassifier, VotingClassifier

05

## Hiperparameter Fine-Tuning and Cross-Validation

GridSearchCV and
RandomizedSearchCV

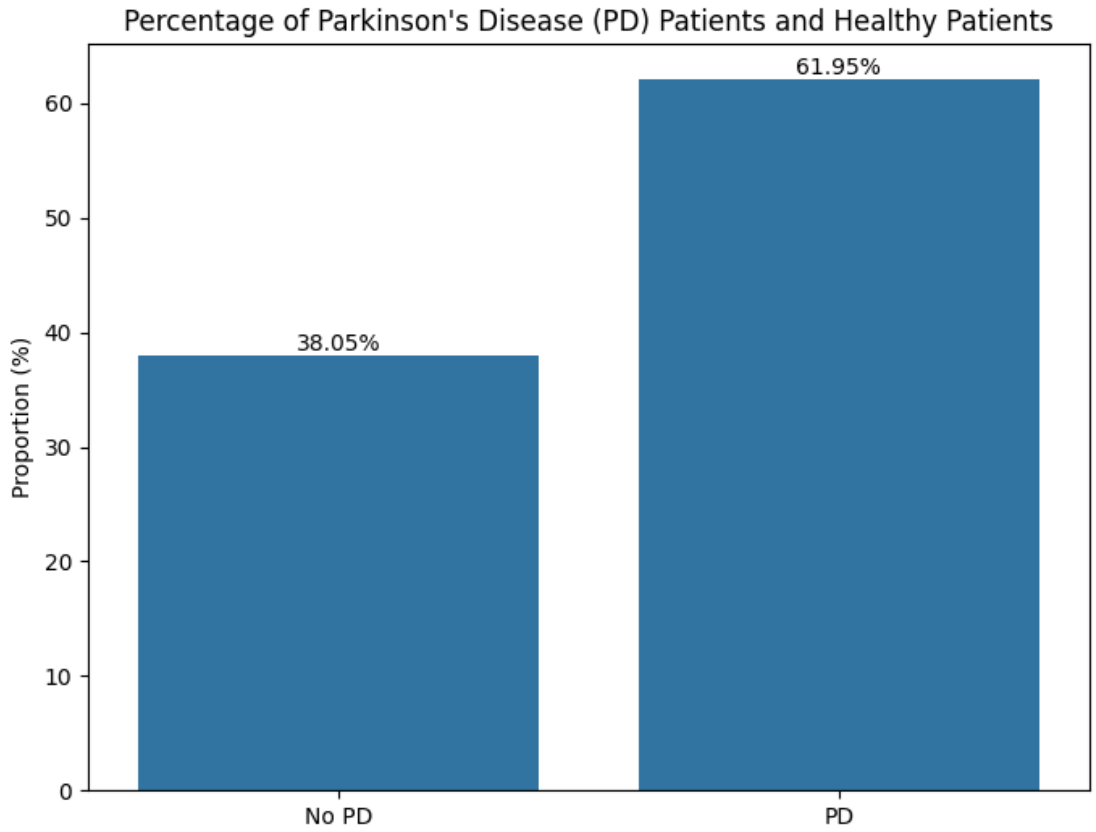
06

## Model Testing and Saving

sklearn.metrics and
pickle

# Exploratory Data Analysis

# Exploratory Data Analysis



# Exploratory Data Analysis

# Exploratory Data Analysis



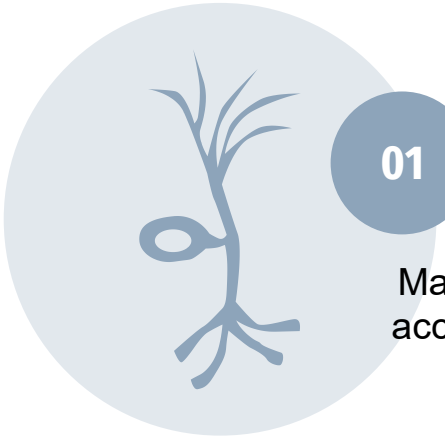

| | Diagnosis |
|---|---|
| UPDRS | 0.411858 |
| Tremor | 0.271641 |
| Bradykinesia | 0.183083 |
| Rigidity | 0.167933 |
| PosturalInstability | 0.159615 |
| Depression | 0.072315 |
| Age | 0.072114 |
| Diabetes | 0.056443 |
| AlcoholConsumption | 0.041170 |
| Stroke | 0.036873 |
| PhysicalActivity | 0.034081 |
| TraumaticBrainInjury | 0.033186 |
| CholesterolTriglycerides | 0.025269 |
| FamilyHistoryParkinsons | 0.024888 |
| BMI | 0.024799 |
| Gender | 0.019451 |
| Smoking | 0.017677 |
| CholesterolLDL | 0.009858 |
| Ethnicity | 0.002865 |
| EducationLevel | -0.003334 |
| SpeechProblems | -0.004429 |
| SleepDisorders | -0.010265 |
| SystolicBP | -0.013242 |
| CholesterolTotal | -0.015405 |
| Hypertension | -0.017471 |
| DiastolicBP | -0.026635 |
| DietQuality | -0.029933 |
| SleepQuality | -0.057237 |
| CholesterolHDL | -0.057320 |
| MoCA | -0.179093 |
| FunctionalAssessment | -0.217524 |

# RESULTS: Final Models

In this project **two reliable predictive models with subtle differences** were developed

# RESULTS: Final Models

In this project **two reliable predictive models with subtle differences** were developed



