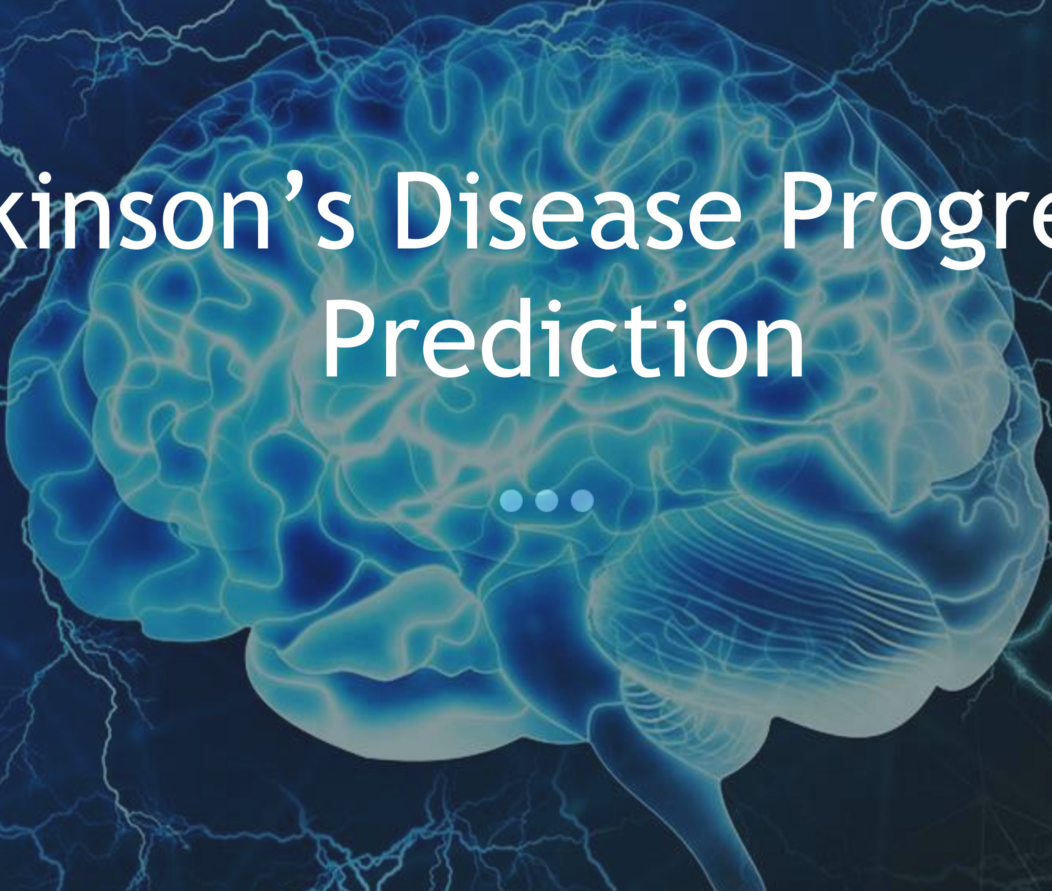
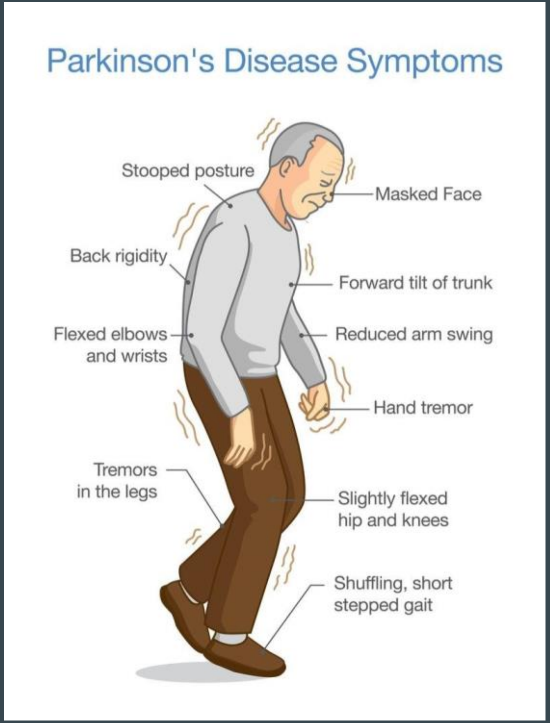
Parkinson’s Disease Progression Prediction



