Parkinson hastalığı veri setindeki ses kayıtları verilerini inceleyerek bu verilerin nasıl işlendiğini ve makine öğrenmesi için nasıl uygun hale getirildiğini açıklayacağım. Genelde bu tür veri setlerinde ses verilerinden elde edilen özellikler, çeşitli biyomedikal analizlerle çıkarılır ve makine öğrenmesi modellerine uygulanabilir hale getirilir.

Bu veri seti, ses kayıtlarından çıkarılan biyomedikal özelliklere dayanır. Parkinson hastalığı üzerinde yapılan analizler için sesin bazı özelliklerinin (frekans, genlik, gürültü oranı gibi) Parkinson hastalığının varlığıyla ilişkilendirilebileceği gösterilmiştir.

**Ses Verisinin İşlenmesi ve Makine Öğrenmesine Uygun Hale Getirilmesi**

1. **Ses Kayıtlarından Özellik Çıkartma**:
   * **MDVP (Multi-Dimensional Voice Program)** verisi, sesin temel frekansları, genlik varyasyonları, gürültü oranları gibi parametreleri içerir.
   * Bu özellikler, sesin biyomedikal analizi ile elde edilir ve genellikle ses kaydının çeşitli zaman dilimlerinden alınan istatistiksel bilgilerle çıkarılır.
   * Bu parametreler arasında **ortalama frekans (Fo)**, **maksimum ve minimum frekans (Fhi ve Flo)**, **jitter (frekansın düzensizlik ölçüsü)**, **shimmer (genlik varyasyonu)** gibi özellikler bulunur.
2. **Veri Hazırlığı ve Özelliklerin Çıkartılması**:
   * Veri setinde her birey için yaklaşık 6 ses kaydı vardır. Bu ses kayıtlarından elde edilen sayısal veriler, her bir kaydın özelliğini temsil eder.
   * **Özellik Çıkartma**: Ses kaydından çıkan her bir özellik, verinin bir sütunu haline gelir. Bu sütunlar, frekans varyasyonları, genlik ölçümleri ve gürültü oranları gibi sesin önemli biyomedikal göstergelerini içerir.
3. **Özniteliklerin Seçimi**:
   * Özellik seçimi, hangi ses özelliklerinin Parkinson hastalığı ile ilişkili olduğunu anlamaya yardımcı olabilir. **Frekans varyasyonları**, **jitter** ve **shimmer** gibi özellikler, Parkinson hastalığının etkilerini yansıtan ses değişikliklerini gözlemleyebilir.
   * Genellikle sesin temel frekansları ve genlik ölçümleri, hastalığın daha erken tespitine yardımcı olan temel özelliklerdir.
4. **Veri Ölçeklendirme (Standardizasyon)**:
   * Makine öğrenmesi modellerinin düzgün çalışabilmesi için verilerin genellikle ölçeklendirilmesi gerekir.
   * Bu, özellikle mesafe tabanlı modeller (örneğin k-NN, SVM) için önemlidir çünkü farklı özelliklerin ölçeklerinin farklı olması modelin başarısını etkileyebilir.
   * **StandardScaler** gibi bir araçla tüm sayısal özellikler, ortalaması 0 ve standart sapması 1 olacak şekilde dönüştürülür.
5. **Veri Bölme (Train/Test Split)**:
   * Veri seti, eğitim ve test verilerine ayrılır. Genellikle, eğitim verisi modelin öğrenmesi için kullanılırken, test verisi modelin genel doğruluğunu test etmek için kullanılır.
   * Bu bölme, modelin eğitim sırasında aşırı öğrenmesini (overfitting) önlemeye yardımcı olur.
6. **Makine Öğrenmesi Modeline Uygun Hale Getirilmesi**:
   * **Hedef değişken (label)** olarak, bu veri setinde Parkinson hastalığı olan bireyler için '1' ve sağlıklı bireyler için '0' kullanılır.
   * Bu, bir **ikili sınıflandırma problemi** haline gelir ve makine öğrenmesi algoritmaları bu sınıflandırma problemini çözmek için uygulanabilir.
7. **Özelliklerin Analizi**:
   * Özellikler (sesin frekans varyasyonları, genlik değişimleri vb.) genellikle **korelasyon matrisi** ve **grafiksel analizler** kullanılarak incelenir. Hangi özelliklerin daha fazla bilgi taşıdığı ve modelin daha iyi tahmin yapmasına yardımcı olacağı belirlenir.