01

## MODEL 1

Maximizes the accuracy score

# RESULTS: Final Models

In this project **two reliable predictive models with subtle differences** were developed

01 **MODEL 1**

Maximizes the accuracy score

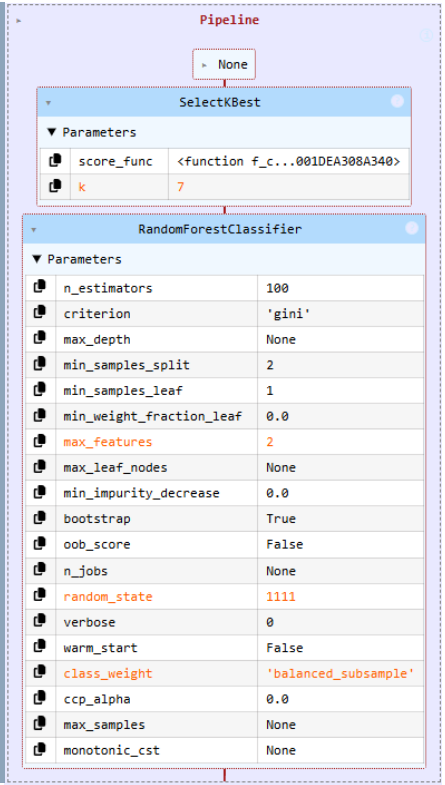
02 **MODEL 2**

Maximazes the sensitivity score

# 01 MODEL 1: Pipeline and Hyperparameters

# 01 MODEL 1: Pipeline and Hyperparameters

Random Forest Classifier

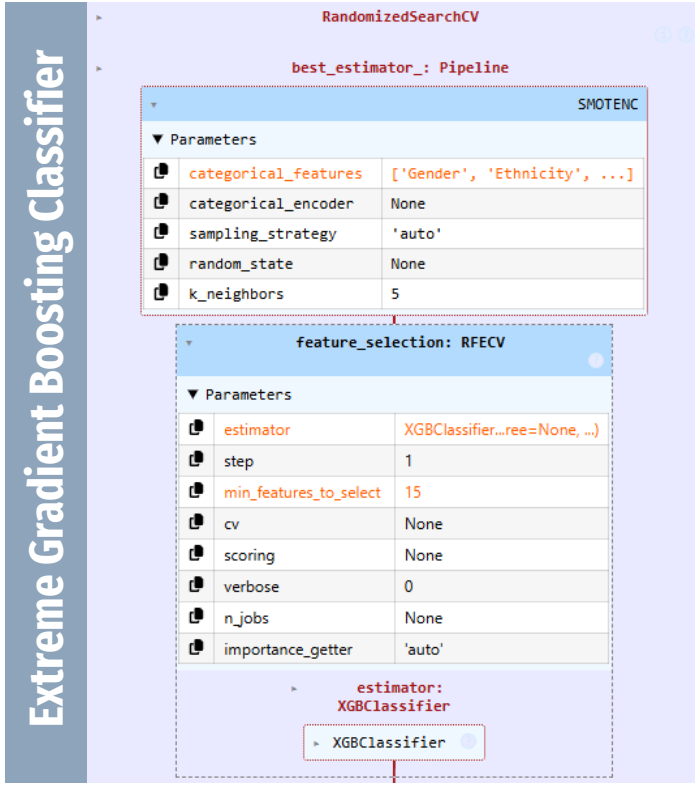


# 01 MODEL 1: Pipeline and Hyperparameters

## Random Forest Classifier




## Extreme Gradient Boosting Classifier



# 01 MODEL 1: Pipeline and Hyperparameters

## Random Forest Classifier

Pipeline

None

SelectKBest

Parameters

| | |
|---|---|
| score_func | <function f_c...001DEA308A340> |
| k | 7 |

RandomForestClassifier

Parameters

| | |
|---|---|
| n_estimators | 100 |
| criterion | 'gini' |
| max_depth | None |
| min_samples_split | 2 |
| min_samples_leaf | 1 |
| min_weight_fraction_leaf | 0.0 |
| max_features | 2 |
| max_leaf_nodes | None |
| min_impurity_decrease | 0.0 |
| bootstrap | True |
| oob_score | False |
| n_jobs | None |
| random_state | 1111 |
| verbose | 0 |
| warm_start | False |
| class_weight | 'balanced_subsample' |
| ccp_alpha | 0.0 |
| max_samples | None |
| monotonic_cst | None |

## Extreme Gradient Boosting Classifier

RandomizedSearchCV

best_estimator_: Pipeline

SMOTENC

Parameters

| | |
|---|---|
| categorical_features | ['Gender', 'Ethnicity', ...] |
| categorical_encoder | None |
| sampling_strategy | 'auto' |
| random_state | None |
| k_neighbors | 5 |

feature_selection: RFECV

Parameters

| | |
|---|---|
| estimator | XGBClassifier...ree=None, ...) |
| step | 1 |
| min_features_to_select | 15 |
| cv | None |
| scoring | None |
| verbose | 0 |
| n_jobs | None |
| importance_getter | 'auto' |