Devansh Singh
Anirudh Gupta

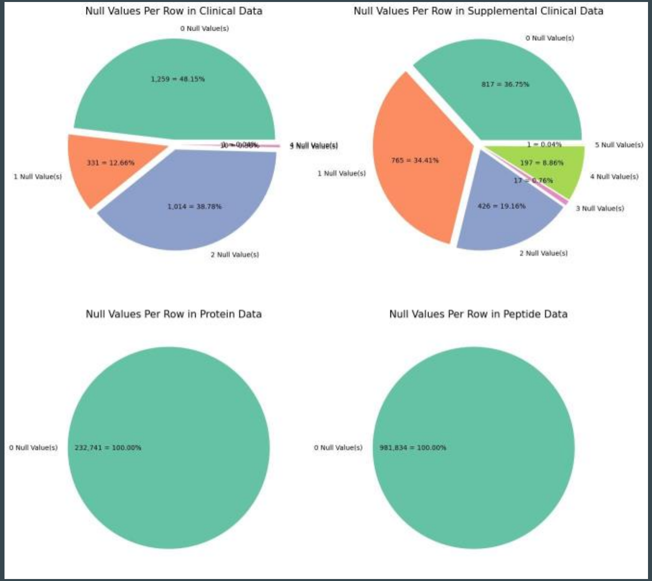
Motivation

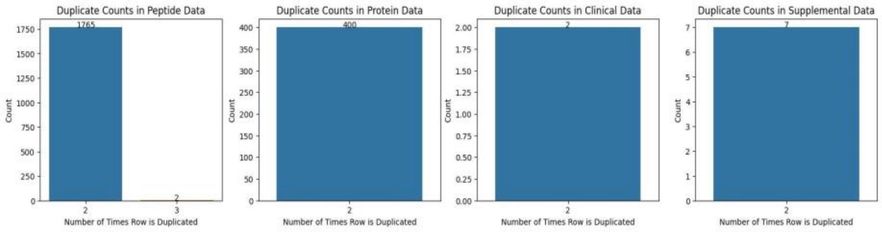
- Parkinson's Disease is a worsening brain disorder with no cure, impacting movement and cognition; protein/peptide abnormalities as key factors.
- Growing prevalence & cost: 1.6 million U.S. cases by 2037 with an $80 billion economic cost.
- Data science potential: Analyzing data from 10,000+ subjects to improve understanding and develop pharmacotherapies.
- Project goal: Identity diagnostic, prognostic, and disease progression biomarkers using MDS-UPDR scores and a model trained on protein/peptide levels.
- Impact: Alleviate patient suffering, reduce medical costs, and provide breakthrough information on molecular changes during PD progression.



Dataset

- The dataset aims to predict Parkinson's disease (PD) progression using protein abundance data obtained from mass spectrometry readings of cerebrospinal fluid (CSF) samples collected from patients over multiple years.
- The dataset includes files: peptides.csv (peptide-level mass spectrometry data), proteins.csv (protein expression frequencies), clinical_data.csv (clinical data including UPDRS scores), and supplemental_clinical_data.csv (additional clinical records without CSF samples).
- Key data attributes include visit_id, visit_month, patient_id, UniProt, Peptide, Peptide Abundance, NPX, updrs_[1-4], and upd23b_clinical_state_on_medication.

