Project Proposal: Predicting No-Shows for Medical Appointments Using Machine Learning

Introduction:

Missed medical appointments cause significant issues, such as resource waste, increased healthcare costs, and scheduling inefficiencies. Predicting no-shows can help healthcare providers optimize their schedules and resources. My project aims to predict no-shows using machine learning techniques. The dataset, containing 49,593 medical records from the University of Vale do Itajaí (CER), includes patient demographics, appointment details, and weather conditions, offering a strong foundation for analysis.

Related Work:

Previous studies have tackled the problem of no-shows using various predictive techniques. Mohammadi et al. (2018) utilized logistic regression and decision trees, while Abdelfotoh et al. (2019) applied neural networks and random forests. These studies focused on factors like age, gender, and type of appointment to enhance scheduling efficiency. We aim to build upon these methods using modern machine learning techniques and additional variables like weather conditions, which may influence attendance.

Requirements:

- **Data Link**: https://data.mendeley.com/datasets/wm6w2fvkfj/1
- **Data Attributes**: This project will use variables such as appointment time, no-show status, age, disability, weather conditions etc.
- **Key Questions**:
  1. How do weather conditions (rain, temperature) influence the likelihood of no-shows?
  2. Do certain appointment times (morning vs. afternoon) have higher no-show rates?
  3. Is there a significant correlation between a patient's age and their likelihood of missing an appointment?
  4. Are patients with disabilities more likely to miss appointments?
  5. How do severe weather conditions (e.g., storms) impact appointment attendance?
  6. Does the gender of a patient influence no-show likelihood?
  7. Are patients from certain cities more prone to missing appointments?
  8. How do different specialties (e.g., physical therapy vs. speech therapy) affect attendance rates?
  9. Do patients attending their first appointment have a higher no-show rate compared to follow-ups?
  10. Are patients who need a companion more likely to miss appointments?

- **Visualizations**:
  1. A heatmap showing the correlation between weather conditions and attendance rates.
  2. A bar chart comparing no-show rates across different age groups.

3. A time-series chart illustrating no-show rates across various appointment times.
4. A map showing no-show patterns by city.
5. A comparison of no-show rates between patients with and without disabilities.

- **Technical Tools**: Python and Jupyter Notebooks will be used for data processing and machine learning. If possible, I will create visualizations using Power BI or Tableau, while managing version control through GitHub.

Milestones:

1. **Data Collection and Cleaning**: Preprocess the dataset by handling missing data, categorizing variables, and ensuring data quality.
2. **Exploratory Data Analysis**: Perform initial data analysis to uncover trends and outliers.
3. **Model Development**: Build predictive models using decision trees, logistic regression, and random forests.
4. **Evaluation**: Assess the models using accuracy, precision, recall, and F1-score metrics.
5. **Visualization and Reporting**: Create the required visualizations and document findings in a final report.

Rationale:

- **Why this project?** Missed appointments lead to inefficiencies in healthcare, wasted time, and additional costs. By predicting no-shows, providers can improve scheduling and resource management, leading to better patient outcomes.

- **Computing Resources**: I will use my personal computer to develop the project using local resources like Python and Jupyter Notebooks for data processing and machine learning. If possible, I will use visualization tools like Power BI or Tableau's desktop versions to create comprehensive reports and dashboards for deeper insights into the data.

- **Project Suitability**: This project aligns with my coursework in data science and allows me to apply machine learning, data analysis, and visualization techniques to real-world healthcare challenges.