estimator: XGBClassifier

XGBClassifier

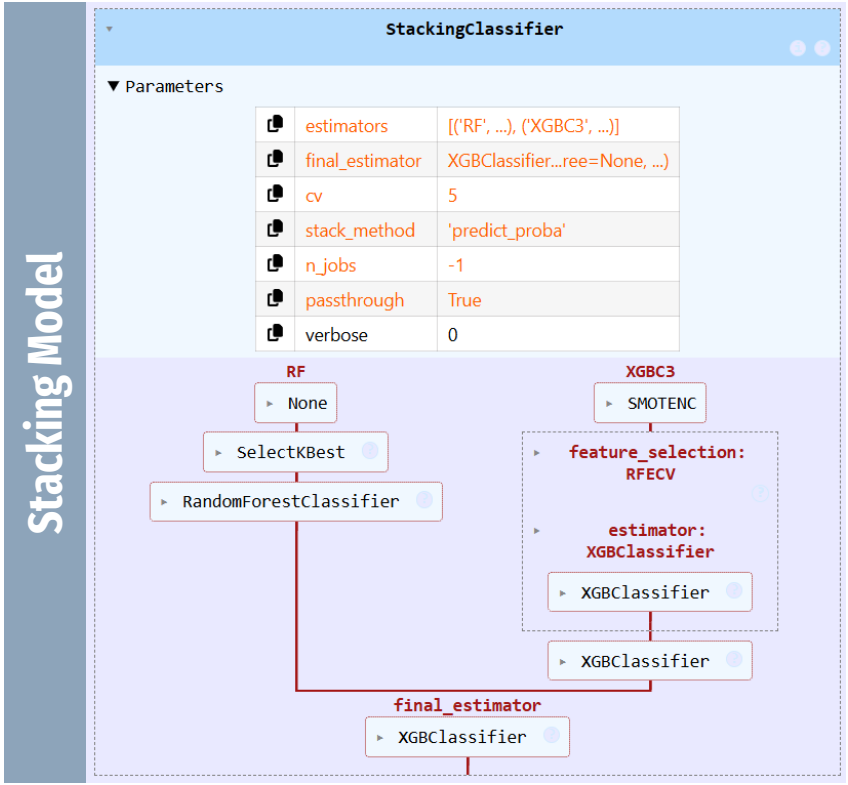
XGBClassifier

Parameters

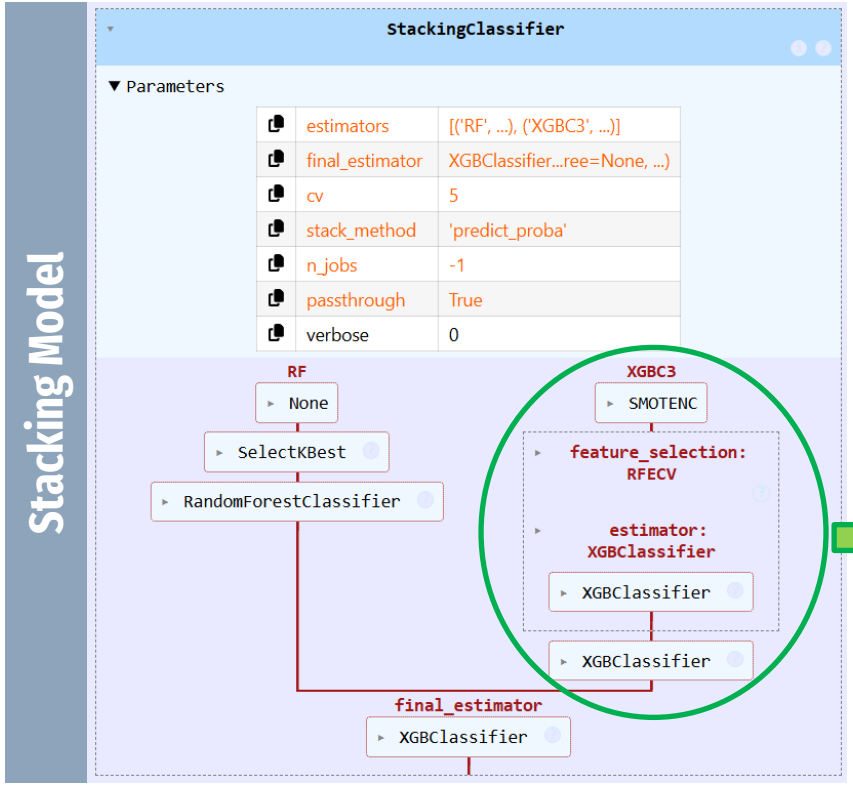
| | |
|---|---|
| objective | 'binary:logistic' |
| base_score | None |
| booster | None |
| callbacks | None |
| colsample_bylevel | None |
| colsample_bynode | None |
| colsample_bytree | 0.7 |
| device | None |
| early_stopping_rounds | None |
| enable_categorical | False |
| eval_metric | None |
| feature_types | None |
| feature_weights | None |
| gamma | None |
| grow_policy | None |
| importance_type | None |
| interaction_constraints | None |
| learning_rate | np.float64(0....7334556031603) |
| max_bin | None |
| max_cat_threshold | None |
| max_cat_to_onehot | None |
| max_delta_step | None |
| max_depth | np.int64(3) |
| max_leaves | None |
| min_child_weight | None |
| missing | nan |
| monotone_constraints | None |
| multi_strategy | None |
| n_estimators | 100 |
| n_jobs | None |
| num_parallel_tree | None |
| random_state | 1111 |
| reg_alpha | None |
| reg_lambda | None |
| sampling_method | None |
| scale_pos_weight | None |
| subsample | 0.9 |
| tree_method | None |
| validate_parameters | None |
| verbosity | None |