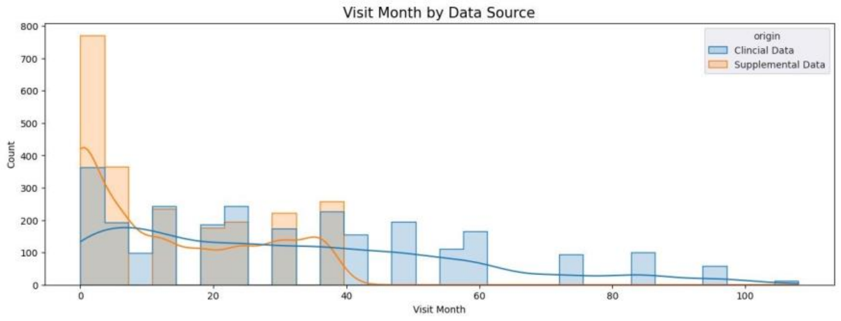


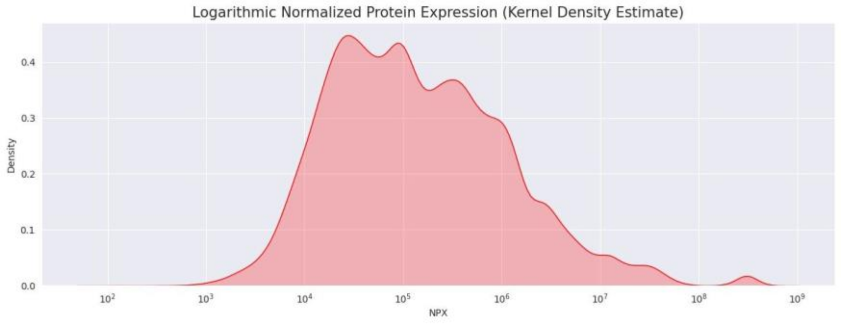


Statistical Analysis

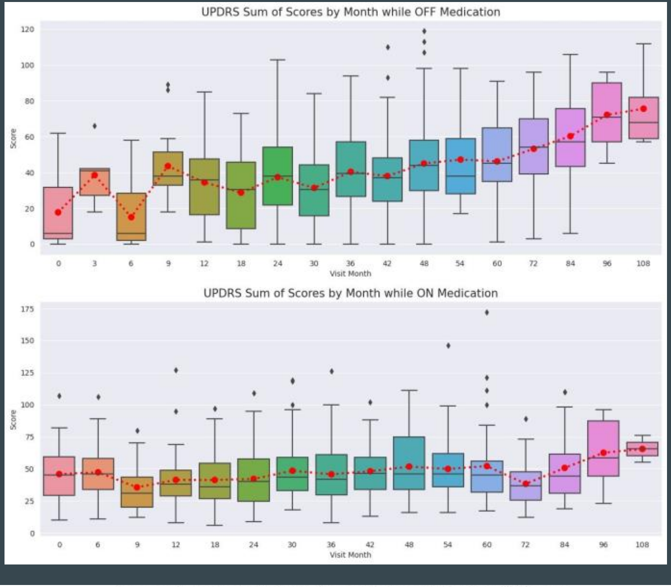
- The supplemental data is concentrated around 0-month visits and extends up to 36 months, while clinical data covers a longer time frame.

- There is significant variability in protein expression frequencies, with normalized protein expression showing high variability as evident from its min, max, and standard deviation values. Further analysis on the distribution of proteins and their relationship to UPDRS scores will be discussed in section 2.





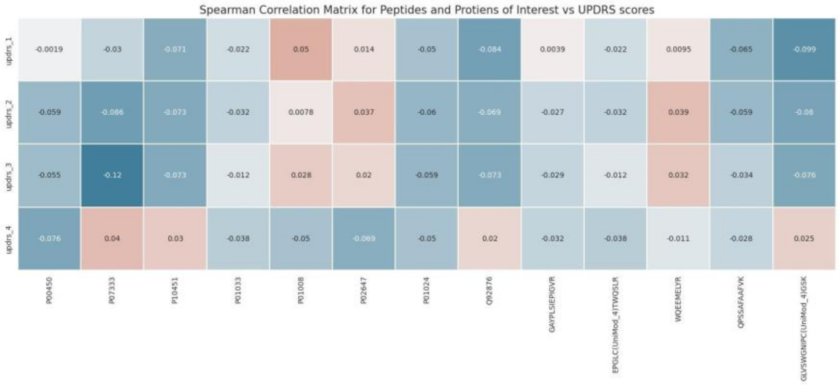
Statistical Analysis


UPDRS Sum of Scores by Month while OFF Medication

UPDRS Sum of Scores by Month while ON Medication

- The overall UPDRS scores show an upward trend when patients are OFF medication, indicating disease progression. The trend remains relatively flat while ON medication until months > 96, where the scores increase, suggesting disease progression as well.

- Only protein P07333 and peptide GLVSWGNIPC(UniMod_4)GSK exhibit weak negative correlations with updrs_3 and updrs_1, respectively. No single protein or peptide demonstrates a clear correlation to UPDRS scores.

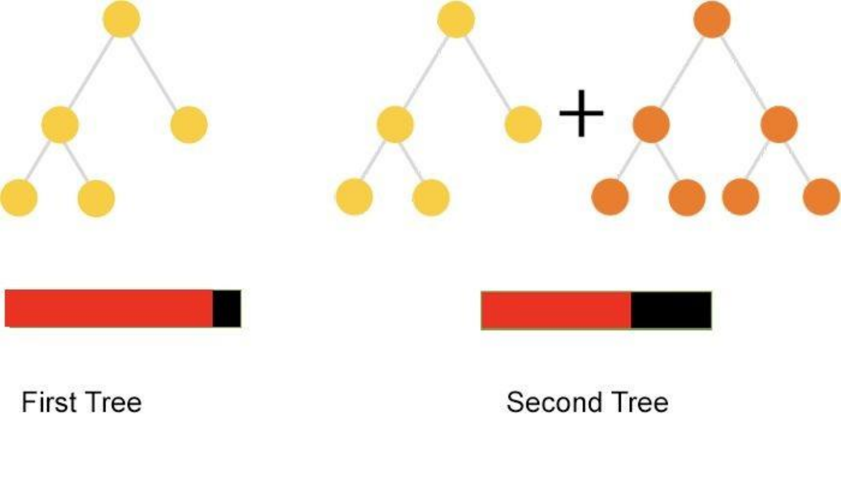

Spearman Correlation Matrix for Peptides and Protiens of Interest vs UPDRS scores

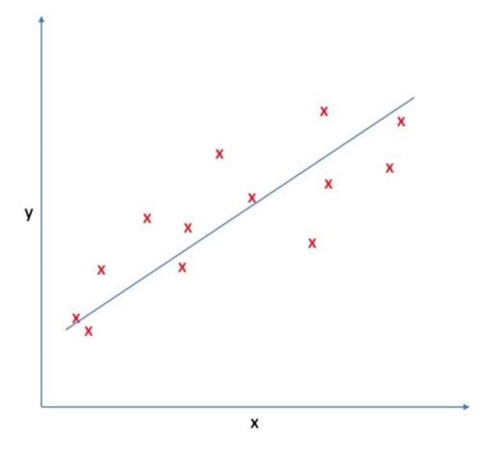
	P00450	P07333	P10451	P01033	P01008	P02647	P01024	Q92876	GAYPLSIEPIGVR	EPGLC(UniMod_4)TWQSLR	WQEEMELYR	QPSSAFAAFVK	GLVSWGNIPC(UniMod_4)GSK
updrs_1	-0.0019	-0.03	-0.071	-0.022	0.05	0.014	-0.05	-0.084	0.0039	-0.022	0.0095	-0.065	-0.099
updrs_2	-0.059	-0.086	-0.073	-0.032	0.0078	0.037	-0.06	-0.069	-0.027	-0.032	0.039	-0.059	-0.08
updrs_3	-0.055	-0.12	-0.073	-0.012	0.028	0.02	-0.059	-0.073	-0.029	-0.012	0.032	-0.034	-0.076
updrs_4	-0.076	0.04	0.03	-0.038	-0.05	-0.069	-0.05	0.02	-0.032	-0.038	-0.011	-0.028	0.025

Methods

- **Baseline CatBoost Model -** CatBoost works by building an ensemble of decision trees using gradient boosting with efficient handling of categorical features, regularization, early stopping, and various optimization techniques. This combination of features makes CatBoost a powerful and versatile machine learning library for solving a wide range of problems.