# 01 MODEL 1: Pipeline and Hyperparameters

# 01 MODEL 1: Pipeline and Hyperparameters



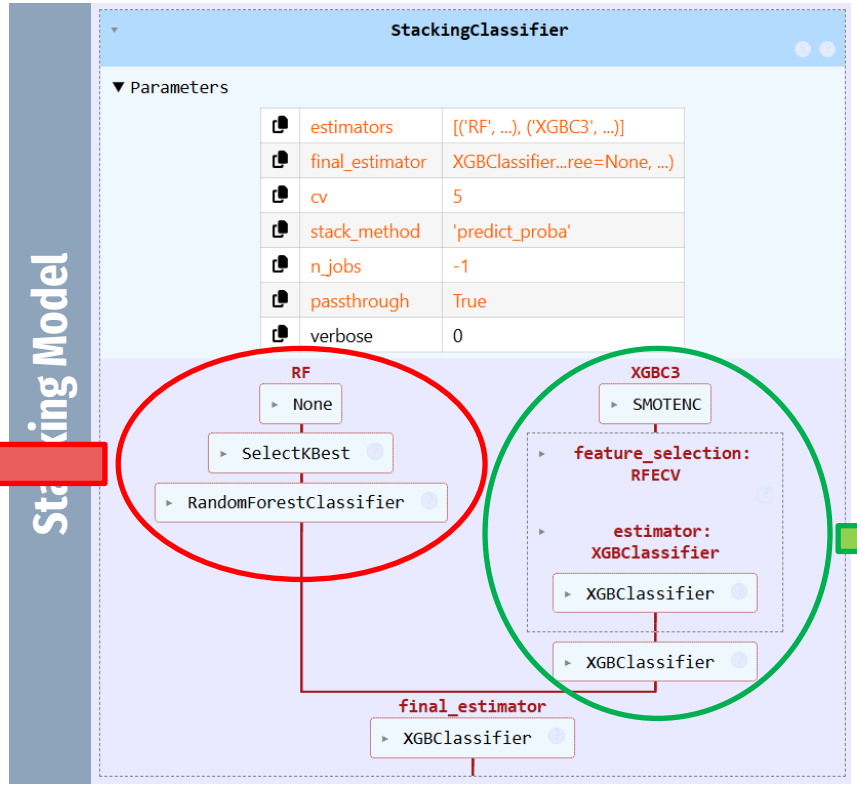
# 01 MODEL 1: Pipeline and Hyperparameters




**Greater Accuracy and Specificity Scores**

# 01 MODEL 1: Pipeline and Hyperparameters

**Greater Recall/Sensitivity Scores**

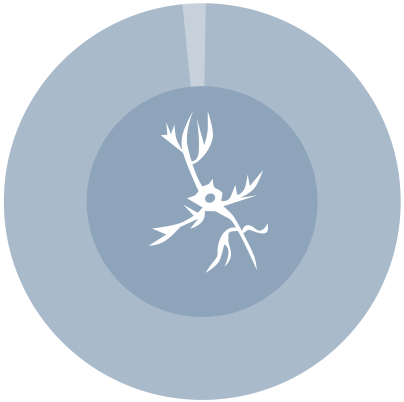



**Greater Accuracy and Specificity Scores**

# 01 MODEL 1: Test Results

01

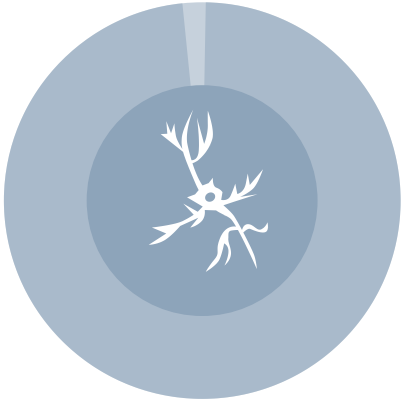
# MODEL 1: Test Results



**96.9% Accuracy**

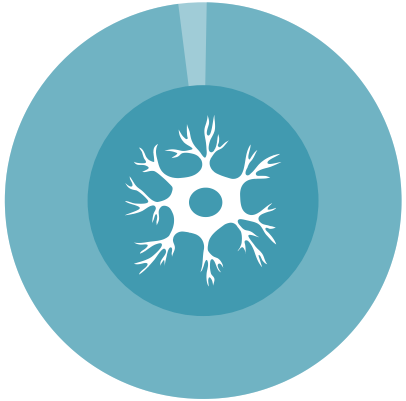
This model predicted correctly 408 of 421 patients' diagnosis

# 01 MODEL 1: Test Results



**96.9% Accuracy**

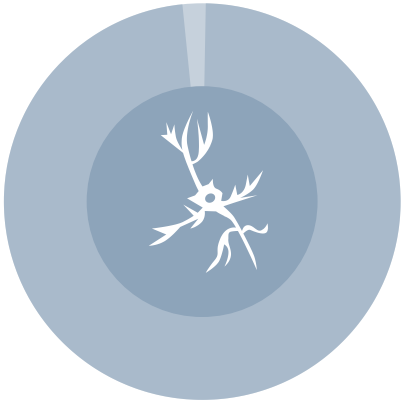
This model predicted correctly 408 of 421 patients' diagnosis



**95.7% Specificity**

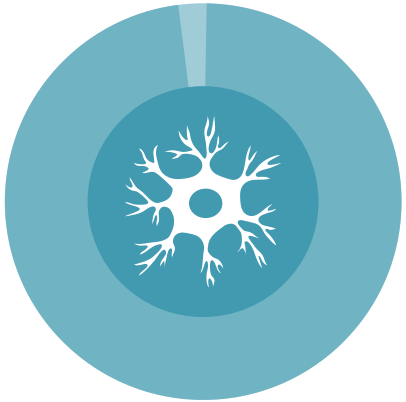
155 of 162 of the heatlhy patients were diagnosed as healthy.

# 01 MODEL 1: Test Results



**96.9% Accuracy**

This model predicted correctly 408 of 421 patients' diagnosis



**95.7% Specificity**

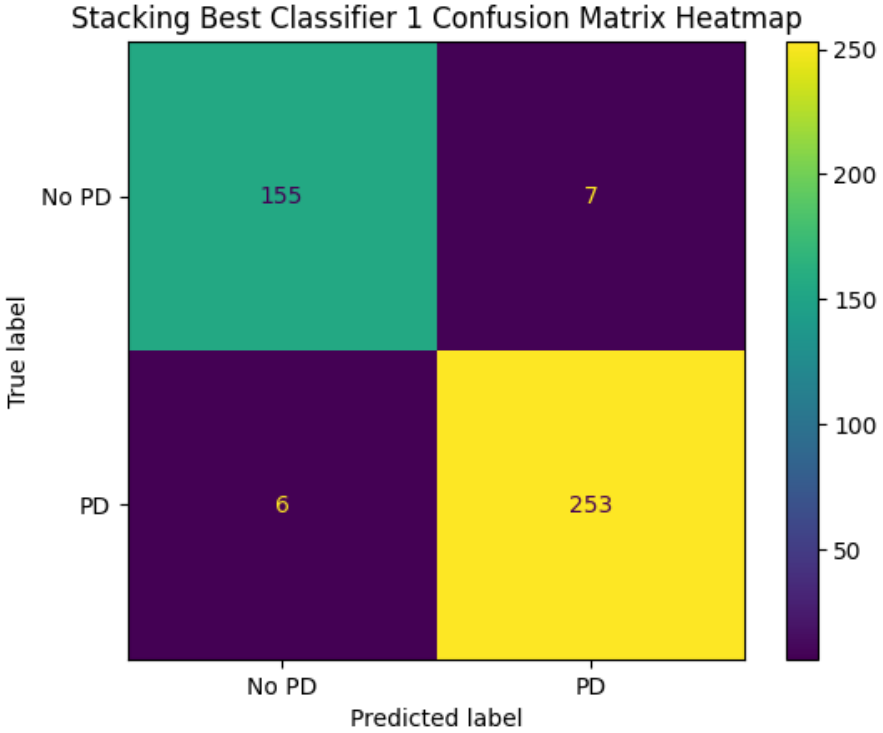
155 of 162 of the heatlhy patients were diagnosed as healthy.