- **Linear Regression Model -** Linear regression model when applied on time series data can be useful to forecast future values of the dependent variable based on the past observations. The linear model for this time series data is trained for each target variable: "updrs_1, updrs_2, updrs_3 and updrs_4", making the linear model as a univariate time series model.

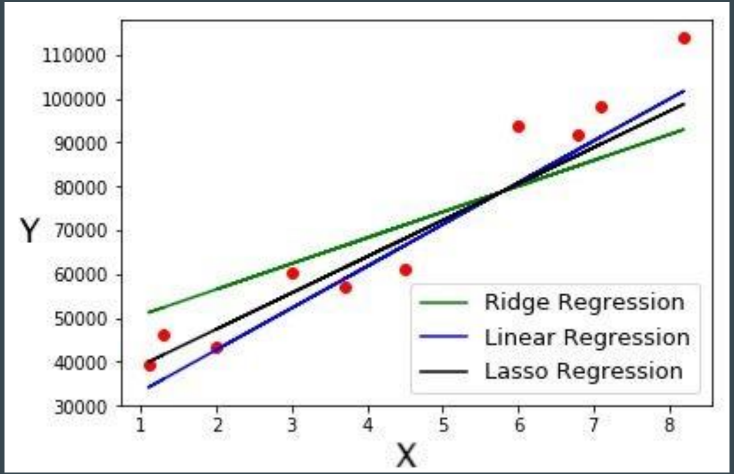


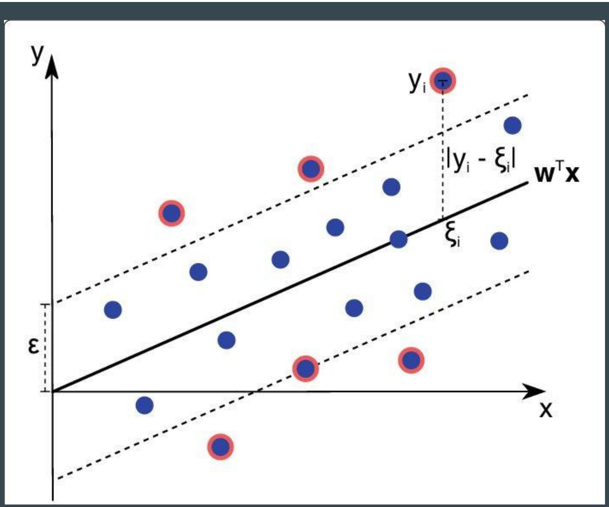


Methods

- **Ridge Regression Model -** Ridge regression is a type of linear regression model that can handle multicollinearity, which is common in time series data. Ridge regression helps to address this issue by introducing a penalty term, which shrinks the coefficient of the predictor variables towards zero and reduces the impact of multicollinearity.

- **Support Vector Regression Model -** SVR has several advantages over traditional time series modelling techniques, including the ability to handle non-linear relationships and the ability to generalize well to unseen data. SVR can be particularly useful when the time series data has complex patterns or relationships that are difficult to model using traditional methods.