**97.7% Sensitivity**

253 of 259 Parkinson's Disease patients were detected

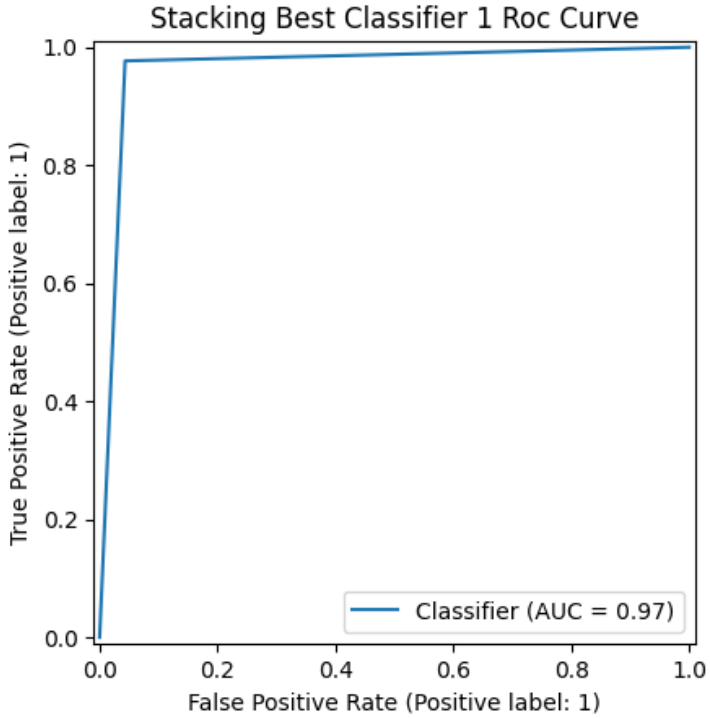
# 01 MODEL 1: Test Results

# 01 MODEL 1: Test Results



# 01 MODEL 1: Test Results



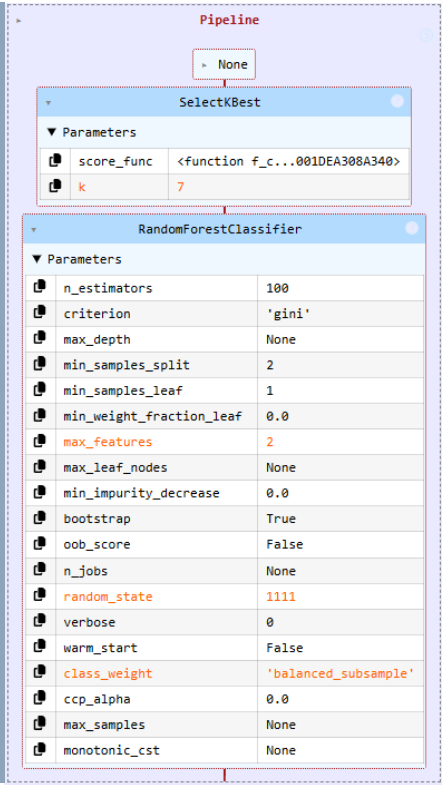



02

# MODEL 2: Pipeline and Hyperparameters

# 02 MODEL 2: Pipeline and Hyperparameters

Random Forest Classifier



# 02 MODEL 2: Pipeline and Hyperparameters

## Random Forest Classifier

Pipeline

None

SelectKBest

▼ Parameters

| | |
|---|---|
| score_func | <function f_c...001DEA308A340> |
| k | 7 |

RandomForestClassifier

▼ Parameters

| | |
|---|---|
| n_estimators | 100 |
| criterion | 'gini' |
| max_depth | None |
| min_samples_split | 2 |
| min_samples_leaf | 1 |
| min_weight_fraction_leaf | 0.0 |
| max_features | 2 |
| max_leaf_nodes | None |
| min_impurity_decrease | 0.0 |
| bootstrap | True |
| oob_score | False |
| n_jobs | None |
| random_state | 1111 |
| verbose | 0 |
| warm_start | False |
| class_weight | 'balanced_subsample' |
| ccp_alpha | 0.0 |
| max_samples | None |
| monotonic_cst | None |

## Extreme Gradient Boosting Classifier

RandomizedSearchCV

best_estimator_: Pipeline

SMOTENC

▼ Parameters

| | |
|---|---|
| categorical_features | ['Gender', 'Ethnicity', ...] |
| categorical_encoder | None |
| sampling_strategy | 'auto' |
| random_state | None |
| k_neighbors | 5 |

feature_selection: RFECV

▼ Parameters

| | |
|---|---|
| estimator | XGBClassifier...ree=None, ...) |
| step | 1 |
| min_features_to_select | 15 |
| cv | None |
| scoring | None |
| verbose | 0 |
| n_jobs | None |
| importance_getter | 'auto' |

estimator: XGBClassifier

XGBClassifier

# 02 MODEL 2: Pipeline and Hyperparameters

## Random Forest Classifier

Pipeline

None

SelectKBest

Parameters

| | |
|---|---|
| score_func | <function f_c...001DEA308A340> |
| k | 7 |

RandomForestClassifier

Parameters

| | |
|---|---|
| n_estimators | 100 |
| criterion | 'gini' |
| max_depth | None |
| min_samples_split | 2 |
| min_samples_leaf | 1 |
| min_weight_fraction_leaf | 0.0 |
| max_features | 2 |
| max_leaf_nodes | None |
| min_impurity_decrease | 0.0 |
| bootstrap | True |
| oob_score | False |
| n_jobs | None |
| random_state | 1111 |
| verbose | 0 |
| warm_start | False |
| class_weight | 'balanced_subsample' |
| ccp_alpha | 0.0 |
| max_samples | None |
| monotonic_cst | None |

## Extreme Gradient Boosting Classifier

RandomizedSearchCV

best_estimator_: Pipeline

SMOTENC

Parameters

| | |
|---|---|
| categorical_features | ['Gender', 'Ethnicity', ...] |
| categorical_encoder | None |
| sampling_strategy | 'auto' |
| random_state | None |
| k_neighbors | 5 |

feature_selection: RFECV

Parameters

| | |
|---|---|
| estimator | XGBClassifier...ree=None, ...) |
| step | 1 |
| min_features_to_select | 15 |
| cv | None |
| scoring | None |
| verbose | 0 |
| n_jobs | None |
| importance_getter | 'auto' |

estimator: XGBClassifier

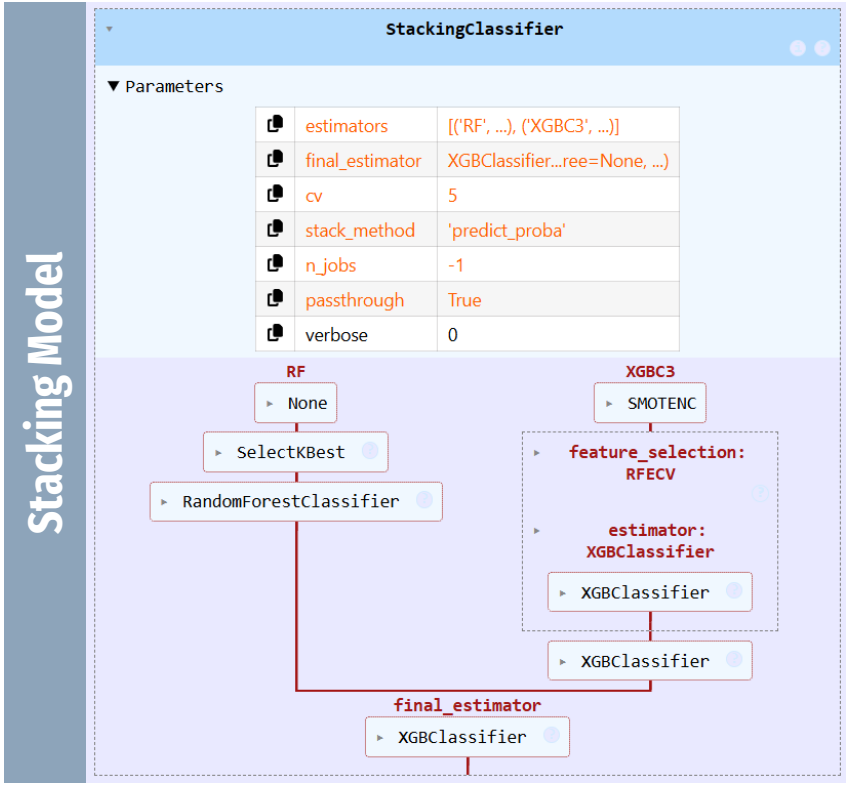
XGBClassifier

XGBClassifier

Parameters

| | |
|---|---|
| objective | 'binary:logistic' |
| base_score | None |
| booster | None |
| callbacks | None |
| colsample_bylevel | None |
| colsample_bynode | None |
| colsample_bytree | 0.7 |
| device | None |
| early_stopping_rounds | None |
| enable_categorical | False |
| eval_metric | None |
| feature_types | None |
| feature_weights | None |
| gamma | None |
| grow_policy | None |
| importance_type | None |
| interaction_constraints | None |
| learning_rate | np.float64(0....7334556031603) |
| max_bin | None |
| max_cat_threshold | None |
| max_cat_to_onehot | None |
| max_delta_step | None |
| max_depth | np.int64(3) |
| max_leaves | None |
| min_child_weight | None |
| missing | nan |
| monotone_constraints | None |
| multi_strategy | None |
| n_estimators | 100 |
| n_jobs | None |
| num_parallel_tree | None |
| random_state | 1111 |
| reg_alpha | None |
| reg_lambda | None |
| sampling_method | None |
| scale_pos_weight | None |
| subsample | 0.9 |
| tree_method | None |
| validate_parameters | None |
| verbosity | None |

# 02 MODEL 2: Pipeline and Hyperparameters

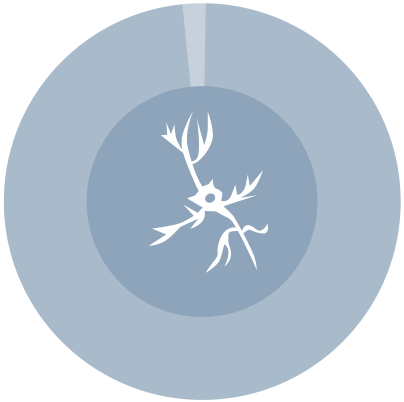


02

# MODEL 2: Test Results

02

# MODEL 2: Test Results

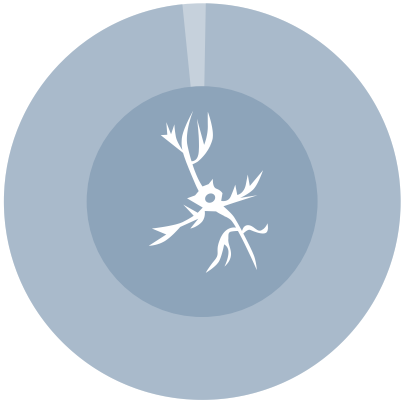


**96.4% Accuracy**

This model predicted correctly 406 of 421 patients' diagnosis

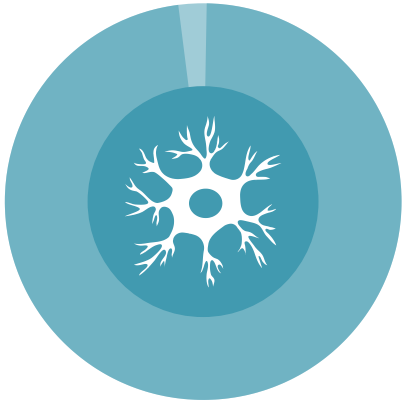
02

# MODEL 2: Test Results



**96.4% Accuracy**

This model predicted correctly 406 of 421 patients' diagnosis

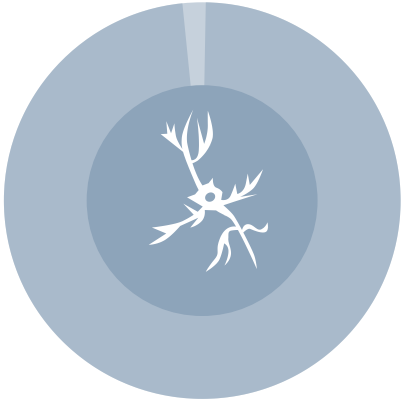


**93.8% Specificity**

152 of 162 of the heatlhy patients were diagnosed as healthy.