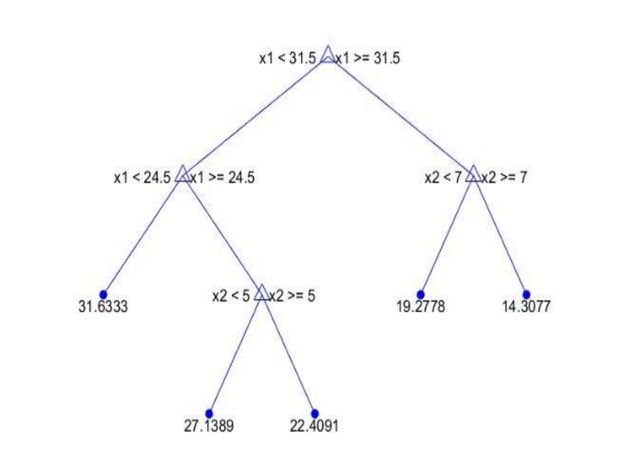


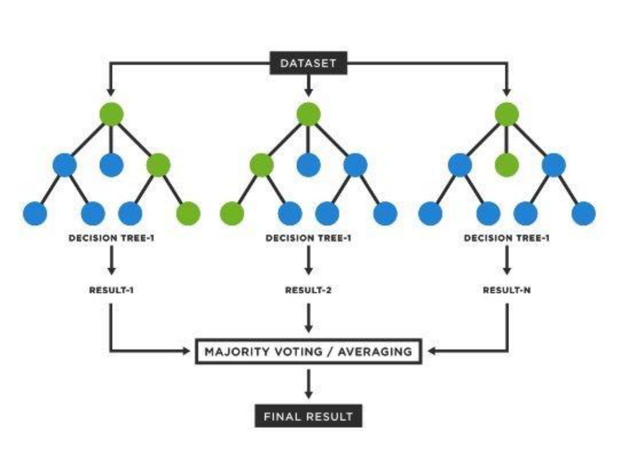


Methods

- **Decision Tree Regression Model -** Decision tree regression is a machine learning algorithm for modeling and predicting time series data by dividing it into smaller subsets based on predictor variables. It offers advantages such as handling non-linear relationships and missing data, outperforming traditional time series models.

- **Random Forest Regression Model -** Random Forest Regression is a machine learning technique that uses multiple decision trees to predict time series data. It effectively handles linear and nonlinear relationships, as well as missing and noisy data, making it a strong tool for time series forecasting.

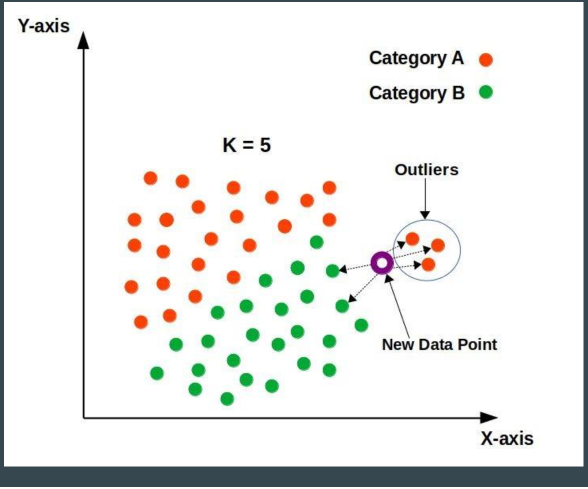


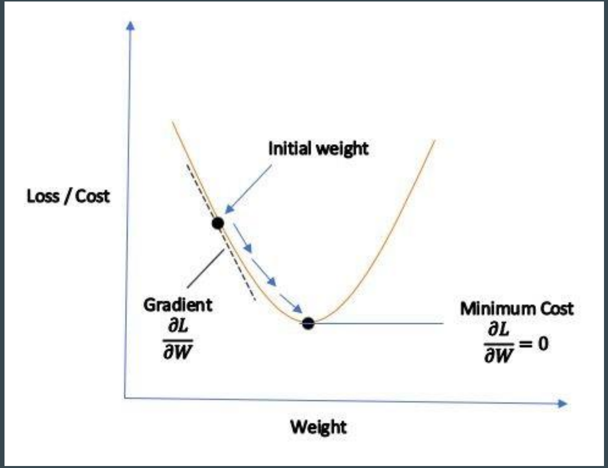


Methods

- **KNN Regression Model -** KNN Regression is a type of machine learning algorithm that can be used to predict time series data by finding the K nearest data points in the training set and using their average to make predictions for new data points. KNN Regression can be particularly useful when the time series data has non-linear relationships and can handle both continuous and categorical data.

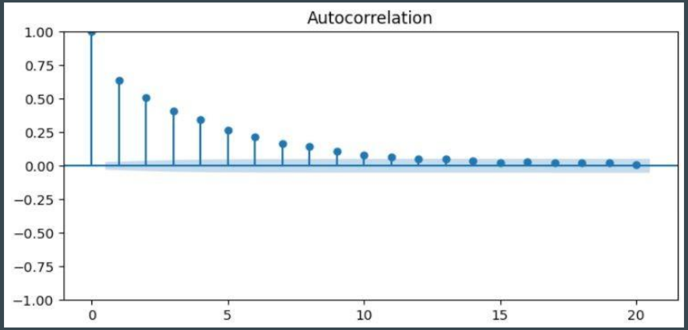
- **Stochastic Gradient Descent Regressor -** SGD Regression is a linear regression method that employs stochastic gradient descent for optimizing model parameters. It is computationally efficient, suitable for large datasets and time series forecasting. The algorithm can manage continuous and categorical data and adapts to dynamic data distributions.

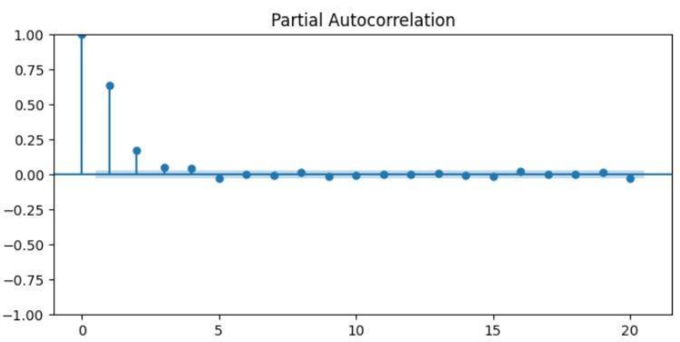



Initial weight
Loss / Cost
Gradient $\frac{\partial L}{\partial W}$
Minimum Cost $\frac{\partial L}{\partial W} = 0$
Weight

Methods

- **ARIMA model -** ARIMA (Autoregressive Integrated Moving Average) is a popular time series forecasting model used to predict future values of a univariate time series. ARIMA models are based on the assumption that the time series is stationary, meaning that the mean, variance, and autocorrelation structure of the series remain constant over time. ARIMA models are denoted as ARIMA(p, d, q), where p, d, and q are the orders of the AR, I, and MA components, respectively. The p and q orders denote the number of lag terms in the AR and MA components, while the d order denotes the number of times the time series needs to be different to become stationary.