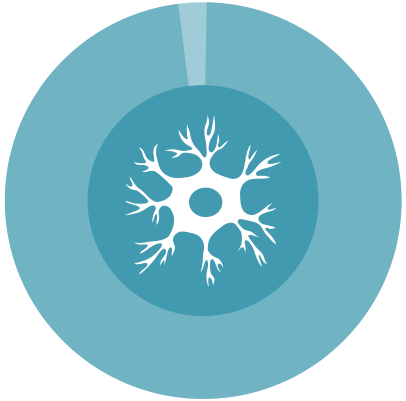
02

# MODEL 2: Test Results



**96.4% Accuracy**

This model predicted correctly 406 of 421 patients' diagnosis



**93.8% Specificity**

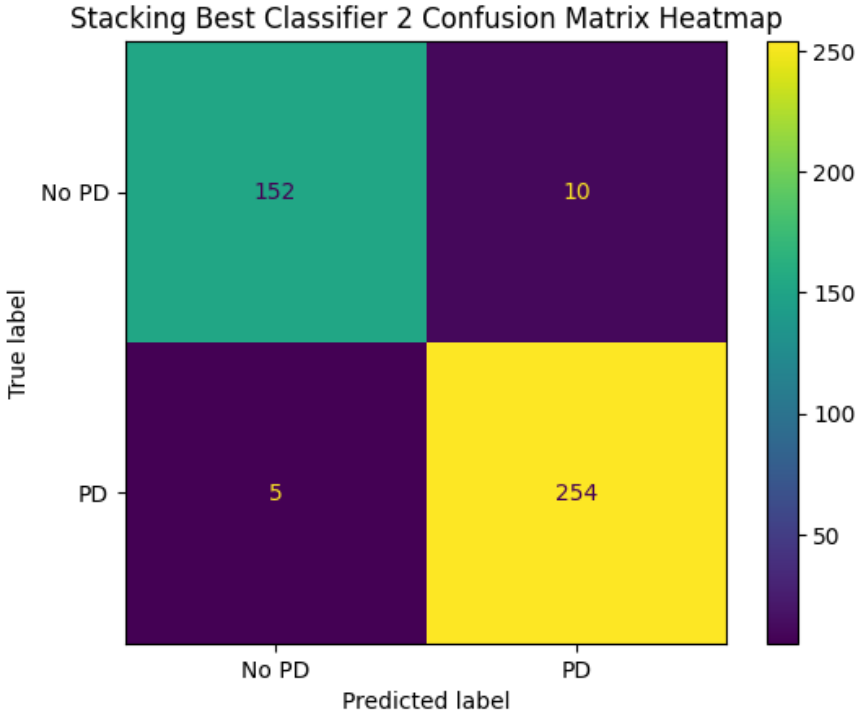
152 of 162 of the heatlhy patients were diagnosed as healthy.

**98.1% Sensitivity**

254 of 259 Parkinson's Disease patients were detected

02

# MODEL 2: Test Results

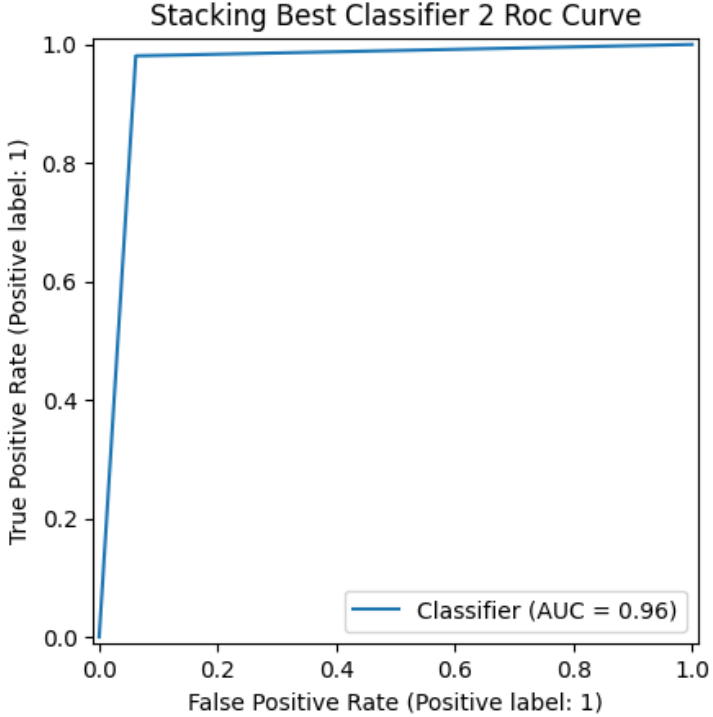


02

# MODEL 2: Test Results

# 02 MODEL 2: Test Results



# Conclusions

# Conclusions

Two machine learning **predictive models with impresive performance were developed sucessfully.**

# Conclusions

Two machine learning **predictive models with impresive performance were developed sucessfully.**

These models can support clinicians in **recognizing subtle signs** that might otherwise go unnoticed, **gain deeper insight** into the **complex interactions that contribute to neurodegenerative disorders** and reach **more personalized and proactive patient care**.

# Conclusions

Two machine learning **predictive models with impresive performance were developed sucessfully.**

These models can support clinicians in **recognizing subtle signs** that might otherwise go unnoticed, **gain deeper insight** into the **complex interactions that contribute to neurodegenerative disorders** and reach **more personalized and proactive patient care**.

**Timely diagnosis** will improve the **effectiveness of available therapeutic interventions**, **influencing patient outcomes and long-term quality of life**.

# Strenghs and Limitations

## Stenghs

- Predictive models obtained impresive sensitivity scores (~98%).
- Components of the model are scalable and fast-performing, allowing fast prediction of huge datasets in seconds.
- Models has great flexibility. Consequently it can be adjusted and re-trained with new data with training times of less than one minute in train dataset of 2000 registers.

# Strenghs and Limitations

## Stenghs

- Predictive models obtained impresive sensitivity scores (~98%).
- Components of the model are scalable and fast-performing, allowing fast prediction of huge datasets in seconds.
- Models has great flexibility. Consequently it can be adjusted and re-trained with new data with training times of less than one minute in train dataset of 2000 registers.

## Limitations

- Although the models obtained impresive sensitivity scores (~98%), specificity was compromised (93-96%), leading to some healthy patients incorrectly diagnosed as Parkinson's Disease patients.
- The model has not been tested in datasets with missing values.

THE BRIDGE

# THANKS FOR YOUR ATTENTION

Github Repository

Streamlit Webpage