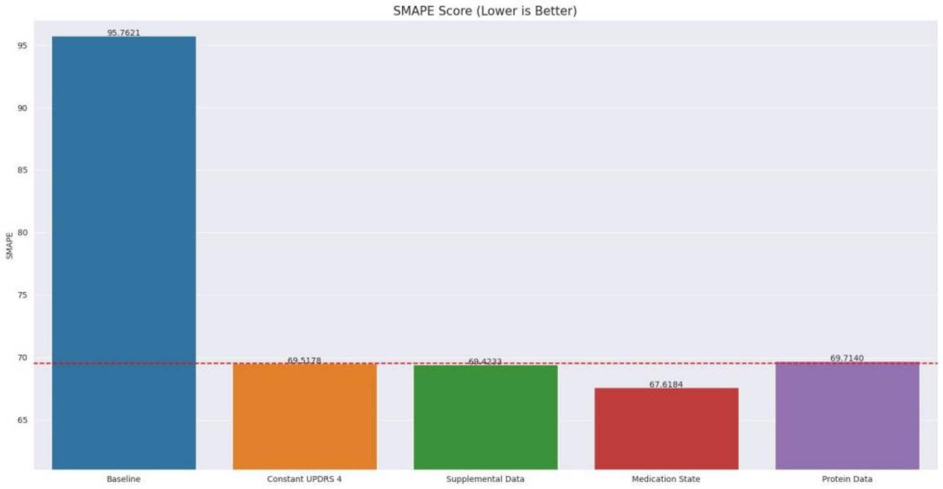


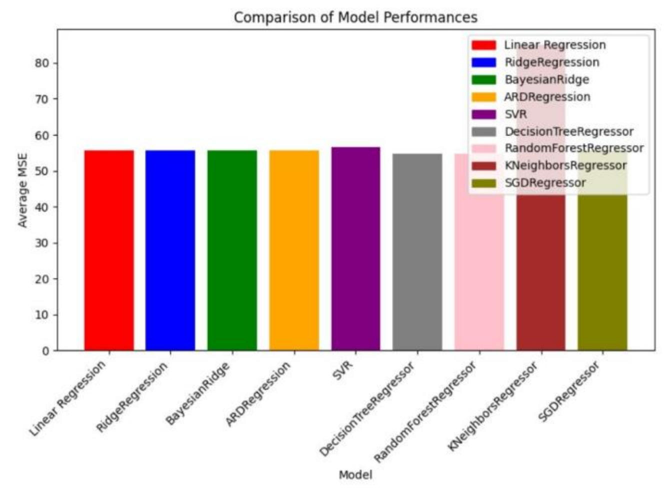


Results

- The performance comparison of the models reveals that all models outperform the baseline SMAPE score of 95.7621. The model using Medication State achieves the lowest SMAPE score of 67.6184, followed by Supplemental Data (69.4233), Constant UPDRS 4 (69.5178), and Protein Data (69.7140). The red dashed line indicates the baseline performance that the models aim to surpass.

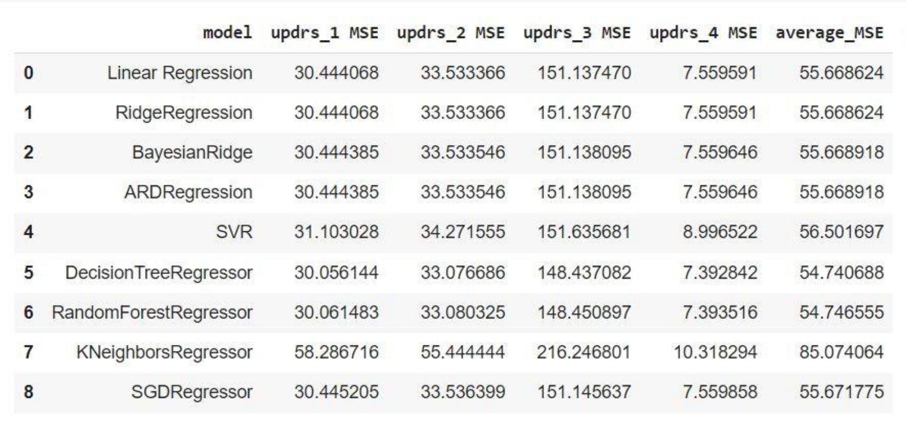
- From the bar chart, we can say that DecisionTreeRegressor model has the lowest average MSE value among all. Therefore, it has the best performance over all other models.





Conclusion

- The dataset consists of 248 patients in clinical data and 771 patients in supplemental data, with no null entries in protein or peptide data but null entries in clinical and supplemental data.
- Null values cannot be assumed to be 0 for UPDRS assessment parts or medication state of the patient, and duplicated data is rare and unlikely to impact model performance.
- Clinical data has a broader visit_month range compared to supplemental data (0-108 months vs. 0-36 months), and both data types have distinct data distributions.
- Protein and peptide samples from the dataset do not show clear indicators of disease progression or severity, according to Shi et al (2015).
- Correlation analysis reveals weak negative correlations: protein P07333 with updrs_3 and peptide GLVSWGNIPC(UniMod_4)GSK with updrs_1.
- DecisionTreeRegressor model has the best performance as it has the lowest average MSE value among all other models when tested.



	model	updrs_1 MSE	updrs_2 MSE	updrs_3 MSE	updrs_4 MSE	average_MSE
0	Linear Regression	30.444068	33.533366	151.137470	7.559591	55.668624
1	RidgeRegression	30.444068	33.533366	151.137470	7.559591	55.668624
2	BayesianRidge	30.444385	33.533546	151.138095	7.559646	55.668918
3	ARDRegression	30.444385	33.533546	151.138095	7.559646	55.668918
4	SVR	31.103028	34.271555	151.635681	8.996522	56.501697
5	DecisionTreeRegressor	30.056144	33.076686	148.437082	7.392842	54.740688
6	RandomForestRegressor	30.061483	33.080325	148.450897	7.393516	54.746555
7	KNeighborsRegressor	58.286716	55.444444	216.246801	10.318294	85.074064
8	SGDRegressor	30.445205	33.536399	151.145637	7.559858	55.671775

Thank You